CrossMark

ORIGINAL PAPER

Likelihood robust optimization for data-driven problems

Zizhuo Wang[1] · **Peter W. Glynn[2]** · **Yinyu Ye[2]**

Abstract We consider optimal decision-making problems in an uncertain environment. In particular, we consider the case in which the distribution of the input is unknown, yet there is some historical data drawn from the distribution. In this paper, we propose a new type of distributionally robust optimization model called the *likelihood robust optimization* (LRO) model for this class of problems. In contrast to previous work on distributionally robust optimization that focuses on certain parameters (e.g., mean, variance, etc.) of the input distribution, we exploit the historical data and define the accessible distribution set to contain only those distributions that make the observed data achieve a certain level of likelihood. Then we formulate the targeting problem as one of optimizing the expected value of the objective function under the worst-case distribution in that set. Our model avoids the over-conservativeness of some prior robust approaches by ruling out unrealistic distributions while maintaining robustness of the solution for any statistically likely outcomes. We present statistical analyses of our model using Bayesian statistics and empirical likelihood theory. Specifically, we prove the asymptotic behavior of our distribution set and establish the relationship between our model and other distributionally robust models. To test the performance of our model, we apply it to the newsvendor problem and the portfolio

✉ Zizhuo Wang
  zwang@umn.edu

  Peter W. Glynn
  glynn@stanford.edu

  Yinyu Ye
  yinyu-ye@stanford.edu

[1] Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, USA

[2] Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, USA

selection problem. The test results show that the solutions of our model indeed have desirable performance.

1 Introduction

The study of decision-making problems in uncertain environments has been a main focus in the operations research community for decades. In such problems, one has a certain objective function to optimize, however, the objective function depends not only on the decision variables, but also on some unknown parameters. Such situations are ubiquitous in practice. For example, in an inventory management problem, the inventory cost is influenced by both the inventory decisions and the random demands. Similarly, in a portfolio selection problem, the realized return is determined by both the choice of the portfolio and the random market fluctuations.

One solution method to such problems is stochastic optimization. In the stochastic optimization approach, one assumes the knowledge of the distribution of the unknown parameters and chooses the decision that optimizes the expected value of the objective function. If the knowledge of the distribution is exact, then this approach is a precise characterization for a risk-neutral decision maker. Much research has been conducted on this topic; we refer the readers to Shapiro et al. (2009) for a comprehensive review on the topic of stochastic optimization.

However, there are several drawbacks of the stochastic optimization approach. First, although stochastic optimization can frequently be formulated as a convex optimization problem, in order to solve it, one often has to resort to the Monte Carlo method, which can be computationally challenging. More importantly, due to the limitation of knowledge, the distribution of the uncertain parameters is rarely known in practice to a precise level. Even if enough data have been gathered in the past to perform statistical analyses for the distribution, the analyses are often based on assumptions (e.g., independence of the observations, or stationarity of the sequence) that are only approximations of the reality. In addition, many decision makers in practice are not risk-neutral. They tend to be risk-averse. A solution approach that can guard them from adverse scenarios is of great practical interest.

One such approach was proposed by Scarf (1958) in a newsvendor context and has been studied extensively in the past decade. It is called the *distributionally robust optimization* (DRO) approach. In the DRO approach, one considers a *set* of distributions for the uncertain parameters and optimizes the worst-case expected value of the objective function among the distributions in that set. Studies have been done by choosing different distribution sets. Among them, most choose the distribution set to contain those distributions with a fixed mean and variance. For example, Scarf (1958) shows that a closed-form solution can be obtained in the newsvendor problem context when such a distribution set is chosen. Other earlier works include Gallego and Moon (1993), Dupačová (1987) and Žáčková (1966). The same form of the distribution set is also used in Calafiore and El Ghaoui (2006), Yue et al. (2006), Zhu et al. (2013) and Popescu Popescu (2007) in which a linear-chance-constrained problem, a minimax regret objective and a portfolio optimization problem are considered, respectively. Other distribution sets beyond the mean and variance have also been proposed in the

literature. For example, Delage and Ye (2008) propose a more general framework with a distribution set formed by moment constraints. A review of the recent developments can be found in Gabrel et al. (2014).[1]

Although the mean-variance DRO approach is intuitive and is tractable under certain conditions, it is unsatisfactory from at least two aspects. First, when constructing the distribution set in such an approach, one only uses the moment information in the sample data, while all the other information is ignored. This procedure may discard important information in the data set. For example, a set of data drawn from an exponential distribution with $\lambda = 1/50$ will have similar mean and variance as a set of data drawn from a normal distribution with $\mu = \sigma = 50$. In the mean-variance DRO, they will result in the same distribution set and the same decision will be chosen. However, these two distributions have very different properties and the optimal decisions may be quite different. Second, in the DRO approach, the worst-case distribution for a decision is often unrealistic. For example, Scarf (1958) shows that the worst-case distribution in the newsvendor context is a two-point distribution. This raises the concern that the decision chosen by this approach is guarding some overly conservative scenarios, while performing poorly in more likely scenarios. Unfortunately, these drawbacks seem to be inherent in the model choice and cannot be remedied easily.

In this paper, we propose another choice of the distribution set in the DRO framework that solves the above two drawbacks of the mean-variance approach. Instead of using the mean and variance to construct the distribution set, we choose to use the *likelihood function*. More precisely, given a set of historical data, we define the distribution set to be the set of distributions that make the observed data achieve a certain level of likelihood. We call this approach the *likelihood robust optimization* (LRO) approach. The goal of this paper is to study the properties of LRO and its performance.

First, we show that the LRO model is tractable. By applying the duality theory, we formulate the robust counterpart of this problem into a single convex optimization problem. In addition, we show that our model is very flexible. We can add any convex constraints (such as moment constraints) to the distribution set while still maintaining its tractability. Two concrete examples (a newsvendor problem and a portfolio selection problem) are discussed in the paper to illustrate the applicability of our framework.

Then we study the statistical theories behind the LRO approach by illustrating the linkage between our approach and the Bayesian statistics and empirical likelihood theory. We show that the distribution set in our approach can be viewed as a confidence region for the distributions given the set of observed data. Then we discuss how to choose the parameter in the distribution set to attain a specified confidence level. Furthermore, we show a connection between the LRO approach and the mean-variance DRO approach. Our analysis shows that the LRO approach is fully data-driven, and it takes advantage of the full strength of the available data while maintaining a certain level of robustness.

---

[1] We note that there is also a vast literature on robust optimization where the worst-case parameter is chosen for each decision made. However, the robust optimization is based on a slightly different philosophy than the distributionally robust optimization and is usually more conservative. It can also be viewed as a special case of the distributionally robust optimization where the distribution set only contains singleton distributions. In view of this, we choose not to include a detailed discussion of this literature in the main text but refer the readers to Ben-Tal et al. (2009) and Bertsimas et al. (2011) for comprehensive reviews.

Finally, we test the performance of the LRO model in two problems, a newsvendor problem and a portfolio selection problem. In the newsvendor problem, we find that our approach produces similar results compared to the mean-variance DRO approach when the underlying distribution is symmetric, while the solution of our approach is much better when the underlying distribution is asymmetric. In the portfolio selection problem, we show by using real historical data that our approach achieves decent returns. Furthermore, the LRO approach will naturally diversify the portfolio. As a result, the returns of the portfolio have a relatively small fluctuation.

Recently, Ben-Tal et al. (2013) studied a distributionally robust optimization model where the distribution set is defined by divergence measures. Their model contains the LRO as a special case. They also discuss solvability and the statistical properties of their models. Besides the paper by Ben-Tal et al. (2013), a recent paper by Bertsimas et al. (2013) also studies DRO with various choices of the distribution set. In particular, they focus on using data and hypothesis-testing tools to construct those sets. In this paper, we focus on the distribution set defined by the likelihood function and further explore the connections of this approach to the empirical likelihood theory. Thus, we derive specific results for the DRO problems with distribution sets defined by the likelihood function.

Two other papers related to this one are Iyengar (2005) and Nilim and El Ghaoui (2005). In these two papers, the authors study the robust Markov Decision Process (MDP) problem in which the transition probabilities can be chosen from a certain set. They mention the likelihood set as one choice. However, they do not further explore the properties of this set or attempt to extend it to general problems.

The remainder of this paper is organized as follows: in Sect. 2, we introduce our likelihood robust optimization framework and discuss its tractability and statistical properties. In Sect. 3, we present numerical tests of our model. Section 4 concludes this paper.

2 Likelihood robust optimization model

In this section, we formulate the *likelihood robust optimization* (LRO) model, discuss its tractability and statistical properties. We only consider the case where the uncertainty parameters have a finite discrete support in this section.

Suppose we want to maximize an objective function $h(x, \xi)$ where x is the decision variable with feasible set D, and ξ is a random variable taking values in $\Xi = \{\xi_1, \xi_2, \ldots, \xi_n\}$. The set Ξ is known in advance. Assume we have observed N independent samples of ξ, with N_i occurrences of ξ_i. We define:

$$\mathbb{D}(\gamma) = \left\{ p = (p_1, p_2, \ldots, p_n) \left| \sum_{i=1}^{n} N_i \log p_i \geq \gamma, \right. \right.$$
$$\left. \sum_{i=1}^{n} p_i = 1, p_i \geq 0, \forall i = 1, \ldots, n \right\}. \tag{1}$$

We call $\mathbb{D}(\gamma)$ the *likelihood robust distribution set* with parameter γ. Note that $\mathbb{D}(\gamma)$ contains all the distributions with support in Ξ such that the observed data achieve an

empirical likelihood of at least $\exp(\gamma)$. At this point, we treat γ as a given constant. Later we will discuss how to choose γ such that (1) has a desirable statistical meaning. We formulate the LRO problem as follows:

$$\text{maximize}_{\boldsymbol{x} \in D} \left\{ \min_{\boldsymbol{p} \in \mathbb{D}(\gamma)} \sum_{i=1}^{n} p_i h(\boldsymbol{x}, \boldsymbol{\xi}_i) \right\}. \tag{2}$$

In (2), we choose the decision variable $\boldsymbol{x}$, such that the expectation of the objective function under the worst-case distribution is maximized, where the worst-case distribution is chosen among the distributions such that the observed data achieve a certain level of likelihood. Prior to this work, researchers have chosen other types of distribution sets for the inner problem in (2), e.g., distributions with moment constraints. Works of that type have been reviewed in Sect. 1.

Here we comment on our assumption about the known support of the random variable $\boldsymbol{\xi}$. In the LRO, the choice of Ξ is important; a different choice of Ξ will result in a different distribution set and a different solution to the decision problem. In practice, sometimes Ξ has a clear definition. For example, if $\boldsymbol{\xi}$ is the transition indicator of a Markov chain (thus those $\boldsymbol{p}$s are transition probabilities), then the set of Ξ is simply all the attainable states in the Markov chain. However, in cases where the choice of Ξ is less clear, e.g., when $\boldsymbol{\xi}$ represents the return of certain assets, the decision maker should choose Ξ to reflect his view of plausible outcomes. Also, as we will show later, one can sometimes add other constraints such as moment constraints into the distribution set. Once such constraints are added, the choice of support often becomes less critical, since the support may be constrained by existing constraints.

2.1 Tractability of the LRO model

In this subsection, we show that the LRO model is easily solvable. To solve (2), we write down the Lagrangian of the inner optimization problem:

$$L(\boldsymbol{p}, \lambda, \mu) = \sum_{i=1}^{n} p_i h(\boldsymbol{x}, \boldsymbol{\xi}_i) + \lambda \left(\gamma - \sum_{i=1}^{n} N_i \log p_i \right) + \mu \left(1 - \sum_{i=1}^{n} p_i \right).$$

Therefore, the dual formulation of the inner problem is

$$\text{maximize}_{\lambda, \mu} \quad \mu + \lambda \left(\gamma + N - \sum_{i=1}^{n} N_i \log N_i \right) - N\lambda \log \lambda$$

$$+ \lambda \sum_{i=1}^{n} N_i \log \left(h(\boldsymbol{x}, \boldsymbol{\xi}_i) - \mu \right)$$

$$\text{s.t.} \quad \lambda \geq 0, \quad h(\boldsymbol{x}, \boldsymbol{\xi}_i) - \mu \geq 0, \quad \forall i. \tag{3}$$

Since the inner problem of (2) is always convex and has an interior solution (given it is feasible), the optimization problem (3) always has the same optimal value as the inner problem of (2) (see, e.g., Boyd and Vandenberghe 2004). Then we combine (3)

and the outer maximization problem. We have that (2) is equivalent to the following concave maximization problem:

$$\text{maximize}_{x,\lambda,\mu} \quad \mu + \lambda \left(\gamma + N - \sum_{i=1}^{n} N_i \log N_i \right) - N\lambda \log \lambda$$

$$+ \lambda \sum_{i=1}^{n} N_i \log (h(x, \xi_i) - \mu)$$

$$\text{s.t.} \quad \lambda \geq 0, \quad h(x, \xi_i) - \mu \geq 0, \ \forall i, \quad x \in D. \tag{4}$$

The following proposition follows immediately from examining the convexity of the problem and studying its KKT conditions. The detailed proof is given in the Appendix.

Proposition 1 *If $h(x, \xi)$ is concave in x, then the likelihood robust optimization problem described in (2) can be solved by solving (4), which is a convex optimization problem. Furthermore, if (x^*, λ^*, μ^*) is an optimal solution for (4), then*

$$p_i = \frac{\lambda^* N_i}{h(x^*, \xi_i) - \mu^*}, \quad \forall i$$

is the corresponding worst-case distribution.

One advantage of the LRO approach is that one can integrate the mean, variance and/or certain other information that is convex in p into the distribution set. For example, if one wants to impose additional linear constraints on p, say $Ap \geq b$ (note that this includes moment constraints as a special case), then the likelihood robust optimization model can be generalized as follows:

$$\text{maximize}_{x \in D} \ \min_{p \in \mathbb{D}(\gamma), Ap \geq b} \sum_{i=1}^{n} p_i h(x, \xi_i). \tag{5}$$

By applying the duality theory again, we can transform (5) into the following problem (μ is the dual variable associated with the constraint $Ap \geq b$, and we assume that $\sum_{i=1}^{n} p_i = 1$ has been incorporated into $Ap \geq b$):

$$\text{maximize}_{x,\lambda,\mu} \ b^T \mu + \lambda \left(\gamma + N - \sum_{i=1}^{n} N_i \log N_i \right) - N\lambda \log \lambda$$

$$+ \lambda \sum_{i=1}^{n} N_i \log (h(x, \xi_i) - a_i^T \mu)$$

$$\text{s.t.} \quad h(x, \xi_i) - a_i^T \mu \geq 0, \ \forall i$$

$$x \in D, \mu \geq 0, \lambda \geq 0,$$

which is again a convex optimization problem and readily solvable.

2.2 Statistical properties of the LRO model

In this subsection, we study the statistical theory behind the LRO model. Specifically, we focus on the likelihood robust distribution set defined in (1). We will address the following questions:

1. What are the statistical meanings of (1)?
2. How does one select a meaningful γ?
3. How does the LRO model relate to other types of distributionally robust optimization models?

We answer the first question in Sect. 2.2.1 by using the Bayesian statistics and empirical likelihood theory to interpret the likelihood constraints. Those interpretations clarify the statistical motivations of the LRO model. Then we answer the second question in Sect. 2.2.2 in which we perform an asymptotic analysis of the likelihood region and point out an asymptotic optimal choice of γ. We study the last question in Sect. 2.2.3 where we present a relationship between our model and the traditional mean robust optimization model.

2.2.1 Bayesian statistics interpretation

Consider a random variable taking values in a finite set Ξ. Without loss of generality, we assume $\Xi = \{1, \ldots, n\}$. Assume the underlying probability distribution of the random variable is $\boldsymbol{p} = \{p_1, \ldots, p_n\}$. We observe historical data $\Psi = \{N_1, \ldots, N_n\}$ in which N_k represents the number of times the random variable takes value k. Then the maximum likelihood estimate (MLE) of $\boldsymbol{p}$ is given by $(N_1/N, N_2/N, \ldots, N_n/N)$ where $N = \sum_{i=1}^n N_i$ is the total number of observations.

Now we examine the set of distributions $\boldsymbol{p}$ such that the likelihood of Ψ under $\boldsymbol{p}$ exceeds a certain threshold. We use the concepts from Bayesian statistics (see e.g., Gelman et al. 1995). Instead of thinking that the data are randomly drawn from the underlying distribution, we treat them as given, and we define a random vector $\boldsymbol{p} = (p_1, \ldots, p_n)$ taking values on the $(n-1)$-dimensional simplex $\Delta = \{\boldsymbol{p} | p_1 + \cdots + p_n = 1, p_i \geq 0\}$ with probability density function proportional to the likelihood function:

$$\prod_{i=1}^n p_i^{N_i}.$$

This distribution of $\boldsymbol{p}$ is known as the Dirichlet distribution. A Dirichlet distribution with parameter Ψ (denoted by $Dir(\Psi)$) has the density function as follows:

$$f(\boldsymbol{p}; \Psi) = \frac{1}{B(\Psi)} \prod_{i=1}^n p_i^{N_i - 1}$$

with

$$B(\Psi) = \int_{\boldsymbol{p} \in \Delta} \prod_{i=1}^n p_i^{N_i - 1} dp_1 \cdots dp_n = \frac{\prod_{i=1}^n \Gamma(N_i)}{\Gamma(\sum_{i=1}^n N_i)},$$

where $\Gamma(\cdot)$ is the Gamma function. The Dirichlet distribution is used to estimate the unknown parameters of a discrete probability distribution given a collection of samples. Intuitively, if the prior distribution is represented as $Dir(\boldsymbol{\alpha})$, then $Dir(\boldsymbol{\alpha} + \boldsymbol{\beta})$ is the posterior distribution following a sequence of observations with histogram $\boldsymbol{\beta}$. For a detailed discussion on the Dirichlet distribution, we refer the readers to Gelman et al. (1995).

In the LRO model, we assume a uniform prior $Dir(\boldsymbol{e})$ on each point in the support of the data, where $\boldsymbol{e}$ is the unit vector. After observing the historical data Ψ, the posterior distribution follows $Dir(\Psi + \boldsymbol{e})$. Note that this process can be adaptive as we observe new data.

Now we turn to examine the likelihood robust distribution set

$$\mathbb{D}(\gamma) = \left\{ \boldsymbol{p} \, \middle| \, \sum_{i=1}^{n} N_i \log p_i \geq \gamma, \, \sum_{i=1}^{n} p_i = 1, \, p_i \geq 0, \forall i = 1, \ldots, n \right\}. \tag{6}$$

This set represents a region in the probability distribution space. In the Bayesian statistics framework, we can compute the probability that $\boldsymbol{p}$ satisfies this constraint:

$$\mathbb{P}^*\left(\boldsymbol{p} \in \mathbb{D}(\gamma)\right) = \frac{1}{B(\Psi + \boldsymbol{e})} \int_{\boldsymbol{p} \in \Delta} \prod_{i=1}^{n} p_i^{N_i} \cdot \mathbb{I}\left(\sum_{i=1}^{n} N_i \log p_i \geq \gamma\right) dp_1 \cdots dp_n,$$

where $\mathbb{P}^*$ is the probability measure of $\boldsymbol{p}$ given the observed data Ψ (thus $\mathbb{P}^*$ is the probability measure under $Dir(\Psi + \boldsymbol{e})$), and $\mathbb{I}(\cdot)$ is the indicator function. When choosing the likelihood robust distribution set, we want to choose γ^* such that

$$\mathbb{P}^*\left(\boldsymbol{p} \in \mathbb{D}(\gamma^*)\right) = 1 - \alpha \tag{7}$$

for some predetermined α. That is, we want to choose γ^* such that (6) is the $1 - \alpha$ confidence region of the probability parameters. The LRO model can then be interpreted as choosing the decision variable to maximize the worst-case objective value where the worst-case distribution is chosen from the confidence region defined by the observed data.

However, in general, trying to find the exact γ^* that satisfies (7) is computationally challenging. In the next subsection, we study the asymptotic properties of γ^*, which help us to approximate it.

2.2.2 Asymptotic analysis of the likelihood robust distribution set

Now we investigate the asymptotic behavior of the likelihood robust distribution set and give an explicit way to choose an appropriate γ in the LRO model. In this section, we assume that the true underlying distribution of the data is $\bar{\boldsymbol{p}} = \{\bar{p}_1, \ldots, \bar{p}_n\}$ with $\bar{p}_i > 0$. We observe N data drawn randomly from the underlying distribution with N_i observations on outcome i ($\sum_{i=1}^{n} N_i = N$). We define γ_N to be the solution such that

$$\mathbb{P}_N\left(\boldsymbol{p} \in \mathbb{D}(\gamma_N)\right) = 1 - \alpha.$$

Here α is a fixed constant and $\mathbb{P}_N$ is the Dirichlet probability measure on $\boldsymbol{p}$ with parameters $N_1, \ldots, N_n$. Clearly γ_N depends on $\{N_i\}_{i=1}^n$ and thus is a random variable. We have the following theorem about the asymptotic properties of γ_N, whose proof is referred to Pardo (2005). (We use $\to_p$ to mean "converge in probability.")

Theorem 1

$$\gamma_N - \sum_{i=1}^n N_i \log \frac{N_i}{N} \to_p -\frac{1}{2} \chi_{n-1,1-\alpha}^2$$

where $\chi_{d,1-\alpha}^2$ is the $1 - \alpha$ quantile of a χ^2 distribution with d degrees of freedom.

Theorem 1 provides a heuristic guideline on how to choose the threshold γ in the LRO model. In particular, one should choose approximately

$$\gamma^* = \sum_{i=1}^n N_i \log \frac{N_i}{N} - \frac{1}{2} \chi_{n-1,1-\alpha}^2$$

in order for the likelihood robust distribution set to have a confidence level of $1-\alpha$. Note that γ^*/N converges to $\sum_i (N_i/N) \log (N_i/N)$ as N grows. This means that when the data set is large, the allowable distributions must be very close to the empirical data. This is unlike some other distributionally robust optimization approaches which construct the distribution sets based on the mean and/or the variance. Even with large data size, the distribution set in those approaches may still contain distributions that are far from the empirical distribution, such as a two-point distribution at the ends of an interval. Therefore our approach does a better job in utilizing the historical data to construct the robust region.

In the following, we prove a theorem about the convergence speed of γ_N. We use $\Rightarrow_d$ to denote "converge in distribution."

Theorem 2

$$\sqrt{N} \left(\frac{\gamma_N}{N} - \sum_{i=1}^n \bar{p}_i \log \bar{p}_i \right) \Rightarrow_d X_0,$$

where $X_0 = \sum_{i=1}^n (1 + \log \bar{p}_i) X_i$ with $(X_1, \ldots, X_n) \sim N(0, \Sigma)$ where

$$\Sigma = \begin{pmatrix} \bar{p}_1(1 - \bar{p}_1) & -\bar{p}_1\bar{p}_2 & \cdots & -\bar{p}_1\bar{p}_n \\ -\bar{p}_1\bar{p}_2 & \bar{p}_2(1 - \bar{p}_2) & \cdots & -\bar{p}_2\bar{p}_n \\ \vdots & \vdots & \ddots & \vdots \\ -\bar{p}_1\bar{p}_n & -\bar{p}_2\bar{p}_n & \cdots & \bar{p}_n(1 - \bar{p}_n) \end{pmatrix}.$$

Proof Given Theorem 1, it suffices to prove that:

$$\sqrt{N} \left(\sum_{i=1}^n \frac{N_i}{N} \log \frac{N_i}{N} - \sum_{i=1}^n \bar{p}_i \log \bar{p}_i \right) \Rightarrow_d X_0.$$

To show this, we note that for fixed N, $(N_1, \ldots, N_n)$ follows a multinomial distribution with parameters $N, \bar{p}_1, \ldots, \bar{p}_n$. By Theorem 14.6 in Wassaman et al. (2009), we have

$$\sqrt{N}\left(\frac{N_1}{N} - \bar{p}_1, \ldots, \frac{N_n}{N} - \bar{p}_n\right) \Rightarrow_d N(0, \Sigma).$$

Define $f(x) = x \log x$. By the mean value theorem,

$$\sqrt{N}\left(\sum_{i=1}^n \frac{N_i}{N} \log \frac{N_i}{N} - \sum_{i=1}^n \bar{p}_i \log \bar{p}_i\right) = \sum_{i=1}^n f'(\eta_i)\sqrt{N}\left(\frac{N_i}{N} - \bar{p}_i\right)$$

where η_i lies between N_i/N and $\bar{p}_i$. By the strong law of large numbers, $\eta_i \to \bar{p}_i$ almost surely. Therefore,

$$\sum_{i=1}^n f'(\eta_i)\sqrt{N}\left(\frac{N_i}{N} - \bar{p}_i\right) \Rightarrow_d X_0,$$

and thus the theorem is proved. $\square$

Theorem 2 precisely states the distribution of γ_N when the sample size is large. In particular, it states that as the number of samples increases, the difference between γ_N/N and the negative entropy of the true underlying distribution ($\sum_{i=1}^n \bar{p}_i \log \bar{p}_i$) shrinks in the speed of $O(1/\sqrt{N})$.

2.2.3 Relation to other types of DRO models

In this section, we show how one can relate other types of DRO models to our LRO model through results in the empirical likelihood theory. Given observed data $X_1, \ldots, X_N$ with the empirical distribution:

$$F_N(t) = \frac{1}{N} \sum_{i=1}^N I(X_i \le t),$$

we define the likelihood ratio of any distribution F by

$$R(F) = \frac{L(F)}{L(F_N)},$$

where $L(F)$ is the likelihood value of the observations $X_1, \ldots, X_N$ under distribution F. Now suppose we are interested in a certain parameter (or certain parameters) $\theta = T(F)$ of the distribution. We can define the profile likelihood ratio function as follows (see Owen 2001):

$$R_T(\theta) = \sup_F \{R(F)|T(F) = \theta\}. \tag{8}$$

Here some common choices of T include the moments or quantiles of a distribution. In DRO, one selects a region Θ and maximizes the worst-case objective value for $\theta \in \Theta$. Using the profile likelihood ratio function in (8), we can define Θ to be of the form

$$\Theta = \{\theta \,|\, R_T(\theta) \geq r_0\} = \{\theta | \exists F, R(F) \geq r_0, T(F) = \theta\}.$$

When θ is the mean of the distribution, the next theorem helps to determine r_0 such that asymptotically Θ is a certain confidence interval for the true mean.

Theorem 3 (Theorem 2.2 in Owen 2001) *Let $X_1, \ldots, X_n$ be independent random variables with common distribution F_0. Let $\mu_0 = \mathbb{E}[X_1]$, and suppose that $0 < Var(X_1) < \infty$. Then $-2\log(R_T(\mu_0))$ converges in distribution to $\chi^2_{(1)}$ as $n \to \infty$.*

Therefore, in order for the set $\{\theta \,|\, R_T(\theta) \geq r_0\}$ to achieve a $1 - \alpha$ confidence level (under the Dirichlet distribution induced by the observed data), one should approximately choose the boundary θ such that

$$R_T(\theta) = e^{-\frac{1}{2}\chi^2_{1,1-\alpha}}.$$

To find such θ, we solve the following convex optimization problems:

$$\text{maximize/minimize}_{\boldsymbol{p},\theta} \qquad \theta$$

$$\text{s.t.} \quad \sum_{i=1}^n N_i \log p_i \geq \log\left(L(F_N)\right) - \tfrac{1}{2}\chi^2_{1,1-\alpha}$$

$$\sum_{i=1}^n p_i = 1, \ \sum_{i=1}^n d_i\, p_i = \theta, \ p_i \geq 0, \forall i. \qquad (9)$$

Note that the optimal values to the maximization and the minimization problems give the exact upper bound and lower bound for θ. We illustrate this with the following example:

Example 1 We consider a data set with support on all integers between 0 and 20 with the histogram shown in Fig. 1. There are 100 data points with sample mean $\bar{X} = 9.88$ and sample standard deviation $s = 3.30$. (Each data is a sample generated by rounding a normal distribution with mean 10 and variance 10.) By using the above procedure while choosing $\alpha = 0.05$, we obtained that the optimal values of the maximization and the minimization problems in (9) are 10.5234 and 9.227, respectively. Therefore, the confidence region generated by this approach is [9.227, 10.523].

We can also directly build a confidence interval for the mean, using the formula

$$\left[\bar{X} - \frac{s \cdot t_{\alpha/2,1-n}}{\sqrt{n}}, \ \bar{X} + \frac{s \cdot t_{\alpha/2,1-n}}{\sqrt{n}}\right] = [9.225, 10.535].$$

As one can see, the two confidence intervals are very close, indicating that when restricting to mean uncertainty, our approach is a reasonable approximation.   □
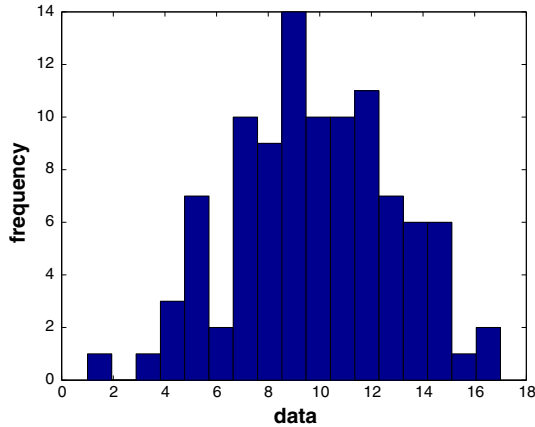
Fig. 1 Histogram for the data in Example 1

Finally, we note that the idea of using the profile likelihood function can also be used to establish relationships between the LRO model and other types of distributionally robust optimization. However, there might not be a closed-form formula for the asymptotic choice of r_0. In those cases, one may need to resort to sampling methods to find an approximate r_0, the procedure of which is beyond the discussion of this paper.

3 Applications and numerical results

In this section, we show two applications of the proposed LRO model and perform numerical tests on them. The two applications are the newsvendor problem (see Sect. 3.1) and the portfolio selection problem (see Sect. 3.2).

3.1 Application 1: newsvendor problem

In this subsection, we apply the LRO model to the newsvendor problem. In such problems, a newsvendor facing an uncertain demand has to decide how many newspapers to stock on a newsstand. If he stocks too many, there will be a per-unit overage cost for each copy that is left unsold; and if he stocks too few, there will be a per-unit underage cost for each unmet demand. The problem is to decide the optimal stocking quantity in order to minimize the expected cost.

The newsvendor problem is one of the most fundamental problems in inventory management and has been studied for more than a century. We refer the readers to Khouja (1999) for a comprehensive review of this problem. In the classical newsvendor problem, one assumes that the distribution of the demand is known (with distribution F). The problem can then be formulated as:

$$\text{minimize}_x \ G_F(x) = b\mathbb{E}_F(d - x)^+ + h\mathbb{E}_F(x - d)^+. \tag{10}$$

In (10), x is the stocking quantity, d is the random demand with a probability distribution F, and $b, h > 0$ are the per-unit underage and overage costs, respectively.

It is well-known that a closed-form solution is available for this problem. However, such a solution relies on the accurate information about the demand distribution F. In practice, one does not always have such information. In many cases, what one has is only some historical data. To deal with this uncertainty, Scarf (1958) proposed a distributionally robust approach in which a decision is selected to minimize the expected cost under the worst-case distribution, where the worst-case distribution is chosen from all distributions that have the same mean and variance as the observed data. However, as we discussed earlier, such an approach does not fully utilize the data. It is also overly conservative as the corresponding worst-case distribution for any given decision is a two-point one, which is implausible from a practical point of view.

On the other hand, researchers have also proposed pure data-driven models to solve the problem, i.e., using the empirical distribution as the true distribution. However, pure data-driven approaches tend to be less robust, since they ignore the potential deviations from the data and do not guard against it.

In the following, we propose the LRO model for the newsvendor problem. We assume that the support of all possible demands is $\Xi = \{0, \ldots, n\}$, i.e., the demand takes integer values between 0 and n. Suppose we are given some historical demand data. We use N_i to denote the number of times that demand is equal to i in the observed historical data. The total number of observations is $N = \sum_{i=0}^{n} N_i$.

The LRO model for the newsvendor problem is given as follows:

$$\text{minimize}_x \quad \max_{\boldsymbol{p}} \quad \sum_{i=0}^{n} p_i \left(b(d_i - x)^+ + h(x - d_i)^+ \right)$$

$$\text{s.t.} \quad \sum_{i=0}^{n} N_i \log p_i \geq \gamma$$

$$\sum_{i=0}^{n} p_i = 1, \quad p_i \geq 0, \ \forall i, \tag{11}$$

where $d_i = i$. Here the outer problem chooses a stocking quantity, while the inner problem finds the worst-case distribution and cost with respect to that decision. Here the first constraint makes sure that the worst-case distribution is chosen among those distributions that make the observed data achieve a chosen level of likelihood γ. By applying the same techniques as in Sect. 2, we can write (11) as a single convex optimization problem:

$$\text{minimize}_{x,\lambda,\mu,\boldsymbol{y}} \quad \mu + \lambda \left(\sum_{i=0}^{n} N_i \log N_i - N - \gamma \right) + N\lambda \log \lambda - \sum_{i=0}^{n} N_i \lambda \log(y_i)$$

$$\text{s.t.} \quad b(d_i - x) + y_i \leq \mu, \ \forall i$$

$$h(x - d_i) + y_i \leq \mu, \ \forall i$$

$$\lambda \geq 0, \quad \boldsymbol{y} \geq 0.$$

This problem is thus easily solvable. Similar approach can be applied to multi-item newsvendor problems.

Now we perform numerical tests for this problem using the LRO model and compare the solution to that of other methods. In the following, we consider a newsvendor problem with unit underage and overage costs $b = h = \$1$. We consider two underlying demand distributions. The first one is a normal distribution with mean $\mu = 50$ and standard deviation $\sigma = 50$. The second one is an exponential distribution with $\lambda = 1/50$. When generating the data, we truncate the demand at 0 and 200 and round the randomly generated values to the nearest integer.

For each underlying distribution, we perform the following procedures for the LRO approach:

1. Generate 1000 historical data from the underlying distribution, and record the number of times the demand is equal to i by N_i.
2. Construct the likelihood robust distribution set $\mathbb{D}(\gamma) = \{\boldsymbol{p} \mid \sum_{i=0}^{n} N_i \log p_i \geq \gamma, \sum_{i=0}^{n} p_i = 1, p_i \geq 0, \forall i\}$. To choose a proper γ, we use the asymptotic result in Theorem 1. We choose γ such that $\mathbb{D}(\gamma)$ covers the true distribution with probability 0.95, where the probability is under the Dirichlet measure with parameters $N_0 + 1, \ldots, N_n + 1$. By Theorem 1, γ should be approximately $\sum_{i=0}^{n} N_i \log \frac{N_i}{N} - \frac{1}{2} \chi^2_{200, 0.95}$.
3. Solve the LRO model (11) with the γ chosen in Step 2.

Using the same group of sample data, we also test the following approaches:

1. LRO with fixed mean and variance: denote the sample mean by $\hat{\mu}$ and the sample variance by $\hat{\sigma}^2$. We add two constraints $\sum_{i=0}^{n} p_i d_i = \hat{\mu}$ and $\sum_{i=0}^{n} p_i d_i^2 = \hat{\mu}^2 + \hat{\sigma}^2$ to the inner problem of (11), that is, we only allow those distributions that have the same mean and variance as the sample mean and variance. As we discussed in (5), this would still be a convex problem and one can obtain the optimal solution easily. We denote this approach by LRO($\hat{\mu}, \hat{\sigma}^2$).
2. The distributionally robust approach with fixed mean and variance: this is the approach proposed by Scarf (1958). In this approach, one minimizes the worst-case expected cost where the worst-case distribution is chosen from all the distributions with a fixed mean (equal to the sample mean) and a fixed variance (equal to the sample variance). We call this approach the Scarf approach. In Scarf (1958), it is shown that the optimal decision for the Scarf approach can be expressed in a closed-form formula:

$$x^*_{scarf} = \hat{\mu} + \frac{\hat{\sigma}}{2} \left(\sqrt{\frac{b}{h}} - \sqrt{\frac{h}{b}} \right).$$

3. The empirical distribution approach: we solve the optimal solution using the empirical distribution as the true distribution.
4. We solve the optimal solution using the true underlying distribution.

Our test results are shown in Tables 1 and 2. All the computations are run on a PC with 1.80GHz CPU and Windows 7 Operating System. We use MATLAB version R2010b to develop the algorithm for solving the optimization problems. In Tables

Table 1 Results when the underlying distribution is normal

	Solution (x^*)	$\mathbb{E}_{F_{true}}[h(x^*, \xi)]$	$h_{\text{LRO}}(x^*)$
LRO	56	35.97 (+0.84 %)	49.32 (+0,00 %)
LRO($\hat{\mu}, \hat{\sigma}^2$)	54	35.80 (+0.37 %)	49.47 (+0.30 %)
Scarf	53	35.74 (+0.20 %)	49.63 (+0.63 %)
Empirical	50	35.79 (+0.34 %)	50.03 (+1.40 %)
Underlying	51	35.67 (+0.00 %)	49.87 (+1.11 %)

Table 2 Results when the underlying distribution is exponential

	Solution (x^*)	$\mathbb{E}_{F_{true}}[h(x^*, \xi)]$	$h_{\text{LRO}}(x^*)$
LRO	46	35.72 (+3.51 %)	53.71 (+0.00 %)
LRO($\hat{\mu}, \hat{\sigma}^2$)	41	34.99 (+1.39 %)	54.45 (+1.38 %)
Scarf	50	36.64 (+6.17 %)	54.27 (+1.04 %)
Empirical	37	34.55 (+0.16 %)	55.75 (+3.80 %)
Underlying	35	34.51 (+0.00 %)	56.20 (+4.64 %)

1 and 2, the second column is the optimal decision computed by each model. The third column shows the expected cost of each decision under the true underlying distribution, which is truncated $N(50, 2500)$ in Table 1 and truncated $\text{Exp}(1/50)$ in Table 2. The last column shows the objective value of each decision under the LRO model, i.e., the worst-case expected objective value when the distribution could be chosen from $\mathbb{D}(\gamma)$. The numbers in the parentheses in both tables are the relative differences between the current solution and the optimal solution under the measure specified in the corresponding column.

In Table 1, one can see that when the underlying demand distribution is normal, the solutions of different models do not differ much from each other. This is mainly because the symmetric property of the normal distribution pushes the solution to the middle in either model. Nevertheless, the worst-case distributions in different cases are quite varied. We plot them in Fig. 2.

In Fig. 2, the x-axis is the demand and the y-axis is the probability mass function over the demands. We can see that the worst-case distribution in the Scarf model is a two-point distribution (with positive mass at 3 and 103). However, a two-point distribution is not a realistic one, which means that Scarf's approach might be guarding some overly conservative scenarios. In contrast, the worst-case distributions when we use LRO and LRO($\hat{\mu}, \hat{\sigma}^2$) models are much closer to the empirical data and look much more plausible. In particular, the LRO with mean and variance constraints results in a worst-case distribution closest to the empirical data among these three approaches.

The situation is different in the second case when the underlying distribution is an exponential distribution. In this case, Scarf's solution is significantly worse than the LRO solutions. This is because it does not use the information that the data are skewed. In fact, it still only takes the mean and variance as the input. In contrast, the LRO and
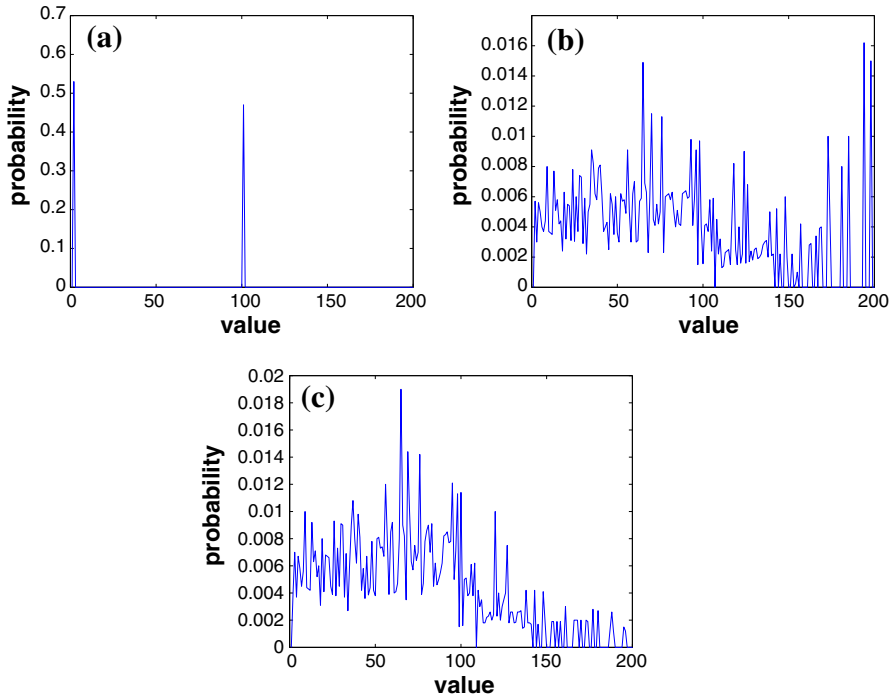
Fig. 2 Worst-case distributions when the underlying demand distribution is normal

LRO($\hat{\mu}, \hat{\sigma}^2$) adapt to the asymmetry of the data and only consider the distributions that are close to the empirical distribution.

In both cases, we observe that the LRO($\hat{\mu}, \hat{\sigma}^2$) seems to work better. Indeed, we find that adding some constraints on p regarding the mean/variance of the distribution usually helps the performance. This is because it helps to further concentrate the distribution set to those that have similar shape as the empirical distribution. Lastly, we find that the empirical distribution works well if we use the true distribution to evaluate the performance. However, it is not robust to potential deviations from the underlying distribution. As shown in the last column in Tables 1 and 2, when we allow the underlying distribution to change to some degree (within a 95 % confidence range), the performance of the solution obtained by using the empirical distribution might be suboptimal.

3.2 Application 2: portfolio selection problem

In this subsection, we apply the LRO model to the portfolio selection problem. In the portfolio selection problem, there are d assets that are available to invest for the next day. The decision maker observes N historical daily returns of the d assets $\{\boldsymbol{\xi}_i\}_{i=1}^{N}$ where each $\boldsymbol{\xi}_i$ is a d-dimensional vector drawn from some underlying unknown distribution. In our case, we consider a certain support Ξ of all possible returns, and

the choice of Ξ can be derived by using statistical models to calculate the boundaries of the possible returns. The decision in this problem is to make a portfolio selection x in a feasible set D (e.g., D may contain some budget constraints) to maximize the expected return. We refer the readers to Luenberger (1997) for a thorough review on this problem.

In this section, we demonstrate how to apply the LRO model to the portfolio selection problem and compare its performance to that of some other methods. First, we formulate the LRO model for the portfolio selection problem as follows:

$$\text{maximize}_{x \in D} \quad \min_p \sum_{i=1}^{N} p_i \cdot \xi_i^T x + \sum_{\sigma \in \Xi \setminus \{\xi_1, \dots, \xi_N\}} p_\sigma \cdot \xi_\sigma^T x$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \log p_i \geq \gamma$$

$$e^T p = 1, \quad p \geq 0. \tag{12}$$

In (12), the first term in the objective function corresponds to the scenarios that have been observed in the data, while the second term corresponds to the unobserved scenarios. Note that theoretically Ξ could be an uncountable set, in which case the second term should be replaced by an integral. However, practically, one can always discretize Ξ to arbitrary accuracy (e.g., use a grid discretization). Therefore, writing the second term as a sum is without loss of practicality. The first constraint in (12) is the likelihood constraint introduced in Sect. 2. For the ease of notation, we assume that each return profile occurs only once in the data, which is reasonable if the set of possible return profiles is huge (e.g., if the set is a result of discretizing a continuous set). Although the number of points in Ξ could be large or even infinite, as long as we can solve

$$g(x) = \min_{\xi \in \Xi} \xi^T x \tag{13}$$

efficiently for any x, the whole problem can be formulated as a convex optimization problem of polynomial size. To see that, we take the dual of the inner problem of (12) and combine it with the outer problem. We have that the entire problem can be written as one single optimization problem as follows:

$$\text{maximize}_{x,\mu,\lambda} \quad \mu + \gamma\lambda + N\lambda - N\lambda \log \lambda + \lambda \sum_{i=1}^{N} \log(\xi_i^T x - \mu)$$

$$\text{s.t.} \quad \mu \leq \xi_\sigma^T x, \quad \forall \sigma \in \Xi \tag{14}$$

$$\lambda \geq 0, \quad x \in D.$$

It appears that the first set of constraints could be of large size, however, if $g(x)$ defined in (13) can be solved efficiently, then this is effectively one single constraint. For example, if $\Xi = \{(r_1, \dots, r_d) | r_i \in [\underline{r}_i, \bar{r}_i]\}$, then the constraint is equivalent to $\mu \leq \underline{r}^T x$. If $\Xi = \{(r_1, \dots, r_d) | \|r - r_0\|_2 \leq \eta\}$, then the constraint is equivalent to $\mu \leq r_0^T x - \eta\|x\|$, which is also convex.

Again, we can add any convex constraints of p into (12) such as constraints on the moments of the returns. Thus our model is quite flexible. Next we perform numerical tests using our model. We adopt the following setup (this setup is also used in other studies of the portfolio selection problem, see, e.g., Delage and Ye 2008):

- We gather the historical data of 30 assets from the S&P 500 Index during the time period from 2001 to 2004. In each experiment, we choose four assets to focus on and the decision is to construct a portfolio using these four assets for each day during this period. We use the past 30 days' data as the observed data in the LRO approach.

To select the γ in (14), we still use the asymptotic result in Theorem 1. In particular, we choose the degree of freedom in the chi-square distribution to equal the number of historical observations. Although this is a heuristic choice, we find it works well in the numerical test. We compare our approach to the following three other approaches:

1. A naive approach in which each day, the stock with the highest past 30 days' average daily return is chosen to be the sole stock in the portfolio. We call this approach the *single stock* (SS) approach.
2. An approach in which each stock is chosen with 1/4 weight (in terms of the capital size) for each day. We call this approach the *equal weights* (EQ) approach.
3. The distributionally robust approach with fixed mean and variance estimated from the past 30 days' observations (see Popescu 2007), called the DRO approach.

We acknowledge that there are many other ways one can choose the portfolio and we do not intend to show that our approach is the best one for this problem. Instead, we aim to illustrate that the LRO approach gives a decent performance with some desired features.

In Table 3, we present the test results. The results are obtained using codes written in MATLAB 2010b. Some of the codes are borrowed from Delage and Ye (2008). We did 100 experiments (each with four different stocks randomly chosen from the 30 to form the portfolio for each day), and the tested period covers 721 days.

In Table 3, the numbers in the second column indicate that among the 100 experiments, the number of times the overall return of the LRO approach outperforms the corresponding method. The numbers in the third column are the improvements of the average return of the LRO method over the corresponding model, and the numbers in the last column are the standard deviations of all daily returns of the portfolio

Table 3 Test results for the portfolio selection problem

Approach	# LRO outperforms (out of 100 cases)	Average gain of LRO over	Std of returns
LRO	N/A	N/A	2.3 %
SS	65	1.8 %	3.3 %
EQ	60	1.1 %	1.5 %
DRO	58	1.3 %	2.4 %

constructed using each method. From Table 3, we observe that among the 100 experiments, LRO outperforms all other tested methods in terms of the number of times it has a higher return, as well as the average overall return during the 721 days. Also, it has a decent standard deviation of the daily returns. In particular, it has a much smaller standard deviation than the SS approach, and a comparable standard deviation to the DRO approach. To investigate further, when we look at the portfolios constructed by the LRO method, we find that 52 % of the time, it chooses a single stock to form the portfolio. And it chooses two stocks, three stocks and all four stocks in its solution for 39 %, 8 % and 1 % of the times. Therefore, we find that LRO approach implicitly achieves diversification, which is a feature that is usually desired in such problems.

4 Conclusion

In this paper, we propose a new distributionally robust optimization framework, the likelihood robust optimization model. The model optimizes the worst-case expected value of a certain objective function where the worst-case distribution is chosen from the set of distributions that make the observed data achieve a certain level of likelihood. The proposed model is easily solvable and has strong statistical meanings. It avoids the over-conservatism of other robust models while protecting the decisions against reasonable deviations from the empirical data. We discuss two applications of our model. The numerical results show that our model might be appealing in several applications.

5 Appendix

Proof of Proposition 1 First, we show that problem (4) is a convex optimization problem. To see this, we first write (4) in a slightly different but equivalent way:

$$\text{maximize}_{\boldsymbol{x},\lambda,\mu,\boldsymbol{y}} \; \mu + \lambda(\gamma + N - \sum_{i=1}^{n} N_i \log N_i) - N\lambda \log \lambda + \lambda \sum_{i=1}^{n} N_i \log y_i$$
$$\text{s.t.} \qquad \lambda \geq 0, \quad h(\boldsymbol{x}, \boldsymbol{\xi}_i) - \mu \geq y_i, \; \forall i, \quad \boldsymbol{x} \in D. \qquad (15)$$

Note that because the objective is to maximize, (15) is an equivalent representation of (4). It is easy to see that when $h(\boldsymbol{x}, \boldsymbol{\xi})$ is concave in $\boldsymbol{x}$, the constraints are all convex constraints. Now we only need to verify that the objective function is jointly concave. Note that the first two terms of the objective function are linear, therefore, we only need to show that the function $-N\lambda \log \lambda + \lambda \sum_{i=1}^{n} N_i \log y_i$ is jointly concave in λ and $\boldsymbol{y}$. We consider the Hessian matrix of the objective function with respect to λ and $\boldsymbol{y}$:

$$H = \begin{pmatrix} -\frac{N}{\lambda} & \frac{N_1}{y_1} & \frac{N_2}{y_2} & \cdots & \frac{N_n}{y_n} \\ \frac{N_1}{y_1} & -\frac{\lambda N_1}{y_1^2} & 0 & \cdots & 0 \\ \frac{N_2}{y_2} & 0 & -\frac{\lambda N_2}{y_2^2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{N_n}{y_n} & 0 & 0 & \cdots & -\frac{\lambda N_n}{y_n^2} \end{pmatrix}.$$

Now consider a diagonal matrix A with diagonal elements $\lambda, y_1, \ldots, y_n$, then we have

$$A^T H A = \begin{pmatrix} -N\lambda & \lambda N_1 & \lambda N_2 & \cdots & \lambda N_n \\ \lambda N_1 & -\lambda N_1 & 0 & \cdots & 0 \\ \lambda N_2 & 0 & -\lambda N_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda N_n & 0 & 0 & \cdots & -\lambda N_n \end{pmatrix}$$

which is a diagonal dominant matrix with all negative diagonal elements. Thus $A^T H A$ is negative semidefinite, which further implies that H is negative semidefinite. Thus the objective function is a concave function and the optimization problem is a convex optimization problem.

Finally, we prove that the worst-case distribution satisfies:

$$p_i = \frac{\lambda^* N_i}{h(\boldsymbol{x}^*, \boldsymbol{\xi}_i) - \mu^*}, \quad \forall i.$$

We consider the derivative of the Lagrangian function $L(\boldsymbol{p}, \lambda, \mu)$ with respect to p_i. By the KKT conditions, we must have $p_i \nabla_{p_i} L(\boldsymbol{p}, \lambda, \mu) = 0$, which implies that $p_i(h(\boldsymbol{x}^*, \boldsymbol{\xi}_i) - \mu^*) = \lambda^* N_i$. Also, we note that in (4), since there is a term of $\log(h(\boldsymbol{x}, \boldsymbol{\xi}_i) - \mu)$, therefore, $h(\boldsymbol{x}^*, \boldsymbol{\xi}_i) - \mu^* \neq 0$. Thus Proposition 1 is proved. $\square$

References

Ben-Tal A, den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. Manage Sci 59(2):341–357

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) Robust optimization. Princeton Series in Applied Mathematics. Princeton University Press, Princeton

Ben-Tal A, Nemirovski A (1999) Robust solutions of uncertain linear programs. Oper Res Lett 25(1):1–13

Bertsimas D, Thiele A (2006) Robust and data-driven optimization: modern decision-making under uncertainty. In: Tutorial on Operations Research, INFORMS

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

Bertsimas D, Gupta V, Kallus N (2013) Data-driven robust optimization. arXiv:1401.0212

Bertsimas D, Sim M (2004) The price of robustness. Oper Res 52(1):35–53

Bertsimas D, Thiele A (2006) A robust optimization approach to inventory theory. Oper Res 54(1):150–168

Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, Cambridge

Calafiore G, El Ghaoui L (2006) On distributionally robust chance-constrained linear programs. J Optim Theory Appl 130(1):1–22

Chen X, Sim M, Sun P (2007) A robust optimization perspective on stochastic programming. Oper Res 55(6):1058–1071

Delage E (2009) Distributionally robust optimization in context of data-driven problems. PhD thesis

Delage E, Ye Y (2008) Distributionally robust optimization under moment uncertainty with application to data-driven problems. Oper Res 58(3):595–612

Dupacová J (1987) The minimax approach to stochastic programming and an illustrative application. Stochastics 20(1):73–88

Gabrel V, Murat C, Thiele A (2014) Recent advances in robust optimization: An overview. Eur J Oper Res 235(3):471–483

Gallego G, Moon I (1993) The distribution free newsboy problem: Review and extension. J Oper Res Soc 44(8):825–834

Gelman A, Carlin J, Stern HS, Rubin DB (1995) Bayesian data analysis. Chapman and Hall Press, New York

Iyengar G (2005) Robust dynamic programming. Math Oper Res 30(2):257–280

Johnson N, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, vol. 1. Wiley Series in Probability and Statistics

Khouja M (1999) The single period newsvendor problem: literature review and suggestions for future research. Omega 27:537–553

Luenberger D (1997) Investment science. Oxford University Press, Oxford

Mason D, Schuenemeyer J (1983) A modified Kolmogorov-Smirnov test sensitive to tail alternatives. Ann Stat 11(3):933–946

Nilim A, El Ghaoui L (2005) Robust control of Markov decision processes with uncertain transition matrices. Oper Res 53(5):780–798

Owen A (2001) Empirical likelihood. Chapman and Hall Press, London

Pardo L (2005) Statistical inference based on divergence measures. Chapman and Hall Press, London

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

Popescu I (2007) Robust mean-covariance solutions for stochastic optimization. Oper Res 55(1):98–112

Scarf H (1958) A min-max solution of an inventory problem. In: Arrow K, Karlin S, Scarf H (eds) Studies in the mathematical theory of inventory and production. Stanford University Press, California, pp 201–209

Scarf H (1959) Bayes solutions of the statistical inventory problem. Ann Math Stat 30(2):490–508

Shapiro A, Dentcheva D, Ruszczynski A (2009) Lectures on stochastic programming: modeling and theory. MPS-SIAM Series on Optimization

Shapiro A, Kleywegt AJ (2002) Minimax analysis of stochastic programs. Optim Methods Softw 17:523–542

So AMS, Zhang J, Ye Y (2009) Stochastic combinatorial optimization with controllable risk aversion level. Math Oper Res 34(3):522–537

Wassaman L (2009) All of statistics: a concise course in statistical inference. Springer Texts in Statistics, New York

Yue J, Chen B, Wang M (2006) Expected value of distribution information for the newsvendor problem. Oper Res 54(6):1128–1136

Ẑácková J (1966) On minimax solutions of stochastic linear programming problems. Casopis pro Pêstování Matematiky 91:423–430

Zhu Z, Zhang J, Ye Y (2013) Newsvendor optimization with limited distribution information. Optim Methods Softw 28(3):640–667