SCIENCE CHINA
Mathematics

• ARTICLES •

https://doi.org/10.1007/s11425-023-xxxx-x

Solving a class of distributionally robust optimization problems with sublinear expectation

Xingbang Cui[1,2] & Xiaojun Chen[2,*]

[1]*School of Mathematics and Statistics, Shaanxi Normal University, Xi'an* 710119*, China;*
[2]*Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon* 999077*, Hong Kong*

Email: cuixingbangsdu@126.com, maxjchen@polyu.edu.hk

Received June 29, 2025; accepted January 12, 2026; published online 2026

Abstract In this paper, we solve a class of distributionally robust optimization (DRO) problems with Lipschitz continuous loss functions and weakly compact ambiguity sets via sublinear expectation introduced by Peng (2009). We reformulate the DRO problem as the minimization problem by using sublinear expectation, and introduce a discrete approximation by grouping samples. We prove that optimal values and optimal solutions of the discrete problem converge to those of the DRO problem with probability 1 under capacity. We show that the discrete form is an asymptotic unbiased estimator for the sublinear expectation of the loss function, and provide the quantification of difference between the discrete problem and the DRO problem with special moment ambiguity set. Numerical experiments of two real life data sets are conducted. Our preliminary numerical results show that the sublinear expectation method outperforms the existing duality method, especially from the perspective of reliability.

Keywords sublinear expectation, distributionally robust optimization, capacity

MSC(2020) 90C15, 60E05

Citation: Cui X, Chen X. Solving a class of distributionally robust optimization problems with sublinear expectation. Sci China Math, 2026, 69, https://doi.org/10.1007/s11425-023-xxxx-x

1 Introduction

In this paper, we focus on applying the sublinear expectation theory to the following distributionally robust optimization (DRO) problem

$$\min_{u\in U}\max_{P\in\mathcal{P}} E_P(\ell(u;\xi)), \tag{1.1}$$

where u is the decision vector taking values from a compact set $U\subset\mathbb{R}^s$, random vector $\xi:\Omega\to\Xi\subset\mathbb{R}^t$ is defined on measurable space $(\Omega,\mathcal{F})$, the loss function $\ell:\mathbb{R}^s\times\mathbb{R}^t\to\mathbb{R}$ is continuous, P is the probability distribution of ξ over Ξ equipped with Borel σ-algebra $\mathcal{B}$, E_P denotes the expectation with respect to probability distribution P, $\mathcal{P}$ denotes a set of probability distributions. In what follows, we give some assumptions that will be used throughout the remainder of this paper.

* Corresponding author

Assumption 1.1. *(a) For any fixed $\xi \in \Xi$, $\ell(\cdot;\xi)$ is Lipschitz continuous on U with Lipschitz modulus being bounded by $\kappa(\xi)$, where $\max_{P\in\mathcal{P}} E_P(\kappa(\xi)) < \infty$.*

(b) There exists $u_0 \in U$ such that $\max_{P\in\mathcal{P}} |E_P(\ell(u_0;\xi))| < \infty$.

(c) $\mathcal{P}$ is nonempty and weakly compact.

(d) For any $u \in U$, $\lim_{r\to\infty} \max_{P\in\mathcal{P}} \int_{\{\xi\in\Xi:|\ell(u;\xi)|\geqslant r\}} |\ell(u;\xi)|P(d\xi) = 0$.

DRO problem (1.1) is to find a decision u that minimizes the worst-case expected loss. Ambiguity set $\mathcal{P}$ is a family of probability distributions characterized through certain known properties of the uncertainty, which contains the true probability distribution. There are many types of ambiguity sets including the moment ambiguity set [5] and Wasserstein ambiguity set [7]. DRO has been successfully applied in many fields such as finance [1,5] and machine learning [2,12,24]. For more details on the DRO problem, readers can refer to [13].

It is challenging to solve DRO problem (1.1) due to the minimax framework of the problem and the involvement of the expectation. Most numerical methods for solving the DRO problem (1.1) address an equivalent semi-infinite reformulation of (1.1) with finitely many variables and infinitely many constraints, which is obtained by dualizing the inner worst-case expectation problem. The numerical methods for the semi-infinite program such as the cutting-plane algorithm are usually computationally expensive. One alternative is to transform the semi-infinite program to finite-dimensional minimization problem, which can be achieved by invoking the results of convex analysis [13, Section 7]. Nevertheless, such reduction has serious limitations due to the specific assumptions on loss function ℓ, ambiguity set $\mathcal{P}$, support Ξ, etc. Further, the dimension of the corresponding finite reformulation may be very large, see [5, p. 13] for an example. Hence, it is necessary to seek a simple and efficient method for solving DRO problem (1.1).

In this paper, we solve DRO problem (1.1) via sublinear expectation theory proposed by Peng [20,21]. Basically, a sublinear expectation is a functional defined on a linear space of random variables that is monotone, constant preserving, positively homogeneous and sub-additive. The sublinear expectation theory includes the classic probability theory as a special case, and is used to characterize the random variables whose distributions are uncertain, which is consistent with the distribution ambiguity in DRO problems. There are fruitful results about sublinear expectation theory such as law of large numbers (LLN) and central limit theorem (CLT) under sublinear expectation [21, Chapter 2]. Sublinear expectation theory has been applied in stochastic differential equations [14,20], Brownian motion [9,20,23], martingale [26], etc. The sublinear expectation theory has also been used extensively in many fields including finance [6,22] and robust statistics [10,15,22].

Another main reason for us to use sublinear expectation in this paper is that we only need the condition that samples are independently identically distributed (i.i.d.) under sublinear expectation (see Definition 2.5). It was pointed out in [21, p. viii] that the i.i.d. condition under sublinear expectation is weaker than the i.i.d condition in classic probability theory. We take the stock market as an example. The classic i.i.d. condition refers to that the means and variances of daily return rates in a period must be the same, while the i.i.d. condition under sublinear expectation refers to that the means and variances can vary within the same range, which is obviously more reasonable.

Given DRO problem (1.1), define $\mathcal{M}(\Xi) := \{\varphi : \Xi \to \mathbb{R} : \varphi \text{ is measurable over } (\Xi, \mathcal{B}) \text{ and } |E_P(\varphi(\xi))| < \infty \text{ for } P \in \mathcal{P}\}$. For any $\varphi \in \mathcal{M}(\Xi)$, let

$$\hat{E}(\varphi(\xi)) := \sup_{P\in\mathcal{P}} E_P(\varphi(\xi)). \tag{1.2}$$

Later, we can show that $\hat{E}$ is a sublinear expectation defined on $\mathcal{M}(\Xi)$. Then DRO problem (1.1) can be transformed to the following model

$$\min_{u\in U} \ \hat{E}(\ell(u;\xi)). \tag{1.3}$$

Like the sample average approximation in classic probability theory [25], a discrete approximation for sublinear expectation has been introduced in [10,15,22]. Let m and n be positive integers, and denote $N = mn$. Given samples $\{\xi^i \in \mathbb{R}^t, i = 1, \cdots, N\}$ that are i.i.d. under sublinear expectation, we divide

these samples sequentially by index into m groups with equal size n, where, for $1 \leqslant k \leqslant m$, the k-th group of samples is $\{\xi^{i+(k-1)n}, i = 1, \cdots, n\}$. Although there exist some other partition schemes such as the partition in the triangle order [10], the partition in the sequential order is the most popular scheme [10,15,22], which can simplify the theoretical analysis and is also more suitable for the time series analysis in the financial applications. According to [10, 15, 22], a discrete approximation for $\hat{E}(\ell(u;\xi))$ is denoted as

$$\max_{1\leqslant k\leqslant m} \frac{1}{n}\sum_{i=1}^{n} \ell(u;\xi^{i+(k-1)n}). \tag{1.4}$$

Therefore, in this paper, we consider investigating the discrete approximation of problem (1.1) denoted as

$$\min_{u\in U} \max_{1\leqslant k\leqslant m} \frac{1}{n}\sum_{i=1}^{n} \ell(u;\xi^{i+(k-1)n}). \tag{1.5}$$

The discrete form (1.4) with $m = 1$ is exactly the sample average approximation in the form of $\frac{1}{N}\sum\limits_{i=1}^{N} \ell(u;\xi^i)$. However, the choice of $m > 1$ is advocated in practice. There are two advantages for the choice of $m > 1$. Firstly, in real life scenarios such as the stock market [5, 22], using a large history data set would make the assumption of i.i.d. samples somewhat unrealistic, so it is more reasonable to group samples in sequence. Secondly, the existing numerical experiments on NASDAQ Composite Index [22, Table 2], S&P 500 Index [22, Tables 5,6,7] and Vendors Database of China [15, Tables 16,17] indicate that discrete form (1.4) with $m > 1$ achieves more reliable performances.

The main contributions of this paper are summarized as follows.

(i) We reformulate DRO problem (1.1) as the minimization problem (1.3) via sublinear expectation, and introduce the discrete problem (1.5).

(ii) We prove that the optimal values and optimal solutions of the problem (1.5) converge to those of the problem (1.1) with probability 1 under capacity if the samples are i.i.d. under sublinear expectation.

(iii) We prove that the discrete form (1.4) is an asymptotic unbiased estimator for objective function $\hat{E}(\ell(u;\xi))$, and provide the quantification of difference between the problem (1.5) and the problem (1.1) with special moment ambiguity set.

The rest of this paper is organized as follows. In Section 2, the knowledge about the sublinear expectation and capacity is introduced. In Section 3, the convergence properties of the problem (1.5) are studied. In Section 4, the statistical properties of the discrete form (1.4) are exhibited. In Section 5, the quantification of difference between the problem (1.5) and the problem (1.1) with special moment ambiguity set is presented. In Section 6, the numerical experiments are conducted. In Section 7, we conclude the paper.

Notations: Let $L^0(\mathbb{R}^k)$ denote the space of Borel measurable functions on $\mathbb{R}^k$, $C(\mathbb{R}^k)$ denote the space of continuous functions on $\mathbb{R}^k$, and $C_{Lip}(\mathbb{R}^k)$ denote the space of Lipschitz continuous functions on $\mathbb{R}^k$. Let $\|\cdot\|$ denote the Euclidean norm. Denote $\mathrm{dist}(x, B) = \inf\limits_{y\in B} \|x - y\|$ and $\mathbb{D}(A, B) = \sup\limits_{x\in A} \mathrm{dist}(x, A)$, for $A, B \subset \mathbb{R}^k$. Denote $A\backslash B = \{x : x \in A, x \notin B\}$, for $A, B \subset \mathbb{R}^k$. Let $\mathcal{N}(\mu, \sigma)$ denote the normal distribution with mean μ and standard deviation σ. Denote $(v)_+ = \max(0, v)$ and $(v)_- = (-v)_+$ for $v \in \mathbb{R}$. For $v^1, v^2 \in \mathbb{R}^k$, $(|v^1|)_i = |(v^1)_i|$, $i = 1, \cdots, k$, and $v^1 \leqslant v^2$ is equivalent to $v^2 - v^1 \in \mathbb{R}^k_+$.

2 Preliminaries

2.1 Sublinear expectation

The following background knowledge can be found in [21]. Let $\tilde{\Omega}$ be a given set and $\mathcal{H}$ be a linear space of real-valued functions defined on $\tilde{\Omega}$. Suppose that $\mathcal{H}$ satisfies that constant function $X_c \in \mathcal{H}$ with $X_c : \omega \in \tilde{\Omega} \mapsto c$ for any given constant $c \in \mathbb{R}$ and $|X| \in \mathcal{H}$ if $X \in \mathcal{H}$.

Definition 2.1. ([21]) A sublinear expectation $\hat{E} : \mathcal{H} \to \mathbb{R}$ is a functional satisfying the following conditions.

(i) Monotonicity: $\hat{E}(X) \geqslant \hat{E}(Y)$ if $X \geqslant Y$.

(ii) Constant preserving: $\hat{E}(c) = c$ for any constant $c \in \mathbb{R}$.

(iii) Sub-additivity: $\hat{E}(X+Y) \leqslant \hat{E}(X) + \hat{E}(Y)$ for $X, Y \in \mathcal{H}$.

(iv) Positive homogeneity: $\hat{E}(\lambda X) = \lambda \hat{E}(X)$ for $\lambda \geqslant 0$.

The triple $(\tilde{\Omega}, \mathcal{H}, \hat{E})$ is called a sublinear expectation space.

If $(\tilde{\Omega}, \mathcal{H}, \hat{E})$ is a sublinear expectation space, then $X \in \mathcal{H}$ is called a random variable, and $Y = (X_1, \cdots, X_k)$ with $X_i \in \mathcal{H}$, for $i = 1, \cdots, k$, is called an k-dimensional random vector in $\mathcal{H}^k$.

Remark 2.2. The sublinear expectation is closely related to coherent risk measure. Let $(\tilde{\Omega}, \tilde{\mathcal{F}})$ be a measurable space, and $\mathcal{H}$ be a linear space of $\tilde{\mathcal{F}}$-measurable functions defined on $\tilde{\Omega}$. Assume that $\hat{E}$ is a sublinear expectation over $\mathcal{H}$. For any $X \in \mathcal{H}$, denote $\rho(X) := \hat{E}(-X)$. Then, from [21, Section 1.6], ρ is a coherent risk measure over $\mathcal{H}$, and the converse conclusion also holds.

Definition 2.3. ([21]) Let X and Y be two k–dimensional random vectors defined on a sublinear expectation space $(\tilde{\Omega}, \mathcal{H}, \hat{E})$. They are called identically distributed if for any $\varphi \in L^0(\mathbb{R}^k)$, $\hat{E}(\varphi(X)) = \hat{E}(\varphi(Y))$, which is denoted by $X \overset{d}{=} Y$.

Definition 2.4. ([21]) Let $(\tilde{\Omega}, \mathcal{H}, \hat{E})$ be a sublinear expectation space. Random vector $Y \in \mathcal{H}^i$ is said to be independent of another random vector $X \in \mathcal{H}^j$ under $\hat{E}$ if for any $\varphi \in L^0(\mathbb{R}^{i+j})$, we have $\hat{E}(\varphi(X, Y)) = \hat{E}(\hat{E}(\varphi(x, Y))|_{x=X})$.

Definition 2.5. ([4]) Let $\{Y^i\}_{i=1}^{\infty}$ be a sequence of k-dimensional random vectors on a sublinear expectation space $(\tilde{\Omega}, \mathcal{H}, \hat{E})$. We say that $\{Y^i\}_{i=1}^{\infty}$ is i.i.d. under sublinear expectation $\hat{E}$ if $Y^{i+1} \overset{d}{=} Y^i$ and Y^{i+1} is independent from $\{Y^1, \cdots, Y^i\}$ for $i = 1, 2, \cdots$.

Remark 2.6. It is claimed in [21, p. viii] that the i.i.d. condition under sublinear expectation is weaker than the i.i.d. condition in the classic probability theory. Firstly, random vectors X and Y may be identically distributed under sublinear expectation even when the true probability distributions of X and Y are different. Secondly, random vector Y can be independent of random vector X even when the true probability distribution of Y is related to the realizations of X.

On the other hand, in DRO problem (1.1), let $\xi^1, \cdots, \xi^N$ be i.i.d. samples of ξ generated by the true probability distribution. Note that the ambiguity set $\mathcal{P}$ in (1.1) is not related to the realization of ξ in general. From Definition 2.5, we can see that $\xi^1, \cdots, \xi^N$ also satisfy the i.i.d. assumption under sublinear expectation $\hat{E}$ in (1.2).

2.2 Capacity

The following definition of capacity can be found in [18].

Definition 2.7. ([18]) Let $\tilde{\Omega} \subset \mathbb{R}^k$ be equipped with σ-algebra $\tilde{\mathcal{F}}$. A capacity on $\tilde{\mathcal{F}}$ is a set function $\mu : \tilde{\mathcal{F}} \to [0, 1]$ such that (i) $\mu(\tilde{\Omega}) = 1$ and $\mu(\varnothing) = 0$ and (ii) $\mu(A) \leqslant \mu(B)$, for any $A \subset B$ and $A, B \in \tilde{\mathcal{F}}$.

Consider sublinear expectation space $(\tilde{\Omega}, \mathcal{H}, \hat{E})$, where $\tilde{\Omega} \subset \mathbb{R}^k$ is equipped with Borel σ-algebra $\tilde{\mathcal{F}}$, and there exists a weakly compact set $\tilde{\mathcal{P}}$ of probability measures on $\tilde{\Omega}$ such that

$$\hat{E}(X) := \sup_{P \in \tilde{\mathcal{P}}} E_P(X), \text{ for } X \in \mathcal{H}. \tag{2.1}$$

For any $A \in \tilde{\mathcal{F}}$, denote

$$\tilde{\mathcal{V}}(A) = \sup_{P \in \tilde{\mathcal{P}}} P(A). \tag{2.2}$$

From [4, Lemma 2.4], we have the following proposition.

Proposition 2.8. *([4]) Let $\tilde{\mathcal{V}}$ be defined in (2.2). Then $\tilde{\mathcal{V}}$ is a capacity on $\tilde{\mathcal{F}}$, and satisfies that $\tilde{\mathcal{V}}(A \cup B) \leqslant \tilde{\mathcal{V}}(A) + \tilde{\mathcal{V}}(B)$, for any $A, B \in \tilde{\mathcal{F}}$.*

We have the following definition related to capacity.

Definition 2.9. ([4, 21]) Consider capacity $\tilde{\mathcal{V}}$ defined in (2.2). Let $\{X^i\}_{i=1}^{\infty}$ and X be $\tilde{\mathcal{F}}$-measurable functions.

(i) We say that $\{X^i\}_{i=1}^{\infty}$ converges to X with probability 1 (w.p.1) under capacity $\tilde{\mathcal{V}}$ if there exists $\Omega_1 \subset \tilde{\Omega}$ such that $\tilde{\mathcal{V}}(\Omega_1) = 1$ and $\lim\limits_{i\to\infty} X^i(\omega) = X(\omega)$ for any $\omega \in \Omega_1$.

(ii) We say that $\{X^i\}_{i=1}^{\infty}$ converges to X quasi-surely (q.s.) if there exists $\Omega_2 \subset \tilde{\Omega}$ such that $\tilde{\mathcal{V}}(\Omega_2) = 0$ and $\lim\limits_{i\to\infty} X^i(\omega) = X(\omega)$ for any $\omega \in \tilde{\Omega}\backslash\Omega_2$.

Remark 2.10. From [4, Lemma 2.4], if $\tilde{\mathcal{V}}(\Omega_2) = 0$, then $\tilde{\mathcal{V}}(\tilde{\Omega}\backslash\Omega_2) = 1$, so the concept of "quasi-surely" is stronger than the concept of "with probability 1 under capacity $\tilde{\mathcal{V}}$". In addition, both concepts are equivalent when $\tilde{\mathcal{P}}$ is a singleton set. The definition of "with probability 1 under capacity $\tilde{\mathcal{V}}$" can be found in [4, 17], while the definition of "quasi-surely" can be found in [21, Section 6.1].

On the other hand, the concept of convergence w.p.1 under capacity is also a natural extension of that of convergence w.p.1 in stochastic programming (see [25, Section 7.2.1]) due to the probability ambiguity. When $\tilde{\mathcal{P}}$ is a singleton set, both concepts are identical.

2.3 LLN under sublinear expectation

In this subsection, we exhibit some LLNs under sublinear expectation, which will be used in our paper. The following theorem is [21, Theorem 2.4.1].

Theorem 2.11. *(Weak LLN, [21]) Let $\{Y^i\}_{i=1}^{\infty}$ be a sequence of i.i.d. k-dimensional random vectors on a sublinear expectation space $(\tilde{\Omega}, \mathcal{H}, \hat{E})$. Assume that there exists some $\alpha > 0$ such that $\hat{E}(|Y^1|^{1+\alpha}) < \infty$. Then*

$$\lim_{n\to\infty} \hat{E}\left(\varphi\left(\frac{Y^1 + \cdots + Y^n}{n}\right)\right) = \max_{\theta\in\Theta} \varphi(\theta),$$

for all functions $\varphi \in C(\mathbb{R}^k)$ satisfying linear growth condition ($|\varphi(x)| \leqslant C(1+|x|)$), where Θ is the unique bounded closed convex subset of R^n satisfying $\max\limits_{\theta\in\Theta} \langle p, \theta\rangle = \hat{E}(\langle p, Y_1\rangle)$, for $p \in \mathbb{R}^k$. Here $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product.

Remark 2.12. If $n = 1$ in Theorem 2.11, then $\Theta = [\underline{v}, \overline{v}]$, where $\underline{v} = -\hat{E}(-Y^1)$ and $\overline{v} = \hat{E}(Y^1)$.

Definition 2.13. ([21]) Let $\tilde{\mathcal{V}}$ be defined in (2.2). A mapping $X : \tilde{\Omega} \to \mathbb{R}$ is said to be quasi-continuous if for any $\epsilon > 0$, there exists open set $\Omega_1 \subset \tilde{\Omega}$ with $\tilde{\mathcal{V}}(\Omega_1) = 0$ such that X is continuous over $\tilde{\Omega}\backslash\Omega_1$.

Consider sublinear expectation space $(\tilde{\Omega}, \mathcal{H}, \hat{E})$ with $\hat{E}$ defined in (2.1). For $p \geqslant 1$, denote $\mathcal{L}^p = \{X \in \mathcal{H} : \hat{E}(|X|^p) < \infty\}$, $\mathcal{L}_b^p = \{X \in \mathcal{L}^p : \lim\limits_{i\to\infty} \hat{E}(|X|^p I_{\{|X|>i\}}) = 0\}$, and $\mathcal{L}_c^p = \{X \in \mathcal{L}_b^p : X \text{ is quasi-continuous}\}$, where $I_{\{|X|>i\}}$ is the indicator function of set $\{|X| > i\}$ with $I_{\{|X|>i\}}(\omega) = 1$ if $\omega \in \{|X| > i\}$ and $I_{\{|X|>i\}}(\omega) = 0$ if $\omega \notin \{|X| > i\}$. The following theorem is [28, Theorem 3.6].

Theorem 2.14. *(Strong LLN, [28]) Let $\{Y^i \in \mathcal{L}_c^{1+\alpha}\}_{i=1}^{\infty}$ for some $\alpha > 0$ be a sequence of i.i.d. random variables on sublinear expectation space $(\tilde{\Omega}, \mathcal{H}, \hat{E})$, where $\hat{E}$ is defined in (2.1) and $\tilde{\mathcal{V}}$ is defined in (2.2). Then there exists $\hat{P} \in \tilde{\mathcal{P}}$ such that*

$$\hat{P}\left(\lim_{n\to\infty} \frac{Y^1 + \cdots + Y^n}{n} = \hat{E}(Y^1)\right) = 1, \tag{2.3}$$

i.e.,

$$\tilde{\mathcal{V}}\left(\lim_{n\to\infty} \frac{Y^1 + \cdots + Y^n}{n} = \hat{E}(Y^1)\right) = 1.$$

Remark 2.15. There exist other strong LLNs under sublinear expectation, see [4] and references therein. Let assumptions of Theorem 2.14 hold. Denote $\overline{\gamma} = \hat{E}(Y^1)$, $\underline{\gamma} = -\hat{E}(-Y^1)$. In [4, Theorem 3.1], it is proved that

$$P\left(\underline{\gamma} \leqslant \liminf_{n\to\infty} \frac{Y^1 + \cdots + Y^n}{n} \leqslant \limsup_{n\to\infty} \frac{Y^1 + \cdots + Y^n}{n} \leqslant \overline{\gamma}\right) = 1, \text{ for } P \in \tilde{\mathcal{P}}. \tag{2.4}$$

One of the major differences between (2.3) and (2.4) is that (2.3) holds for some $\hat{P} \in \tilde{\mathcal{P}}$, whereas (2.4) holds for all $P \in \tilde{\mathcal{P}}$.

Consider a sublinear space $(\tilde{\Omega}, \mathcal{H}, \hat{E})$. Let $\{X^i\}_{i=1}^{\infty}$ be a sequence of i.i.d. samples with $X^i \overset{d}{=} X$. For integers $m, n > 0$, define $\hat{\overline{v}} = \max\limits_{1 \leqslant k \leqslant m} \frac{1}{n} \sum\limits_{i=1}^{n} X^{i+n(k-1)}$. Let $\overline{v} = \hat{E}(X)$, $\underline{v} = -\hat{E}(-X)$ and $\bar{\sigma}^2 = \sup\limits_{P \in \mathcal{P}} E_P((X - E_P(X))^2)$. The following proposition is [8, Proposition 2.1].

Proposition 2.16. *([8]) Let X be a random variable on sublinear space $(\tilde{\Omega}, \mathcal{H}, \hat{E})$. Let $\{X^i\}_{i=1}^{\infty}$ be a sequence of i.i.d. samples with $X^i \overset{d}{=} X$. Assume that $\hat{E}((X)^2) < \infty$. There exists constant $C > 0$ such that*

$$\hat{E}(((\hat{\overline{v}} - \overline{v})_+)^2) \leqslant \frac{Cm}{n} \text{ and } \hat{E}(((\hat{\overline{v}} - \underline{v})_-)^2) \leqslant \frac{Cm}{n}, \tag{2.5}$$

for integers $m, n > 0$, where $C = 2(\bar{\sigma}^2 + (\overline{v} - \underline{v})^2)$.

Remark 2.17. In [8, Proposition 2.1], $\hat{\overline{v}}$ in the second inequality of (2.5) is $\min\limits_{1 \leqslant k \leqslant m} \frac{1}{n} \sum\limits_{i=1}^{n} X^{i+n(k-1)}$. Nonetheless, since $\hat{\overline{v}} \geqslant \min\limits_{1 \leqslant k \leqslant m} \frac{1}{n} \sum\limits_{i=1}^{n} X^{i+n(k-1)}$, (2.5) holds obviously. This minor modification is aimed at focusing on discrete form (1.4).

The following proposition is [27, Theorem 5.1].

Proposition 2.18. *([27]) Let X be a random variable on sublinear space $(\tilde{\Omega}, \mathcal{H}, \hat{E})$. Let $\{X^i\}_{i=1}^{\infty}$ be a sequence of i.i.d. samples with $X^i \overset{d}{=} X$. Assume that $\hat{E}((X)^2) < \infty$. Then there exists constant $\hat{C} > 0$ only depending on $\hat{E}((X)^2)$ such that*

$$\left| \hat{E}\left(\phi\left(\frac{X^1 + \cdots + X^n}{n} \right) \right) - \max_{\theta \in [\underline{v}, \overline{v}]} \phi(\theta) \right| \leqslant \hat{C} n^{-\frac{1}{2}}, \tag{2.6}$$

for integers $n > 0$, where $\phi : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant not greater than 1.

3 Convergence properties of the discrete problem (1.5)

3.1 Reformulation of DRO problem (1.1) via sublinear expectation

Firstly, we show that functional $\hat{E}$ defined in (1.2) is a sublinear expectation over $\mathcal{M}(\Xi)$.

Theorem 3.1. *Define $\hat{E}$ in (1.2). Then $\hat{E}$ is a sublinear expectation defined on $\mathcal{M}(\Xi)$, and $(\Xi, \mathcal{M}(\Xi), \hat{E})$ is a sublinear expectation space.*

Proof. Obviously, for any constant c, we have $c \in \mathcal{M}(\Xi)$, which also implies the nonemptiness of $\mathcal{M}(\Xi)$. Moreover, if $|E_P(\varphi(\xi))| < \infty$, then $|E_P(|\varphi(\xi)|)| < \infty$, which indicates that $|\varphi| \in \mathcal{M}(\Xi)$.

Then we verify (i)-(iv) in Definition 2.1. For any $\varphi_1, \varphi_2 \in \mathcal{M}(\Xi)$, if $\varphi_1 \geqslant \varphi_2$ for any $P \in \mathcal{P}$, we have

$$\hat{E}(\varphi_1(\xi)) = \sup_{P \in \mathcal{P}} E_P(\varphi_1(\xi)) \geqslant \sup_{P \in \mathcal{P}} E_P(\varphi_2(\xi)) = \hat{E}(\varphi_2(\xi)),$$

so (i) holds. In addition, for any constant c, we have $\hat{E}(c) = \sup\limits_{P \in \mathcal{P}} E_P(c) = c$, which indicates (ii). Besides, for any $\varphi_1, \varphi_2 \in \mathcal{M}(\Xi)$, we can deduce that

$$\begin{aligned}
\hat{E}(\varphi_1(\xi) + \varphi_2(\xi)) &= \sup_{P \in \mathcal{P}} E_P(\varphi_1(\xi) + \varphi_2(\xi)) \\
&= \sup_{P \in \mathcal{P}} (E_P(\varphi_1(\xi)) + E_P(\varphi_2(\xi))) \\
&\leqslant \sup_{P \in \mathcal{P}} E_P(\varphi_1(\xi)) + \sup_{P \in \mathcal{P}} E_P(\varphi_2(\xi)) \\
&= \hat{E}(\varphi_1(\xi)) + \hat{E}(\varphi_2(\xi)).
\end{aligned}$$

Thus (iii) holds. Finally, for any $\varphi \in \mathcal{M}(\Xi)$ and $\lambda \geqslant 0$, we have

$$\hat{E}(\lambda\varphi(\xi)) = \sup_{P\in\mathcal{P}} E_P(\lambda\varphi(\xi)) = \sup_{P\in\mathcal{P}} \lambda E_P(\varphi(\xi)) = \lambda \sup_{P\in\mathcal{P}} E_P(\varphi(\xi)) = \lambda\hat{E}(\varphi(\xi)),$$

which indicates (iv). So we have proved the conclusion. □

Let $\Phi(u) = \max_{P\in\mathcal{P}} E_P(\ell(u;\xi))$. Then we have the following proposition.

Proposition 3.2. *Let Assumption 1.1 hold. Then $\ell(u;\cdot) \in \mathcal{M}(\Xi)$ for any $u \in U$, and DRO problem (1.1) is equivalent to problem (1.3).*

Proof. According to [29, Proposition 2], we have $|\Phi(u)| < \infty$ for $u \in U$, which indicates the conclusion. In addition, we also know that Φ is Lipschitz continuous over U and the solution set for inner problem of (1.1) is nonempty from [29, Proposition 2], which justifies the well definedness of inner problem of (1.1). □

Remark 3.3. It is often the case that Assumption 1.1(c)(d) hold under moderate conditions. For example, consider the moment ambiguity set in [5] and Wasserstein ambiguity set in [7]. If Ξ is compact and $\ell(u;\cdot)$ is continuous over Ξ for any $u \in U$, then Assumption 1.1(c)(d) hold for the above ambiguity sets (see [29, Remark 3 and Proposition 7]).

3.2 Convergence properties

In this subsection, we will show the convergence properties of discrete problem (1.5) under capacity. Consider $\hat{E}$ in (1.2). For any $A \in \mathcal{B}$, denote

$$\mathcal{V}(A) = \sup_{P\in\mathcal{P}} P(A). \tag{3.1}$$

The following assumption is added.

Assumption 3.4. *Consider capacity $\mathcal{V}$ defined in (3.1). For a finite number of sets $A_j \subset \Xi$ with $\mathcal{V}(A_j) = 1$, $j = 1, \cdots, J$, $\mathcal{V}(\bigcap_{j=1}^{J} A_j) = 1$.*

Remark 3.5. The above assumption holds under various settings of ambiguity set $\mathcal{P}$. When ambiguity set $\mathcal{P}$ is a singleton set, Assumption 3.4 holds obviously. In addition, denote $\mathcal{P}_{\mathcal{N}} = \{\mathcal{N}(\mu,\sigma) : \mu \in [\mu_1,\mu_2], \sigma \in [\sigma_1,\sigma_2]\}$, where $\mu_1,\mu_2,\sigma_1,\sigma_2 \in \mathbb{R}$. Note that the sets of measure 0 with respect to different probability distributions in $\mathcal{P}_{\mathcal{N}}$ are same, so Assumption 3.4 is satisfied automatically.

Denote

$$\Phi_{k,n}(u) := \frac{1}{n}\sum_{i=1}^{n} \ell(u;\xi^{i+(k-1)n}),\ \Phi_n(u) := \max_{1\leqslant k\leqslant m} \Phi_{k,n}(u).$$

Then problems (1.1) and (1.5) can be respectively recast as

$$\min_{u\in U}\ \Phi(u) \tag{3.2}$$

and

$$\min_{u\in U}\ \Phi_n(u). \tag{3.3}$$

Let v^* and v_n denote the optimal values of problems (3.2) and (3.3) respectively, and S^* and S_n denote the optimal solution sets of problems (3.2) and (3.3) respectively. We say that $\Phi_{k,n}(u)$ converges to $\Phi(u)$ w.p.1 under capacity $\mathcal{V}$ (3.1) uniformly on U as $n \to \infty$ if for any $\epsilon > 0$ and $\xi \in \tilde{\Xi}$ with $\tilde{\Xi} \subset \Xi$ and $\mathcal{V}(\tilde{\Xi}) = 1$, there exists $\bar{n}(\epsilon,\xi)$ such that for all $n \geqslant \bar{n}(\epsilon,\xi)$, $\max_{u\in U} |\Phi_{k,n}(u) - \Phi(u)| < \epsilon$.

Based on Theorem 2.14, we have the following lemma.

Lemma 3.6. *Let $\xi^1, \cdots, \xi^N$ be i.i.d. samples of ξ under sublinear expectation $\hat{E}$ in (1.2). Let Assumptions 1.1 and 3.4 hold. Assume that there exists $\alpha > 0$ such that $\hat{E}(|\ell(u;\xi)|^{1+\alpha}) < \infty$ for any*

$u \in U$. Then given any $1 \leqslant k \leqslant m$, $\Phi_{k,n}(u)$ converges to $\Phi(u)$ w.p.1 under capacity $\mathcal{V}$ (3.1) uniformly on U as $n \to \infty$, where $m > 0$ is fixed.

Proof. Given $\bar{u} \in U$ and a sequence ϵ_s of positive numbers converging to 0, define

$$V_s = \{u \in U : \|u - \bar{u}\| \leqslant \epsilon_s\},\ \delta_s(\xi) = \max_{u \in V_s} |\ell(u;\xi) - \ell(\bar{u};\xi)|.$$

Because of Assumption 1.1(a), we have $0 \leqslant \delta_s(\xi) \leqslant \kappa(\xi)\epsilon_s$ for $\xi \in \Xi$, which implies that

$$\lim_{s \to \infty} \hat{E}(\delta_s(\xi)) = 0. \tag{3.4}$$

Obviously, we have

$$\max_{u \in V_s} |\Phi_{k,n}(u) - \Phi_{k,n}(\bar{u})| \leqslant \frac{1}{n} \sum_{i=1}^{n} \delta_s(\xi^{i+(k-1)n}). \tag{3.5}$$

By virtue of Assumption 1.1 and Theorem 2.14, we know that the right-hand side of (3.5) converges to $\hat{E}(\delta_s(\xi))$ as $n \to \infty$ w.p.1 under capacity. Together with (3.4), we can find that for any $\epsilon > 0$, there exists a neighborhood W of $\bar{u}$ such that for sufficiently large n, and $\xi \in \tilde{\Xi}$ with $\tilde{\Xi} \subset \Xi$ and $\mathcal{V}(\tilde{\Xi}) = 1$, $\max\limits_{u \in W \bigcap U} |\Phi_{k,n}(u) - \Phi_{k,n}(\bar{u})| < \epsilon$. Since U is compact, there exists a finite number of points $u^1, \cdots, u^{\bar{m}}$ and corresponding neighborhoods $W^1, \cdots, W^{\bar{m}}$ covering U such that

$$\max_{u \in W^j \bigcap U} |\Phi_{k,n}(u) - \Phi_{k,n}(u^j)| < \epsilon,\ j = 1, \cdots, \bar{m}, \tag{3.6}$$

for sufficiently large n, and $\xi \in \Xi_j$ with $\mathcal{V}(\Xi_j) = 1$ respectively. According to the proof of Proposition 3.2, Φ is continuous on U, so these neighborhoods can be chosen such that

$$\max_{u \in W^j \bigcap U} |\Phi(u) - \Phi(u^j)| < \epsilon,\ j = 1, \cdots, \bar{m}. \tag{3.7}$$

Again by Theorem 2.14 and the hypothesis that $\hat{E}(|\ell(u;\xi)|^{1+\alpha}) < \infty$ for any $u \in U$, we have, for n large enough and $\xi \in \hat{\Xi}_j$ with $\mathcal{V}(\hat{\Xi}_j) = 1$ repectively,

$$|\Phi_{k,n}(u) - \Phi(u)| < \epsilon,\ j = 1, \cdots, \bar{m}. \tag{3.8}$$

Denote $\hat{\Xi} = \bigcap_{j=1}^{\bar{m}} (\Xi_j \bigcap \hat{\Xi}_j)$. Based on Assumption 3.4, we have $\mathcal{V}(\hat{\Xi}) = 1$. Combining (3.6), (3.7) and (3.8), for any given $\epsilon > 0$, we can prove that, for n large enough and $\xi \in \hat{\Xi}$, $\max\limits_{u \in U} |\Phi_{k,n}(u) - \Phi(u)| < 3\epsilon$. So we have completed the proof. □

Lemma 3.7. *Let assumptions of Lemma 3.6 hold. Then for any fixed $m > 0$, $\Phi_n(u)$ converges to $\Phi(u)$ w.p.1 under capacity $\mathcal{V}$ (3.1) uniformly on U as $n \to \infty$.*

Proof. According to Lemma 3.6, for any $\epsilon > 0$ and $1 \leqslant k \leqslant m$, we have $\max\limits_{u \in U} |\Phi_{k,n}(u) - \Phi(u)| < \epsilon$ for n large enough and $\xi \in \Xi_k$ with $\mathcal{V}(\Xi_k) = 1$ respectively, which implies that, for n large enough and $\xi \in \bigcap_{k=1}^{m} \Xi_k$,

$$\begin{aligned}
\max_{u \in U} |\Phi_n(u) - \Phi(u)| &\leqslant \max_{u \in U} |\max_{1 \leqslant k \leqslant m} \Phi_{k,n}(u) - \Phi(u)| \\
&\leqslant \max_{u \in U} \max_{1 \leqslant k \leqslant m} |\Phi_{k,n}(u) - \Phi(u)| \\
&= \max_{1 \leqslant k \leqslant m} \max_{u \in U} |\Phi_{k,n}(u) - \Phi(u)| \\
&< \epsilon.
\end{aligned}$$

So we have derived the conclusion. □

Theorem 3.8. *Let assumptions of Lemma 3.6 hold. Then for any fixed $m > 0$, $v_n \to v^*$ and $\mathbb{D}(S_n, S^*) \to 0$ w.p.1 under capacity $\mathcal{V}$ (3.1) as $n \to \infty$.*

Proof. Via Lemma 3.7, we know that $\Phi_n(u)$ converges to $\Phi(u)$ w.p.1 under capacity $\mathcal{V}$ uniformly on U. That is, for any given $\epsilon > 0$, we have $\max\limits_{u\in U}|\Phi_n(u) - \Phi(u)| < \epsilon$ for n large enough and $\xi \in \hat{\Xi}$ with $\mathcal{V}(\hat{\Xi}) = 1$. It follows that $|v_n - v^*| \leqslant \epsilon$ for n large enough and $\xi \in \hat{\Xi}$, which indicates that $v_n \to v^*$ w.p.1 under capacity $\mathcal{V}$ as $n \to \infty$.

We prove the other conclusion by contradiction. Suppose that $\mathbb{D}(S_n, S^*)$ does not converge to 0 w.p.1 under capacity as $n \to \infty$. Since U is compact, for some given $\epsilon > 0$, we can find a sequence of points $u^n \in S_n$ such that $\mathrm{dist}(u^n, S^*) \geqslant \epsilon$ for $n \geqslant 1$ and u^n tends to a point $u^* \in U$. We do not take a subsequence for the ease of statement. Due to the hypothesis, $u^* \notin S^*$, so we have $\Phi(u^*) > v^*$. Via [25, Proposition 5.1], we have $\lim\limits_{n\to\infty} \Phi_n(u^n) = \Phi(u^*) > v^*$. This is a contradiction since $v_n \to v$ w.p.1 under capacity $\mathcal{V}$ as $n \to \infty$ and $\Phi_n(u^n) = v_n$. So we have proved the conclusion. □

4 Statistical properties of the discrete problem (1.5)

In this section, we will explore the statistical properties of the discrete form (1.5). Let $\overline{\mu} = \hat{E}(\ell(u;\xi))$ and $\underline{\mu} = -\hat{E}(-\ell(u;\xi))$. The following proposition appeared in [10] without a proof. Here we attempt to prove the conclusion.

Proposition 4.1. *Let $\xi^1, \cdots, \xi^N$ be i.i.d. samples of ξ under sublinear expectation in (1.2). Suppose that Assumption 1.1 holds. Assume that $\hat{E}((\ell(u;\xi))^2) < \infty$ for any $u \in U$. Then for any fixed $m > 0$ and $u \in U$, we have*

$$\lim_{n\to\infty} \hat{E}\left(\max_{1\leqslant k\leqslant m} \frac{1}{n}\sum_{i=1}^{n} \ell(u;\xi^{i+(k-1)n})\right) = \hat{E}(\ell(u;\xi)). \tag{4.1}$$

Proof. The case that $m = 1$ follows from Theorem 2.11 readily, so we focus on the case where $m \geqslant 2$. Denote $\tilde{\varphi} : \mathbb{R} \to \mathbb{R}$ with $\tilde{\varphi}(v) = \max(\overline{\mu}, v)$ for $v \in \mathbb{R}$. We firstly prove that

$$\lim_{n\to\infty} \hat{E}\left(\tilde{\varphi}\left(\max_{1\leqslant k\leqslant m} \Phi_{k,n}(u)\right)\right) = \overline{\mu}. \tag{4.2}$$

Let $m = 2$. Then

$$\begin{aligned}
&\left|\hat{E}\left(\tilde{\varphi}\left(\max_{1\leqslant k\leqslant 2} \Phi_{k,n}(u)\right)\right) - \overline{\mu}\right| \\
=&\left|\hat{E}\left(\hat{E}\left(\tilde{\varphi}\left(\max\left(x, \Phi_{2,n}(u)\right)\right)\right)|_{x=\Phi_{1,n}(u)}\right) - \overline{\mu}\right| \\
\leqslant&\left|\hat{E}\left(\hat{E}\left(\tilde{\varphi}\left(\max\left(x, \Phi_{2,n}(u)\right)\right)\right)|_{x=\Phi_{1,n}(u)}\right) - \hat{E}\left(\max_{\theta\in[\underline{\mu},\overline{\mu}]}\left(\tilde{\varphi}\left(\max(x,\theta)\right)\right)|_{x=\Phi_{1,n}(u)}\right)\right| \\
&\qquad + \left|\hat{E}\left(\max_{\theta\in[\underline{\mu},\overline{\mu}]}\left(\tilde{\varphi}\left(\max(x,\theta)\right)\right)|_{x=\Phi_{1,n}(u)}\right) - \overline{\mu}\right| \\
\leqslant&\hat{E}\left(\left|\hat{E}\left(\tilde{\varphi}\left(\max\left(x, \Phi_{2,n}(u)\right)\right)\right) - \max_{\theta\in[\underline{\mu},\overline{\mu}]}\left(\tilde{\varphi}\left(\max(x,\theta)\right)\right)\right|_{x=\Phi_{1,n}(u)}\right) + \left|\hat{E}\left(\tilde{\varphi}\left(\Phi_{1,n}(u)\right)\right) - \overline{\mu}\right| \\
\leqslant&\hat{E}\left(\left|C_1 n^{-\frac{1}{2}}\right|_{x=\Phi_{1,n_1}(u)}\right) + C_2 n^{-\frac{1}{2}} \\
=&(C_1 + C_2)n^{-\frac{1}{2}},
\end{aligned}$$

where the first equality is due to Definition 2.4, the second inequality is due to Definition 2.1, the third inequality is due to Proposition 2.18, and C_1, C_2 are constants depending only on $\hat{E}((\ell(u;\xi))^2)$. So (4.2)

holds for $m=2$. Now we assume that (4.2) holds for $m-1$. Then

$$\begin{aligned}
&\left|\hat{E}\left(\tilde{\varphi}\left(\max_{1\leqslant k\leqslant m}\Phi_{k,n}(u)\right)\right)-\overline{\mu}\right| \\
\leqslant&\hat{E}\left(\left|\hat{E}\left(\tilde{\varphi}\left(\max\left(x,\Phi_{m,n}(u)\right)\right)\right)-\max_{\theta\in[\underline{\mu},\overline{\mu}]}\left(\tilde{\varphi}\left(\max(x,\theta)\right)\right)\right|_{x=\max\limits_{1\leqslant k\leqslant m-1}\Phi_{k,n}(u)}\right) \\
&+\left|\hat{E}\left(\max_{\theta\in[\underline{\mu},\overline{\mu}]}\left(\tilde{\varphi}\left(\max(x,\theta)\right)\right)|_{x=\max\limits_{1\leqslant k\leqslant m-1}\Phi_{k,n}(u)}\right)-\overline{\mu}\right| \\
=&C_3n^{-\frac{1}{2}}+\left|\hat{E}\left(\tilde{\varphi}\left(\max_{1\leqslant k\leqslant m-1}\Phi_{k,n}(u)\right)\right)-\overline{\mu}\right|,
\end{aligned}$$

where C_3 is a constant depending only on $\hat{E}((\ell(u;\xi))^2)$. From the inductive assumption, (4.2) holds. Similar to above deductions, we have

$$\begin{aligned}
&\left|\hat{E}\left(\max_{1\leqslant k\leqslant m}\Phi_{k,n}(u)\right)-\overline{\mu}\right| \\
=&\left|\hat{E}\left(\hat{E}\left(\max\left(x,\Phi_{m,n}(u)\right)\right)|_{x=\max\limits_{1\leqslant k\leqslant m-1}\Phi_{k,n}(u)}\right)-\hat{E}\left(\max_{\theta\in[\underline{\mu},\overline{\mu}]}\left(\max(x,\theta)\right)|_{x=\max\limits_{1\leqslant k\leqslant m-1}\Phi_{k,n}(u)}\right)\right. \\
&\left.+\hat{E}\left(\max_{\theta\in[\underline{\mu},\overline{\mu}]}\left(\max(x,\theta)\right)|_{x=\max\limits_{1\leqslant k\leqslant m-1}\Phi_{k,n}(u)}\right)-\overline{\mu}\right| \\
\leqslant&\hat{E}\left(\left|\hat{E}\left(\max\left(x,\Phi_{m,n}(u)\right)\right)-\max_{\theta\in[\underline{\mu},\overline{\mu}]}\left(\max(x,\theta)\right)\right|_{x=\max\limits_{1\leqslant k\leqslant m-1}\Phi_{k,n}(u)}\right) \\
&+\left|\hat{E}\left(\tilde{\varphi}\left(\max_{1\leqslant k\leqslant m-1}\Phi_{k,n}(u)\right)\right)-\overline{\mu}\right| \\
=&C_4n^{-\frac{1}{2}}+\left|\hat{E}\left(\tilde{\varphi}\left(\max_{1\leqslant k\leqslant m-1}\Phi_{k,n}(u)\right)\right)-\overline{\mu}\right|,
\end{aligned}$$

where C_4 is a constant depending only on $\hat{E}((\ell(u;\xi))^2)$. Together with (4.2), we can prove the desirable results. □

The following proposition follows from Proposition 2.16.

Proposition 4.2. *Let $\xi^1,\cdots,\xi^N$ be i.i.d. samples of ξ under sublinear expectation $\hat{E}$ in (1.2). Let Assumption 1.1 hold. Assume that Ξ is compact. Then there exists constant $\tilde{C}>0$ not related to u such that*

$$\hat{E}(((\Phi_n(u)-\overline{\mu})_+)^2)\leqslant\frac{\tilde{C}m}{n},\ \hat{E}(((\Phi_n(u)-\underline{\mu})_-)^2)\leqslant\frac{\tilde{C}m}{n}, \tag{4.3}$$

for any $u\in U$ and $n>0$.

Proof. Let $\tilde{\sigma}^2=\max\limits_{P\in\mathcal{P}}E_P((\ell(u;\xi)-E_P(\ell(u;\xi)))^2)$. Now we prove that $\overline{\mu}$, $\underline{\mu}$ and $\tilde{\sigma}$ are bounded for $u\in U$. Actually, according to [29, Proposition 2], we can obtain that $\hat{E}(-\ell(\cdot;\xi))$ and $\hat{E}((\ell(\cdot;\xi))^2)$ are continuous over U. Noting that U is compact, we know that $\hat{E}(\ell(u;\xi))$, $\hat{E}(-\ell(u;\xi))$ and $\hat{E}((\ell(u;\xi))^2)$ are bounded for $u\in U$. Furthermore, since

$$\begin{aligned}
\max_{P\in\mathcal{P}}E_P((\ell(u;\xi)-E_P(\ell(u;\xi)))^2)&=\max_{P\in\mathcal{P}}\left(E_P((\ell(u;\xi))^2)-(E_P(\ell(u;\xi)))^2\right) \\
&\leqslant\max_{P\in\mathcal{P}}E_P((\ell(u;\xi))^2)+\max_{P\in\mathcal{P}}(E_P(\ell(u;\xi)))^2 \\
&=\hat{E}((\ell(u;\xi))^2)+(\hat{E}(\ell(u;\xi)))^2,
\end{aligned}$$

we know that $\tilde{\sigma}$ is bounded for $u \in U$. Let $\tilde{C}$ be a positive constant such that $\tilde{C} \geqslant \max\limits_{u\in U} 2(\tilde{\sigma}^2 + (\overline{\mu} - \underline{\mu})^2)$. Due to Proposition 2.16, we know that (4.3) holds. □

From Proposition 4.2, it is very likely that estimator $\Phi_n(u)$ is concentrated inside $[\underline{\mu}, \overline{\mu}]$ if $\lim\limits_{n\to\infty} \frac{m}{n} = 0$, where $m > 0$ is fixed.

5 Quantification of difference between the problems (1.5) and (1.1)

In this section, we investigate the quantification of difference between the problem (1.5) and the problem (1.1) with special moment ambiguity set. Let $\mathcal{P}(\Xi)$ be the set of probability distributions over Ξ. In the following, ambiguity set $\mathcal{P}$ is assumed to be $\mathcal{P} = \{P \in \mathcal{P}(\Xi) : |E_P\Psi(\xi)| \leqslant \gamma\}$, where $\Psi : \Xi \to \mathbb{R}^\nu$ is $\mathcal{B}$-measurable, and $\gamma \in \mathbb{R}^\nu$ is a positive vector. The following assumption is needed, where the third assumption is the Slater type assumption (see [11, 29]).

Assumption 5.1. *(i) Ψ is Lipschitz continuous.*

(ii) Ξ is a compact set.

(iii) There exists $P_0 \in \mathcal{P}(\Xi)$ and $\alpha > 0$ such that $|E_{P_0}(\Psi(\xi)) + \alpha b| \leqslant \gamma$, where $b \in \mathbb{B}$ with $\mathbb{B} \subset \mathbb{R}^\nu$ being the closed unit ball.

We consider the following problem:

$$\min_{u\in U} \max_{1\leqslant k\leqslant m, q\in\mathcal{P}^{k,n}} \sum_{i=1}^{n} q_i \ell(u; \xi^{i+(k-1)n}), \tag{5.1}$$

where $\mathcal{P}^{k,n} := \left\{q \in \mathbb{R}^n : q \geqslant 0, \sum\limits_{i=1}^{n} q_i = 1, \left|\sum\limits_{i=1}^{n} q_i \Psi(\xi^{i+(k-1)n})\right| \leqslant \gamma\right\}$, $k = 1, \cdots, m$. Problem (5.1) is extended from problem (1.5) by considering more discrete probability distributions.

Denote

$$\mathcal{P}_m^N := \left\{p \in \mathbb{R}^N : p \geqslant 0, \sum_{i=1}^{N} p_i = 1, \left|\sum_{i=1}^{N} p_i \Psi(\xi^i)\right| \leqslant \gamma, \text{there is } k \text{ so that} \sum_{i=1}^{n} p_{i+(k-1)n} = 1\right\}.$$

Then we have the following proposition.

Proposition 5.2. *For any $u \in \mathbb{R}^s$, we have*

$$\max_{1\leqslant k\leqslant m, q\in\mathcal{P}^{k,n}} \sum_{i=1}^{n} q_i \ell(u; \xi^{i+(k-1)n}) = \max_{p\in\mathcal{P}_m^N} \sum_{i=1}^{N} p_i \ell(u; \xi^i).$$

Proof. Given $q \in \mathcal{P}^{k,n}$, let p be an N-dimensional vector with $p_{i+(k-1)n} = q_i$ for $i = 1, \cdots, n$ and the other components being 0. Obviously, $p \in \mathcal{P}_m^N$. So, for any fixed u, $\max\limits_{1\leqslant k\leqslant m, q\in\mathcal{P}^{k,n}} \sum\limits_{i=1}^{n} q_i \ell(u; \xi^{i+(k-1)n}) \leqslant \max\limits_{P\in\mathcal{P}_m^N} \sum\limits_{i=1}^{N} p_i \ell(u; \xi^i)$. On the other hand, given $p \in \mathcal{P}_m^N$, there must exist k such that $(p_{1+(k-1)n}, \cdots, p_{kn})^\top \in \mathcal{P}^{k,n}$, which indicates $\max\limits_{1\leqslant k\leqslant m, q\in\mathcal{P}^{k,n}} \sum\limits_{i=1}^{n} q_i \ell(u; \xi^{i+(k-1)n}) \geqslant \max\limits_{P\in\mathcal{P}_m^N} \sum\limits_{i=1}^{N} p_i \ell(u; \xi^i)$, for any fixed u. Thus the conclusion can be obtained from above two inequalities. □

From above proposition, problem (5.1) can be recast as the following problem:

$$\min_{u\in U} \max_{p\in\mathcal{P}_m^N} E_p(\ell(u; \xi)), \tag{5.2}$$

where $E_p(\ell(u; \xi)) = \sum\limits_{i=1}^{N} p_i \ell(u; \xi^i)$. Now we investigate the quantification of difference between $\mathcal{P}$ and $\mathcal{P}_m^N$.

Let $\hat{\mathcal{H}} = \{h : \Xi \to \mathbb{R} : |h(\xi^1) - h(\xi^2)| \leqslant \|\xi^1 - \xi^2\|\}$. For $Q_1, Q_2 \in \mathcal{P}$, denote the Wasserstein metric by $D_W(Q_1, Q_2) = \sup_{h\in\hat{\mathcal{H}}} |E_{Q_1}(h(\xi)) - E_{Q_2}(h(\xi))|$. For two sets $\mathcal{Q}_1, \mathcal{Q}_2 \subset \mathcal{P}$, denote $D_W(\mathcal{Q}_1, \mathcal{Q}_2) = \sup_{Q^1\in\mathcal{Q}_1} \inf_{Q^2\in\mathcal{Q}_2} D_W(Q^1, Q^2)$. More infomation about the Wasserstein metric between probability measures can be found in [11]. Let $\beta_{k,n} := \max_{\xi\in\Xi} \min_{1\leqslant i\leqslant n} \|\xi - \xi^{i+(k-1)n}\|$, $k = 1, \cdots, m$, where $\beta_{k,n}$ is indeed the Hausdorff distance between Ξ and discrete sample space $\{\xi^{i+(k-1)n}, i = 1, \cdots, n\}$. The following proposition is [11, Proposition 2].

Proposition 5.3. *([11]) Let $m > 0$ be fixed. Suppose that Assumptions 1.1 and 5.1 hold and $\beta_{k,n} \to 0$ when $n \to \infty$ for $k = 1, \cdots, m$. Then, for sufficiently large n, we have*

$$D_W(\mathcal{P}, \mathcal{P}^{k,n}) \leqslant L_k \beta_{k,n}, \tag{5.3}$$

for some $L_k > 0$, $k = 1, \cdots, m$.

Remark 5.4. The assumption that $\beta_{k,n} \to 0$ when $n \to \infty$ can be found in [11, 16, 29], which holds in many cases. For example, it is shown in [16, Proposition 9] that the above assumption holds true if Ξ is bounded, the true probability distribution of ξ denoted by P is continuous, and there exist positive constants τ, ν, and δ_0 such that $P(\|\xi - \xi^0\| \leqslant \delta) > \tau\delta^\nu$, for any fixed point $\xi^0 \in \Xi$ and $\delta \in (0, \delta_0)$.

Via Proposition 5.3, we have the following theorem.

Theorem 5.5. *Let assumptions of Proposition 5.3 hold. Then, for sufficiently large n, we have*

$$D_W(\mathcal{P}, \mathcal{P}_m^N) \leqslant L \min_{1\leqslant k\leqslant m} \beta_{k,n}, \tag{5.4}$$

for some $L > 0$.

Proof. Given $\tilde{p} \in \mathcal{P}$ and $p \in \mathcal{P}_m^N$, there exists k such that $q = (p_{1+(k-1)n}, \cdots, p_{kn})^\top \in \mathcal{P}^{k,n}$. Then

$$\begin{aligned}
D_W(\tilde{p}, p) &= \sup_{h\in\mathcal{H}} |E_{\tilde{p}}(h(\xi)) - E_p(h(\xi))| \\
&= \sup_{h\in\mathcal{H}} |E_{\tilde{p}}(h(\xi)) - \sum_{i=1}^N p_i h(\xi^i)| \\
&= \sup_{h\in\mathcal{H}} |E_{\tilde{p}}(h(\xi)) - \sum_{i=1}^n p_{i+n(k-1)} h(\xi^{i+n(k-1)})| \\
&= \sup_{h\in\mathcal{H}} |E_{\tilde{p}}(h(\xi)) - \sum_{i=1}^n q_i h(\xi^{i+n(k-1)})| \\
&= \sup_{h\in\mathcal{H}} |E_{\tilde{p}}(h(\xi)) - E_q(h(\xi))| \\
&= D_W(\tilde{p}, q).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
D_W(\mathcal{P}, \mathcal{P}_m^N) &= \sup_{p^1\in\mathcal{P}} \inf_{p^2\in\mathcal{P}_m^N} D_W(p^1, p^2) \\
&= \sup_{p^1\in\mathcal{P}} \min_{1\leqslant k\leqslant m} \inf_{q\in\mathcal{P}^{k,n}} D_W(p^1, q) \\
&\leqslant \min_{1\leqslant k\leqslant m} \sup_{p^1\in\mathcal{P}} \inf_{q\in\mathcal{P}^{k,n}} D_W(p^1, q) \\
&\leqslant \min_{1\leqslant k\leqslant m} D_W(\mathcal{P}, \mathcal{P}^{k,n}) \\
&\leqslant L \min_{1\leqslant k\leqslant m} \beta_{k,n},
\end{aligned}$$

where the second equality can be proved via the same way in the proof of Proposition 5.2, and the last

inequality follows from Proposition 5.3 with $L := \max\limits_{1\leqslant k\leqslant m} L_k$. □

Let v and $\tilde{v}_n$ denote the optimal value of the problems (1.1) and (5.1) respectively, and S^* and $\tilde{S}_n$ denote the optimal solution set of the problems (1.1) and (5.1) respectively. The following theorem follows from [11, Theorem 4]. The proof is similar and thus is omitted.

Theorem 5.6. *Let assumptions of Proposition 5.3 hold. Then there exists constant $\tilde{L} > 0$ such that $|v - \tilde{v}_n| \leqslant \tilde{L} D_W(\mathcal{P}, \mathcal{P}_m^N)$ and $\mathbb{D}(\tilde{S}_n, S^*) \leqslant \mathcal{T}(2\tilde{L} D_W(\mathcal{P}, \mathcal{P}_m^N))$, where function $\mathcal{T} : \mathbb{R}_+ \to \mathbb{R}_+$ is defined as $\mathcal{T}(t) = \sup\{\tau \in \mathbb{R}_+ : \mathcal{R}(\tau) \leqslant t\}$ with $\mathcal{R} : \mathbb{R}_+ \to \mathbb{R}_+$ defined as $\mathcal{R}(\tau) = \min\left\{\hat{E}(\ell(u;\xi)) - v : \operatorname{dist}(u, S^*) \geqslant \tau, u \in U\right\}$.*

6 Numerical Experiments

In this section, we apply sublinear expectation method to solve DRO problem (1.1) with two real life data sets involving portfolio optimization and recurrent neural networks (RNNs), and the results will be compared with those obtained via duality methods in [5]. The numerical results are obtained by using Matlab R2021b on a desktop (Windows 11 with 2.30 GHz Inter Core i7-12700H CPU and 16GB RAM).

6.1 DRO in portfolio optimization

Given that s investment options are available, $\xi \in \mathbb{R}^s$ is a random vector of return rates for the different options, a portfolio is encoded by a vector of weights $u \in \mathbb{R}^s$ ranging over the simplex $U = \{u \in \mathbb{R}_+^s : e_s^\top u = 1\}$ with all elements of $e_s \in \mathbb{R}^s$ equal to 1. Following [5, Section 4.1], the utility of portfolio can be denoted as $h(u^\top \xi)$, where $h(x) = \min\limits_{1\leqslant k\leqslant K} a_k v + b_k$ with $a_k \geqslant 0$, $k = 1, \cdots, K$, for $v \in \mathbb{R}$. Let $\hat{\mu}$ and $\hat{\Sigma}$ be estimates of the true mean and variance of ξ, respectively. Then the DRO problem for portfolio optimization can be formulated as

$$\min_{u\in U} \max_{P\in\mathcal{P}} \; E_P(-h(u^\top \xi)), \tag{6.1}$$

where

$$\mathcal{P} = \left\{ P \in \mathcal{P}(\Xi) : \begin{array}{c} E_P(\xi - \hat{\mu})^\top \hat{\Sigma}^{-1} E_P(\xi - \hat{\mu}) \leqslant \gamma_1, \\ E_P((\xi - \hat{\mu})(\xi - \hat{\mu})^\top) \preceq \gamma_2 \hat{\Sigma} \end{array} \right\}, \tag{6.2}$$

with $\gamma_1, \gamma_2 \geqslant 0$ and $\Xi = \{\xi : \|\xi\|^2 \leqslant 1\}$. From the structures of U and h, it is not difficult to see that Assumption 1.1(a)(b) hold. Furthermore, note that Ξ is compact and h is continuous with respect to ξ for any fixed $u \in U$. According to [29, Remark 3 and Proposition 7], Assumption 1.1(c)(d) are satisfied. Thus the sublinear expectation can be empolyed to solve problem (6.1). The corresponding discrete form under sublinear expectation is

$$\min_{u\in U} \max_{1\leqslant k\leqslant m} \; \frac{1}{n} \sum_{i=1}^{n} -h(u^\top \xi^{i+n(k-1)}). \tag{6.3}$$

As a further example, we consider the 30 assets composing the Dow Jones Industrial Average Index (DJI) and collect 280 historical daily return rate scenarios from August 22, 2023 to October 1, 2024. Based on the experimental procedure in [5], we divide the entire data into investing periods of length 5 days. That is, we have 56 periods of data, and each period of data is denoted as $G_t = \{\xi^{1+5(t-1)}, \xi^{2+5(t-1)}, \cdots, \xi^{5t}\}$, $t = 1, \cdots, 56$. At any given period, we use 6 periods of data (30 days) from the most recent history to assign the portfolio. Specifically, to obtain a portfolio for given period of data G_t, we solve problem (6.1) via data from $G_{t-5} \cup G_{t-4} \cup G_{t-3} \cup G_{t-2} \cup G_{t-1}$. Then the average utility for data G_t is $\frac{1}{5} \sum\limits_{i=1}^{5} -h(u^\top \xi^{i+5(t-1)})$. Since the sample size is 280, we can obtain utilities for 50 periods of data.

From [19, Table 4.1], we set $K = 10$ in utility function h, and denote the parameters as follows:

Table 1 Parameters of the utility function

k	1	2	3	4	5
a_k	1.3521	1.1070	0.8848	0.6891	0.5367
b_k	0.0002	0	0	0.0002	0.0006
k	6	7	8	9	10
a_k	0.4179	0.3178	0.2355	0.1626	0.1037
b_k	0.0011	0.0016	0.0021	0.0027	0.0033

For ambiguity set defined as in (6.2), we let $\gamma_1 = 0.1$ and $\gamma_2 = 1.1$, and let $\hat{\mu}$ and $\hat{\Sigma}$ be the mean and covariance matrix calculated through historical data. In discrete approximation (6.3), we consider four combinations of m and n: $m = 1, n = 30$; $m = 2, n = 15$; $m = 3, n = 10$; $m = 6, n = 5$. We use CVX solver to solve problem (6.3). We compare our approach to the one proposed in [5] (denoted by "DY Model").

We conduct two types of experiments which differ in the number of assets.

- We consider all 30 assets in DJI. The numerical results are presented in Fig. 1 and Table 2.
- We consider 10 assets in DJI, which include BA, CAT, JNJ, MMM, MRK, MSFT, PG, SHW, TRV, UNH. The numerical results are presented in Fig. 2 and Table 3.

In the tables, we list the mean, 25% quantile, 75% quantile and the median over utilities of 50 periods of data for each setting. In the figures, we exhibit the means and the quantiles respectively. Moreover, in the right figures of Fig. 1 and 2, the central mark indicates the median, and the bottom and top edges of the box indicate the 25% and 75% quantiles, respectively.

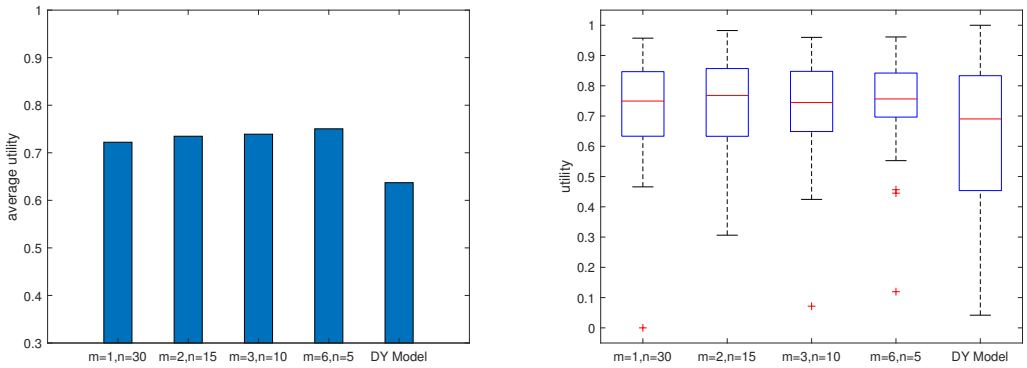

Figure 1 (Color online) Numerical results for 30 assets

Table 2 Numerical results for 30 assets

(m,n)	(1,30)	(2,15)	(3,10)	(6,5)	DY Model
mean	0.7222	0.7348	0.7391	0.7504	0.6371
25% quantile	0.6333	0.6330	0.6488	0.6965	0.4537
median	0.7495	0.7680	0.7446	0.7565	0.6904
75% quantile	0.8467	0.8569	0.8477	0.8420	0.8333

Compared with DY model in [5], the sublinear expectation method with $m > 1$ achieves the largest utility and most reliable numerical performances.

6.2 DRO for training RNNs

Denote a series of inputs by $\{x_t, t = 1, \cdots, T\} \subset \mathbb{R}^{N_0}$ and outputs by $\{y_t, t = 1, \cdots, T\} \subset \mathbb{R}^{N_1}$. Let $W \in \mathbb{R}^{N_2 \times N_2}$, $U \in \mathbb{R}^{N_2 \times N_0}$, $V \in \mathbb{R}^{N_1 \times N_2}$, $b \in \mathbb{R}^{N_2}$ and $c \in \mathbb{R}^{N_1}$. Denote $x = ((x_1)^\top, \cdots, (x_T)^\top)^\top$, $y =$

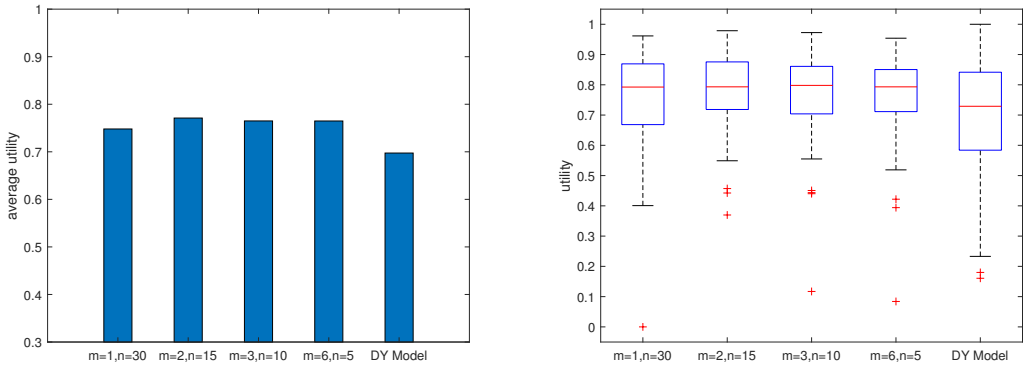

Figure 2 (Color online) Numerical results for 10 assets

Table 3 Numerical results for 10 assets

(m,n)	(1,30)	(2,15)	(3,10)	(6,5)	DY Model
mean	0.7480	0.7710	0.7649	0.7648	0.6974
25% quantile	0.6684	0.7184	0.7039	0.7114	0.5839
median	0.7925	0.7933	0.7981	0.7932	0.7290
75% quantile	0.8692	0.8756	0.8609	0.8504	0.8416

$((y_1)^\top, \cdots, (y_T)^\top)^\top$, $\xi = (x^\top, y^\top)^\top, u = (\mathrm{vec}(W)^T, \mathrm{vec}(U)^T, \mathrm{vec}(V)^T, b^\top, c^\top)^\top$, where vec denotes the column expansion of the corresponding matrix. Then the loss function for training RNNs is denoted as follows:

$$l(u;\xi) = \frac{1}{T}\sum_{t=1}^{T} \|y_t - (V\sigma(W(\cdots\sigma(Ux_1+b)\cdots)+Ux_t+b)+c)\|^2, \tag{6.4}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable activation function.

Motivated by [2,24], we employ the DRO problem to train RNNs. Consider the ambiguity set proposed in [7] via the Wasserstein metric, which is defined as $\mathcal{P} = \{P \in \mathcal{P}(\Xi) : D_W(P, \tilde{P}) \leqslant \epsilon\}$. Here $\tilde{P}$ is the nominal probability distribution. Let $s = N_2^2 + N_0N_2 + N_1N_2 + N_1 + N_2$, $r = T(N_0 + N_1)$, and $U \subset \mathbb{R}^s$ and $\Xi \subset \mathbb{R}^r$ be compact sets.

The DRO problem for training RNNs can be formulated as

$$\min_{u\in U}\max_{P\in\mathcal{P}} E_P(\ell(u;\xi)). \tag{6.5}$$

Note that U, Ξ are compact and $\ell(u;\cdot)$ is continuous for any fixed $u \in U$. Using same arguments in the last subsection, we can show that Assumption 1.1 also holds for problem (6.5). Then the corresponding discrete form under sublinear expectation is

$$\min_{u\in U}\max_{1\leqslant k\leqslant m} \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} \|y_t^{i+n(k-1)} - \left(V\sigma\left(W\left(\cdots\sigma\left(Ux_1^{i+n(k-1)}+b\right)\cdots\right)+Ux_t^{i+n(k-1)}+b\right)+c\right)\|^2, \tag{6.6}$$

where x_t^i, y_t^i denote the t-th input and output in the i-th scenario respectively.

Consider the real life data like Volatility of S&P Index [30]. We collect 435 scenarios in this data set from February 1973 to April 2009, where the monthly realized volatility of S&P index is appointed as the output variable, and 11 exogenous variables are input variables. For training the RNNs, we first standardize the dataset as zero mean and unit variance. We let $N_0 = 11$, $N_1 = 1$ and $N_2 = 20$. We also set $T = 3$, so there are 145 groups of scenarios, where we allocate 120 groups as the training set, and the remaining 25 groups are allocated as the test set. The *tanh* function is set as the activation function, that is, $\sigma(w) = \frac{e^w - e^{-w}}{e^w + e^{-w}}$, for $w \in \mathbb{R}$. We set $U = [-10^3 * e_s, 10^3 * e_s]$ and $\Xi = \{\xi : \|\xi\|^2 \leqslant 100\}$. The

training error is denoted as

$$\text{TrainErr} := \frac{1}{360}\sum_{i=1}^{120}\sum_{t=1}^{3}\|y_t^i - \left(V\sigma\left(W\left(\cdots\sigma\left(Ux_1^i+b\right)\cdots\right)+Ux_t^i+b\right)+c\right)\|^2.$$

The testing error is denoted as

$$\text{TestErr} := \frac{1}{75}\sum_{i=121}^{145}\sum_{t=1}^{3}\|y_t^i - \left(V\sigma\left(W\left(\cdots\sigma\left(Ux_1^i+b\right)\cdots\right)+Ux_t^i+b\right)+c\right)\|^2.$$

Note that $\Phi_{k,n}(\cdot) = \frac{1}{n}\sum_{i=1}^{n}\ell(\cdot;\xi^{i+(k-1)n})$, $k = 1,\cdots,m$, is continuously differentiable, whereas $\Phi_n(\cdot) = \max_{1\leqslant k\leqslant m}\Phi_{k,n}(\cdot)$ is nonsmooth when $m \geqslant 2$. Thus we employ the exponential smoothing function proposed in [3] for Φ_n, which is defined as $\Phi_n(u,\mu) = \mu\ln\left(\sum_{k=1}^{m}e^{\Phi_{k,n}(u)/\mu}\right)$, for $\mu > 0$. Via smoothing function $\Phi_n(\cdot,\mu)$, problem (6.6) can be solved via smoothing projected gradient method in [31]. As far as we know, problem (6.5) has not been investigated before, so we only present the numerical results of our method based on (6.6).

For the smoothing projected gradient method, the initial points are generated 100 times from the uniform distribution over $[0,1]$, and the numerical results are exhibited as follows.

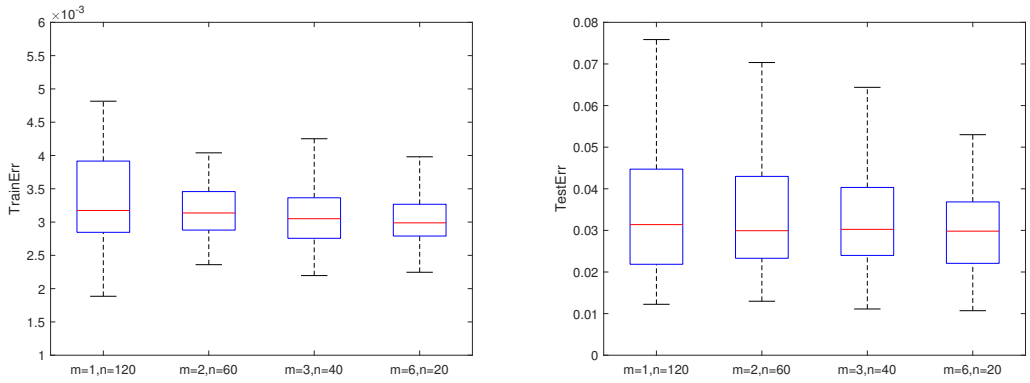

Figure 3 (Color online) Numerical results

Table 4 Numerical results

	(m,n)	(1,120)	(2,60)	(3,40)	(6,20)
TrainErr	mean	0.0039	0.0037	0.0035	0.0034
	25% quantile	0.0028	0.0029	0.0028	0.0028
	median	0.0032	0.0031	0.0031	0.0030
	75% quantile	0.0039	0.0035	0.0034	0.0033
TestErr	mean	0.0362	0.0352	0.0355	0.0343
	25% quantile	0.0219	0.0233	0.0240	0.0221
	median	0.0314	0.0299	0.0302	0.0298
	75% quantile	0.0447	0.0430	0.0403	0.0368

From the numerical results, we can see that the sublinear expectation method based on (6.6) with $m = 6$ achieves the lowest average training error and testing error, as well as the most reliable numerical performances.

7 Conclusion

Sublinear expectation theory proposed by Peng [20, 21] is aimed to deal with the probability model uncertainty, and has been proved to be very effective in finance and statistics. Considering the serious limitations of the existing methods for solving DRO problem (1.1), we attempt to employ the sublinear expectation to solve DRO problems. In particular, the DRO problem is recast as a minimization problem using the sublinear expectation, and a discrete approximation based on grouped samples is introduced. We have shown that optimal values and optimal solutions of the discrete problem converge to those of the DRO problem w.p.1 under capacity (3.1). We have also proved that the discrete form is an asymptotic unbiased estimator for $\hat{E}(\ell(u;\xi))$, and presented the quantification of difference between problem (1.5) and problem (1.1) with special moment ambiguity set. From the numerical results of two real life data sets, the sublinear expectation method performs better than the existing duality method in [5], especially in terms of reliability.

Acknowledgements This work is supported by the Post-Doctoral Fellowship of CAS AMSS-PolyU Joint Laboratory in Applied Mathematics and the Hong Kong Research Grant Council PolyU15300123, 15300124. We would like to thank Professor Shige Peng, Professor Zengjing Chen and Professor Yongsheng Song for their helpful discussions and comments.

References

1 Ben-Tal A, Den Hertog D, De Waegenaere A, et al. Robust solutions of optimization problems affected by uncertain probabilities. Manage Sci, 2013, 59: 341–357
2 Chen R, Paschalidis I C. A robust learning approach for regression models based on distributionally robust optimization. J Mach Learn Res, 2018, 19: 1–48
3 Chen X. Smoothing methods for nonsmooth, nonconvex minimization. Math Program, 2012, 134: 71–99
4 Chen Z. Strong laws of large numbers for sub-linear expectations. Sci China Math, 2016, 59: 945–954
5 Delage E, Ye Y. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Oper Res, 2010, 58: 595–612
6 Epstein L G, Ji S. Ambiguous volatility and asset pricing in continuous time. Rev Financ Stud, 2013, 26: 1740–1786
7 Esfahani P M, Kuhn D. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. Math Program, 2018, 171: 115–166
8 Fang X, Peng S, Shao Q, et al. Limit theorems with rate of convergence under sublinear expectations. Bernoulli, 2019, 25: 2564–2596
9 Ji X, Peng S. Spatial and temporal white noises under sublinear G-expectation. Sci China Math, 2020, 63: 61–82
10 Jin H, Peng S. Optimal unbiased estimation for maximal distribution. arXiv:1611.07994, 2016
11 Jiang J, Chen X. Pure characteristics demand models and distributionally robust mathematical programs with stochastic complementarity constraints. Math Program, 2023, 198: 1449–1484
12 Kuhn D, Esfahani P M, Nguyen V A, et al. Wasserstein distributionally robust optimization: Theory and applications in machine learning. Oper Res & Manage Sci in the Age of Analytics(INFORMS), 2019, 130–166
13 Kuhn D, Shafiee S, Wiesemann W. Distributionally robust optimization. arXiv:2411.02549, 2025
14 Li H, Peng S, Soumana Hima A. Reflected solutions of backward stochastic differential equations driven by G-Brownian motion. Sci China Math, 2018, 61: 1–26
15 Lin L, Liu Y, Lin C. Mini-max-risk and mini-mean-risk inferences for a partially piecewise regression. Stat, 2017, 51: 745–765
16 Liu Y C, Pichler A, Xu H F. Discrete approximation and quantification in distributionally robust optimization. Math Oper Res, 2019, 44: 19–37
17 Maccheroni F, Marinacci M. A strong law of large numbers for capacities. Ann Probab, 2005, 33: 1171–1178
18 Marinacci M. Limit laws for non-additive probabilities and their frequentist interpretation. J Econ Theory, 1999, 84: 145–195
19 Natarajan K, Sim M, Uichanco J. Tractable robust expected utility and risk models for portfolio optimization. Math Financ, 2010, 20: 695–731
20 Peng S. Survey on normal distributions, central limit theorem, Brownian motion and the related stochastic calculus under sublinear expectations. Sci China Ser A: Math, 2009, 52: 1391–1411
21 Peng S. Nonlinear Expectations and Stochastic Calculus under Uncertainty. Berlin: Springer, 2019
22 Peng S, Yang S, Yao J. Improving value-at-risk prediction under model uncertainty. J Financ Econ, 2023, 21: 228–259

23 Peng S, Zhang H. Stochastic calculus with respect to G-Brownian motion viewed through rough paths. Sci China Math, 2017, 60: 1–20
24 Shafieezadeh-Abadeh S, Kuhn D, Esfahani P M. Regularization via mass transportation. J Mach Learn Res, 2019, 20: 1–68
25 Shapiro A, Dentcheva D, Ruszczynski A. Lectures on Stochastic Programming: Modeling and Theory. Philadelphia: SIAM, 2009
26 Song Y. Some properties on G-evaluation and its applications to G-martingale decomposition. Sci China Math, 2011, 54: 287—300
27 Song Y. Stein's method for law of large numbers under sublinear expectations. Probab Uncertaint Quant Risk, 2021, 6: 199–212
28 Song Y. A strong law of large numbers under sublinear expectations. Probab Uncertain Quant Risk, 2023, 8: 333–350
29 Sun H, Xu H. Convergence analysis for distributionally robust optimization and equilibrium problems. Math Oper Res, 2016, 41: 377–401
30 Wang Y, Zhang C, Chen X. An augmented Lagrangian method for training recurrent neural networks. SIAM J Sci Comput, 2025, 47: C22–C51
31 Zhang C, Chen X. Smoothing projected gradient method and its application to stochastic linear complementarity problems. SIAM J Optim, 2009, 20: 627–649