Rates of convergence of sample average approximation under heavy tailed distributions

Jie Jiang [†*]   Zhiping Chen [‡]   Xinmin Yang [§]

Abstract

In this paper, we consider the rate of convergence of the sample average approximation (SAA) under heavy tailed distributions and quantify it under both independent identically distributed (iid) sampling and non-iid sampling. We first develop the uniform polynomial rates of convergence for both random functions and random set-valued mappings under iid sampling. Further, we extend them to the non-iid sampling case by using a Gärtner–Ellis type theorem. Finally, we apply the obtained theoretical results to the SAA analysis of several kinds of stochastic optimization problems, which show the applicability of our results in the discretization of stochastic optimization. This study is motivated by the fact that data distributions in many real applications are significantly heavy tailed, which cause many existing results obtained under light tailed distribution assumption not applicable.

Keywords: polynomial rate   SAA   heavy tailed distributions   uniform convergence   non-iid sampling   stochastic optimization

1 Introduction

Stochastic optimization is usually intractable numerically in a straightforward way. For instance, it may involve expectations (multidimensional integrals), which cannot be calculated explicitly or even with high accuracy, the wait-and-see constraints in stochastic optimization usually mean there are semi-infinite constraints. Different sampling methods have thus been proposed to pretreat and numerically solve stochastic optimization problems. Through sampling, expectations are replaced by finite summations and the semi-infinite wait-and-see constraints are reduced to finite constraints, which, therefore, ease numerical difficulty.

When the sampling approach is adopted to derive the discretization problem, a subsequent and important question is: whether the discretization problem approximates the true counterpart well or not? Plenty of literature focused on this topic, see e.g. monographs [1, 2] and papers [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15], to name a few. In general, these results can be roughly divided into two classes: the qualitative ones and the quantitative ones. The former ones usually use the uniform law of large numbers (LLN for short) (see e.g. [2]), central limit theorem (see e.g. [14]) or epi-convergence (see e.g. [16, 17]) to obtain asymptotic convergence assertions. Nevertheless, they fail to tell how fast the convergence could be or how large the sample size should be to achieve

---

[*]Corresponding Author

[†]College of Mathematics and Statistics, Chongqing University, Chongqing, China, jiangjiecq@163.com

[‡]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shannxi, China, zchen@xjtu.edu.cn

[§]School of Mathematical Science, Chongqing Normal University, Chongqing, China, xmyang@cqnu.edu.cn

a certain accuracy. The latter ones aim to give an estimation of the necessary sample size for fixed accuracy and confidence level. The quantitative estimation can provide more practical guidance in numerical solution process. Here the commonly-used tool[*] is the large deviation theory, which has been proved to be a crucial tool to handle many problems in statistics, engineering, statistical mechanics and applied probability [18]. Through large deviation theory, the uniform exponential rate of convergence with increase of the sample size can be derived, see [1, 3, 2] for more details.

In view of the tremendous literature in this respect, we review here some typical researches. As for qualitative convergence analysis, early works on the asymptotic behaviour of stochastic programs are mainly based on the M-estimators ([19]), see e.g. [16, 20, 21]. Based on Monte Carlo (i.e., iid sampling) method, Kleywegt et al. considered in [14] the convergence of the optimal value and optimal solution set of the SAA problem. The uniform LLN is often needed to derive convergence for the optimal value and optimal solution set. The uniform LLN for real-valued random functions has been discussed by Rubinstein and Shapiro in [22] and a uniform version of LLN for random set-valued mappings was studied by Shapiro and Xu in [4]. By uniform LLN, the convergence assertions for different kinds of stochastic optimization models have been investigated in papers, like [6, 17, 23, 24, 1, 10].

As for quantitative convergence analysis, Kaniovski et al. first adopted in [25] the large derivation theory to derive the exponential rate of convergence. Shapiro and Homem-de-Mello [12], Kleywegt et al. [14] and Shapiro and Xu [3] employed the Monte Carlo simulation technique (iid sampling) to derive the SAA problem, and established the exponential rate of convergence by large derivation. On the basis of the relationship between support function and deviation distance, Xu considered in [9] the exponential rate of convergence for random set-valued mappings. Ralph and Xu studied in [6] the SAA of general two-stage stochastic minimization problems and obtained the exponential convergence assertions of stationary points by utilizing the uniform exponential convergence of random set-valued mappings. For the non-iid sampling, the well-known Gärtner–Ellis theorem (see e.g. [18, Theorem 2.3.6]) is usually employed to derive the exponential rate of convergence. Dai et al. considered in [7] the empirical two-stage stochastic programming problem under non-iid sampling, and established the exponential convergence for the probability of deviation of the empirical optimum from the true optimum by using large deviation techniques. Homem-de-Mello investigated in [8] the rates of convergence of estimators of the optimal solution set and optimal value with respect to the sample size under general sampling, which contains Latin hypercube sampling and quasi-Monte Carlo as special cases. Xu [9] considered the uniform exponential rate of convergence under H-calmness and general sampling. Further, Sun and Xu [13] extended the results in [9] to a more general setting than H-calmness.

To establish the exponential rate of convergence, some strong assumptions are imposed. One of the most restricted assumptions is the light tailed distribution assumption. We say that a probability distribution is a light tailed distribution if its tails damp exponentially [26, Chapter 7]; otherwise, we call the probability distribution heavy tailed. Therefore, if the underlying probability distribution is heavy tailed, the existing results, such as [3, 6, 4, 12, 14], cannot quantify the rate of convergence. However, in many practical applications, such as financial risk management ([27]), insurance policy selection and analysis of resident income, the relevant data distributions are significantly skewed with high leptokurtosis. For these problems, the heavy tailed distribution becomes quite important for properly describing the underlying probability distribution.

Considering the above issues, we investigate in this paper the rate of convergence

---

[*]The Hoeffding's inequality can also be used to give the exponential rate of convergence. However, it requires the boundedness of the support set

under heavy tailed distributions, for both iid and non-iid sampling cases. The main contributions of this paper can be summarized as follows.

- We study the SAAs of random functions and random set-valued mappings, respectively, and establish their uniform polynomial rates of convergence under iid sampling and heavy tailed distributions.

- We extend the above results to the general (non-iid) sampling case by establishing a Gärtner–Ellis type theorem. To the best of our knowledge, there are not such results in the current literature.

- Finally, we apply our quantitative results about rates of convergence under heavy tailed distributions to several kinds of stochastic optimization problems. By this way, we obtain many novel results of these models compared with the existing works.

The rest of this paper is organized as follows. In Section 2, under iid sampling, some preliminaries from Chebyshev's inequality are discussed. Then the pointwise polynomial rate of convergence under heavy tailed distributions is presented. In Section 3, we develop the uniform polynomial rates of convergence for both random functions and random set-valued mappings under iid sampling and heavy tailed distributions. In Section 4, we reestablish the rates of convergence in Section 3 under general sampling framework. In Section 5, the obtained theoretical results are applied to several types of stochastic optimization problems to illustrate their applicability and effectiveness. Finally, we conclude the whole paper in Section 6.

We use the following notations and terminologies throughout the paper. Denote by $\mathbb{R}_+^n$ the collection of all nonnegative elements in $\mathbb{R}^n$; $\|\cdot\|$ denotes the Euclidean norm in the corresponding space; For $A, B \subseteq \mathbb{R}^n$, $\|A\| := \sup_{a \in A} \|a\|$; The distance from a point $a \in A$ to B is denoted by $d(a, B) = \inf_{b \in B} \|a - b\|$; $\mathbb{D}(A, B)$ denotes the deviation distance between A and B, i.e., $\mathbb{D}(A, B) := \sup_{a \in A} d(a, B)$; $\mathbb{H}(A, B)$ denotes the Hausdorff distance between A and B, i.e., $\mathbb{H}(A, B) = \max\{\mathbb{D}(A, B), \mathbb{D}(B, A)\}$; We use '$\top$' to denote the transposition of vectors or matrices; $A + B := \{a + b : a \in A, b \in B\}$ and $\lambda A := \{\lambda a : a \in A\}$ for $\lambda \in \mathbb{R}$; Denote by $\mathcal{B}$ the unit closed ball in the corresponding space according to the context; We use '$\rightrightarrows$' to denote set-valued mappings.

2 Preliminaries from Chebyshev's inequality

In this section, we will present some preliminary results from Chebyshev's inequality, which are important for our later discussion. Chebyshev's inequality (see e.g. [28, Theorem 1.6.4] and [29, Lemma 3.1]) is a well-known and powerful tool in probability theory. We first give a trivial variant of the classical Chebyshev's inequality, which can also be found in [26, Section 6.1].

Lemma 2.1 ([26]). *Let X be a nonnegative random variable defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $g : \mathbb{R}_+ \to \mathbb{R}_+$ be a nondecreasing function. Then*

$$\mathbb{P}\{X \geq \epsilon\} \leq \frac{\mathbb{E}[g(X)]}{g(\epsilon)} \tag{1}$$

holds for any $\epsilon \geq 0$ with $g(\epsilon) > 0$.

Proof. Note that

$$\mathbb{P}\{X \geq \epsilon\} \leq \int_{g(X) \geq g(\epsilon)} 1 \mathrm{d}\mathbb{P} \overset{(a)}{\leq} \int_{g(X) \geq g(\epsilon)} \frac{g(X)}{g(\epsilon)} \mathrm{d}\mathbb{P}$$

$$= \int_{\Omega} \frac{g(X(\omega)) \mathbf{1}_{\{g(X) \geq g(\epsilon)\}}(\omega)}{g(\epsilon)} \mathrm{d}\mathbb{P}$$

$$\overset{(b)}{\leq} \frac{\mathbb{E}[g(X)]}{g(\epsilon)},$$

where (a) follows from $g(X) \geq g(\epsilon)$ and $g(\epsilon) > 0$; $\mathbf{1}_{\{g(X) \geq g(\epsilon)\}}(\omega)$ denotes the indicator function, that is

$$\mathbf{1}_{\{g(X) \geq g(\epsilon)\}}(\omega) = \begin{cases} 1, & g(X(\omega)) \geq g(\epsilon), \\ 0, & \text{otherwise;} \end{cases}$$

(b) follows from the nonnegativeness of X and g. $\qquad\square$

It is easy to see that Lemma 2.1 reduces to the classical Markov's inequality when $g(x) = x$. Moreover, the random variable X can take values over $\mathbb{R}$, that is, $X : \Omega \to \mathbb{R}$. In that case, $g : \mathbb{R} \to \mathbb{R}_+$ is nondecreasing. Moreover, we can choose $\epsilon \in \mathbb{R}$ as long as $g(\epsilon) > 0$. Then (1) holds by using a similar proof.

Let $X_1, \cdots, X_N$ be N iid random samples of X over probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Denote by $\mu = \mathbb{E}[X]$ and

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i.$$

It is known from the strong LLN that $\bar{X} \to \mu$ as $N \to \infty$ with probability 1. However, it fails to tell us how fast this convergence could be. This is very important in practical calculation for specified accuracy and confidence level. In view of this, the rate of convergence has been widely discussed. In the relevant works, the Cramér's large deviation theorem is used to develop the exponential rate of convergence. In fact, the Cramér's large deviation theorem can be derived from Lemma 2.1 by letting g be an exponential function; see e.g. [2, Chapter 7] for more details. The Cramér's large deviation theorem asserts that

$$\mathbb{P}\left\{\left|\bar{X} - \mu\right| \geq \epsilon\right\} \leq \exp(-N \min\{I(\epsilon), I(-\epsilon)\}), \qquad (2)$$

where $I(\epsilon) := \sup_{t \in \mathbb{R}} \{\epsilon t - \ln M(t)\}$ is the large deviation rate function of X and $M(t) = \mathbb{E}[\exp(tX)]$ is the moment generating function of X. Under certain light tailed distributions assumption, both $I(\epsilon)$ and $I(-\epsilon)$ can be positive and bounded. Then the exponential rate of convergence with sample size N follows directly from (2).

Surely, the exponential rate of convergence is a quite fast convergence rate. To derive the exponential rate of convergence, a light tailed distribution (the tail parts of the distribution should damp exponentially) assumption is usually necessary, which is restrictive sometimes in practical applications. Are there relatively weak conditions imposed on the underlying distributions, such that the rate of convergence under more general distributions can be derived? In what follows, we try to settle this problem. To this end, we give the following vital lemma.

Lemma 2.2 ([30]). *Suppose that: (i) $Y_1, \cdots, Y_N$ are N iid random samples of the random variable Y defining on probability space $(\Omega, \mathcal{F}, \mathbb{P})$; (ii) $\mathbb{E}[Y] = 0$ and the first p moments of Y are finite for some $p \geq 2$. Then*

$$\mathbb{E}\left[\left(\sum_{i=1}^{N} Y_i\right)^p\right] \leq C N^{\frac{p}{2}}$$

for sufficiently large N, where C is a positive constant depending only on p and the first p moments of Y.

Based on Lemma 2.2, we immediately have the following conclusion.

Theorem 2.3 ([30]). *Let $\mathbb{E}[X] = \mu$ and the first p moments of X be finite with some $p \geq 2$. Then,*

$$\mathbb{P}\left\{\left|\bar{X} - \mu\right| \geq \epsilon\right\} \leq \frac{C}{N^{\frac{p}{2}}\epsilon^p}$$

holds for any $\epsilon > 0$ and sufficiently large N, where C is a positive constant depending on p and the first p moments of $X - \mu$.

In this section, we have presented the pointwise polynomial rate of convergence (Theorem 2.3) under iid sampling. Compared with the existing results about the exponential rate of convergence, our polynomial rate of convergence does not require the light tailed assumption. Instead, it only needs some finite moment information, which can be easily satisfied in many applications. All these lay the foundation for the discussion in the sequel.

3 Uniform polynomial rate of convergence under iid sampling

In this section, we will derive the uniform polynomial rate of convergence under iid sampling. The uniform convergence is quite important in the SAA convergence analysis of stochastic optimization problems [2]. Specifically, we examine two cases in this section. The first case is for random functions. The second case is for random set-valued mappings.

3.1 Uniform polynomial rate of convergence for random functions

Consider a random vector ξ defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the support set being $\Xi \subseteq \mathbb{R}^s$. Let $\mathcal{X} \subseteq \mathbb{R}^n$ and $\xi^1, \cdots, \xi^N$ be N iid samples for ξ. For a random function $F : \mathcal{X} \times \Xi \to \mathbb{R}$, let $f(x) := \mathbb{E}[F(x, \xi)]$ and $\hat{f}_N(x) := \frac{1}{N}\sum_{i=1}^{N} F(x, \xi^i)$.

For fixed $x \in \mathcal{X}$, if $F(x, \xi)$ has the finite first p moments, we know from Theorem 2.3 that

$$\mathbb{P}\left\{\left|\hat{f}_N(x) - f(x)\right| \geq \epsilon\right\} \leq \frac{C_x}{N^{\frac{p}{2}}\epsilon^p}$$

for sufficiently large N. However, this estimation depends on x. For some applications, we need a uniform rate of convergence with respect to $x \in \mathcal{X}$, especially for the SAA convergence of stochastic optimization problems. To this end, we first introduce the concept of H-calmness introduced in [9, Definition 2.3].

Definition 3.1 (H-calmness, [9]). *Let $F : \mathcal{X} \times \Xi \to \mathbb{R}$. Then*

(i) *F is H-calm at x from above with modulus $\kappa(\xi)$ and order γ if $F(x, \xi)$ is finite and there exist $\kappa : \Xi \to \mathbb{R}_+$, $\gamma > 0$ and $\delta > 0$ such that*

$$F(x', \xi) - F(x, \xi) \leq \kappa(\xi)\left\|x' - x\right\|^\gamma$$

for all $x' \in \mathcal{X}$ with $\|x' - x\| \leq \delta$ and $\xi \in \Xi$;

(ii) *F is H-calm at x from below with modulus $\kappa(\xi)$ and order γ if $F(x, \xi)$ is finite and there exist $\kappa : \Xi \to \mathbb{R}_+$, $\gamma > 0$ and $\delta > 0$ such that*

$$F(x', \xi) - F(x, \xi) \geq -\kappa(\xi)\left\|x' - x\right\|^\gamma$$

for all $x' \in \mathcal{X}$ with $\|x' - x\| \leq \delta$ and $\xi \in \Xi$;

5

(iii) F is H-calm at x with modulus $\kappa(\xi)$ and order γ if $F(x,\xi)$ is finite and there exist $\kappa : \Xi \to \mathbb{R}_+$, $\gamma > 0$ and $\delta > 0$ such that

$$\left| F(x',\xi) - F(x,\xi) \right| \leq \kappa(\xi) \left\| x' - x \right\|^\gamma$$

for all $x' \in \mathcal{X}$ with $\|x' - x\| \leq \delta$ and $\xi \in \Xi$.

F is said to be H-calm from above, H-calm from below, H-calm over $\mathcal{X}$ if the corresponding properties stated above hold at every point of $\mathcal{X}$.

If F is H-calm from above, H-calm from below and H-calm at x with modulus $\kappa(\xi)$ and order 1, F is also called calm from above, calm from below and calm, respectively (see [32, Chapter 8]). From the above definition, we know that: if F is locally Lipschitz continuous at x with Lipschitz modulus $\kappa(\xi)$, then it must be H-calm at x with modulus $\kappa(\xi)$ and order 1 (i.e., F is calm at x with modulus $\kappa(\xi)$), but not vise versa. From that point of view, the concept of H-calmness is somehow an extension of calmness and locally Lipschitz continuity. One can refer to [9, Example 2.1] for an example of H-calmness.

In the light of Theorem 2.3 and H-calmness, we have the following uniform polynomial rate of convergence.

Theorem 3.2. *Suppose that: (i) $\mathcal{X}$ is compact; (ii) $\mathbb{E}[F(x,\xi)^p] < +\infty$ for each $x \in \mathcal{X}$ and some $p \geq 2$; (iii) $f(x)$ is continuous on $\mathcal{X}$; (iv) $\xi^1, \cdots, \xi^N$ are iid samples. Let $\kappa : \Xi \to \mathbb{R}_+$ be a measurable function with $\mathbb{E}[\kappa^p(\xi)] < +\infty$ and γ be a positive scalar.*

(a) If $F(\cdot,\xi)$ is H-calm from above over $\mathcal{X}$ with modulus $\kappa(\xi)$ and order γ, for arbitrary $\epsilon > 0$, there exists a positive scalar C_1, independent of N, such that

$$\mathbb{P}\left\{ \sup_{x \in \mathcal{X}} \left(\hat{f}_N(x) - f(x) \right) \geq \epsilon \right\} \leq \frac{C_1}{N^{\frac{p}{2}} \epsilon^p}$$

for sufficiently large N;

(b) If $F(\cdot,\xi)$ is H-calm from below over $\mathcal{X}$ with modulus $\kappa(\xi)$ and order γ, for arbitrary $\epsilon > 0$, there exists a positive scalar C_2, independent of N, such that

$$\mathbb{P}\left\{ \inf_{x \in \mathcal{X}} \left(\hat{f}_N(x) - f(x) \right) \leq -\epsilon \right\} \leq \frac{C_2}{N^{\frac{p}{2}} \epsilon^p}$$

for sufficiently large N;

(c) If $F(\cdot,\xi)$ is H-calm over $\mathcal{X}$ with modulus $\kappa(\xi)$ and order γ, for arbitrary $\epsilon > 0$, there exists a positive scalar C_3, independent of N, such that

$$\mathbb{P}\left\{ \sup_{x \in \mathcal{X}} \left| \hat{f}_N(x) - f(x) \right| \geq \epsilon \right\} \leq \frac{C_3}{N^{\frac{p}{2}} \epsilon^p}$$

for sufficiently large N.

Proof. This proof is similar to that of [2, Theorem 7.73] and [9, Theorem 3.1]. For completeness, we give a simple proof. Since (a) and (b) can be verified in a similar way, and (c) holds automatically from (a) and (b), we just give the proof of (a) in what follows.

Since F is H-calm from above on $\mathcal{X}$ with modulus $\kappa(\xi)$ and order γ, we have from Definition 3.1 that: for any $x \in \mathcal{X}$, there exists a positive scalar δ_x depending on x such that

$$F(x',\xi) - F(x,\xi) \leq \kappa(\xi) \left\| x' - x \right\|^\gamma \tag{3}$$

holds for all $x' \in \mathcal{X}$ with $\|x' - x\| \leq \delta_x$ and $\xi \in \Xi$.

Let ϵ be an arbitrary positive scalar. Since $\mathcal{X}$ is compact and f is continuous on $\mathcal{X}$, we have a finite η-net of $\mathcal{X}$, denoted by $\{\bar{x}_1, \cdots, \bar{x}_\nu\}$, such that for any $x \in \mathcal{X}$, there exists an $i(x) \in \{1, \cdots, \nu\}$ such that

$$\left\| x - \bar{x}_{i(x)} \right\| \leq \eta \leq \delta_{\bar{x}_{i(x)}}, \tag{4}$$

$$\frac{\epsilon}{3\eta^\gamma} - \bar{\kappa} \geq \epsilon \tag{5}$$

and

$$\left| f(\bar{x}_{i(x)}) - f(x) \right| \leq \frac{\epsilon}{3}, \tag{6}$$

where $\bar{\kappa} = \mathbb{E}[\kappa(\xi)]$. (3) together with (4) imply that

$$F(x, \xi) - F(\bar{x}_{i(x)}, \xi) \leq \kappa(\xi) \left\| x - \bar{x}_{i(x)} \right\|^\gamma. \tag{7}$$

Note that

$$\hat{f}_N(x) - f(x) \leq \hat{f}_N(x) - \hat{f}_N(\bar{x}_{i(x)}) + \left| \hat{f}_N(\bar{x}_{i(x)}) - f(\bar{x}_{i(x)}) \right| + \left| f(\bar{x}_{i(x)}) - f(x) \right|$$

$$\leq \frac{1}{N} \sum_{i=1}^N \kappa(\xi^i) \eta^\gamma + \left| \hat{f}_N(\bar{x}_{i(x)}) - f(\bar{x}_{i(x)}) \right| + \frac{\epsilon}{3},$$

where the last inequality follows from (6) and (7). Therefore,

$$\mathbb{P}\left\{ \hat{f}_N(x) - f(x) \geq \epsilon \right\} \leq \mathbb{P}\left\{ \frac{1}{N} \sum_{i=1}^N \kappa(\xi^i) \eta^\gamma + \left| \hat{f}_N(\bar{x}_{i(x)}) - f(\bar{x}_{i(x)}) \right| + \frac{\epsilon}{3} \geq \epsilon \right\}$$

$$= \mathbb{P}\left\{ \frac{1}{N} \sum_{i=1}^N \kappa(\xi^i) \eta^\gamma + \left| \hat{f}_N(\bar{x}_{i(x)}) - f(\bar{x}_{i(x)}) \right| \geq \frac{2}{3} \epsilon \right\}$$

$$\leq \mathbb{P}\left\{ \frac{1}{N} \sum_{i=1}^N \kappa(\xi^i) \geq \frac{\epsilon}{3\eta^\gamma} \right\} + \mathbb{P}\left\{ \left| \hat{f}_N(\bar{x}_{i(x)}) - f(\bar{x}_{i(x)}) \right| \geq \frac{\epsilon}{3} \right\}.$$

As for $\mathbb{P}\left\{ \frac{1}{N} \sum_{i=1}^N \kappa(\xi^i) \geq \frac{\epsilon}{3\eta^\gamma} \right\}$, we have

$$\mathbb{P}\left\{ \frac{1}{N} \sum_{i=1}^N \kappa(\xi^i) \geq \frac{\epsilon}{3\eta^\gamma} \right\} = \mathbb{P}\left\{ \frac{1}{N} \sum_{i=1}^N \kappa(\xi^i) - \bar{\kappa} \geq \frac{\epsilon}{3\eta^\gamma} - \bar{\kappa} \right\}$$

$$\leq \mathbb{P}\left\{ \left| \frac{1}{N} \sum_{i=1}^N \kappa(\xi^i) - \bar{\kappa} \right| \geq \frac{\epsilon}{3\eta^\gamma} - \bar{\kappa} \right\}$$

$$\leq \mathbb{P}\left\{ \left| \frac{1}{N} \sum_{i=1}^N \kappa(\xi^i) - \bar{\kappa} \right| \geq \epsilon \right\}.$$

Theorem 2.3 concludes that

$$\mathbb{P}\left\{ \left| \frac{1}{N} \sum_{i=1}^N \kappa(\xi^i) - \bar{\kappa} \right| \geq \epsilon \right\} \leq \frac{\bar{C}_1}{N^{\frac{p}{2}} \epsilon^p} \tag{8}$$

for sufficiently large N, where $\bar{C}_1$ is a positive constant depending on p and the first p moments of $\kappa(\xi) - \bar{\kappa}$.

As for $\mathbb{P}\left\{\left|\hat{f}_N(\bar{x}_{i(x)}) - f(\bar{x}_{i(x)})\right| \geq \frac{\epsilon}{3}\right\}$, we have directly from Theorem 2.3 that

$$\mathbb{P}\left\{\left|\hat{f}_N(\bar{x}_{i(x)}) - f(\bar{x}_{i(x)})\right| \geq \frac{\epsilon}{3}\right\} \leq \frac{C_{\bar{x}_{i(x)}}}{N^{\frac{p}{2}}\epsilon^p}$$

holds for sufficiently large N, where $C_{\bar{x}_{i(x)}}$ is a positive constant depending on p and the first p moments of $F(\bar{x}_{i(x)}, \xi) - f(\bar{x}_{i(x)})$. Because there are finite points in the η-net, we can select a uniform positive scalar $\hat{C}_1$, such that

$$\mathbb{P}\left\{\left|\hat{f}_N(\bar{x}_i) - f(\bar{x}_i)\right| \geq \frac{\epsilon}{3}\right\} \leq \frac{\hat{C}_1}{N^{\frac{p}{2}}\epsilon^p} \tag{9}$$

for sufficiently large N and all $i \in \{1, \cdots, \nu\}$.

Finally, (8) together with (9) conclude that

$$\mathbb{P}\left\{\hat{f}_N(x) - f(x) \geq \epsilon\right\} \leq \frac{C_1}{N^{\frac{p}{2}}\epsilon^p}$$

for sufficiently large N, where $C_1 := \bar{C}_1 + \hat{C}_1$. $\qquad\square$

Theorem 3.2 gives the polynomial rate of convergence under mild conditions. We have some comments as follows.

Remark 3.3. *It knows from the proof that C_1, C_2 and C_3 (independent of the sample size N) depends on the η-net we choose. The net is dependent on the accuracy parameter ϵ. Thus, C_1, C_2 and C_3 are also dependent on ϵ.*

The assumptions in Theorem 3.2 is much weaker than those in [2, 9, 3]. In general, to ensure the exponential rate of convergence, $F(x, \xi)$ for each $x \in \mathcal{X}$ should have a light tailed distribution and $\kappa(\xi)$ also needs to obey a light tailed distribution. That is, both $\mathbb{E}[\exp\{t(F(x, \xi) - f(x))\}]$ for every $x \in \mathcal{X}$ and $\mathbb{E}[\exp\{t(\kappa(\xi) - \bar{\kappa})\}]$ are finitely valued for all t in a neighbourhood of zero (see e.g. [3, Section 5, C1-C3] and [2, Theorem 7.73] for more details). We know from [26, Chapter 7] that these conditions stand for light tailed assumptions, which are difficult to verify to some extent. Compared with those limited conditions, Theorem 3.2 might be more friendly to provide some quantification assertions for the rate of convergence, because it just requires certain finite moment information.

For arbitrarily fixed accuracy parameter ϵ, Theorem 3.2 can help us to estimate how large the sample size N should be to ensure the confidence level $\rho \in (0, 1)$. Take (c) in Theorem 3.2 as an example. To ensure

$$\mathbb{P}\left\{\sup_{x \in \mathcal{X}}\left|\hat{f}_N(x) - f(x)\right| \geq \epsilon\right\} \leq \rho,$$

a sufficient condition is $C_3/N^{\frac{p}{2}}\epsilon^p \leq \rho$, which is equivalent to $N \geq (C_3)^{\frac{2}{p}}/(\rho^{\frac{2}{p}}\epsilon^2)$.

Generally, our estimation of the sample size N will be larger than those in [2, 9, 3] which estimate the sample size by the exponential rate of convergence. However, as what we stated, our assumptions are much weaker than theirs. Thus, our results are expected to give some quantification assertions under certain harsh conditions, like heavy tailed distributions.

3.2 Uniform polynomial rate of convergence for random set-valued mappings

In this subsection, we consider the uniform polynomial rate of convergence for random set-valued mappings. For this purpose, denote by $\Phi : \mathcal{X} \times \Xi \rightrightarrows \mathbb{R}^n$ a set-valued mapping

where $\mathcal{X}$ and Ξ are defined in Section 3.1. Let $\xi^1, \cdots, \xi^N$ be N iid random samples. Then its SAA approximation is $\frac{1}{N}\sum_{i=1}^{N}\Phi(x, \xi^i)$. For the simplicity of notation, we denote

$$\hat{\Phi}_N(x) = \frac{1}{N}\sum_{i=1}^{N}\Phi(x, \xi^i)$$

and

$$\hat{\Phi}(x) = \mathbb{E}[\text{conv}\{\Phi(x, \xi)\}],$$

where 'conv' stands for the convex hull and the expectation of a set-valued mapping is defined by the Aumann integral [33]. Then we consider the rate of convergence between $\hat{\Phi}_N(x)$ and $\hat{\Phi}(x)$ in what follows.

Recall that $\Phi(x, \xi)$ is called (convex-)compact-valued if $\Phi(x, \xi)$ is a (convex)compact set for every $x \in \mathcal{X}$ and $\xi \in \Xi$. The strong LLN for set-valued mappings (see [34]) tells us: if $\Phi(x, \xi)$ is compact-valued and $\mathbb{E}[\|\|\Phi(x, \xi)\|\|] < \infty$, we have

$$\mathbb{H}(\hat{\Phi}_N(x), \hat{\Phi}(x)) \to 0$$

as $N \to \infty$, where $\mathbb{E}[\|\|\Phi(x, \xi)\|\|] := \sup_{\phi(x,\xi)\in\Phi(x,\xi)}\mathbb{E}[\|\|\phi(x, \xi)\|\|]$ and $\mathbb{H}(\cdot, \cdot)$ denotes the Hausdorff distance between two sets. Recall that for probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a set $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$ is an atom of $\mathbb{P}$ if for all $B \in \mathcal{F}$ and $B \subseteq A$, either $\mathbb{P}(B) = 0$ or $\mathbb{P}(B) = \mathbb{P}(A)$. $(\Omega, \mathcal{F}, \mathbb{P})$ or $\mathbb{P}$ is called nonatomic if it has no atoms. It follows from the definition of Aumann integral [33] that: if the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is nonatomic, we have

$$\mathbb{E}[\text{conv}\{\Phi(x, \xi)\}] = \mathbb{E}[\Phi(x, \xi)].$$

Thus, if probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is nonatomic, we have the following consistent estimation in the sense of Hausdorff distance

$$\mathbb{H}(\hat{\Phi}_N(x), \mathbb{E}[\Phi(x, \xi)]) \to 0$$

as $N \to \infty$. For more discussions in this aspect, one can refer to [9, Proposition 4.1].

Since the above convergence is in pointwise sense, which fails to analyze convergence of stochastic optimization problems, the uniform LLN for random set-valued mappings was investigated in [4]. However, note that this convergence analysis does not quantify the rate of convergence as $N \to \infty$. The uniform exponential rate of convergence with respect to x has been conducted under light tailed distributions [9, 6]. Here, we extend them to heavy tailed distributions. For this purpose, we first address that the deviation of a set $A \subseteq \mathbb{R}^n$ from a set $B \subseteq \mathbb{R}^n$, denoted by $\mathbb{D}(A, B)$, has a close relationship with the support function of A, denoted by $\sigma(u, A) := \sup_{a\in A} u^\top a$, and the support function of B, denoted by $\sigma(u, B) := \sup_{b\in B} u^\top b$. We summarize it as the following lemma (see also [35, Theorem II-18] and [6, Lemma 3.1]).

Lemma 3.4 ([35]). *Let A, B be two nonempty and compact subsets of $\mathbb{R}^n$. Then*

$$\mathbb{D}(A, B) \geq \max_{\|u\|\leq 1} (\sigma(u, A) - \sigma(u, B)) \tag{10}$$

and

$$\mathbb{H}(A, B) \geq \max_{\|u\|\leq 1} |\sigma(u, A) - \sigma(u, B)|. \tag{11}$$

If, moreover, A and B are convex, the equalities in (10) and (11) hold.

Lemma 3.4 is a commonly-used tool for the analysis of the rate of convergence for random set-valued mappings, see [9, (4.31)] and [6]. For $\Phi(x, \xi)$ being nonempty, compact and convex, we have

$$\mathbb{D}(\hat{\Phi}_N(x), \hat{\Phi}(x)) = \max_{\|u\| \le 1} \left(\sigma\left(u, \hat{\Phi}_N(x)\right) - \sigma\left(u, \hat{\Phi}(x)\right) \right)$$

$$= \max_{\|u\| \le 1} \left(\sigma\left(u, \frac{1}{N} \sum_{i=1}^{N} \Phi(x, \xi^i)\right) - \sigma\left(u, \hat{\Phi}(x)\right) \right)$$

$$= \max_{\|u\| \le 1} \left(\frac{1}{N} \sum_{i=1}^{N} \sigma\left(u, \Phi(x, \xi^i)\right) - \sigma\left(u, \mathbb{E}[\Phi(x, \xi)]\right) \right)$$

$$= \max_{\|u\| \le 1} \left(\frac{1}{N} \sum_{i=1}^{N} \sigma\left(u, \Phi(x, \xi^i)\right) - \mathbb{E}\left[\sigma\left(u, \Phi(x, \xi)\right)\right] \right).$$

Then for any positive scalar ϵ, we have

$$\mathbb{P}\left\{ \mathbb{D}(\hat{\Phi}_N(x), \hat{\Phi}(x)) \ge \epsilon \right\} = \mathbb{P}\left\{ \max_{\|u\| \le 1} \left(\frac{1}{N} \sum_{i=1}^{N} \sigma\left(u, \Phi(x, \xi^i)\right) - \mathbb{E}\left[\sigma\left(u, \Phi(x, \xi)\right)\right] \right) \ge \epsilon \right\}.$$

Analogously, we have

$$\mathbb{P}\left\{ \mathbb{D}(\hat{\Phi}(x), \hat{\Phi}_N(x)) \ge \epsilon \right\} = \mathbb{P}\left\{ \max_{\|u\| \le 1} \left(\mathbb{E}\left[\sigma\left(u, \Phi(x, \xi)\right)\right] - \frac{1}{N} \sum_{i=1}^{N} \sigma\left(u, \Phi(x, \xi^i)\right) \right) \ge \epsilon \right\}$$

and

$$\mathbb{P}\left\{ \mathbb{H}(\hat{\Phi}_N(x), \hat{\Phi}(x)) \ge \epsilon \right\} = \mathbb{P}\left\{ \max_{\|u\| \le 1} \left| \frac{1}{N} \sum_{i=1}^{N} \sigma\left(u, \Phi(x, \xi^i)\right) - \mathbb{E}\left[\sigma\left(u, \Phi(x, \xi)\right)\right] \right| \ge \epsilon \right\}.$$

The above equalities mean that: investigating the rate of convergence for random set-valued mappings can be transformed into the corresponding random functions. Denote

$$G(x, u, \xi) := \sigma\left(u, \Phi(x, \xi)\right).$$

We are now ready to present the main results of this subsection.

Theorem 3.5. *Suppose that: (i) $\mathcal{X}$ is compact; (ii) $\Phi(x, \xi)$ is nonempty, compact and convex for each pair $(x, \xi) \in \mathcal{X} \times \Xi$; (iii) $\mathbb{E}[G(x, u, \xi)^p] < +\infty$ for each pair $(x, u) \in \mathcal{X} \times \mathcal{B}$ and some $p \ge 2$; (iv) $\mathbb{E}[G(\cdot, \cdot, \xi)]$ is continuous on $\mathcal{X} \times \mathcal{B}$; (v) $\xi^1, \cdots, \xi^N$ are iid samples. Let $\kappa : \Xi \to \mathbb{R}_+$ be a measurable function with $\mathbb{E}[\kappa^p(\xi)] < +\infty$ and γ be a positive scalar.*

(a) If $G(\cdot, \cdot, \xi)$ is H-calm from above over $\mathcal{X} \times \mathcal{B}$ with modulus $\kappa(\xi)$ and order γ, for arbitrary $\epsilon > 0$, there exists an $L_1 > 0$, independent of N, such that

$$\mathbb{P}\left\{ \sup_{x \in \mathcal{X}} \mathbb{D}(\hat{\Phi}_N(x), \hat{\Phi}(x)) \ge \epsilon \right\} \le \frac{L_1}{N^{\frac{p}{2}} \epsilon^p}$$

for sufficiently large N;

(b) If $G(\cdot, \cdot, \xi)$ is H-calm from below over $\mathcal{X} \times \mathcal{B}$ with modulus $\kappa(\xi)$ and order γ, for arbitrary $\epsilon > 0$, there exists an $L_2 > 0$, independent of N, such that

$$\mathbb{P}\left\{ \sup_{x \in \mathcal{X}} \mathbb{D}(\hat{\Phi}(x), \hat{\Phi}_N(x)) \ge \epsilon \right\} \le \frac{L_2}{N^{\frac{p}{2}} \epsilon^p}$$

for sufficiently large N;

(c) If $G(\cdot, \cdot, \xi)$ is H-calm over $\mathcal{X} \times \mathcal{B}$ with modulus $\kappa(\xi)$ and order γ, for arbitrary $\epsilon > 0$, there exists an $L_3 > 0$, independent of N, such that

$$\mathbb{P}\left\{\sup_{x \in \mathcal{X}} \mathbb{H}(\hat{\Phi}_N(x), \hat{\Phi}(x)) \geq \epsilon\right\} \leq \frac{L_3}{N^{\frac{p}{2}}\epsilon^p}$$

for sufficiently large N.

This theorem can be similarly proved as that of Theorem 3.2, and thus we omit it here. Now we give sufficient conditions for $G(x, u, \xi)$ to be H-calm from above and H-calm over $\mathcal{X} \times \mathcal{B}$, respectively.

Proposition 3.6. *Let $\kappa_1 : \Xi \to \mathbb{R}_+$, $\kappa_2 : \Xi \to \mathbb{R}_+$ and $\gamma > 0$. Suppose that $\Phi(x, \xi)$ is nonempty, compact and convex for each pair $(x, \xi) \in \mathcal{X} \times \Xi$, and $\|\Phi(x, \xi)\| \leq \kappa_1(\xi)$.*

(i) If for every $x \in \mathcal{X}$, there exists a $\delta_x > 0$ such that $\mathbb{D}(\Phi(x', \xi), \Phi(x, \xi)) \leq \kappa_2(\xi)\|x' - x\|^{\gamma}$ for all $x' \in \mathcal{X}$ with $\|x' - x\| \leq \delta_x$, then

$$G(x', u', \xi) - G(x, u, \xi) \leq \kappa_1(\xi)\|u' - u\| + \kappa_2(\xi)\|x' - x\|^{\gamma}$$

for all $x' \in \mathcal{X}$ with $\|x' - x\| \leq \delta_x$, $u', u \in \mathcal{B}$ and $\xi \in \Xi$;

(ii) If for every $x \in \mathcal{X}$, there exists a $\delta_x > 0$ such that $\mathbb{H}(\Phi(x', \xi), \Phi(x, \xi)) \leq \kappa_2(\xi)\|x' - x\|^{\gamma}$ for all $x' \in \mathcal{X}$ with $\|x' - x\| \leq \delta_x$, then

$$\left|G(x', u, \xi) - G(x, u, \xi)\right| \leq \kappa_1(\xi)\|u' - u\| + \kappa_2(\xi)\|x' - x\|^{\gamma}$$

for all $x' \in \mathcal{X}$ with $\|x' - x\| \leq \delta_x$, $u', u \in \mathcal{B}$ and $\xi \in \Xi$.

Proof. We only give the proof of (i), and (ii) can be proved similarly. For any $x \in \mathcal{X}$, considering $x' \in \mathcal{X}$ with $\|x' - x\| \leq \delta_x$, $u', u \in \mathcal{B}$ and $\xi \in \Xi$, we have

$$
\begin{aligned}
G(x', u', \xi) - G(x, u, \xi) &= \sigma\left(u', \Phi(x', \xi)\right) - \sigma\left(u, \Phi(x, \xi)\right) \\
&= \sup_{\phi(x', \xi) \in \Phi(x', \xi)} (u')^{\top}\phi(x', \xi) - \sup_{\phi(x, \xi) \in \Phi(x, \xi)} u^{\top}\phi(x, \xi) \\
&= \sup_{\phi(x', \xi) \in \Phi(x', \xi)} (u')^{\top}\phi(x', \xi) - \sup_{\phi(x', \xi) \in \Phi(x', \xi)} u^{\top}\phi(x', \xi) \\
&\quad + \sup_{\phi(x', \xi) \in \Phi(x', \xi)} u^{\top}\phi(x', \xi) - \sup_{\phi(x, \xi) \in \Phi(x, \xi)} u^{\top}\phi(x, \xi).
\end{aligned}
$$

Since

$$
\begin{aligned}
&\sup_{\phi(x', \xi) \in \Phi(x', \xi)} (u')^{\top}\phi(x', \xi) - \sup_{\phi(x', \xi) \in \Phi(x', \xi)} u^{\top}\phi(x', \xi) \\
&\leq \sup_{\phi(x', \xi) \in \Phi(x', \xi)} (u' - u)^{\top}\phi(x', \xi) \\
&\leq \|u' - u\|\|\Phi(x', \xi)\| \\
&\leq \kappa_1(\xi)\|u' - u\|
\end{aligned}
$$

and

$$
\begin{aligned}
&\sup_{\phi(x', \xi) \in \Phi(x', \xi)} u^{\top}\phi(x', \xi) - \sup_{\phi(x, \xi) \in \Phi(x, \xi)} u^{\top}\phi(x, \xi) \\
&= \sup_{\phi(x', \xi) \in \Phi(x', \xi)} \inf_{\phi(x, \xi) \in \Phi(x, \xi)} u^{\top}(\phi(x', \xi) - \phi(x, \xi)) \\
&\leq \sup_{\phi(x', \xi) \in \Phi(x', \xi)} \inf_{\phi(x, \xi) \in \Phi(x, \xi)} \|\phi(x', \xi) - \phi(x, \xi)\| \\
&= \mathbb{D}(\Phi(x', \xi), \Phi(x, \xi)) \\
&\leq \kappa_2(\xi)\|x' - x\|^{\gamma},
\end{aligned}
$$

11

we have

$$G(x', u', \xi) - G(x, u, \xi) \leq \kappa_1(\xi) \left\| u' - u \right\| + \kappa_2(\xi) \left\| x' - x \right\|^\gamma$$

for $x' \in \mathcal{X}$ with $\left\| x' - x \right\| \leq \delta_x$, $u', u \in \mathcal{B}$ and $\xi \in \Xi$. $\qquad\square$

Obviously, part (i) of Proposition 3.6 implies that $G(\cdot, \cdot, \xi)$ is H-calm from above over $\mathcal{X} \times \mathcal{B}$, and part (ii) of Proposition 3.6 implies that $G(\cdot, \cdot, \xi)$ is H-calm over $\mathcal{X} \times \mathcal{B}$. They provide some sufficient conditions for the assumptions in Theorem 3.5.

In this section, we have considered the uniform polynomial rate of convergence under iid sampling. With the concept of H-calmness, we first establish the uniform polynomial rate of convergence for random functions in Theorem 3.2. Based on Theorem 3.2 and Lemma 3.4, we establish the uniform polynomial rate of convergence for random set-valued mappings in Theorem 3.5.

4 Uniform rate of convergence under non-iid sampling

Besides usually iid sampling techniques, e.g., Monte Carlo sampling method [25, 12, 14], there exist many non-iid sampling approaches, like Latin Hypercube sampling [11, 8], quasi-Monte Carlo sampling [36] etc. Some discrete approximation schemes through probability metrics [37] are also corresponding to non-iid sampling. Therefore, it is worthy studying the rate of convergence under non-iid sampling for heavy tailed distributions.

Gärtner–Ellis theorem is usually adopted to establish the rate of convergence under non-iid sampling, see, e.g. [7, 8, 9, 18, 13]. In that case, the Gärtner–Ellis theorem replaces Cramér's large deviation theorem, and is used to establish the exponential convergence of the SAA. Considering this, we first establish a Gärtner–Ellis type theorem under heavy tailed distributions. To this end, we need the following assumption.

Suppose that $Y_1, \cdots, Y_N$ are N general random samples (probably non-iid) generated according to some random variable Y with $\mathbb{E}[Y] = 0$.

Assumption 1. *There exist two mappings $\alpha : \mathbb{R} \rightarrow \mathbb{R}_+$ being non-decreasing, $\beta : \mathbb{N} \rightarrow \mathbb{R}_+$ with $\beta(N) \rightarrow \infty$ as $N \rightarrow \infty$ and a scalar $C > 0$, independent of N, such that*

$$\limsup_{N \rightarrow \infty} \mathbb{E}\left[\beta(N) \alpha \left(\frac{1}{N} \sum_{i=1}^N Y_i \right) \right] \leq C$$

and

$$\limsup_{N \rightarrow \infty} \mathbb{E}\left[\beta(N) \alpha \left(-\frac{1}{N} \sum_{i=1}^N Y_i \right) \right] \leq C.$$

For convenience, we simply call $Y_1, \cdots, Y_N$ satisfy Assumption 1 in what follows if Assumption 1 holds.

Remark 4.1. *Assumption 1 has a close relationship with Theorem 2.3. It means that the following terms*

$$\mathbb{E}\left[\alpha \left(\frac{1}{N} \sum_{i=1}^N Y_i \right) \right] \;\; and \;\; \mathbb{E}\left[\alpha \left(-\frac{1}{N} \sum_{i=1}^N Y_i \right) \right]$$

converge to zero at least as fast as $\frac{1}{\beta(N)}$ converges to zero for sufficiently large sample size N. However, Gärtner–Ellis theorem assumes that these convergence are at exponential rates, see e.g. [9, Assumption 3.1], [7, Theorem 3.7] and [8]. In Assumption 1, β

controls the rate of convergence. It allows us to examine the rate of convergence under heavy tailed distributions, which is the focus of this paper. Especially, when $Y_1, \cdots, Y_N$ are iid random samples, we know from Lemma 2.2 that: if Y has the first p moments for some positive number $p \geq 2$, Assumption 1 holds with $\beta(N) = N^{\frac{p}{2}}$ and $\alpha(\epsilon) = \epsilon^p$.

Theorem 4.2 (Gärtner–Ellis type theorem). *Let Assumption 1 hold and $\epsilon > 0$ with $\alpha(\epsilon) > 0$. Then we have*

$$\limsup_{N \to \infty} \beta(N) \mathbb{P} \left\{ \frac{1}{N} \sum_{i=1}^{N} Y_i \geq \epsilon \right\} \leq \frac{C}{\alpha(\epsilon)} \tag{12}$$

and

$$\limsup_{N \to \infty} \beta(N) \mathbb{P} \left\{ \frac{1}{N} \sum_{i=1}^{N} Y_i \leq -\epsilon \right\} \leq \frac{C}{\alpha(\epsilon)}, \tag{13}$$

where C, α and β are defined in Assumption 1.

Proof. By Chebyshev's inequality (Lemma 2.1), we have

$$\mathbb{P} \left\{ \frac{1}{N} \sum_{i=1}^{N} Y_i \geq \epsilon \right\} \leq \frac{\mathbb{E} \left[\alpha \left(\frac{1}{N} \sum_{i=1}^{N} Y_i \right) \right]}{\alpha(\epsilon)}$$

$$= \frac{\mathbb{E} \left[\beta(N) \alpha \left(\frac{1}{N} \sum_{i=1}^{N} Y_i \right) \right]}{\beta(N) \alpha(\epsilon)}.$$

Then we obtain

$$\beta(N) \mathbb{P} \left\{ \frac{1}{N} \sum_{i=1}^{N} Y_i \geq \epsilon \right\} \leq \frac{\mathbb{E} \left[\beta(N) \alpha \left(\frac{1}{N} \sum_{i=1}^{N} Y_i \right) \right]}{\alpha(\epsilon)}.$$

Taking the upper limit with respect to N on both sides, we obtain

$$\limsup_{N \to \infty} \beta(N) \mathbb{P} \left\{ \frac{1}{N} \sum_{i=1}^{N} Y_i \geq \epsilon \right\} \leq \frac{\limsup_{N \to \infty} \mathbb{E} \left[\beta(N) \alpha \left(\frac{1}{N} \sum_{i=1}^{N} Y_i \right) \right]}{\alpha(\epsilon)}$$

$$\leq \frac{C}{\alpha(\epsilon)},$$

which verify (12).

For (13), we have

$$\mathbb{P} \left\{ \frac{1}{N} \sum_{i=1}^{N} Y_i \leq -\epsilon \right\} = \mathbb{P} \left\{ -\frac{1}{N} \sum_{i=1}^{N} Y_i \geq \epsilon \right\}.$$

Then similar procedures can be conducted. □

The following theorem can be viewed as an extension of Theorem 2.3 from iid sampling to non-iid sampling.

Theorem 4.3. *Let $X_1, \cdots, X_N$ be N general (probably non-iid) samples of random variable X with $\mathbb{E}[X] = \mu$, and $\bar{X} := \frac{1}{N} \sum_{i=1}^{N} X_i$. Denote $Y_i = X_i - \mu$ for $i = 1, \cdots, N$. Suppose that Assumption 1 holds. Then for any $\epsilon > 0$ and $\Delta > 0$,*

$$\mathbb{P} \left\{ |\bar{X} - \mu| \geq \epsilon \right\} \leq \frac{C + \Delta}{\beta(N) \alpha(\epsilon)}$$

holds for sufficiently large N, where C, α and β are defined in Assumption 1.

Proof. Since $Y_i = X_i - \mu$, for $i = 1, \cdots, N$, satisfies Assumption 1, we have from (12) and (13) that

$$\limsup_{N \to \infty} \beta(N) \mathbb{P} \left\{ \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu) \geq \epsilon \right\} \leq \frac{C}{\alpha(\epsilon)}$$

and

$$\limsup_{N \to \infty} \beta(N) \mathbb{P} \left\{ \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu) \leq -\epsilon \right\} \leq \frac{C}{\alpha(\epsilon)}.$$

Thus,

$$\limsup_{N \to \infty} \beta(N) \mathbb{P} \left\{ \left| \bar{X} - \mu \right| \geq \epsilon \right\} \leq \frac{C}{\alpha(\epsilon)}.$$

According to the definition of superior limit, for any $\Delta > 0$ we have

$$\beta(N) \mathbb{P} \left\{ \left| \bar{X} - \mu \right| \geq \epsilon \right\} \leq \frac{C + \Delta}{\alpha(\epsilon)}$$

for sufficiently large N. $\qquad\square$

With similar processes, we can prove the following theorem about the uniform rate of convergence for random functions. All the notations, such as $\xi, \Xi, \mathcal{X}, F(x, \xi), f(x), \hat{f}_N(x)$ etc., are defined in Section 3.1. $\xi^1, \cdots, \xi^N$ are general samples of the random vector ξ.

Theorem 4.4. *Suppose that: (i) $\mathcal{X}$ is compact; (ii) $F(x, \xi^i) - f(x)$ for $i = 1, \cdots, N$ satisfy Assumption 1 for each $x \in \mathcal{X}$; (iii) $f(x)$ is continuous on $\mathcal{X}$. Let $\kappa : \Xi \to \mathbb{R}_+$ be a measurable function with $\kappa(\xi^i) - \mathbb{E}[\kappa(\xi)]$ for $i = 1, \cdots, N$ satisfying Assumption 1 and γ be a positive scalar.*

(a) If $F(\cdot, \xi)$ is H-calm from above over $\mathcal{X}$ with modulus $\kappa(\xi)$ and order γ, for arbitrary $\epsilon > 0$, there exists a $C_1 > 0$, independent of N, such that

$$\mathbb{P} \left\{ \sup_{x \in \mathcal{X}} \left(\hat{f}_N(x) - f(x) \right) \geq \epsilon \right\} \leq \frac{C_1}{\beta(N) \alpha(\epsilon)}$$

for sufficiently large N;

(b) If $F(\cdot, \xi)$ is H-calm from below over $\mathcal{X}$ with modulus $\kappa(\xi)$ and order γ, for arbitrary $\epsilon > 0$, there exists a $C_2 > 0$, independent of N, such that

$$\mathbb{P} \left\{ \inf_{x \in \mathcal{X}} \left(\hat{f}_N(x) - f(x) \right) \leq -\epsilon \right\} \leq \frac{C_2}{\beta(N) \alpha(\epsilon)}$$

for sufficiently large N;

(c) If $F(\cdot, \xi)$ is H-calm over $\mathcal{X}$ with modulus $\kappa(\xi)$ and order γ, for arbitrary $\epsilon > 0$, there exists a $C_3 > 0$, independent of N, such that

$$\mathbb{P} \left\{ \sup_{x \in \mathcal{X}} \left| \hat{f}_N(x) - f(x) \right| \geq \epsilon \right\} \leq \frac{C_3}{\beta(N) \alpha(\epsilon)}$$

for sufficiently large N.

We skip the proof, which can be similarly given as that of Theorem 3.2 based on Theorem 4.3.

Based on Theorem 3.5 and Theorem 4.4, we immediately have the following theorem about the uniform rate of convergence for the random set-valued mapping case. We adopt the same notation, such as $G(x, u, \xi), \mathcal{B}, \hat{\Phi}_N(x), \hat{\Phi}(x)$, as those in Section 3.2.

14

Theorem 4.5. *Suppose that: (i) $\mathcal{X}$ is compact; (ii) $\Phi(x, \xi)$ is nonempty, compact and convex for all $(x, \xi) \in \mathcal{X} \times \Xi$; (iii) $G(x, u, \xi^i) - \mathbb{E}[G(x, u, \xi)]$ for $i = 1, \cdots, N$ satisfy Assumption 1 for each $(x, u) \in \mathcal{X} \times \mathcal{B}$; (iv) $\mathbb{E}[G(\cdot, \cdot, \xi)]$ is continuous on $\mathcal{X} \times \mathcal{B}$. Let $\kappa : \Xi \to \mathbb{R}_+$ be a measurable function with $\kappa(\xi^i) - \mathbb{E}[\kappa(\xi)]$ for $i = 1, \cdots, N$ satisfying Assumption 1 and γ be a positive scalar.*

(a) If $G(\cdot, \cdot, \xi)$ is H-calm from above over $\mathcal{X} \times \mathcal{B}$ with modulus $\kappa(\xi)$ and order γ, for arbitrary $\epsilon > 0$, independent of N, there exists an $L_1 > 0$ such that

$$\mathbb{P}\left\{\sup_{x \in \mathcal{X}} \mathbb{D}(\hat{\Phi}_N(x), \hat{\Phi}(x)) \geq \epsilon\right\} \leq \frac{L_1}{\beta(N)\alpha(\epsilon)}$$

for sufficiently large N;

(b) If $G(\cdot, \cdot, \xi)$ is H-calm from below over $\mathcal{X} \times \mathcal{B}$ with modulus $\kappa(\xi)$ and order γ, for arbitrary $\epsilon > 0$, there exists an $L_2 > 0$, independent of N, such that

$$\mathbb{P}\left\{\sup_{x \in \mathcal{X}} \mathbb{D}(\hat{\Phi}(x), \hat{\Phi}_N(x)) \geq \epsilon\right\} \leq \frac{L_2}{\beta(N)\alpha(\epsilon)}$$

for sufficiently large N;

(c) If $G(\cdot, \cdot, \xi)$ is H-calm over $\mathcal{X} \times \mathcal{B}$ with modulus $\kappa(\xi)$ and order γ, for arbitrary $\epsilon > 0$, there exists an $L_3 > 0$, independent of N, such that

$$\mathbb{P}\left\{\sup_{x \in \mathcal{X}} \mathbb{H}(\hat{\Phi}_N(x), \hat{\Phi}(x)) \geq \epsilon\right\} \leq \frac{L_3}{\beta(N)\alpha(\epsilon)}$$

for sufficiently large N.

In this section, we have investigated the convergence rate under general sampling. We first establish a Gärtner–Ellis type theorem (Theorem 4.2), from which we derive the rate of convergence in the pointwise sense (Theorem 4.3). Finally, by analogous procedures as those in Theorems 3.2 and 3.5, we obtain the uniform rates of convergence for random functions (Theorem 4.4) and random set-valued mappings (Theorem 4.5), respectively.

5   SAA of stochastic optimization

Discretization is quite important because it concerns numerical solvability of stochastic optimization problems. It can help to avoid high dimensional integrations and infinite constraints. The crucial question is how large the sample size should be to ensure certain approximation accuracy. The general methodology is to employ the Cramér's large deviation theorem, so that the exponential rate of convergence with respect to the sample size can be derived. However, this relies on the light tailed distribution assumption, which limits the scope of application. In this section, we will show how the results got in previous sections can be applied to establish the convergence rate conclusions of SAA for several kinds of stochastic optimization problems with iid or non-iid samplings under heavy tailed distributions.

5.1   Stochastic convex optimization problems

Consider $F : \mathcal{X} \times \Xi \to \mathbb{R}$, here $F(\cdot, \xi)$ is convex for almost everywhere (a.e. for short) $\xi \in \Xi$. Assume that $\mathcal{X} \subseteq \mathbb{R}^n$ is compact. Then we consider the following stochastic optimization problem (see [2, 38, 39]):

$$\min_{x \in \mathcal{X}} \mathbb{E}[F(x, \xi)]. \tag{14}$$

Problem (14) is a generic form which includes many specific stochastic optimization models, such as two-stage stochastic linear programming problems. In that case, we have

$$F(x, \xi) = c^\top x + Q(x, \xi) \tag{15}$$

and

$$Q(x, \xi) := \inf\{\langle q(\xi), y(\xi) \rangle : W(\xi)y(\xi) + T(\xi)x = h(\xi), y(\xi) \geq 0\}, \tag{16}$$

where $c \in \mathbb{R}^n$, $q : \Xi \to \mathbb{R}^m$, $W : \Xi \to \mathbb{R}^{r \times m}$, $T : \Xi \to \mathbb{R}^{r \times n}$ and $h : \Xi \to \mathbb{R}^r$. Problem (14)-(16) is called the full random two-stage stochastic linear programming problem because the recourse matrix $W(\xi)$ is also random, see [40, 41] for more details. When $W(\xi) = W$ is fixed, the resulting two-stage stochastic linear programming problem has been widely discussed, see e.g. [42, 2]. For both the deterministic recourse matrix W and the random recourse matrix $W(\xi)$, it is easy to verify that $F(\cdot, \xi)$ is convex with respect to x. That is, problem (14) covers both the fixed recourse case and the random recourse case.

Assume that we have random samples $\xi^1, \cdots, \xi^N$. Then we obtain the SAA problem of problem (14):

$$\min_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^{N} F(x, \xi^i). \tag{17}$$

Due to the convexity of $F(\cdot, \xi)$, both the objective functions in problems (14) and (17) are convex.

We denote by S^* and S_N the optimal solution sets of problems (14) and (17), and v^* and v_N the optimal values of problems (14) and (17), respectively. In order to quantify the upper semicontinuity property, we define the general growth function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ as follows:

$$\psi(\tau) := \inf\{\mathbb{E}[F(x, \xi)] - v^* : d(x, S^*) \geq \tau, x \in \mathcal{X}\}.$$

Its inverse function is defined as

$$\psi^{-1}(\eta) := \sup\{\tau \in \mathbb{R}_+ : \psi(\tau) \leq \eta\}.$$

Then we define the conditioning function $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ as

$$\Psi(\eta) := \eta + \psi^{-1}(2\eta).$$

Obviously, ψ is lower semicontinuous on $\mathbb{R}_+$, nondecreasing and vanish at 0. Ψ is lower semicontinuous on $\mathbb{R}_+$, increasing and vanish at 0. For more details, one can refer to [43].

Theorem 5.1. *Suppose that: (i) $\mathcal{X}$ is compact; (ii) $F(\cdot, \xi)$ is H-calm over $\mathcal{X}$ with modulus $\kappa(\xi)$ and order $\gamma > 0$ where $\mathbb{E}[\kappa(\xi)] < \infty$.*

(a) If $\xi^1, \cdots, \xi^N$ are iid, $\mathbb{E}[F(x, \xi)^p] < \infty$ for each $x \in \mathcal{X}$ and $\mathbb{E}[\kappa(\xi)^p] < \infty$ with some $p \geq 2$, then for any $\epsilon > 0$, there exists a positive C, independent of N, such that

$$\mathbb{P}\{|v_N - v^*| \geq \epsilon\} \leq \frac{C}{N^{\frac{p}{2}}\epsilon^p},$$

$$\mathbb{P}\{\mathbb{D}(S_N, S^*) \geq \Psi(\epsilon)\} \leq \frac{C}{N^{\frac{p}{2}}\epsilon^p}$$

for sufficiently large N;

(b) If $\xi^1, \cdots, \xi^N$ are non-iid, $F(x, \xi^i) - \mathbb{E}[F(x, \xi)]$, $i = 1, \cdots, N$, for each $x \in \mathcal{X}$ and $\kappa(\xi^i) - \mathbb{E}[\kappa(\xi)]$ for $i = 1, \cdots, N$ satisfy Assumption 1, then for any $\epsilon > 0$, there exists a positive constant C, independent of N, such that

$$\mathbb{P}\left\{|v_N - v^*| \geq \epsilon\right\} \leq \frac{C}{\beta(N)\alpha(\epsilon)},$$

$$\mathbb{P}\left\{\mathbb{D}(S_N, S^*) \geq \Psi(\epsilon)\right\} \leq \frac{C}{\beta(N)\alpha(\epsilon)}$$

for sufficiently large N, where $\alpha(\cdot)$ and $\beta(\cdot)$ are defined in Assumption 1.

Proof. We only give the proof of part (a) since part (b) can be similarly verified by Theorem 4.4. Notice that

$$|v_N - v^*| = \left| \min_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^{N} F(x, \xi^i) - \min_{x \in \mathcal{X}} \mathbb{E}[F(x, \xi)] \right|$$

$$\leq \max_{x \in \mathcal{X}} \left| \frac{1}{N} \sum_{i=1}^{N} F(x, \xi^i) - \mathbb{E}[F(x, \xi)] \right|.$$

It is known from (c) of Theorem 3.2 that

$$\mathbb{P}\left\{ \max_{x \in \mathcal{X}} \left| \frac{1}{N} \sum_{i=1}^{N} F(x, \xi^i) - \mathbb{E}[F(x, \xi)] \right| \geq \epsilon \right\} \leq \frac{C}{N^{\frac{p}{2}} \epsilon^p} \tag{18}$$

for some constant C independent of N and sufficiently large N. Noticing that

$$\mathbb{P}\left\{|v_N - v^*| \geq \epsilon\right\} \leq \mathbb{P}\left\{ \max_{x \in \mathcal{X}} \left| \frac{1}{N} \sum_{i=1}^{N} F(x, \xi^i) - \mathbb{E}[F(x, \xi)] \right| \geq \epsilon \right\},$$

we have

$$\mathbb{P}\left\{|v_N - v^*| \geq \epsilon\right\} \leq \frac{C}{N^{\frac{p}{2}} \epsilon^p}.$$

For any $\tilde{x} \in S_N$, we have

$$2 \max_{x \in \mathcal{X}} \left| \frac{1}{N} \sum_{i=1}^{N} F(x, \xi^i) - \mathbb{E}[F(x, \xi)] \right| \geq \max_{x \in \mathcal{X}} \left| \frac{1}{N} \sum_{i=1}^{N} F(x, \xi^i) - \mathbb{E}[F(x, \xi)] \right| + v_N - v^*$$

$$\geq \mathbb{E}[F(\tilde{x}, \xi)] - \frac{1}{N} \sum_{i=1}^{N} F(\tilde{x}, \xi^i) + v_N - v^*$$

$$= \mathbb{E}[F(\tilde{x}, \xi)] - v^* \geq \psi(d(\tilde{x}, S^*)).$$

We obtain

$$d(\tilde{x}, S^*) \leq \psi^{-1}\left(2 \max_{x \in \mathcal{X}} \left| \frac{1}{N} \sum_{i=1}^{N} F(x, \xi^i) - \mathbb{E}[F(x, \xi)] \right| \right)$$

$$\leq \Psi\left(\max_{x \in \mathcal{X}} \left| \frac{1}{N} \sum_{i=1}^{N} F(x, \xi^i) - \mathbb{E}[F(x, \xi)] \right| \right).$$

Due to the arbitrariness of $\tilde{x} \in S_N$, we actually have

$$\mathbb{D}(S_N, S^*) \leq \Psi\left(\max_{x \in \mathcal{X}} \left| \frac{1}{N} \sum_{i=1}^{N} F(x, \xi^i) - \mathbb{E}[F(x, \xi)] \right| \right).$$

Then

$$\mathbb{P}\left\{\mathbb{D}(S_N, S^*) \geq \Psi(\epsilon)\right\}$$

$$\leq \mathbb{P}\left\{\Psi\left(\max_{x \in \mathcal{X}}\left|\frac{1}{N}\sum_{i=1}^{N}F(x,\xi^i) - \mathbb{E}[F(x,\xi)]\right|\right) \geq \Psi(\epsilon)\right\}$$

$$= \mathbb{P}\left\{\max_{x \in \mathcal{X}}\left|\frac{1}{N}\sum_{i=1}^{N}F(x,\xi^i) - \mathbb{E}[F(x,\xi)]\right| \geq \epsilon\right\}$$

$$\leq \frac{C}{N^{\frac{p}{2}}\epsilon^p},$$

where the last inequality follows from (18). □

Different from the existing results, such as [9, 6, 3, 15], where the exponential rates of convergence are derived, Theorem 5.1 establishes the polynomial rate of convergence. The main advantage is that our results do not require the light tailed distribution assumption.

5.2   Stochastic nonconvex optimization problems

Quite often, the stochastic optimization problem (14) may be a nonconvex optimization problem in practice. In that case, we can hardly find a global optimal solution, or even a local optimal solution. Instead, only some stationary points can be derived through the first order optimality condition. Stationary points may not be the optimal solution. However, under some constraint qualifications, the optimal solution must be a stationary point. Therefore, in the nonconvex case, stationary points stand for the optima to some extent. In view of these, we consider the nonconvex case and study the limiting behavior of stationary points of the corresponding SAA problem in this section.

For convenience, we assume in this section that $F(\cdot,\xi)$ is locally Lipschitz continuous (probably nonconvex) for a.e. $\xi \in \Xi$, and the feasible set $\mathcal{X}$ is closed and convex. Thus, $\mathbb{E}[F(\cdot,\xi)]$ is also locally Lipschitz, and the normal cone of $\mathcal{X}$ with respect to $x \in \mathcal{X}$ is unique, denoted by $\mathcal{N}_\mathcal{X}(x) := \{v : v^\top(y-x) \leq 0, \ \forall y \in \mathcal{X}\}$. Then, according to [32, Theorem 10.1], we have the following first order optimality condition for problem (14):

$$0 \in \partial\mathbb{E}[F(x,\xi)] + \mathcal{N}_\mathcal{X}(x), \tag{19}$$

where $\partial\mathbb{E}[F(x,\xi)]$ stands for the Clarke subdifferential [44]. Without any confusion, we write ∂ in short of ∂_x hereinafter. We call x a stationary point of problem (14) if it satisfies (19). Obviously, when $x^* \in \mathcal{X}$ is an optimal solution of problem (14), it must be a stationary point of problem (14), but not vice versa. It knows from [9, Theorem 2.1] that

$$\partial\mathbb{E}[F(x,\xi)] \subseteq \mathbb{E}[\partial F(x,\xi)].$$

The equality holds when $F(x,\xi)$ is Clarke regular [44, Definition 2.3.4] at x. In that case, $\mathbb{E}[F(x,\xi)]$ is also Clarke regular at x.

Generally, it is more interesting for us to consider the following weaker optimality condition:

$$0 \in \mathbb{E}[\partial F(x,\xi)] + \mathcal{N}_\mathcal{X}(x). \tag{20}$$

We say an x satisfying (20) is a weak stationary point of problem (14). Obviously, a stationary point of problem (14) is a weak stationary point of problem (14), but not vice versa. If a weak stationary point of problem (14) satisfies the Clarke regularity, it is a stationary point of problem (14).

Then SAA problem of (20) can be written as follows:

$$0 \in \frac{1}{N} \sum_{i=1}^{N} \partial F(x, \xi^i) + \mathcal{N}_{\mathcal{X}}(x). \tag{21}$$

In what follows, we consider the convergence rate of the solution set to (21) to that of (20) by employing our results in last sections. To this end, we first investigate the convergence relationship between $\frac{1}{N} \sum_{i=1}^{N} \partial F(x, \xi^i)$ and $\mathbb{E}[\partial F(x, \xi)]$ based on Theorems 3.5 and 4.5.

Proposition 5.2. *Suppose that: (i) $\mathcal{X}$ is convex and compact; (ii) $F(\cdot, \xi)$ is Lipschitz continuous over $\mathcal{X}$ with modulus $\kappa_1(\xi)$ and $\mathbb{E}[\kappa_1(\xi)] < \infty$; (iii) for each given $u \in \mathcal{B}$, $\mathbb{E}[\sigma(u, \partial F(\cdot, \xi))]$ is continuous with respect to x; (iv) for each given $u \in \mathcal{B}$, $\sigma(u, \partial F(\cdot, \xi))$ is H-calm from above over $\mathcal{X}$ with modulus $\kappa_2(\xi)$ and order $\gamma > 0$. Let $\kappa(\xi) := \kappa_1(\xi) + \kappa_2(\xi)$.*

(a) If $\xi^1, \cdots, \xi^N$ are iid, $\mathbb{E}[\sigma(u, \partial F(x, \xi))^p] < \infty$ for each $x \in \mathcal{X}$ and $\mathbb{E}[\kappa(\xi)^p] < \infty$ for some $p \geq 2$, then for any $\epsilon > 0$, there exists a positive C, independent of N, such that

$$\mathbb{P}\left\{ \sup_{x \in \mathcal{X}} \mathbb{D}\left(\frac{1}{N} \sum_{i=1}^{N} \partial F(x, \xi^i), \mathbb{E}[\partial F(x, \xi)] \right) \geq \epsilon \right\} \leq \frac{C}{N^{\frac{p}{2}} \epsilon^p}$$

for sufficiently large N;

(b) If $\xi^1, \cdots, \xi^N$ are non-iid, both $F(x, \xi^i) - \mathbb{E}[F(x, \xi)], i = 1, \cdots, N$, for all $x \in \mathcal{X}$ and $\kappa(\xi^i) - \mathbb{E}[\kappa(\xi)], i = 1, \cdots, N$ satisfy Assumption 1, then for any $\epsilon > 0$, there exists a positive C, independent of N, such that

$$\mathbb{P}\left\{ \sup_{x \in \mathcal{X}} \mathbb{D}\left(\frac{1}{N} \sum_{i=1}^{N} \partial F(x, \xi^i), \mathbb{E}[\partial F(x, \xi)] \right) \geq \epsilon \right\} \leq \frac{C}{\beta(N)\alpha(\epsilon)}$$

for sufficiently large N, where $\alpha(\cdot)$ and $\beta(\cdot)$ are defined in Assumption 1.

Proof. Part (a): Since $\partial F(x, \xi)$ is nonempty, convex and compact ([44, Proposition 2.1.2]), we know from Section 3.2 that

$$\mathbb{P}\left\{ \mathbb{D}\left(\frac{1}{N} \sum_{i=1}^{N} \partial F(x, \xi^i), \mathbb{E}[\partial F(x, \xi)] \right) \geq \epsilon \right\}$$

$$= \mathbb{P}\left\{ \max_{\|u\| \leq 1} \left(\frac{1}{N} \sum_{i=1}^{N} \sigma\left(u, \partial F(x, \xi^i)\right) - \mathbb{E}\left[\sigma\left(u, \partial F(x, \xi)\right)\right] \right) \geq \epsilon \right\}.$$

Note that

$$\sigma\left(u', \partial F(x', \xi)\right) - \sigma\left(u, \partial F(x, \xi)\right)$$

$$= \sup_{\phi(x', \xi) \in \partial F(x', \xi)} (u')^\top \phi(x', \xi) - \sup_{\phi(x, \xi) \in \partial F(x, \xi)} u^\top \phi(x, \xi)$$

$$= \sup_{\phi(x', \xi) \in \partial F(x', \xi)} (u')^\top \phi(x', \xi) - \sup_{\phi(x', \xi) \in \partial F(x', \xi)} u^\top \phi(x', \xi) \tag{22}$$

$$+ \sup_{\phi(x', \xi) \in \partial F(x', \xi)} u^\top \phi(x', \xi) - \sup_{\phi(x, \xi) \in \partial F(x, \xi)} u^\top \phi(x, \xi).$$

Then, we obtain from (22) that

$$\left| \mathbb{E}\left[\sigma\left(u', \partial F(x', \xi)\right)\right] - \mathbb{E}\left[\sigma\left(u, \partial F(x, \xi)\right)\right] \right|$$

$$\leq \mathbb{E}[\kappa_1(\xi)] \left\| u' - u \right\| + \left| \mathbb{E}\left[\sigma\left(u, \partial F(x', \xi)\right)\right] - \mathbb{E}\left[\sigma\left(u, \partial F(x, \xi)\right)\right] \right|.$$

19

Since $\mathbb{E}[\sigma(u, \partial F(\cdot, \xi))]$ is continuous with respect to x for each $u \in \mathcal{B}$, $\mathbb{E}[\sigma(\cdot, \partial F(\cdot, \xi))]$ is continuous with respect to (u, x).

Moreover, we can also know from (22) that

$$
\begin{aligned}
\sigma\left(u', \partial F(x', \xi)\right) &- \sigma\left(u, \partial F(x, \xi)\right) \\
&\leq \kappa_1(\xi)\left\|u' - u\right\| + \sigma\left(u, \partial F(x', \xi)\right) - \sigma\left(u, \partial F(x, \xi)\right),
\end{aligned}
$$

which implies that $\sigma(u, \partial F(\cdot, \xi))$ is H-calm from above over $\mathcal{X}$ with modulus $\kappa_1(\xi) + \kappa_2(\xi)$ and order $\min\{1, \gamma\}$. Then the assertion directly follows from Theorem 3.5.

Part (b): Since the continuity of $\mathbb{E}[\sigma(\cdot, \partial F(\cdot, \xi))]$ with respect to (u, x) and the H-calmness of $\sigma(u, \partial F(\cdot, \xi))$ from above over $\mathcal{X}$, we can conclude the assertion from Theorem 4.5. $\qquad\square$

Denote the solution sets of problems (20) and (21) by S^* and S_N, respectively. To establish the relationships between S^*, S_N and $\mathbb{E}[\partial F(x, \xi)], \frac{1}{N}\sum_{i=1}^N \partial F(x, \xi^i)$, we need the following general growth function of (20), denoted by $\mathcal{G} : \mathbb{R}_+ \to \mathbb{R}_+$. Specifically,

$$
\mathcal{G}(\tau) := \inf\left\{d\left(0, \mathbb{E}[\partial F(x, \xi)] + \mathcal{N}_{\mathcal{X}}(x)\right) : x \in \mathcal{X}, d(x, S^*) \geq \tau\right\}.
$$

Its inverse function is

$$
\mathcal{G}^{-1}(\eta) := \sup\left\{\tau : \mathcal{G}(\tau) \leq \eta\right\}.
$$

Then we denote

$$
\theta(\eta) = \eta + \mathcal{G}^{-1}(\eta).
$$

Thus, $\theta(\eta) \geq \mathcal{G}^{-1}(\eta)$ and $\theta(\eta)$ is increasing. Both $\theta(\eta)$ and $\mathcal{G}^{-1}(\eta)$ vanish at 0.

We have for any $\tilde{x} \in S_N$ the following key inequality:

$$
\begin{aligned}
\sup_{x \in \mathcal{X}} &\mathbb{D}\left(\frac{1}{N}\sum_{i=1}^N \partial F(x, \xi^i), \mathbb{E}[\partial F(x, \xi)]\right) \\
&\geq \sup_{x \in \mathcal{X}} \mathbb{D}\left(\frac{1}{N}\sum_{i=1}^N \partial F(x, \xi^i) + \mathcal{N}_{\mathcal{X}}(x), \mathbb{E}[\partial F(x, \xi)] + \mathcal{N}_{\mathcal{X}}(x)\right) \\
&\geq \mathbb{D}\left(\frac{1}{N}\sum_{i=1}^N \partial F(\tilde{x}, \xi^i) + \mathcal{N}_{\mathcal{X}}(\tilde{x}), \mathbb{E}[\partial F(\tilde{x}, \xi)] + \mathcal{N}_{\mathcal{X}}(\tilde{x})\right).
\end{aligned}
$$

Consider two closed sets $A, B \subseteq \mathbb{R}^n$ and $\bar{a} \in A$ such that $d(0, \bar{a}) = d(0, A)$. There exists $\bar{b} \in B$ such that $d(\bar{a}, B) = d(\bar{a}, \bar{b})$. Then we have

$$
\begin{aligned}
\mathbb{D}(A, B) &\geq d(\bar{a}, B) = d(\bar{a}, \bar{b}) \\
&\geq d(0, \bar{b}) - d(0, \bar{a}) \\
&= d(0, \bar{b}) - d(0, A) \\
&\geq d(0, B) - d(0, A).
\end{aligned}
$$

With the aid of the above fact, we have

$$
\begin{aligned}
\mathbb{D}&\left(\frac{1}{N}\sum_{i=1}^N \partial F(\tilde{x}, \xi^i) + \mathcal{N}_{\mathcal{X}}(\tilde{x}), \mathbb{E}[\partial F(\tilde{x}, \xi)] + \mathcal{N}_{\mathcal{X}}(\tilde{x})\right) \\
&\geq d\left(0, \mathbb{E}[\partial F(\tilde{x}, \xi)] + \mathcal{N}_{\mathcal{X}}(\tilde{x})\right) - d\left(0, \frac{1}{N}\sum_{i=1}^N \partial F(\tilde{x}, \xi^i) + \mathcal{N}_{\mathcal{X}}(\tilde{x})\right) \\
&= d\left(0, \mathbb{E}[\partial F(\tilde{x}, \xi)] + \mathcal{N}_{\mathcal{X}}(\tilde{x})\right) \\
&\geq \mathcal{G}\left(d(\tilde{x}, S^*)\right).
\end{aligned}
$$

Due the arbitrariness of $\tilde{x} \in S_N$, we obtain

$$
\mathbb{D}(S_N, S^*) \leq \mathcal{G}^{-1}\left(\sup_{x \in \mathcal{X}} \mathbb{D}\left(\frac{1}{N}\sum_{i=1}^{N}\partial F(x, \xi^i), \mathbb{E}[\partial F(x, \xi)]\right)\right)
$$

$$
\leq \theta\left(\sup_{x \in \mathcal{X}} \mathbb{D}\left(\frac{1}{N}\sum_{i=1}^{N}\partial F(x, \xi^i), \mathbb{E}[\partial F(x, \xi)]\right)\right).
$$

Then, for any $\epsilon > 0$,

$$
\mathbb{P}\left\{\mathbb{D}(S_N, S^*) \geq \theta(\epsilon)\right\} \leq \mathbb{P}\left\{\theta\left(\sup_{x \in \mathcal{X}} \mathbb{D}\left(\frac{1}{N}\sum_{i=1}^{N}\partial F(x, \xi^i), \mathbb{E}[\partial F(x, \xi)]\right)\right) \geq \theta(\epsilon)\right\}
$$

$$
= \mathbb{P}\left\{\sup_{x \in \mathcal{X}} \mathbb{D}\left(\frac{1}{N}\sum_{i=1}^{N}\partial F(x, \xi^i), \mathbb{E}[\partial F(x, \xi)]\right) \geq \epsilon\right\}. \tag{23}
$$

With the quantitative relationship (23) and Proposition 5.2, we have the following theorem.

Theorem 5.3. *Let assumptions in Proposition 5.2 hold.*

(i) *If $\xi^1, \cdots, \xi^N$ are iid, $\mathbb{E}[\sigma(u, \partial F(x, \xi))^p] < \infty$ for each $x \in \mathcal{X}$ and $\mathbb{E}[\kappa(\xi)^p] < \infty$ with some $p \geq 2$, then for any $\epsilon > 0$, there exists a positive C, independent of N, such that*

$$
\mathbb{P}\left\{\mathbb{D}(S_N, S^*) \geq \theta(\epsilon)\right\} \leq \frac{C}{N^{\frac{p}{2}}\epsilon^p}
$$

for sufficiently large N;

(ii) *If $\xi^1, \cdots, \xi^N$ are non-iid, both $F(x, \xi^i) - \mathbb{E}[F(x, \xi)], i = 1, \cdots, N$, for all $x \in \mathcal{X}$ and $\kappa(\xi^i) - \mathbb{E}[\kappa(\xi)], i = 1, \cdots, N$, satisfy Assumption 1, then for any $\epsilon > 0$, there exists a positive C, independent of N, such that*

$$
\mathbb{P}\left\{\mathbb{D}(S_N, S^*) \geq \theta(\epsilon)\right\} \leq \frac{C}{\beta(N)\alpha(\epsilon)}
$$

for sufficiently large N, where $\alpha(\cdot)$ and $\beta(\cdot)$ are defined in Assumption 1.

5.3 Two-stage stochastic variational inequalities

Two-stage stochastic variational inequalities can be reformulated as (see e.g. [15, 45, 46]):

$$
\begin{aligned}
&0 \in \mathbb{E}[\Phi_1(x, y(\xi), \xi)] + \mathcal{N}_{\mathcal{X}}(x), \\
&0 \in \Phi_2(x, y(\xi), \xi) + \mathcal{N}_{\mathcal{Y}(\xi)}(y(\xi)), \text{ for a.e. } \xi \in \Xi,
\end{aligned} \tag{24}
$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ is convex; $\mathcal{Y} : \Xi \rightrightarrows \mathbb{R}^m$ is a convex set-valued mapping; $\Phi_1 : \mathbb{R}^n \times \mathbb{R}^m \times \Xi \to \mathbb{R}^n$, $\Phi_2 : \mathbb{R}^n \times \mathbb{R}^m \times \Xi \to \mathbb{R}^m$.

Problem (24) is a quite general form, which includes many applications, such as game theory, optimality condition for two-stage stochastic programming, Cournot-Nash equilibrium [15, 47, 48, 49, 50]. Especially, when $\mathcal{X} = \mathbb{R}^n_+$ and $\mathcal{Y}(\xi) = \mathbb{R}^m_+$, the resulting problem is the two-stage stochastic complementarity problem (see [47] for the linear case):

$$
\begin{aligned}
&0 \leq x \perp \mathbb{E}[\Phi_1(x, y(\xi), \xi)] \geq 0, \\
&0 \leq y(\xi) \perp \Phi_2(x, y(\xi), \xi) \geq 0, \text{ for a.e. } \xi \in \Xi.
\end{aligned}
$$

21

Based on samples $\xi^1, \cdots, \xi^N$, the SAA approximation problem of (24) can be written as

$$0 \in \frac{1}{N} \sum_{i=1}^{N} \Phi_1(x, y(\xi^i), \xi^i) + \mathcal{N}_{\mathcal{X}}(x), \ x \in \mathcal{X}$$

$$0 \in \Phi_2(x, y(\xi^i), \xi^i) + \mathcal{N}_{\mathcal{Y}(\xi^i)}(y(\xi^i)), \ \text{for } i = 1, \cdots, N. \tag{25}$$

In what follows we focus on the convergence analysis between problems (24) and (25). By convention, we make the following strong monotonicity assumption for the second stage problem (see [51, Definition 2.3.1]).

Assumption 2 (strong monotonicity). *For each fixed $x \in \mathcal{X}$ and a.e. $\xi \in \Xi$, $\Phi_2(x, \cdot, \xi)$ is strongly monotone, that is, there exists a positive measurable function $\kappa_y(\xi)$ such that for all $y_1, y_2 \in \mathbb{R}^m$*

$$\langle y_1 - y_2, \Phi_2(x, y_1, \xi) - \Phi_2(x, y_2, \xi) \rangle \geq \kappa_y(\xi) \|y_1 - y_2\|^2.$$

It is noteworthy that: although the second stage problem is strongly monotone under Assumption 2, the two-stage stochastic variational inequalities (24) could be nonmonotone.

Immediately, we have from [51, Theorem 2.3.3] the following lemma.

Lemma 5.4. *Let Assumption 2 hold and $\Phi_2(x, \cdot, \xi)$ be continuous. For $x \in \mathcal{X}$ and a.e. $\xi \in \Xi$, the second stage problem*

$$0 \in \Phi_2(x, y(\xi), \xi) + \mathcal{N}_{\mathcal{Y}(\xi)}(y(\xi)) \tag{26}$$

has a unique solution.

To derive the Lipschitz continuity property, for any fixed $x \in \mathcal{X}$ and $\xi \in \Xi$, we introduce the several concepts of regularity. The first one is the strong regularity proposed by Robinson [52].

Definition 5.5 (strong regularity, [52]). *For any fixed $x \in \mathcal{X}$ and $\xi \in \Xi$, the second stage problem (26) is said to be strongly regular at a solution $\bar{y}$ if $\Phi_2(x, \cdot, \xi)$ is differentiable at $\bar{y}$ and there exist neighborhoods $\mathcal{U}$ of $\bar{y}$ and $\mathcal{V}$ of 0 such that for every $\delta \in \mathcal{V}$, the perturbed (partially) linearized problem*

$$\delta \in \Phi_2(x, \bar{y}, \xi) + \nabla_y \Phi_2(x, \bar{y}, \xi)(y - \bar{y}) + \mathcal{N}_{\mathcal{Y}(\xi)}(y)$$

has in $\mathcal{U}$ a unique solution $y(\delta)$, and the mapping $\delta \to y(\delta) : \mathcal{V} \to \mathcal{U}$ is Lipschitz continuous.

Izmailov extended the strong regularity from the smooth case to the locally Lipschitz continuous case in [53], which is called the CD-regularity.

Definition 5.6 (CD-regularity, [53]). *For any fixed $x \in \mathcal{X}$ and $\xi \in \Xi$, a solution $\bar{y}$ of the second stage problem (26) is said to be CD-regular if $\Phi_2(x, \cdot, \xi)$ is locally Lipschitz continuous at a neighborhood of $\bar{y}$ and for each $J \in \partial_y \Phi_2(x, \bar{y}, \xi)$ where ∂_y denotes the Clarke subdifferential with respect to y, the following linear variational inequality*

$$0 \in \Phi_2(x, \bar{y}, \xi) + J(y - \bar{y}) + \mathcal{N}_{\mathcal{Y}(\xi)}(y)$$

is strongly regular at the solution $\bar{y}$.

We use the notation Π_y to denote the projection onto y-space and $\partial_{(x,y)}$ to denote the Clarke subdifferential with respect to (x, y).

Definition 5.7 (parametrically CD-regularity, [53]). *For any fixed $\xi \in \Xi$, a solution $\bar{y}$ of the second stage problem*

$$0 \in \Phi_2(x, y, \xi) + \mathcal{N}_{\mathcal{Y}(\xi)}(y)$$

for $x = \bar{x}$ is said to be parametrically CD-regular if $\Phi_2(\cdot, \cdot, \xi)$ is locally Lipschitz continuous at a neighborhood of $(\bar{x}, \bar{y})$ and for each $J \in \Pi_y \partial_{(x,y)} \Phi_2(\bar{x}, \bar{y}, \xi)$ the solution $\bar{y}$ of

$$0 \in \Phi_2(\bar{x}, \bar{y}, \xi) + J(y - \bar{y}) + \mathcal{N}_{\mathcal{Y}(\xi)}(y)$$

is strongly regular.

Proposition 5.8. *Suppose that: (i) $\Phi_2(\cdot, \cdot, \xi)$ is Lipschitz continuous on $\mathbb{R}^n \times \mathbb{R}^m$; (ii) Assumption 2 holds; (iii) $\mathcal{Y}(\xi)$ is nonempty and polyhedral for a.e. $\xi \in \Xi$. Then the unique solution of problem (26), denoted by $\hat{y}(x, \xi)$, is parametrically CD-regular for $x \in \mathcal{X}$ and a.e. $\xi \in \Xi$. Further, $\hat{y}(\cdot, \xi)$ is Lipschitz continuous on $\mathcal{X} \cap \mathcal{C}$ for a.e. $\xi \in \Xi$ where $\mathcal{C} \subseteq \mathbb{R}^n$ is an arbitrary compact set such that $\mathcal{X} \cap \mathcal{C} \neq \emptyset$.*

Proof. We know from Lemma 5.4 that for any fixed $x \in \mathcal{X}$ and a.e. $\xi \in \Xi$, there exists a unique solution for (26). Due to the Lipschitz continuity of $\Phi_2(\cdot, \cdot, \xi)$ over $\mathbb{R}^n \times \mathbb{R}^m$, $\Phi_2(\cdot, \cdot, \xi)$ is differentiable densely over $\mathbb{R}^n \times \mathbb{R}^m$. We denote by $\mathcal{D} \subseteq \mathbb{R}^n \times \mathbb{R}^m$ the collection of all differentiable points. For any fixed $\xi \in \Xi$, $\bar{x} \in \mathcal{X}$ and $\bar{y} = \hat{y}(\bar{x}, \xi)$, we consider the following linear variational inequality:

$$0 \in \Phi_2(\bar{x}, \bar{y}, \xi) + J(y - \bar{y}) + \mathcal{N}_{\mathcal{Y}(\xi)}(y), \tag{27}$$

where $J \in \Pi_y \partial_{(x,y)} \Phi_2(\bar{x}, \bar{y}, \xi)$. According to the definition of Clarke subdifferential, there exists a sequence $\{(x_k, y_k)\}_k \subseteq \mathcal{D}$ with $(x_k, y_k) \to (\bar{x}, \bar{y})$ as $k \to \infty$ such that

$$J = \lim_{k \to \infty} \Pi_y \nabla_{(x,y)} \Phi_2(x_k, y_k, \xi).$$

Since $\Phi_2(x, \cdot, \xi)$ is strongly monotone with constant $\kappa_y(\xi)$ for each $x \in \mathcal{X}$ and a.e. $\xi \in \Xi$ due to Assumption 2, we know from [51, Proposition 2.3.2] that $\Pi_y \nabla_{(x,y)} \Phi_2(x_k, y_k, \xi)$ is strongly positive definite with constant $\kappa_y(\xi)$. This implies that any $J \in \Pi_y \partial_{(x,y)} \Phi_2(\bar{x}, \bar{y}, \xi)$ is strongly positive definite with constant $\kappa_y(\xi)$.

The positive definiteness of J together with [54, Theorem 1 and Corollary 1] mean that (27) is strongly regular at $\bar{y}$. Then we obtain that (26) is parametrically CD-regular at $\bar{y}$ with $x = \bar{x}$. The locally Lipschitz continuity of $\hat{y}(\cdot, \xi)$ at a neighborhood of $\bar{x}$ follows from [53, Theorem 4]. Finally, due to the arbitrariness of $\bar{x}$ and the compactness of $\mathcal{X} \cap \mathcal{C}$, $\hat{y}(\cdot, \xi)$ is Lipschitz continuous over $\mathcal{X} \cap \mathcal{C}$. $\qquad \square$

Although a similar result can be found in [15, Theorem 3.6], Proposition 5.8 differs from [15, Theorem 3.6] in two aspects. First, [15, Theorem 3.6] considers a nonlinear complementarity problem, and here we consider a nonlinear variational inequality, which of course is more general. Second, we do not assume the continuous differentiability of $\Phi_2(\cdot, \cdot, \xi)$ over $\mathbb{R}^n \times \mathbb{R}^m$, which was employed in [15, Theorem 3.6].

Now, under Assumption 2 and Lemma 5.4, we can equivalently rewrite problems (24) and (25) as

$$0 \in \mathbb{E}[\Phi_1(x, \hat{y}(x, \xi), \xi)] + \mathcal{N}_{\mathcal{X}}(x), \tag{28}$$

and

$$0 \in \frac{1}{N} \sum_{i=1}^{N} \Phi_1(x, \hat{y}(x, \xi^i), \xi^i) + \mathcal{N}_{\mathcal{X}}(x), \tag{29}$$

respectively. Therefore, the convergence analysis between problems (24) and (25) is equivalent to the convergence analysis between problems (28) and (29). We use S^* and

S_N to denote the solution sets of problems (28) and (29), respectively. To quantify the rate of convergence between S^* and S_N, we first consider the rate of convergence between $\frac{1}{N}\sum_{i=1}^{N}\Phi_1(x,\hat{y}(x,\xi^i),\xi^i)$ and $\mathbb{E}[\Phi_1(x,\hat{y}(x,\xi),\xi)]$.

We know from Proposition 5.8 that if $\mathcal{X}$ is a compact and convex set, and $\Phi_1(\cdot,\cdot,\xi)$ is Lipschitz continuous over $\mathbb{R}^n\times\mathbb{R}^m$, $\Phi_1(\cdot,\hat{y}(\cdot,\xi),\xi)$ is Lipschitz continuous with respect to x over $\mathcal{X}$. Denote by $\kappa(\xi)$ the Lipschitz modulus in what follows.

Proposition 5.9. *Let assumptions in Proposition 5.8 hold, and $\mathcal{X}$ be a compact and convex set.*

(a) *Suppose that: (i) $\mathbb{E}[\Phi_1(x,\hat{y}(x,\xi),\xi)^p]<+\infty$ for each $x\in\mathcal{X}$ and some $p\geq 2$; (ii) $\mathbb{E}[\kappa^p(\xi)]<+\infty$. Then for arbitrary $\epsilon>0$, there exists a $C>0$, independent of N, such that*

$$\mathbb{P}\left\{\sup_{x\in\mathcal{X}}\left|\frac{1}{N}\sum_{i=1}^{N}\Phi_1(x,\hat{y}(x,\xi^i),\xi^i)-\mathbb{E}[\Phi_1(x,\hat{y}(x,\xi),\xi)]\right|\geq\epsilon\right\}\leq\frac{C}{N^{\frac{p}{2}}\epsilon^p}$$

for sufficiently large N;

(b) *Suppose that: (i) $\Phi_1(x,\hat{y}(x,\xi^i),\xi^i)-\mathbb{E}[\Phi_1(x,\hat{y}(x,\xi),\xi)], i=1,\cdots,N$ satisfy Assumption 1 for each $x\in\mathcal{X}$; (ii) $\mathbb{E}[\kappa(\xi)]<+\infty$; (iii) $\kappa(\xi^i)-\mathbb{E}[\kappa(\xi)], i=1,\cdots,N$ satisfy Assumption 1. Then for arbitrary $\epsilon>0$, there exists a $C>0$, independent of N, such that*

$$\mathbb{P}\left\{\sup_{x\in\mathcal{X}}\left|\frac{1}{N}\sum_{i=1}^{N}\Phi_1(x,\hat{y}(x,\xi^i),\xi^i)-\mathbb{E}[\Phi_1(x,\hat{y}(x,\xi),\xi)]\right|\geq\epsilon\right\}\leq\frac{C}{\beta(N)\alpha(\epsilon)}$$

for sufficiently large N, where $\alpha(\cdot)$ and $\beta(\cdot)$ are defined in Assumption 1.

Due to the Lipschitz continuity of $\Phi_1(\cdot,\hat{y}(\cdot,\xi),\xi)$ with respect to x over $\mathcal{X}$, which of course implies the H-calmness of $\Phi_1(\cdot,\hat{y}(\cdot,\xi),\xi)$ over $\mathcal{X}$ with $\kappa(\xi)$ and order 1, the assertions directly follow from Theorems 3.2 and 4.4. Thus we neglect the proof here.

To derive the relationship between $\mathbb{E}[\Phi_1(x,\hat{y}(x,\xi),\xi)]$ and its solution set, we introduce the growth function of (28), denoted by $\mathcal{R}:\mathbb{R}_+\to\mathbb{R}_+$. Specifically,

$$\mathcal{R}(\tau):=\inf\left\{d\left(0,\mathbb{E}[\Phi_1(x,\hat{y}(x,\xi),\xi)]+\mathcal{N}_{\mathcal{X}}(x)\right):x\in\mathcal{X},d(x,S^*)\geq\tau\right\}.$$

Its inverse function is

$$\mathcal{R}^{-1}(\eta):=\sup\left\{\tau:\mathcal{R}(\tau)\leq\eta\right\}.$$

Then we denote

$$\varsigma(\eta)=\eta+\mathcal{R}^{-1}(\eta).$$

Obviously, $\varsigma(\eta)\geq\mathcal{R}^{-1}(\eta)$; both $\varsigma(\eta)$ and $\mathcal{R}^{-1}(\eta)$ vanish at 0; $\mathcal{R}$ and $\mathcal{R}^{-1}$ are nondecreasing and ς is strictly increasing.

We have for any $\tilde{x}\in S_N$ that the following key result:

$$\sup_{x\in\mathcal{X}}\left|\frac{1}{N}\sum_{i=1}^{N}\Phi_1(x,\hat{y}(x,\xi^i),\xi^i)-\mathbb{E}[\Phi_1(x,\hat{y}(x,\xi),\xi)]\right|$$

$$\geq\left|\frac{1}{N}\sum_{i=1}^{N}\Phi_1(\tilde{x},\hat{y}(\tilde{x},\xi^i),\xi^i)-\mathbb{E}[\Phi_1(\tilde{x},\hat{y}(\tilde{x},\xi),\xi)]\right|$$

$$\geq d\left(0,\mathbb{E}[\Phi_1(\tilde{x},\hat{y}(\tilde{x},\xi),\xi)]+\mathcal{N}_{\mathcal{X}}(\tilde{x})\right)-d\left(0,\frac{1}{N}\sum_{i=1}^{N}\Phi_1(\tilde{x},\hat{y}(\tilde{x},\xi^i),\xi^i)+\mathcal{N}_{\mathcal{X}}(\tilde{x})\right)$$

$$= d\left(0,\mathbb{E}[\Phi_1(\tilde{x},\hat{y}(\tilde{x},\xi),\xi)]+\mathcal{N}_{\mathcal{X}}(\tilde{x})\right)$$

$$\geq\mathcal{R}\left(d(\tilde{x},S^*)\right).$$

24

Thus we have

$$d(\tilde{x}, S^*) \le \mathcal{R}^{-1} \left(\sup_{x \in \mathcal{X}} \left| \frac{1}{N} \sum_{i=1}^{N} \Phi_1(x, \hat{y}(x, \xi^i), \xi^i) - \mathbb{E}[\Phi_1(x, \hat{y}(x, \xi), \xi)] \right| \right)$$

$$\le \varsigma \left(\sup_{x \in \mathcal{X}} \left| \frac{1}{N} \sum_{i=1}^{N} \Phi_1(x, \hat{y}(x, \xi^i), \xi^i) - \mathbb{E}[\Phi_1(x, \hat{y}(x, \xi), \xi)] \right| \right)$$

for any $\tilde{x} \in S_N$. Finally, we obtain

$$\mathbb{P}\left(\mathbb{D}(S_N, S^*) \ge \varsigma(\epsilon)\right) \le \mathbb{P}\left\{ \sup_{x \in \mathcal{X}} \left| \frac{1}{N} \sum_{i=1}^{N} \Phi_1(x, \hat{y}(x, \xi^i), \xi^i) - \mathbb{E}[\Phi_1(x, \hat{y}(x, \xi), \xi)] \right| \ge \epsilon \right\}. \tag{30}$$

According to (30) and Proposition 5.9, we can easily obtain the following assertions.

Theorem 5.10. *Let assumptions in Proposition 5.8 hold, and $\mathcal{X}$ be a compact and convex set.*

(a) *Suppose that: (i) $\mathbb{E}[\Phi_1(x, \hat{y}(x, \xi), \xi)^p] < +\infty$ for each $x \in \mathcal{X}$ and some $p \ge 2$; (ii) $\mathbb{E}[\kappa^p(\xi)] < +\infty$. Then for arbitrary $\epsilon > 0$, there exists a $C > 0$, independent of N, such that*

$$\mathbb{P}\left\{ \mathbb{D}(S_N, S^*) \ge \varsigma(\epsilon) \right\} \le \frac{C}{N^{\frac{p}{2}} \epsilon^p}$$

for sufficiently large N;

(b) *Suppose that: (i) $\Phi_1(x, \hat{y}(x, \xi^i), \xi^i) - \mathbb{E}[\Phi_1(x, \hat{y}(x, \xi), \xi)], i = 1, \cdots, N$ satisfy Assumption 1 for each $x \in \mathcal{X}$; (ii) $\mathbb{E}[\kappa(\xi)] < +\infty$; (iii) $\kappa(\xi^i) - \mathbb{E}[\kappa(\xi)], i = 1, \cdots, N$ satisfy Assumption 1. Then for arbitrary $\epsilon > 0$, there exists a $C > 0$, independent of N, such that*

$$\mathbb{P}\left\{ \mathbb{D}(S_N, S^*) \ge \varsigma(\epsilon) \right\} \le \frac{C}{\beta(N)\alpha(\epsilon)}$$

for sufficiently large N, where $\alpha(\cdot)$ and $\beta(\cdot)$ are defined in Assumption 1.

We have a few comments on Theorem 5.10. To the best of our knowledge, there is no SAA convergence analysis for two-stage nonlinear stochastic variational inequalities under nonmonotonicity or non-iid sampling situations. Although the second stage problem is assumed to be strongly monotone, the whole two-stage problem can be nonmonotone. [15] considered a mixed two-stage nonlinear stochastic variational inequality where the first stage problem is a nonlinear variational inequality and the second stage problem is a nonlinear complementarity problem. However, they assume that the whole two-stage problem is strongly monotone (see [15, Assumption 3.2]). Our model (24) is thus more general. Furthermore, we consider the rate of convergence under heavy tailed distributions and non-iid sampling which are obviously different from those in [15]. Our results also extend some other related works. For example, [47] considered a two-stage stochastic linear complementarity problem under strong monotonicity, and [49] investigated a mixed two-stage linear stochastic variational inequality under nonmonotonicity.

6 Conclusions

In this paper, we investigate the rate of convergence of the SAA scheme under heavy tailed distributions. Based on the pointwise convergence, we extend it to uniform cases under random functions and random set-valued mappings, respectively. Further, by

establishing a Gärtner–Ellis type theorem, we extend the above results to the general (possibly non-iid) sampling. Finally, as an application of our results, we consider several stochastic optimization models and derive the rates of convergence of SAA approximations to these problems.

Different from the existing results in convergence rate estimation, our results can be applied to more general situations, especially to heavy tailed distributions. Thus, it is expected to have wider applications, which will be examined in details in our future work.

References

[1] Alexander Shapiro. Monte carlo sampling methods. *Handbooks in Operations Research and Management Science*, 10:353–425, 2003.

[2] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory.* SIAM, Philadelphia, 2009.

[3] Alexander Shapiro and Huifu Xu. Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation. *Optimization*, 57(3):395–418, 2008.

[4] Alexander Shapiro and Huifu Xu. Uniform laws of large numbers for set-valued mappings and subdifferentials of random functions. *Journal of Mathematical Analysis and Applications*, 325(2):1390–1399, 2007.

[5] Alan J King and R Tyrrell Rockafellar. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Mathematics of Operations Research*, 18(1):148–162, 1993.

[6] Daniel Ralph and Huifu Xu. Convergence of stationary points of sample average two-stage stochastic programs: A generalized equation approach. *Mathematics of Operations Research*, 36(3):568–592, 2011.

[7] L Dai, CH Chen, and JR Birge. Convergence properties of two-stage stochastic programming. *Journal of Optimization Theory and Applications*, 106(3):489–509, 2000.

[8] Tito Homem-de-Mello. On rates of convergence for stochastic optimization problems under non–independent and identically distributed sampling. *SIAM Journal on Optimization*, 19(2):524–551, 2008.

[9] Huifu Xu. Uniform exponential convergence of sample average random functions under general sampling with applications in stochastic programming. *Journal of Mathematical Analysis and Applications*, 368(2):692–710, 2010.

[10] Zhengling Qi, Ying Cui, Yufeng Liu, and Jong-Shi Pang. Statistical analysis of stationary solutions of coupled nonconvex nonsmooth empirical risk minimization. *To preprint on arXiv*, 2019.

[11] Shane S Drew and Tito Homem-de-Mello. Some large deviations results for latin hypercube sampling. In *Proceedings of the Winter Simulation Conference, 2005.* IEEE, 2005.

[12] Alexander Shapiro and Tito Homem-de-Mello. On the rate of convergence of optimal solutions of monte carlo approximations of stochastic programs. *SIAM journal on Optimization*, 11(1):70–86, 2000.

[13] Hailin Sun and Huifu Xu. A note on uniform exponential convergence of sample average approximation of random functions. *Journal of Mathematical Analysis and Applications*, 385(2):698–708, 2012.

[14] Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

[15] Xiaojun Chen, Alexander Shapiro, and Hailin Sun. Convergence analysis of sample average approximation of two-stage stochastic generalized equations. *SIAM Journal on Optimization*, 29(1):135–161, 2019.

[16] Jitka Dupacová and Roger Wets. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *The Annals of Statistics*, pages 1517–1549, 1988.

[17] Alexander Shapiro. Consistency of sample estimates of risk averse stochastic programs. *Journal of Applied Probability*, 50(2):533–541, 2013.

[18] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications.* Springer, New York, 1998.

[19] Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

[20] Alexander Shapiro. Asymptotic properties of statistical estimators in stochastic programming. *The Annals of Statistics*, pages 841–858, 1989.

[21] Alexander Shapiro. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1):169–186, 1991.

[22] Reuven Y Rubinstein and Alexander Shapiro. *Discrete event systems: Sensitivity analysis and stochastic optimization by the score function method.* John Wiley & Sons Inc, New York, 1993.

[23] Alexander Shapiro. On complexity of multistage stochastic programs. *Operations Research Letters*, 34(1):1–8, 2006.

[24] Alexander Shapiro. Stochastic programming with equilibrium constraints. *Journal of Optimization Theory and Applications*, 128(1):221–243, 2006.

[25] Yuri M Kaniovski, Alan J King, and Roger JB Wets. Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems. *Annals of Operations Research*, 56(1):189–208, 1995.

[26] Zhengyan Lin and Zhidong Bai. *Probability inequalities*. Springer Science & Business Media, 2011.

[27] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton University Press, 2015.

[28] Rick Durrett. *Probability: Theory and Examples*, volume 49. Cambridge University Press, 2019.

[29] Olav Kallenberg. *Foundations of Modern Probability*. Springer Science & Business Media, 2006.

[30] David R Brillinger. A note on the rate of convergence of a mean. *Biometrika*, 49(3/4):574–576, 1962.

[31] Napat Rujeerapaiboon, Daniel Kuhn, and Wolfram Wiesemann. Chebyshev inequalities for products of random variables. *Mathematics of Operations Research*, 43(3):887–918, 2018.

[32] R Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[33] Robert J Aumann. Integrals of set-valued functions. *Journal of Mathematical Analysis and Applications*, 12(1):1–12, 1965.

[34] Zvi Artstein and Richard A Vitale. A strong law of large numbers for random compact sets. *The Annals of Probability*, pages 879–882, 1975.

[35] Charles Castaing and Michel Valadier. *Convex Analysis and Measurable Multifunctions*, volume 580. Springer, 2006.

[36] Harald Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*, volume 63. SIAM, 1992.

[37] Jitka Dupacová, Nicole Gröwe-Kuska, and Werner Römisch. *Scenario Reduction in Stochastic Programming: An Approach Using Probability Metrics*. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät . . . , 2000.

[38] Peter Kall and Stein W Wallace. *Stochastic Programming*. John Wiley & Sons, Chichester, 1994.

[39] Andrzej Ruszczyński and Alexander Shapiro. *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*. Elsevier, 2003.

[40] Jie Jiang and Zhiping Chen. Quantitative stability analysis of two-stage stochastic linear programs with full random recourse. *Numerical Functional Analysis and Optimization*, 40(16):1847–1876, 2019.

[41] Zhiping Chen and Jie Jiang. Stability analysis of optimization problems with k th order stochastic and distributionally robust dominance constraints induced by full random recourse. *SIAM Journal on Optimization*, 28(2):1396–1419, 2018.

[42] Svetlozar T Rachev and Werner Römisch. Quantitative stability in stochastic programming: The method of probability metrics. *Mathematics of Operations Research*, 27(4):792–818, 2002.

[43] Werner Römisch. Stability of stochastic programming problems. *Handbooks in Operations Research and Management Science*, 10:483–554, 2003.

[44] Frank H Clarke. *Optimization and Nonsmooth Analysis*, volume 5. SIAM, 1990.

[45] R Tyrrell Rockafellar and Roger JB Wets. Stochastic variational inequalities: single-stage to multistage. *Mathematical Programming*, 165(1):331–360, 2017.

[46] Min Zhang, Jie Sun, and Honglei Xu. Two-stage quadratic games under uncertainty and their solution by progressive hedging algorithms. *SIAM Journal on Optimization*, 29(3):1799–1818, 2019.

[47] Xiaojun Chen, Hailin Sun, and Huifu Xu. Discrete approximation of two-stage stochastic and distributionally robust linear complementarity problems. *Mathematical Programming*, 177(1-2):255–289, 2019.

[48] Xiaojun Chen, Ting Kei Pong, and Roger JB Wets. Two-stage stochastic variational inequalities: an erm-solution procedure. *Mathematical Programming*, 165(1):71–111, 2017.

[49] Jie Jiang, Xiaojun Chen, and Zhiping Chen. Quantitative analysis for a class of two-stage stochastic linear variational inequality problems. *Computational Optimization and Applications*, 76(2):431—-460, 2020.

[50] Jie Jiang, Yun Shi, Xiaozhou Wang, and Xiaojun Chen. Regularized two-stage stochastic variational inequalities for Cournot-Nash equilibrium under uncertainty. *Journal of Computational Mathematics*, 37(6):813–842, 2019.

[51] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional Variational Inequalities and Complementarity Problems*. Springer, 2007.

[52] Stephen M Robinson. Strongly regular generalized equations. *Mathematics of Operations Research*, 5(1):43–62, 1980.

[53] Alexey F Izmailov. Strongly regular nonsmooth generalized equations. *Mathematical Programming*, 147(1-2):581–590, 2014.

[54] Asen L Dontchev and R Tyrrell Rockafellar. Characterizations of strong regularity for variational inequalities over polyhedral convex sets. *SIAM Journal on Optimization*, 6(4):1087–1105, 1996.