Journal of Machine Learning Research 14 (2013) 351-376 Submitted 11/11; Revised 11/12; Published 2/13

Risk Bounds of Learning Processes for Lévy Processes

Chao Zhang[*] ZHANGCHAO1015@GMAIL.COM
Center for Evolutionary Medicine and Informatics
Biodesign Institute, Arizona State University
Tempe, AZ 85287, U.S.A.

Dacheng Tao DACHENG.TAO@GMAIL.COM
Centre for Quantum Computation & Intelligent Systems
FEIT, University of Technology, Sydney
NSW 2007, Australia

Editor: Mehryar Mohri

Abstract

Lévy processes refer to a class of stochastic processes, for example, Poisson processes and Brownian motions, and play an important role in stochastic processes and machine learning. Therefore, it is essential to study risk bounds of the learning process for time-dependent samples drawn from a Lévy process (or briefly called learning process for Lévy process). It is noteworthy that samples in this learning process are not independently and identically distributed (i.i.d.). Therefore, results in traditional statistical learning theory are not applicable (or at least cannot be applied directly), because they are obtained under the sample-i.i.d. assumption. In this paper, we study risk bounds of the learning process for time-dependent samples drawn from a Lévy process, and then analyze the asymptotical behavior of the learning process. In particular, we first develop the deviation inequalities and the symmetrization inequality for the learning process. By using the resultant inequalities, we then obtain the risk bounds based on the covering number. Finally, based on the resulting risk bounds, we study the asymptotic convergence and the rate of convergence of the learning process for Lévy process. Meanwhile, we also give a comparison to the related results under the sample-i.i.d. assumption.

Keywords: Lévy process, risk bound, deviation inequality, symmetrization inequality, statistical learning theory, time-dependent

1. Introduction

In statistical learning theory, one of the major concerns is the risk bound, which explains the asymptotic behavior of the probability that a function produced by an algorithm has a sufficiently small error. Generally, there are three essential parts in the process of obtaining risk bounds: deviation or concentration inequalities, symmetrization inequalities and complexity measures of function classes. For example, Van der Vaart and Wellner (1996) showed risk bounds based on the Rademacher complexity and the covering number by using Hoeffding's inequality. Vapnik (1998) gave risk bounds based on the annealed Vapnik-Chervonenkis (VC) entropy and the VC dimension, respectively. In Vapnik (1998), Vapnik applied some classical inequalities, for example, Chernoff's inequality and Hoeffding's inequality, but also developed specific concentration inequalities

∗. This work was partly completed when the author was with the School of Computer Engineering, Nanyang Technological University, 639798, Singapore.

to study the asymptotic behavior of the i.i.d. empirical process. Bartlett et al. (2005) proposed the local Rademacher complexity and obtained a sharp risk bound for a particular function class $\{f \in \mathcal{F} | \mathrm{E}f^2 < \beta \mathrm{E}f, \beta > 0\}$ by using Talagrand's inequality. Moreover, there are also other investigations to statistical learning theory (see Cesa-Bianchi and Gentile, 2008; Mendelson, 2008, 2002; Koltchinskii, 2001). However, all of these results are built under the sample-i.i.d. assumption.

Samples are not always i.i.d. in practice, for example, some financial and physical behaviors are temporally dependent, and the aforementioned research results are not applicable (or at least cannot be applied directly) to most cases. Thus, it is essential to study the risk bounds in the scenario of non-i.i.d. samples. The scenario of non-i.i.d. samples contains a wide variety of cases and it is impossible to find a unified form to cover all the cases. Instead, one feasible scheme is to find some representative processes, such as the Lévy process and the mixing process that cover several useful cases in the scenario of non-i.i.d. samples, and then we study the theoretical properties of each process individually.

Recently, Mohri and Rostamizadeh (2010) obtained risk bounds for stationary β-mixing sequences based on the Rademacher complexity. Mixing sequences can be deemed as a transition between the i.i.d. scenario and the non-i.i.d. scenario, where the dependence between samples diminishes along time. Especially, by adopting a technique of independent blocks (Yu, 1994), samples drawn from a β-mixing sequence can be transformed to an i.i.d. scenario and thus some classical results under the sample-i.i.d. assumption can be applied to obtain the risk bounds. Jiang (2009) extended Hoeffding's inequality to handle the situations with unbounded loss and dependent data, and then provided probability bounds for uniform deviations in a general framework involving discrete decision rules, unbounded loss and a dependence structure. Moreover, there are also some works about the uniform laws for dependent processes (Nobel and Dembo, 1993).

Lévy processes are the stochastic processes with stationary and independent increments and cover a large class of stochastic processes, for example, Brownian motions, Poisson processes, stable processes and subordinators (see Kyprianou, 2006). Moreover, Lévy processes have been regarded as prototypes of semimartingales and Feller-Markov processes (Applebaum, 2004b; Sato, 2004). Lévy processes have been successfully applied to practical applications in finance (Cont and Tankov, 2006), physics (Applebaum, 2004a), signal processing (Duncan, 2009), image processing (Pedersen et al., 2005) and actuarial science (Barndorff-Nielsen et al., 2001). Figueroa-López and Houdré (2006) used projection estimators to estimate the Lévy density, and then gave a bound to exhibit the discrepancy between a projection estimator and the orthogonal projection by using the concentration inequalities for functionals of Poisson integrals. In this paper, we extend the existing works on the infinitely divisible distribution (see Houdré, 2002; Houdré et al., 1998) to develop the deviation inequalities for Lévy processes and then obtain the risk bounds by using the resulted deviation inequalities. Next, we summarize the main results of this paper.

1.1 Overview of Main Results

This paper is mainly concerned with the theoretical analysis of the learning process for the time-dependent samples drawn from a Lévy process. There are four major concerns in this paper: the deviation inequality for Lévy process; the symmetrization inequality of the learning process; the risk bounds and the asymptotical behavior of the learning process.

Generally, in order to obtain the risk bounds of a certain learning process, it is necessary to first obtain the corresponding concentration (or deviation) inequalities for the learning process. Thus, we

extend the previous works (Houdré, 2002; Houdré et al., 1998) to develop the deviation inequalities for the Lévy process, which are suitable for the sequence of random variables at different time points. We then present the symmetrization inequality of the learning process for Lévy process. By applying the derived deviation and symmetrization inequalities, we obtain the risk bounds of the learning process, which is based on the covering number. Finally, we use the resulted risk bounds to analyze the asymptotical convergence and the rate of convergence of the learning process for Lévy process, respectively. Meanwhile, we also give a comparison with the learning process for i.i.d. samples.

Zhang and Tao (2010) discussed risk bounds for Lévy process with *zero* Gaussian component, but their results are based on some specific assumptions to function classes. The current results do not require any conditions of function classes except the boundedness and the Lipschitz continuity and are valid for a more general scenario where the considered Lévy process has non-zero Gaussian component, so they are more general than the previous results.

1.2 Organization of the Paper

The rest of this paper is organized as follows. In Section 2, we formalize the main research of this paper. Section 3 introduces some preliminaries of the infinitely divisible (ID) distribution and the Lévy process. We present the deviation inequalities and the symmetrization inequality of the learning process for Lévy process in Section 4. Section 5 gives the risk bounds of the learning process. In Section 6, we analyze the asymptotic behavior of the learning process and the last section concludes the paper. The proofs of main results are given in the appendices including Theorem 8, Theorem 11, Theorem 12 and Theorem 15.

2. Problem Setup

Denote $\mathcal{X} \subset \mathbb{R}^I$ as an input space and $\mathcal{Y} \subset \mathbb{R}^J$ as its corresponding output space. Let $\mathbf{Z} = (\mathcal{X}, \mathcal{Y}) \subset \mathbb{R}^K$ $(K = I + J)$ and $\{Z_t\}_{t\geq 0}$ be an undetermined Lévy process. Assume that $\mathbf{Z} = \{Z_t\}_{t\geq 0}$ with $Z_t = (\mathbf{x}_t, \mathbf{y}_t)$. Let $\mathcal{G} \subset \mathcal{Y}^{\mathcal{X}}$ be a function class with the domain $\mathcal{X}$ and the range $\mathcal{Y}$. Given a loss function $\ell : \mathcal{Y}^2 \to \mathbb{R}$ and a time interval $[T_1, T_2]$, it is expected to find a function $g^* \in \mathcal{G} : \mathcal{X} \to \mathcal{Y}$ that minimizes the expected risk

$$\mathrm{E}(\ell \circ g) := \frac{1}{T} \int_{T_1}^{T_2} \int \ell(g(\mathbf{x}_t), \mathbf{y}_t) dP_t dt, \;\; g \in \mathcal{G}, \tag{1}$$

where $T = T_2 - T_1$, P_t stands for the distribution of $Z_t = (\mathbf{x}_t, \mathbf{y_t})$ at time t and $\ell(g(x), y)$ is denoted as $(\ell \circ g)(x, y)$.

Generally, if P_t $(t \in [T_1, T_2])$ are unknown, the target function g^* usually cannot be directly obtained by minimizing (1). Instead, we can apply the empirical risk minimization (ERM) principle to handle this issue. Given a function class $\mathcal{G}$ and a sample set $\mathbf{Z}_1^N := \{Z_{t_n}\}_{n=1}^N$ drawn from $\mathbf{Z}$ in the time interval $[T_1, T_2]$ with $T_1 \leq t_1 < \cdots < t_N \leq T_2$, we define the empirical risk of $g \in \mathcal{G}$ as

$$\mathrm{E}_N(\ell \circ g) := \frac{1}{N} \sum_{n=1}^{N} \ell(g(\mathbf{x}_{t_n}), \mathbf{y}_{t_n}), \tag{2}$$

which is considered as an approximation to the expected risk (1). Let $g_N \in \mathcal{G}$ be the function that minimizes the empirical risk (2) over $\mathcal{G}$ and we deem g_N as an estimate to g^* with respect to $\mathbf{Z}_1^N$.

It is noteworthy that such learning process covers many kinds of practical applications, for example, the predicting for time series (Mukherjee et al., 1997; Kim, 2003) and the estimation of channel state information (Biguesh and Gershman, 2006; Love et al., 2008; Tulino et al., 2005). We take the estimation of channel state information for example.

In the estimation of channel state information, $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^I$ and $\mathbf{y} \in \mathcal{Y} \subset \mathbb{R}^J$ are regarded as the transmit and the receive vectors, respectively. The following are the reasons why we suppose that $\mathbf{Z}$ is a segment of an undetermined Lévy process:

- In fact, the tasks of the estimation of channel state information are time-dependent and can be regarded as the approximation to unknown stochastic processes.
- The Lévy process is one of representative processes and covers a large body of stochastic processes, that is, Brownian motions, Poisson processes, compound Poisson processes, Gamma processes and inverse Gaussian processes (see Kyprianou, 2006).
- Many kinds of signals have the Poisson property, the martingale property or both of them.

One of the most frequently used models is $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$, where $\mathbf{H}$ and $\mathbf{n}$ are the channel matrix and the noise vector, respectively (see Love et al., 2008; Tulino et al., 2005). The corresponding function class $\mathcal{G}$ can be formalized as $\mathcal{G} := \{\mathbf{x} \mapsto \mathbf{H}\mathbf{x} + \mathbf{n} : \mathbf{H} \in \mathbb{R}^J \times \mathbb{R}^I,\ \mathbf{n} \in \mathbb{R}^J\}$. The loss function ℓ is selected as the mean square error function, and then the least-square estimation is used to find the function that minimizes the empirical risk (2). Moreover, there are also other ERM-based methods proposed for the estimation of channel state information (see Sanchez-Fernandez et al., 2004; Sutivong et al., 2005).

In the aforementioned learning process, we are mainly interested in the asymptotic behavior of the quantity $(\mathrm{E}(\ell \circ g^*) - \mathrm{E}_N(\ell \circ g_N))$, when the sample number N goes to the *infinity*. Since $\mathrm{E}_N(\ell \circ g^*) - \mathrm{E}_N(\ell \circ g_N) \geq 0$, we have

$$\begin{aligned}
\mathrm{E}(\ell \circ g_N) =& \mathrm{E}(\ell \circ g_N) - \mathrm{E}(\ell \circ g^*) + \mathrm{E}(\ell \circ g^*) \\
\leq& \mathrm{E}_N(\ell \circ g^*) - \mathrm{E}_N(\ell \circ g_N) + \mathrm{E}(\ell \circ g_N) - \mathrm{E}(\ell \circ g^*) + \mathrm{E}(\ell \circ g^*) \\
\leq& 2 \sup_{g \in \mathcal{G}} \left|\mathrm{E}(\ell \circ g) - \mathrm{E}_N(\ell \circ g)\right| + \mathrm{E}(\ell \circ g^*),
\end{aligned}$$

and thus

$$0 \leq \mathrm{E}(\ell \circ g_N) - \mathrm{E}(\ell \circ g^*) \leq 2 \sup_{g \in \mathcal{G}} \left|\mathrm{E}(\ell \circ g) - \mathrm{E}_N(\ell \circ g)\right|.$$

The supremum

$$\sup_{g \in \mathcal{G}} \left|\mathrm{E}(\ell \circ g) - \mathrm{E}_N(\ell \circ g)\right| \tag{3}$$

is the so-called risk bound of the learning process for a Lévy process $\{Z_t\}_{t \geq 0}$.

Then, we define the loss function class

$$\mathcal{F} := \{Z \mapsto \ell(g(\mathbf{x}), \mathbf{y}) : g \in \mathcal{G}\}, \tag{4}$$

and call $\mathcal{F}$ the function class in the rest of this paper. Given a sample set $\{Z_{t_n}\}_{n=1}^N$ drawn from $\{Z_t\}_{t \geq 0}$, we shortly denote for any $f \in \mathcal{F}$,

$$\mathrm{E}_t f := \int f(Z) dP_t\ ,\ \ t > 0, \tag{5}$$

and

$$\mathrm{E}_N f := \frac{1}{N}\sum_{n=1}^{N} f(Z_{t_n}), \tag{6}$$

where E_t stands for the expectation taken with respect to Z_t.

According to (3), (4), (5) and (6), we have

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} |\mathrm{E}f - \mathrm{E}_N f| \\
=&\sup_{g\in\mathcal{G}} \left|\mathrm{E}(\ell\circ g) - \mathrm{E}_N(\ell\circ g)\right| \\
=&\sup_{g\in\mathcal{G}} \left|\frac{1}{T}\int_{T_1}^{T_2}\int \ell(g(\mathbf{x}_t),\mathbf{y}_t)dP_t dt - \frac{1}{N}\sum_{n=1}^{N}\ell(g(\mathbf{x}_{t_n}),\mathbf{y}_{t_n})\right| \\
\leq& 2 \sup_{\substack{g\in\mathcal{G}\\ t\in[T_1,T_2]}} \left|\int \ell(g(\mathbf{x}_t),\mathbf{y}_t)dP_t - \frac{1}{N}\sum_{n=1}^{N}\ell(g(\mathbf{x}_{t_n}),\mathbf{y}_{t_n})\right| \\
=& 2 \sup_{\substack{f\in\mathcal{F}\\ t\in[T_1,T_2]}} \left|\mathrm{E}_t f - \mathrm{E}_N f\right|.
\end{aligned}$$

Therefore, the supremum

$$\sup_{\substack{f\in\mathcal{F}\\ t\in[T_1,T_2]}} \left|\mathrm{E}_t f - \mathrm{E}_N f\right|$$

plays an important role in studying the risk bound $\sup_{f\in\mathcal{F}}|\mathrm{E}f - \mathrm{E}_N f|$ of the learning process for Lévy process.

3. Infinitely Divisible Distributions and Lévy Processes

Since the infinitely divisible (ID) distribution is strongly related to the Lévy process, this section first introduces the ID distribution and then briefs the Lévy process for the subsequent discussion.

3.1 ID Distributions

A probability distribution is said to be infinitely divisible if and only if it can be represented by the distribution of the sum of an arbitrary number of i.i.d. random variables. Many probability distributions have the infinite divisibility, for example, Poisson, geometric, lognormal, noncentral chi-square, exponential, Gamma, Pareto and Cauchy (see Bose et al., 2002). The ID distribution can be defined based on the characteristic function.

Definition 1 *Let $\phi(\theta)$ be the characteristic function of a random variable Z, that is*

$$\phi(\theta) := \mathrm{E}\left\{\mathrm{e}^{i\theta Z}\right\} = \int_{-\infty}^{+\infty}\mathrm{e}^{i\theta Z}dP(Z). \tag{7}$$

Then, the distribution of Z is infinitely divisible if and only if for any $N\in\mathbb{N}$, there exists a characteristic function $\phi_N(\theta)$ such that

$$\phi(\theta) = \underbrace{\phi_N(\theta) * \cdots * \phi_N(\theta)}_{N},$$

where "$*$" stands for multiplication.

By (7), given a characteristic function $\phi(\theta)$, we define the corresponding characteristic exponent as

$$\psi(\theta) := \ln \phi(\theta) = \ln \left(\mathrm{E} e^{i\theta Z} \right).$$

Afterward, we will show that the characteristic exponent of any ID distribution has a unified form (see Sato, 2004). Before the formal presentation, we need to give a definition of the Lévy measure (see Applebaum, 2004a).

Definition 2 *Let ν be a Borel measure defined on $\mathbb{R}^K \setminus \{0\}$. This ν will be a Lévy measure if*

$$\int_{\mathbb{R}^K \setminus \{0\}} \min\{\|u\|^2, 1\} \nu(du) < \infty,$$

and $\nu(\{0\}) = 0$.

The Lévy measure describes the expected number of a certain height jump in a time interval of the unit length. Define the indicator function for the event $\mathcal{E}$ as

$$\mathbf{1}_{\mathcal{E}} = \begin{cases} 1, & \text{the event } \mathcal{E} \text{ appears;} \\ 0, & \text{otherwise,} \end{cases}$$

and for any ID random variable, its characteristic exponent takes the following form (see Sato, 2004).

Theorem 3 (Lévy-Khintchine) *A Borel probability measure μ of a random variable $Z \in \mathbb{R}^K$ is infinitely divisible if and only if there exists a triplet $(\mathbf{a}, \mathbf{A}, \nu)$ such that for all $\theta \in \mathbb{R}^K$, the characteristic exponent ψ_μ is of the form*

$$\psi_\mu(\theta) = i\langle \mathbf{a}, \theta \rangle - \frac{1}{2} \langle \theta, \mathbf{A}\theta \rangle + \int_{\mathbb{R}^K \setminus \{0\}} \left(e^{i\langle \theta, u \rangle} - 1 - i\langle \theta, u \rangle \mathbf{1}_{\|u\| \le 1} \right) \nu(du), \tag{8}$$

where $\mathbf{a} \in \mathbb{R}^K$, $\mathbf{A}$ is a $K \times K$ positive-definite symmetric matrix, ν is a Lévy measure on $\mathbb{R}^K \setminus \{0\}$, and $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ stand for the inner product and the norm in $\mathbb{R}^K$, respectively.

Theorem 3 shows that an ID distribution can be completely determined by a triplet $(\mathbf{a}, \mathbf{A}, \nu)$, where $\mathbf{a}$ is a drift, $\mathbf{A}$ is a Gaussian component and ν is a Lévy measure. Thus, we call $(\mathbf{a}, \mathbf{A}, \nu)$ the generating triplet of an ID distribution.

3.2 Lévy Processes

First, we give a rigorous definition of Lévy processes.

Definition 4 *A stochastic process $\{Z_t\}_{t \ge 0}$ on $\mathbb{R}^K$ is a Lévy process if it satisfies the following conditions:*

1. *$Z_0 = 0$, almost surely.*

2. *For any $n \geq 1$ and $0 \leq t_0 \leq t_1 \leq \cdots \leq t_n$, the random variables*

$$Z_{t_0}, Z_{t_1} - Z_{t_0}, \cdots, Z_{t_n} - Z_{t_{n-1}}$$

are independent.

3. *The increments are stationary, that is, the distribution of $Z_{s+t} - Z_s$ is independent of s.*

4. *The process is right continuous, that is, for any $0 \leq t \leq s$ and $\varepsilon > 0$, we have*

$$\lim_{s \to t} \Pr\Big\{ |Z_t - Z_s| > \varepsilon \Big\} = 0.$$

According to Theorem 7.10 of Sato (2004), a Lévy process $\{Z_t\}_{t \geq 0}$ can be distinguished by the distribution of Z_1, which has an ID distribution with the generating triplet $(\mathbf{a}_1, \mathbf{A}_1, \nu_1)$, and at any time $t > 0$, $Z_t \in \{Z_t\}_{t \geq 0}$ has an ID distribution with the generating triplet $(\mathbf{a}_t, \mathbf{A}_t, \nu_t)$. Therefore, we call $(\mathbf{a}_1, \mathbf{A}_1, \nu_1)$ the characteristic triplet of the Lévy process $\{Z_t\}_{t \geq 0}$. For any $t > 0$, there also holds that

$$(\mathbf{a}_t, \mathbf{A}_t, \nu_t) := (\mathbf{a}_1 t, \mathbf{A}_1 t, \nu_1 t).$$

Next, we introduce the Lévy-Ito decomposition to discuss the relationship between the path of a Lévy process and its characteristic triplet $(\mathbf{a}_1, \mathbf{A}_1, \nu_1)$. The details are referred to Kyprianou (2006); Sato (2004).

Theorem 5 (Lévy-Ito Decomposition) *Consider a triplet $(\mathbf{a}_1, \mathbf{A}_1, \nu_1)$ where $\mathbf{a}_1 \in \mathbb{R}^K$, $\mathbf{A}_1$ is a $K \times K$ positive-definite symmetric matrix, ν_1 is a Lévy measure on $\mathbb{R}^K \setminus \{0\}$. Then, there exist four independent Lévy processes, $L^{(1)}$, $L^{(2)}$, $L^{(3)}$ and $L^{(4)}$, where $L^{(1)}$ is a constant drift, $L^{(2)}$ is a Brownian motion, $L^{(3)}$ is a compound Poisson process and $L^{(4)}$ is a square integrable (pure jump) martingale with an a.s. countable number of jumps of magnitude less than 1 on each finite time interval. Taking $L = L^{(1)} + L^{(2)} + L^{(3)} + L^{(4)}$, there then exists a Lévy process $L = \{Z_t\}_{0 \leq t \leq T}$ with characteristic exponent in the form of* (8).

The proof of this theorem has been given by Chapter 4 in Sato (2004) or Chapter 2 in Kyprianou (2006), so we omit it here. We only go through some steps of the proof to reveal the relationship between the path of a Lévy process and its characteristic triplet $(\mathbf{a}_1, \mathbf{A}_1, \nu_1)$. Recalling (8), we can split the characteristic exponent ψ into four parts:

$$\psi = \psi^{(1)} + \psi^{(2)} + \psi^{(3)} + \psi^{(4)}$$

with

$$\begin{aligned}
&\psi^{(1)}(\theta) = i\langle \mathbf{a}_1, \theta \rangle; \quad \psi^{(2)}(\theta) = -\frac{1}{2}\langle \theta, \mathbf{A}_1 \theta \rangle; \\
&\psi^{(3)}(\theta) = \int_{\|u\| \geq 1} \big(e^{i\langle \theta, u \rangle} - 1 \big) \nu_1(du); \\
&\psi^{(4)}(\theta) = \int_{\|u\| < 1} \big(e^{i\langle \theta, u \rangle} - 1 - i\langle \theta, u \rangle \big) \nu_1(du),
\end{aligned}$$

which correspond to $L^{(1)}$, $L^{(2)}$, $L^{(3)}$ and $L^{(4)}$, respectively. We also refer to Jacobsen (2005) for the knowledge on jump processes as well as Lévy processes. At the end of this section, we give two examples of Lévy processes in addition to the corresponding Lévy-Khintchine representations and Lévy-Ito decompositions:

- A Poisson process $\{N_t\}_{t\geq 0}$ is a Lévy process that has a Poisson distribution with parameter pt at any time $t > 0$. In the Lévy-Khintchine representation, we find that $\mathbf{a}_1$ and $\mathbf{A}_1$ are both *zero* and $\nu_1 = p\delta_1$, where δ_1 is the Dirac measure supported on $\{1\}$. In the Lévy-Ito decomposition, its characteristic exponent is expressed as $\psi(\theta) = \psi^{(3)}(\theta) = p(e^{i\theta} - 1)$.

- A scaled Brownian motion with linear drift is also a Lévy process with the characteristic triplet $(\mathbf{a}_1, \mathbf{A}_1, 0)$ in the Lévy-Khintchine representation. In the Lévy-Ito decomposition, its characteristic exponent is expressed as $\psi(\theta) = \psi^{(1)}(\theta) + \psi^{(2)}(\theta)$ with $\psi^{(1)}(\theta) = i\langle \mathbf{a}_1, \theta\rangle$ and $\psi^{(2)}(\theta) = -\frac{1}{2}\langle \theta, \mathbf{A}_1\theta\rangle$.

4. Deviation Inequalities and Symmetrization Inequalities

In this section, we present the deviation inequalities and symmetrization inequality of the learning process for Lévy process.

4.1 Preliminaries

Firstly, we need to introduce some notations and conditions for the following discussion.

4.1.1 NOTATIONS

Assume that $\mathcal{F}$ is a function class consisting of λ-Lipschitz functions and $\{Z_t\}_{t\geq 0} \subset \mathbb{R}^K$ is a Lévy process with the characteristic triplet $(\mathbf{a}_1, \mathbf{A}_1, \nu_1)$. Let $\mathbf{Z}_1^N = \{Z_{t_n}\}_{n=1}^N$ be a sample set drawn from $\{Z_t\}_{t\geq 0}$ in the time interval $[T_1, T_2]$. For any $t \in [T_1, T_2]$, we give the following definitions:

$$
\begin{aligned}
&(D1) \quad \Sigma_N^{(*)} := \sup_{t\in[T_1,T_2]} \frac{1}{N}\sum_{n=1}^{N} \sup_{f\in\mathcal{F}} |\mathrm{E}_{t_n} f - \mathrm{E}_t f|;\\
&(D2) \quad \varphi(\alpha) := \sum_{n=1}^{N} \lambda^2 \pi K^2 \alpha t_n + \int_{\mathbb{R}^K} \lambda \|u\| \left(e^{\lambda\alpha\|u\|} - 1\right) \nu_{t_n}(du);\\
&(D3) \quad V^{(n)} := \int_{\mathbb{R}^K} \|u\|^2 \nu_{t_n}(du) = t_n \int_{\mathbb{R}^K} \|u\|^2 \nu_1(du);\\
&(D4) \quad \Gamma(x) := x - (x+1)\ln(x+1).
\end{aligned}
$$

Note that the quantity $\sup_{f\in\mathcal{F}} |\mathrm{E}_{t_n} f - \mathrm{E}_t f|$ is called the integral probability metric and has been widely used to measure the difference between two probability distributions (see Zolotarev, 1984; Rachev, 1991; Müller, 1997; Reid and Williamson, 2011). Recently, Sriperumbudur et al. (2012) gave the further investigation and proposed the empirical method to compute the integral probability metric. As mentioned by Müller (1997), the quantity $\sup_{f\in\mathcal{F}} |\mathrm{E}_{t_n} f - \mathrm{E}_t f|$ is a (semi)metric to measure the difference between the distributions of $\{Z_t\}_{t\geq 0}$ at two time points t and t_n. In fact, given a non-trivial function class $\mathcal{F}$, the quantity $\sup_{f\in\mathcal{F}} |\mathrm{E}_{t_n} f - \mathrm{E}_t f|$ is equal to *zero* if the distributions at the two time points match or the two time points coincide, that is, $t = t_n$.

4.1.2 CONDITIONS

In order to achieve the desired risk bounds, some necessary conditions need to be introduced to specify the behavior of Lévy processes.

$(C1)$ The f is a partially differentiable function on $\mathbb{R}^K$ and there exists a constant $\lambda > 0$ such that for any $Z = (z_1, \cdots, z_K)^T \in \mathbb{R}^K$,

$$\max_{1 \le k \le K} \left| \frac{\partial f(Z)}{\partial z_k} \right| \le \lambda.$$

$(C2)$ Denoting $\mathbf{A}_1 = \{a_{ij}\}_{K \times K}$, there exists a constant $\pi > 0$ such that

$$\max_{1 \le i,j \le K} |a_{ij}| \le \pi.$$

$(C3)$ The ν_1 has a bounded support with

$$R = \inf\{\rho > 0 : \nu_1(\{u : \|u\| > \rho\}) = 0\}.$$

Condition $(C1)$ implies that f has bounded partial derivatives and holds for many kinds of functions, for example, quadratic functions with bounded domains and trigonometric functions. The constant λ is determined by the selected function and thus it is manipulatable. Condition $(C2)$ implies that all entries of $\mathbf{A}_1$ are bounded. Condition $(C3)$ implies the Lévy measure ν_1 has a bounded support. To take an example of Conditions $(C2)$-$(C3)$, we refer to Poisson processes whose characteristic triplet is $(0,0,\nu_1)$ with ν_1 supporting on $\{1\}$. Afterwards, we come up with the deviation inequalities of the learning process for Lévy process.

4.2 Deviation Inequalities

Deviation inequalities play an essential role in obtaining risk bounds of a certain learning process. Generally, specific deviation inequalities need to be developed for different learning processes. There are a lot of popular concentration inequalities and deviation inequalities, for example, Hoeffding's inequality, McDiarmid's inequality, Bennett's inequality, Bernstein's inequality and Talagrand's inequality, which are all valid under the sample-i.i.d. assumption. Moreover, there have been the deviation inequalities for ID distributions and Lévy processes both with *zero* Gaussian components proposed by Houdré (2002); Houdré and Marchal (2008), respectively. Here, we extend the deviation results in Houdré (2002) to develop the deviation inequalities of the learning process for Lévy process, which is related to a sequence of random variables taking values from a Lévy process at different time points.

Based on a fact that the vector formed by N independent ID random vectors is itself infinitely divisible, the following theorem and corollary can be derived from Theorem 1 & Corollary 1 of Houdré (2002) and Proposition 2 of Houdré et al. (1998). We also refer to Zhang and Tao (2011a,b) for the related discussions.

Theorem 6 *Assume that f is a function satisfying Condition $(C1)$ and $\{Z_t\}_{t \ge 0} \subset \mathbb{R}^K$ is a Lévy process with the characteristic triplet $(\mathbf{a}_1, \mathbf{A}_1, \nu_1)$ satisfying Condition $(C2)$. Let $\mathbf{Z}_1^N = \{Z_{t_n}\}_{n=1}^N$ $(t_1 < t_2 < \cdots < t_N)$ be a set of time-dependent samples drawn from $\{Z_t\}_{t \ge 0}$ in the time interval $[T_1, T_2]$. Define a function $F : \mathbb{R}^{NK} \to \mathbb{R}$ as*

$$F\left(\mathbf{Z}_1^N\right) := \sum_{n=1}^{N} f(Z_{t_n}). \tag{9}$$

If Condition $(C1)$ *is valid and* $\mathrm{E}e^{\alpha\|\mathbf{z}_t\|}\big|_{t=1} < +\infty$ *holds for some* $\alpha > 0$*, then we have for any* $0 < \xi < \varphi((M/\lambda)^-)$*,*

$$\Pr\left\{\left|F\left(\mathbf{Z}_1^N\right) - \overline{\mathrm{E}}F\right| > \xi\right\} \leq 2\exp\left\{-\int_0^{\xi} \varphi^{-1}(s)ds\right\}, \tag{10}$$

where the expectation $\overline{\mathrm{E}}$ *is taken on all* $\{Z_{t_1}, \cdots, Z_{t_N}\}$*,* φ *is given in Definition* $(D2)$*,* $\varphi(a^-)$ *is the left-hand limit of* φ *at* a*,* $M = \sup\left\{\alpha > 0 : \mathrm{E}e^{\alpha\|\mathbf{z}_t\|}\big|_{t=1} < +\infty\right\}$ *and* φ^{-1} *is the inverse of* $\varphi(\alpha)$ *with the domain of* $0 < \alpha < M/\lambda$*.*

In Theorem 6, we present a deviation inequality of the learning process for the Lévy process satisfying Condition $(C2)$. However, there are two drawbacks of this result that will bring some difficulties to the future theoretical analysis of asymptotic behavior.

- The deviation inequality (10) is represented by the integral of φ^{-1}, and thus the inequality cannot explicitly reflect the asymptotic behavior as N goes to the *infinity*.
- Recalling Definition $(D2)$, there is an integral term in the expression of the function φ. Thus, given a certain $\xi > 0$, it may be difficult to justify whether the ξ satisfies the condition $\xi < \varphi((M/\lambda)^-)$.

In order to overcome these drawbacks, we add Condition $(C3)$ to achieve another deviation inequality for the learning process.

Corollary 7 *Follow notations in Theorem 6. If Conditions* $(C1)$*-*$(C3)$ *are all valid, then we have for any* $\xi > 0$*,*

$$\begin{aligned}
&\Pr\left\{\left|F\left(\mathbf{Z}_1^N\right) - \overline{\mathrm{E}}F\right| > \xi\right\} \\
\leq& 2\exp\left\{\frac{\sum_{n=1}^N (\lambda^2\pi K^2\alpha t_n + V^{(n)})}{(\lambda R)^2}\Gamma\left(\frac{\lambda R\xi}{\sum_{n=1}^N (\lambda^2\pi K^2\alpha t_n + V^{(n)})}\right)\right\} \\
\leq& 2\exp\left\{\frac{NT_1(\lambda^2\pi K^2\alpha + V)}{(\lambda R)^2}\Gamma\left(\frac{\lambda R\xi}{NT_2(\lambda^2\pi K^2\alpha + V)}\right)\right\},
\end{aligned} \tag{11}$$

where Γ *is given in Definition* $(D4)$ *and*

$$V := \int_{\mathbb{R}^K} \|u\|^2 \nu_1(du).$$

The second inequality of the above result is derived from the facts that there holds that $V^{(n)} \leq T_2 V$ for any $1 \leq n \leq N$ and the function $\Gamma(x)$ is a monotonically decreasing function when $x > 0$ as shown in Figure 1.

Compared to the result (10), the deviation inequality (11) holds for any $\xi > 0$ and its right-hand-side is represented by using the function $\Gamma(x)$ $(x > 0)$. Therefore, we can directly analyze the asymptotic behavior as N goes to the *infinity*. In fact, since the function $\Gamma(x)$ is smaller than *zero* when $x > 0$, the right-hand-side of (11) will go to *zero* for any $\xi > 0$ when N approaches to the *infinity*. Next, we present the symmetrization inequality of the learning process for Lévy process.

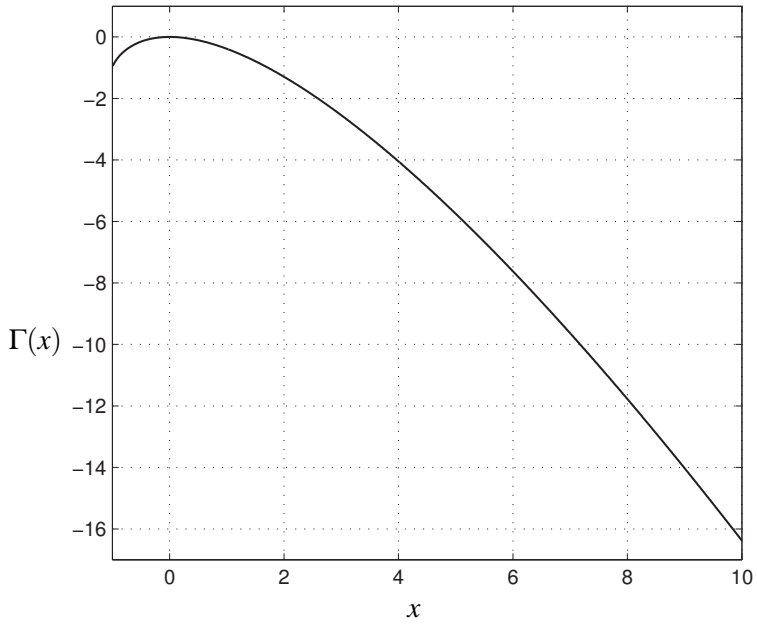

Figure 1: The Function Curve of $\Gamma(x)$

4.3 Symmetrization Inequality

Symmetrization inequalities are mainly used to replace the expected risk by an empirical risk computed on another sample set that is independent of the given sample set but has the identical distribution. In this manner, risk bounds can be achieved by using some kinds of complexity measures, for example, the covering number and the VC dimension. However, the classical symmetrization results (see Bousquet et al., 2004) are only valid under the sample-i.i.d. assumption. Afterward, we propose the symmetrization inequality of the learning process for Lévy process.

For clarity of presentation, we give a notation that will be used in the rest of the paper. Given a sample set $\mathbf{Z}_1^N = \{Z_{t_n}\}_{n=1}^N$ $(t_1 < t_2 < \cdots < t_N)$, we denote $\mathbf{Z}'^N_1 := \{Z'_{t_n}\}_{n=1}^N$ as the ghost sample set of $\mathbf{Z}_1^N$, where Z'_{t_n} has the same distribution as Z_{t_n} for any $1 \le n \le N$. Then, the following theorem presents the symmetrization inequality of the learning process.

Theorem 8 *Assume that $\mathcal{F}$ is a function class with the range $[a,b]$ and $\{Z_t\}_{t\ge 0} \subset \mathbb{R}^K$ is a Lévy process. Let $\mathbf{Z}_1^N$ and $\mathbf{Z}'^N_1$ be drawn from $\{Z_t\}_{t\ge 0}$ in the time interval $[T_1,T_2]$. Then, given any $\xi > \Sigma_N^{(*)}$, we have for any $N \ge \frac{8(b-a)^2}{(\xi')^2}$ with $\xi' = \xi - \Sigma_N^{(*)}$,*

$$\Pr\left\{\sup_{\substack{f\in\mathcal{F}\\ t\in[T_1,T_2]}} \left|\mathrm{E}_t f - \mathrm{E}_N f\right| > \xi\right\} \le 2\Pr\left\{\sup_{f\in\mathcal{F}} \left|\mathrm{E}'_N f - \mathrm{E}_N f\right| > \frac{\xi'}{2}\right\}. \tag{12}$$

This theorem shows that given $\xi > 0$, the probability of the event:

$$\sup_{\substack{f\in\mathcal{F}\\ t\in[T_1,T_2]}} \left|\mathrm{E}_t f - \mathrm{E}_N f\right| > \xi$$

can be bounded by using the probability of the event:

$$\left|\mathrm{E}'_N f - \mathrm{E}_N f\right| > \frac{\xi'}{2}$$

that is only determined by the characteristics of the sample sets $\mathbf{Z}_1^N$ and $\mathbf{Z}'^N_1$, when $N \geq \frac{8(b-a)^2}{(\xi')^2}$ for any given $\xi' > 0$ with $\xi' = \xi - \Sigma_N^{(*)}$. Compared with the classical symmetrization result under the sample-i.i.d. assumption (see Bousquet et al., 2004), the derived symmetrization inequality (12) incorporated a discrepancy term $\Sigma_N^{(*)}$ and the two results coincide when the time interval $[T_1, T_2]$ shrinks to one time point that match to t, that is, $T_1 = T_2 = t$ that results in $\Sigma_N^{(*)} = 0$.

In the next section, we use the resulted deviation inequalities and symmetrization inequality to achieve the risk bounds of the learning process for Lévy process.

5. Risk Bounds of Learning Processes for Lévy Processes

In this section, we present the risk bounds of the learning process for Lévy process. Since the resulting bounds are based on the covering number, we first introduce the definition of the cover and then present the definition of the covering number of $\mathcal{F}$.

Definition 9 *Let $\mathcal{N}$ be a collection of sets. Then, the collection $\mathcal{N}$ is said to be a cover of a given set Ω, if for any $\mathbf{x} \in \Omega$, there always exists an element of $\mathcal{N}$ that contains the point $\mathbf{x}$.*

Next, we define the the covering number of $\mathcal{F}$ as follows.

Definition 10 *Let $\mathbf{Z}_1^N$ be a sample set drawn from a distribution $\mathbf{Z}$. For any $1 \leq p \leq \infty$ and $\xi > 0$, the covering number of $\mathcal{F}$ at radius ξ, with respect to $\ell_p(\mathbf{Z}_1^N)$, denoted by $\mathcal{N}(\mathcal{F}, \xi, \ell_p(\mathbf{Z}_1^N))$ is the minimum size of a cover of radius ξ.*

Subsequently, we come up with the main results of this paper.

Theorem 11 *Assume that $\mathcal{F}$ is a function class composed of functions satisfying Condition $(C1)$ with the range $[a,b]$ and $\{Z_t\}_{t\geq 0} \subset \mathbb{R}^K$ is a Lévy process with the characteristic triplet $(\mathbf{a}_1, \mathbf{A}_1, \nu_1)$ satisfying Condition $(C2)$. Let $\mathbf{Z}_1^N$ and $\mathbf{Z}'^N_1$ be drawn from $\{Z_t\}_{t\geq 0}$ in the time interval $[T_1, T_2]$, and denote $\mathbf{Z}_1^{2N} := \{\mathbf{Z}_1^N, \mathbf{Z}'^N_1\}$. Given any $\Sigma_N^{(*)} < \xi < \Sigma_N^{(*)} + \frac{8\varphi((M/\lambda)^-)}{N}$, if $\mathrm{E}e^{\alpha\|\mathbf{z}_t\|}\big|_{t=1} < +\infty$ holds for*

some $\alpha > 0$*, then we have for any* $N \geq \frac{8(b-a)^2}{(\xi')^2}$ *with* $\xi' = \xi - \Sigma_N^{(*)}$,

$$\begin{aligned}
&\Pr\left\{\sup_{f\in\mathcal{F}}\frac{1}{2}\left|\mathrm{E}f-\mathrm{E}_N f\right|>\xi\right\}\\
\leq&\Pr\left\{\sup_{\substack{f\in\mathcal{F}\\ t\in[T_1,T_2]}}\left|\mathrm{E}_t f-\mathrm{E}_N f\right|>\xi\right\}\\
\leq&8\mathrm{E}\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1(\mathbf{Z}_1^{2N})\right)\exp\left\{-\int_0^{\frac{N\xi'}{8}}\varphi^{-1}(s)ds\right\},
\end{aligned}\tag{13}$$

where $\varphi(a^-)$ *denotes the left-hand limit of* φ *at the point a,* $M = \sup\left\{\alpha>0 : \mathrm{E}e^{\alpha\|\mathbf{z}_t\|}\big|_{t=1} < +\infty\right\}$ *and* φ^{-1} *is the inverse of* $\varphi(\alpha)$ *with the domain of* $0<\alpha<M/\lambda$.

The result shown in this theorem has the same drawbacks as those of Theorem 6. The right-hand-side of the inequality (13) is represented by using the integrals of φ^{-1}, so it is difficult to find the asymptotic behavior of the risk bound as N goes to the *infinity*. The range $0<\xi'<\frac{8\varphi((M/\lambda)^-)}{N}$ of ξ' is expressed by incorporating the function φ that contains an integral term [see Definition $(D2)$]. These will bring difficulties to the future theoretical analysis of asymptotic convergence. To overcome the two drawbacks, we develop another risk bound of the learning process for Lévy process by adding a mild condition $(C3)$ that requires that the Lévy measure ν have a bounded support.

Theorem 12 *Follow notations in Theorem 11. Given any* $\xi > \Sigma_N^{(*)}$*, if Conditions* $(C1)$*-*$(C3)$ *are valid, then we have for any* $N \geq \frac{8(b-a)^2}{(\xi')^2}$ *with* $\xi' = \xi - \Sigma_N^{(*)}$,

$$\begin{aligned}
&\Pr\left\{\sup_{f\in\mathcal{F}}\frac{1}{2}\left|\mathrm{E}f-\mathrm{E}_N f\right|>\xi\right\}\\
\leq&\Pr\left\{\sup_{\substack{f\in\mathcal{F}\\ t\in[T_1,T_2]}}\left|\mathrm{E}_t f-\mathrm{E}_N f\right|>\xi\right\}\\
\leq&8\mathrm{E}\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1(\mathbf{Z}_1^{2N})\right)\exp\left\{\frac{NT_1(\lambda^2\pi K^2\alpha+V)}{(\lambda R)^2}\Gamma\left(\frac{\lambda R(\xi-\Sigma_N^{(*)})}{8T_2(\lambda^2\pi K^2\alpha+V)}\right)\right\},
\end{aligned}\tag{14}$$

where Γ *is given in Definition* $(D4)$.

This theorem shows that under Conditions $(C1)$*-*$(C3)$, given any $\xi > \Sigma_N^{(*)}$, the probability of the event that for any $N \geq \frac{8(b-a)^2}{(\xi')^2}$ with $\xi' = \xi - \Sigma_N^{(*)}$,

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}f-\mathrm{E}_N f\right|>2\xi$$

can be bounded by the last term of (14). Until now, we have achieved the risk bound (3) and the result (14) can explicitly reflect the asymptotic behavior as N goes to the *infinity*. Following this result, the next section will discuss the asymptotical behavior of the learning process for Lévy process.

6. Convergence Analysis

Based on the risk bound (14), this section presents a detailed theoretical analysis to asymptotic convergence and the rate of convergence of the learning process for Lévy process. Meanwhile, we also give a comparison with the related results of the learning process for i.i.d. samples.

6.1 Asymptotic Convergence

In statistical learning theory, it is well-known that the complexity of function classes is the main factor to the asymptotic convergence of the learning process for i.i.d. samples (see Vapnik, 1998; Van der Vaart and Wellner, 1996; Mendelson, 2003).

Based on Theorem 12, we show that the asymptotic convergence of the learning process for Lévy process is affected by two factors: the covering number $\mathcal{N}(\mathcal{F},\xi'/8,\ell_1(\mathbf{Z}_1^{2N}))$ and the quantity $\Sigma_N^{(*)}$.

Recalling Definition $(D4)$, it is noteworthy that there is only one solution $x=0$ to the equation $\Gamma(x)=0$ and $\Gamma(x)$ is monotonically decreasing when $x\geq 0$ (see Figure 1). Thus, according to Theorem 12, we can obtain the following result that describes the asymptotic convergence of the learning process for Lévy process.

Theorem 13 *Assume that $\mathcal{F}$ is a function class composed of functions satisfying Condition $(C1)$ with the range $[a,b]$ and $\{Z_t\}_{t\geq 0}\subset\mathbb{R}^K$ is a Lévy process with the characteristic triplet $(\mathbf{a}_1,\mathbf{A}_1,\nu_1)$ satisfying Conditions $(C2)$ and $(C3)$. Let $\mathbf{Z}_1^N=\{Z_{t_n}\}_{n=1}^N$ and $\mathbf{Z}'^N_1=\{Z'_{t_n}\}_{n=1}^N$ be drawn from $\{Z_t\}_{t\geq 0}$ in the time interval $[T_1,T_2]$, and denote $\mathbf{Z}_1^{2N}:=\{\mathbf{Z}_1^N,\mathbf{Z}'^N_1\}$. If the following condition holds:*

$$\lim_{N\to+\infty}\frac{\ln \mathrm{E}\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1(\mathbf{Z}_1^{2N})\right)}{N}<+\infty, \tag{15}$$

then we have for any $\xi>\Sigma_N^{(*)}$ with $\xi'=\xi-\Sigma_N^{(*)}$,

$$\lim_{N\to+\infty}\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E}f-\mathrm{E}_Nf\right|>2\xi\right\}=0, \tag{16}$$

where $\mathrm{E}f$ and E_Nf are defined in (1) *and* (2)*, respectively.*

As shown in Theorem 13, if the covering number $\mathcal{N}(\mathcal{F},\xi'/8,\ell_1(\mathbf{Z}_1^{2N}))$ satisfies the condition (15), the probability of the event

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}f-\mathrm{E}_Nf\right|>2\xi$$

will converge to *zero* for any $\xi>\Sigma_N^{(*)}$, when the sample number N goes to the *infinity*. This is partially in accordance with the classical result given by Theorem 2.3 of Mendelson (2003): the probability of the event

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}f-\mathrm{E}_Nf\right|>\xi \tag{17}$$

will converge to *zero* for any $\xi > 0$, if the covering number $\mathcal{N}\left(\mathcal{F}, \xi, \ell_1(\mathbf{Z}_1^N)\right)$ satisfies the following condition:

$$\lim_{N \to +\infty} \frac{\ln \mathrm{E}\left\{\mathcal{N}\left(\mathcal{F}, \xi, \ell_1(\mathbf{Z}_1^N)\right)\right\}}{N} < +\infty. \tag{18}$$

Note that in the learning process for Lévy process, the uniform convergence of the empirical risk $\mathrm{E}_N f$ to the expected risk $\mathrm{E}f$ may not hold, because the limit (16) does not hold for any $\xi > 0$ but for any $\xi > \Sigma_N^{(*)}$. By contrast, the inequality (17) holds for all $\xi > 0$ in the learning process for i.i.d. samples, if the condition (18) is satisfied. Again, these two results coincide when the time interval $[T_1, T_2]$ shrinks to one single time point that matches to t, that is, $T_1 = T_2 = t$ that results in $\Sigma_N^{(*)} = 0$.

Interestingly, we show below that by ignoring the quantity $\Sigma_N^{(*)}$, the learning process for Lévy process has a faster rate of convergence than the classical result (see Mendelson, 2003, Theorem 2.3) in the large-derivation case.

6.2 Rate of Convergence

The classical result (see Mendelson, 2003, Theorem 2.3) is actually derived from Hoeffding's inequality. Thus, it is said to be of Hoeffding-type and can directly lead to its alternative expression as follows:

$$\sup_{f \in \mathcal{F}} \left|\mathrm{E}_N f - \mathrm{E}f\right| \leq O\left(\left(\frac{\ln \mathrm{E}\left\{\mathcal{N}\left(\mathcal{F}, \xi, \ell_1(\mathbf{Z}_1^N)\right)\right\} - \ln(\varepsilon/8)}{N}\right)^{\frac{1}{2}}\right), \tag{19}$$

which implies that the rate of convergence of the i.i.d. learning process is up to $O(1/\sqrt{N})$.

Recalling the classical Bennett's inequality (Bennett, 1962; Bousquet, 2002), we can find that the expression of the risk bound (14) is similar to that of Bennett's inequality, that is, both of them are in the form of $\mathrm{e}^{\Gamma(x)}$ with $\Gamma(x) = x - (x+1)\ln(x+1)$. For convenience, this form is said to be of Bennett-type. Differing from the Hoeffding-type result (see Mendelson, 2003, Theorem 2.3), it is difficult to directly achieve the alternative expression of the Bennett-type result (14), because it is difficult to obtain the analytical expression of the inverse function of $\Gamma(x)$. Instead, one generally uses the term $\frac{-x^2}{2+(2x/3)}$ to approximate the function $\Gamma(x)$ and then get the so-called Bernstein's inequality. In this way, we can obtain the following alternative expression of the Bennett-type result (14):[1]

$$\begin{aligned}\sup_{f \in \mathcal{F}} \left|\mathrm{E}_N f - \mathrm{E}f\right| \leq & 2\Sigma_N^{(*)} + \frac{64\lambda R T_2 \left(\ln \mathrm{E}\left\{\mathcal{N}\left(\mathcal{F}, \xi'/8, \ell_1(\mathbf{Z}_1^N)\right)\right\} - \ln(\varepsilon/8)\right)}{3NT_1} \\ & + \frac{16 T_2 \sqrt{2(\lambda^2 \pi K^2 \alpha + V)\left(\ln \mathrm{E}\left\{\mathcal{N}\left(\mathcal{F}, \xi'/8, \ell_1(\mathbf{Z}_1^N)\right)\right\} - \ln(\varepsilon/8)\right)}}{\sqrt{NT_1}},\end{aligned} \tag{20}$$

which implies that the rate of convergence of the learning process for Lévy process is also up to $O(1/\sqrt{N})$, which is in accordance with the classical result (19), if the discrepancy term $\Sigma_N^{(*)}$ is ignored.

1. The details are referred to `http://ocw.mit.edu/courses/mathematics/18-465-topics-in-statistics-statistical-learning-theory-spring-2007/lecture-notes/l6.pdf`.

Here, we adopt a new method to obtain another alternative expression of the Bennett-type risk bound (14) and show that the rate of convergence of the learning process can be up to $o(1/N^{\frac{1}{1.3}})$ in the large-deviation case.

Remark 14 *Here, "large-deviation" means that the discrepancy between the empirical risk and the expected risk is large (or not small). Given any $\xi > \Sigma_N^{(*)}$ with $\xi' := \xi - \Sigma_N^{(*)}$, one of our major concerns is the probability $\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E}_N f - \mathrm{E}f\right| > \xi\right\}$, and then we say that the case that $\frac{\lambda R\xi'}{8T_2(\lambda^2\pi K^2\alpha+V)} > 1.719$ is of large-deviation, that is, $\xi > \frac{13.752T_2(\lambda^2\pi K^2\alpha+V)}{\lambda R} + \Sigma_N^{(*)}$.*

Theorem 15 *Follow the notations and conditions of Theorem 12. Then, given any $\xi > \frac{13.752T_2(\lambda^2\pi K^2\alpha+V)}{\lambda R} + \Sigma_N^{(*)}$ and for any $N \geq \frac{8(b-a)^2}{(\xi')^2}$ with $\xi' = \xi - \Sigma_N^{(*)}$, we have with probability at least $1-\varepsilon$,*

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}_N f - \mathrm{E}f\right| \leq 2\Sigma_N^{(*)} + \frac{27.504T_2(\lambda^2\pi K^2\alpha+V)}{\lambda R} + \frac{16T_2(\lambda^2\pi K^2\alpha+V)}{\lambda R} \times \left(\frac{(\lambda R)^2\left(\ln \mathrm{E}\left\{\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1(\mathbf{Z}_1^{2N})\right)\right\} - \ln(\varepsilon/8)\right)}{NT_1(\lambda^2\pi K^2\alpha+V)}\right)^{\frac{1}{\gamma}},$$

where

$$\varepsilon := 8\mathrm{E}\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1(\mathbf{Z}_1^{2N})\right)\exp\left\{\frac{NT_1(\lambda^2\pi K^2\alpha+V)}{(\lambda R)^2}\Gamma\left(\frac{\lambda R(\xi-\Sigma_N^{(*)})}{8T_2(\lambda^2\pi K^2\alpha+V)}\right)\right\},$$

and $0 < \gamma \leq \gamma\left(\frac{\lambda R\xi'}{8T_2(\lambda^2\pi K^2\alpha+V)}\right) < 1.3$ with

$$\gamma(x) := \frac{\ln\left((x+1)\ln(x+1)-x\right)}{\ln x}.$$

The above theorem provides another upper bound of the risk bound $\sup_{f\in\mathcal{F}}\left|\mathrm{E}_N f - \mathrm{E}f\right|$ in the large-deviation case, where 1.719 is the numerical solution to the equation $\gamma(x) = 0$. Compared to the classical result (19), there is a discrepancy quantity $\Sigma_N^{(*)}$ that also appears in the Bernstein-type result (20). Interestingly, in the large-deviation case, the risk bound (14) can provide a faster rate $o\left(\frac{1}{N^{1/1.3}}\right)$ of convergence than the rate $O\left(\frac{1}{N^{1/2}}\right)$ of the classical result (19) and the Bernstein-type result (20). Note that the rate $o\left(\frac{1}{N^{1/1.3}}\right)$ will not hold if the large-deviation case is not valid (that is, $0 < \frac{\lambda R\xi'}{8T_2(\lambda^2\pi K^2\alpha+V)} \leq 1.719$), while the Bernstein-type result (20) for the learning process performs well and provides the rate $O\left(\frac{1}{N^{1/2}}\right)$ regardless of whether the large-deviation case is valid.

7. Conclusion

In this paper, we study the risk bounds of the learning process for time-dependent samples drawn from a Lévy process. We first provide the deviation inequalities and the symmetrization inequality of the learning process, respectively. We then use the resulted deviation inequalities and symmetrization inequality to derive the risk bounds based on the covering number.

By using the risk bound shown in Theorem 12, we analyze the asymptotic convergence and the rate of convergence of the learning process for Lévy process. We point out that the asymptotic convergence of such learning process is affected by two factors: the complexities of the function class $\mathcal{F}$ measured by the covering number and the quantity $\Sigma_N^{(*)}$. This is partially in accordance with the classical result on the asymptotic convergence of the learning process for i.i.d. samples (see Mendelson, 2003). Due to the quantity $\Sigma_N^{(*)}$, the uniform convergence of the learning process for Lévy process may not be valid. We also show that the rate of convergence of the learning process is up to $O(1/\sqrt{N})$, which matches with the the classical result under the sample-i.i.d. assumption. Furthermore, we adopt a new method to obtain another alternative expression of the risk bound (14) and then find that the rate of convergence of the learning process can reach $o(1/N^{\frac{1}{1.3}})$ in the large-deviation case. Note that as stated in Sections 3 & 5, the faster rate of convergence is actually provided by the specific deviation inequality (17) which is of Bennett-type (that is, its expression is similar to that of Bennett's inequality), while the classical result (19) is derived from Hoeffding's inequality (see Mendelson, 2003).

In our future work, we will attempt to study risk bounds for other stochastic processes via some specific concentration or deviation inequalities, for example, stochastic processes with exchangeable increments that are a well-known generalization of stochastic processes with independent increments (Kallenberg, 1973; Kallenberg et al., 1975). Then, we will develop the risk bounds of the learning process for Lévy process by using other complexity measures, for example, the Rademacher complexity and the fat-shattering dimension.

Acknowledgments

We are grateful to the anonymous reviewers and the editors for their valuable comments and suggestions. This project was supported by Australian Research Council Discovery Project with number ARC DP-120103730.

Appendix A. Proof of Theorem 8

Proof of Theorem 8. Let $\widehat{f}$ and $\widehat{t}$ be the function and the time achieving the supremum

$$\sup_{\substack{f\in\mathcal{F}\\ t\in[T_1,T_2]}} \left|\mathrm{E}_t f - \mathrm{E}_N f\right|$$

with respect to $\mathbf{Z}_1^N$, respectively. According to Definition (D1), we have

$$\begin{aligned}\left|\mathrm{E}_{\widehat{t}}\widehat{f} - \mathrm{E}_N\widehat{f}\right| =& \left|\mathrm{E}_{\widehat{t}}\widehat{f} - \frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f} + \frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f} - \mathrm{E}_N\widehat{f}\right| \\ \leq& \Sigma_N^{(*)} + \left|\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f} - \mathrm{E}_N\widehat{f}\right|,\end{aligned}$$

which can lead to for any $\xi > \Sigma_N^{(*)}$ with $\xi' = \xi - \Sigma_N^{(*)}$,

$$\Pr\left\{\left|\mathrm{E}_{\widehat{t}}\widehat{f} - \mathrm{E}_N\widehat{f}\right| > \xi\right\} \leq \Pr\left\{\left|\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f} - \mathrm{E}_N\widehat{f}\right| > \xi'\right\}.$$

According to the triangle inequality, we have

$$\left|\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f}-\mathrm{E}_N\widehat{f}\right|-\left|\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f}-\mathrm{E}'_N\widehat{f}\right|\leq\left|\mathrm{E}'_N\widehat{f}-\mathrm{E}_N\widehat{f}\right|. \tag{21}$$

Let $\mathcal{A}$ stand for an event and denote the indicator function of the event $\mathcal{A}$ as

$$\mathbf{1}_{\mathcal{A}}=\begin{cases}1, & \text{if } \mathcal{A} \text{ occurs;}\\ 0, & \text{otherwise.}\end{cases}$$

By denoting $\wedge$ as the conjunction of two events, it is followed from (21) that

$$\begin{aligned}&\left(\mathbf{1}_{\left|\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f}-\mathrm{E}_N\widehat{f}\right|>\xi'}\right)\left(\mathbf{1}_{\left|\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f}-\mathrm{E}'_N\widehat{f}\right|<\frac{\xi'}{2}}\right)\\ =&\mathbf{1}_{\left\{\left|\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f}-\mathrm{E}_N\widehat{f}\right|>\xi'\right\}\wedge\left\{\left|\mathrm{E}'_N\widehat{f}-\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f}\right|<\frac{\xi'}{2}\right\}}\\ \leq&\mathbf{1}_{\left|\mathrm{E}'_N\widehat{f}-\mathrm{E}_N\widehat{f}\right|>\frac{\xi'}{2}}.\end{aligned}$$

Then, taking the expectation with respect to $\mathbf{Z}'^{N}_{1}$ gives

$$\begin{aligned}&\left(\mathbf{1}_{\left|\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f}-\mathrm{E}_N\widehat{f}\right|>\xi'}\right)\Pr'\left\{\left|\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f}-\mathrm{E}'_N\widehat{f}\right|<\frac{\xi'}{2}\right\}\\ \leq&\Pr'\left\{\left|\mathrm{E}'_N\widehat{f}-\mathrm{E}_N\widehat{f}\right|>\frac{\xi'}{2}\right\}.\end{aligned} \tag{22}$$

By Chebyshev's inequality, we have for any $\xi'>0$,

$$\begin{aligned}\Pr'\left\{\left|\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f}-\mathrm{E}'_N\widehat{f}\right|\geq\frac{\xi'}{2}\right\}=&\Pr'\left\{\left|\sum_{n=1}^{N}(\mathrm{E}_{t_n}\widehat{f}-\widehat{f}(Z'_{t_n})\right|\geq\frac{N\xi'}{2}\right\}\\ \leq&\frac{4\mathrm{E}\left\{\sum_{n=1}^{N}\left(\mathrm{E}_{t_n}\widehat{f}-\widehat{f}(Z'_{t_n})\right)^2\right\}}{N^2(\xi')^2}\\ \leq&\frac{4N(b-a)^2}{N^2(\xi')^2}=\frac{4(b-a)^2}{N(\xi')^2}.\end{aligned} \tag{23}$$

Subsequently, according to (22) and (23), we have for any $\xi'>0$,

$$\Pr'\left\{\left|\mathrm{E}'_N\widehat{f}-\mathrm{E}_N\widehat{f}\right|>\frac{\xi'}{2}\right\}\geq\left(\mathbf{1}_{\left|\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f}-\mathrm{E}_N\widehat{f}\right|>\xi'}\right)\left(1-\frac{4(b-a)^2}{N(\xi')^2}\right).$$

Let

$$\frac{4(b-a)^2}{N(\xi')^2}\leq\frac{1}{2}$$

and take the expectation with respect to $\mathbf{Z}_1^N$. Given any $\xi>\Sigma_N^{(*)}$, we then have for any $N\geq\frac{8(b-a)^2}{(\xi')^2}$ with $\xi'=\xi-\Sigma_N^{(*)}$,

$$\begin{aligned}\Pr\left\{\left|\mathrm{E}_{\widehat{t}}\widehat{f}-\mathrm{E}_N\widehat{f}\right|>\xi\right\}\leq&2\Pr\left\{\left|\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_{t_n}\widehat{f}-\mathrm{E}_N\widehat{f}\right|>\frac{\xi'}{2}\right\}\\ \leq&2\Pr\left\{\left|\mathrm{E}'_N\widehat{f}-\mathrm{E}_N\widehat{f}\right|>\frac{\xi'}{2}\right\}.\end{aligned}$$

This completes the proof. ■

Appendix B. Proofs of Theorems 11 & 12

Proof of Theorem 11. Consider $\{\varepsilon_n\}_{n=1}^N$ as independent Rademacher random variables, that is, independent $\{-1,1\}$-valued random variables with equal probability of taking either value. Given $\{\varepsilon_n\}_{n=1}^N$ and $\mathbf{Z}_1^{2N}$, denote

$$\overrightarrow{\varepsilon} := (\varepsilon_1, \cdots, \varepsilon_N, -\varepsilon_1, \cdots, -\varepsilon_N)^T, \tag{24}$$

and for any $f \in \mathcal{F}$,

$$\overrightarrow{f}(\mathbf{Z}_1^{2N}) := \big(f(Z'_{t_1}), \cdots, f(Z'_{t_N}), f(Z_{t_1}), \cdots, f(Z_{t_N})\big)^T. \tag{25}$$

According to (6) and Theorem 8, given any $\xi > \Sigma_N^{(*)}$, we have for any $N \geq \frac{8(b-a)^2}{(\xi')^2}$ with $\xi' = \xi - \Sigma_N^{(*)}$,

$$\begin{aligned}
&\Pr\left\{\sup_{\substack{f\in\mathcal{F}\\ t\in[T_1,T_2]}} \big|\mathrm{E}_t f - \mathrm{E}_N f\big| > \xi\right\}\\
\leq& 2\Pr\left\{\sup_{f\in\mathcal{F}} \big|\mathrm{E}'_N f - \mathrm{E}_N f\big| > \frac{\xi'}{2}\right\} \qquad \text{(by Theorem 8)}\\
=& 2\Pr\left\{\sup_{f\in\mathcal{F}} \Big|\frac{1}{N}\sum_{n=1}^N \big(f(Z'_{t_n}) - f(Z_{t_n})\big)\Big| > \frac{\xi'}{2}\right\}\\
=& 2\Pr\left\{\sup_{f\in\mathcal{F}} \Big|\frac{1}{N}\sum_{n=1}^N \varepsilon_n\big(f(Z'_{t_n}) - f(Z_{t_n})\big)\Big| > \frac{\xi'}{2}\right\} \qquad \text{(since } Z'_{t_n} \text{ and } Z_{t_n} \text{ are i.i.d.)}\\
=& 2\Pr\left\{\sup_{f\in\mathcal{F}} \Big|\frac{1}{2N}\big\langle\overrightarrow{\varepsilon}, \overrightarrow{f}(\mathbf{Z}_1^{2N})\big\rangle\Big| > \frac{\xi'}{4}\right\}. \qquad \text{(by (24) and (25))}
\end{aligned} \tag{26}$$

Fix a realization of $\mathbf{Z}_1^{2N}$ and let Λ be a $\xi'/8$-radius cover of $\mathcal{F}$ with respect to the $\ell_1(\mathbf{Z}_1^{2N})$ norm. Since $\mathcal{F}$ is composed of the λ-Lipschitz functions with the range $[a,b]$, we assume that the same holds for any $h \in \Lambda$. If $\widehat{f}$ is the function that achieves $\sup_{f\in\mathcal{F}} \frac{1}{2N}\big|\big\langle\overrightarrow{\varepsilon}, \overrightarrow{f}(\mathbf{Z}_1^{2N})\big\rangle\big| > \frac{\xi'}{4}$, there must be an $\widehat{h} \in \Lambda$ that satisfies that

$$\frac{1}{2N}\sum_{n=1}^N \Big(|\widehat{f}(Z'_{t_n}) - \widehat{h}(Z'_{t_n})| + |\widehat{f}(Z_{t_n}) - \widehat{h}(Z_{t_n})|\Big) < \frac{\xi'}{8},$$

and meanwhile,

$$\sup_{h\in\Lambda} \frac{1}{2N}\big|\big\langle\overrightarrow{\varepsilon}, \overrightarrow{h}(\mathbf{Z}_1^{2N})\big\rangle\big| > \frac{\xi'}{8}.$$

Therefore, for the realization of $\mathbf{Z}_1^{2N}$, we arrive at

$$\Pr\left\{\sup_{f\in\mathcal{F}} \Big|\frac{1}{2N}\big\langle\overrightarrow{\varepsilon}, \overrightarrow{f}(\mathbf{Z}_1^{2N})\big\rangle\Big| > \frac{\xi'}{4}\right\} \leq \Pr\left\{\sup_{h\in\Lambda} \Big|\frac{1}{2N}\big\langle\overrightarrow{\varepsilon}, \overrightarrow{h}(\mathbf{Z}_1^{2N})\big\rangle\Big| > \frac{\xi'}{8}\right\}. \tag{27}$$

Moreover, we denote the event

$$A := \left\{ \Pr\left\{ \sup_{h\in\Lambda} \left| \frac{1}{2N} \langle \overrightarrow{\varepsilon}, \overrightarrow{h}(\mathbf{Z}_1^{2N}) \rangle \right| > \frac{\xi'}{8} \right\} \right\},$$

and let $\mathbf{1}_A$ be the characteristic function of the event A. By Fubini's Theorem, we have

$$\Pr\{A\} = \mathrm{E}\left\{ \mathrm{E}_{\overrightarrow{\varepsilon}}\{\mathbf{1}_A\} \,\middle|\, \mathbf{Z}_1^{2N} \right\} = \mathrm{E}\left\{ \Pr\left\{ \sup_{h\in\Lambda} \left| \frac{1}{2N} \langle \overrightarrow{\varepsilon}, \overrightarrow{h}(\mathbf{Z}_1^{2N}) \rangle \right| > \frac{\xi'}{8} \right\} \,\middle|\, \mathbf{Z}_1^{2N} \right\}. \tag{28}$$

Fix a realization of $\mathbf{Z}_1^{2N}$ again. According to (24), (25) and Theorem 6, we have

$$\begin{aligned}
&\Pr\left\{ \sup_{h\in\Lambda} \left| \frac{1}{2N} \langle \overrightarrow{\varepsilon}, \overrightarrow{h}(\mathbf{Z}_1^{2N}) \rangle \right| > \frac{\xi'}{8} \right\} \\
\leq & |\Lambda| \max_{h\in\Lambda} \Pr\left\{ \left| \frac{1}{2N} \langle \overrightarrow{\varepsilon}, \overrightarrow{h}(\mathbf{Z}_1^{2N}) \rangle \right| > \frac{\xi'}{8} \right\} \\
= & \mathcal{N}\left(\mathcal{F}, \xi'/8, \ell_1(\mathbf{Z}_1^{2N})\right) \max_{h\in\Lambda} \Pr\left\{ \left| \mathrm{E'}_N h - \mathrm{E}_N h \right| > \frac{\xi'}{4} \right\} \\
\leq & \mathcal{N}\left(\mathcal{F}, \xi'/8, \ell_1(\mathbf{Z}_1^{2N})\right) \max_{h\in\Lambda} \Pr\left\{ \left| \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}_{t_n} h - \mathrm{E'}_N h \right| + \left| \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}_{t_n} h - \mathrm{E}_N h \right| > \frac{\xi'}{4} \right\} \\
\leq & 2\mathcal{N}\left(\mathcal{F}, \xi'/8, \ell_1(\mathbf{Z}_1^{2N})\right) \max_{h\in\Lambda} \Pr\left\{ \left| \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}_{t_n} h - \mathrm{E}_N h \right| > \frac{\xi'}{8} \right\} \\
\leq & 4\mathcal{N}\left(\mathcal{F}, \xi'/8, \ell_1(\mathbf{Z}_1^{2N})\right) \exp\left\{ -\int_0^{\frac{N\xi'}{8}} \varphi^{-1}(s) ds \right\}.
\end{aligned} \tag{29}$$

The combination of (26), (27), (28) and (29) leads to the result (13). This completes the proof. ■

In the similar way, we can also prove Theorem 12.

Proof of Theorem 12. Similarly, by (11), we have

$$\begin{aligned}
&\Pr\left\{ \sup_{h\in\Lambda} \left| \frac{1}{2N} \langle \overrightarrow{\varepsilon}, \overrightarrow{h}(\mathbf{Z}_1^{2N}) \rangle \right| > \frac{\xi'}{8} \right\} \\
\leq & 2\mathcal{N}\left(\mathcal{F}, \xi'/8, \ell_1(\mathbf{Z}_1^{2N})\right) \max_{h\in\Lambda} \Pr\left\{ \left| \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}_{t_n} h - \mathrm{E}_N h \right| > \frac{\xi'}{8} \right\} \\
\leq & 4\mathcal{N}\left(\mathcal{F}, \xi'/8, \ell_1(\mathbf{Z}_1^{2N})\right) \exp\left\{ \frac{NT_1(\lambda^2\pi K^2\alpha + V)}{(\lambda R)^2} \Gamma\left(\frac{\lambda R(\xi - \Sigma_N^{(*)})}{8T_2(\lambda^2\pi K^2\alpha + V)} \right) \right\}.
\end{aligned} \tag{30}$$

Then, the combination of (26), (27), (28) and (30) can lead to the result (14). This completes the proof. ■

B.1 Proof of Theorem 15

Proof of Theorem 15. Given any $x > 1$, consider the following equation with respect to $\gamma > 0$

$$x - (x+1)\ln(x+1) = -x^{\gamma}, \tag{31}$$

and denote its solution as

$$\gamma(x) := \frac{\ln\big((x+1)\ln(x+1)-x\big)}{\ln(x)}. \tag{32}$$

It is evident that $\gamma(x)$ is a continuously differentiable function with respect to $x > 1$ and there is only one solution to the equation $\gamma(x) = 0$. Its numerical solution is $\bar{x} \approx 1.719$ and $\gamma(x) > 0$ holds for all $x > \bar{x} \approx 1.719$. Then, given any $x > 1.719$, we have for any $0 < \widetilde{\gamma} \le \gamma(x)$,

$$x - (x+1)\ln(x+1) \le -x^{\widetilde{\gamma}}. \tag{33}$$

By combining Theorem 12, (31), (32) and (33), we can straightforwardly show an upper bound of the risk bound $\sup_{f\in\mathcal{F}} \big|\mathrm{E}_N f - \mathrm{E}f\big|$ in the large-deviation case: letting

$$\varepsilon := 8\mathrm{E}\mathcal{N}\big(\mathcal{F}, \xi'/8, \ell_1(\mathbf{Z}_1^{2N})\big) \exp\left\{ \frac{NT_1(\lambda^2\pi K^2\alpha + V)}{(\lambda R)^2} \Gamma\left(\frac{\lambda R(\xi - \Sigma_N^{(*)})}{8T_2(\lambda^2\pi K^2\alpha + V)} \right) \right\}.$$

and with probability at least $1-\varepsilon$,

$$\begin{aligned} \sup_{f\in\mathcal{F}} \big|\mathrm{E}_N f - \mathrm{E}f\big| \le{}& 2\Sigma_N^{(*)} + \frac{27.504T_2(\lambda^2\pi K^2\alpha + V)}{\lambda R} + \frac{16T_2(\lambda^2\pi K^2\alpha + V)}{\lambda R} \\ & \times \left(\frac{(\lambda R)^2 \left(\ln \mathrm{E}\left\{\mathcal{N}\big(\mathcal{F}, \xi'/8, \ell_1(\mathbf{Z}_1^{2N})\big)\right\} - \ln(\varepsilon/8)\right)}{NT_1(\lambda^2\pi K^2\alpha + V)} \right)^{\frac{1}{\gamma}}, \end{aligned}$$

where $0 < \gamma \le \gamma\left(\frac{\lambda R\xi'}{8T_2(\lambda^2\pi K^2\alpha+V)}\right)$ with $\frac{\lambda R\xi'}{8T_2(\lambda^2\pi K^2\alpha+V)} > 1.719$. Thus, we only need to find the upper bound of the function $\gamma(x)$ when $x > 1.719$.

According to (32), for any $x > 1.719$, we consider the derivative of $\gamma(x)$

$$\gamma'(x) = \frac{\ln(x+1)}{\ln(x)\big((x+1)\ln(x+1)-x\big)} - \frac{\ln\big((x+1)\ln(x+1)-x\big)}{x(\ln x)^2}, \tag{34}$$

and draw the function curve of $\gamma'(x)$ in Figure 2.

Figure 2 shows that there is only one solution to the equation $\gamma'(x) = 0$ $(x > 1.719)$. Letting the solution be $\widehat{x}$, we then have $\gamma'(x) > 0$ $(1.719 < x < \widehat{x})$ and $\gamma'(x) < 0$ $(x > \widehat{x})$, that is, $\gamma(x)$ is monotonically decreasing when $x > \widehat{x}$. Meanwhile, by (34), there holds that

$$\lim_{x\to+\infty} \gamma'(x) = 0. \tag{35}$$

Furthermore, we study the second derivative of $\gamma''(x)$

$$\begin{aligned} \gamma''(x) = {}& \frac{\ln\big((x+1)\ln(x+1)-x\big)}{x^2(\ln x)^2} - \frac{1}{(x+1)(x-(x+1)\ln(x+1))\ln x} \\ & + \frac{2\ln\big((x+1)\ln(x+1)-x\big)}{x^2(\ln x)^3} - \frac{(\ln(x+1))^2}{(x-(x+1)\ln(x+1))^2\ln x} \\ & + \frac{2\ln(x+1)}{x(\ln x)^2(x-(x+1)\ln(x+1))}, \end{aligned} \tag{36}$$

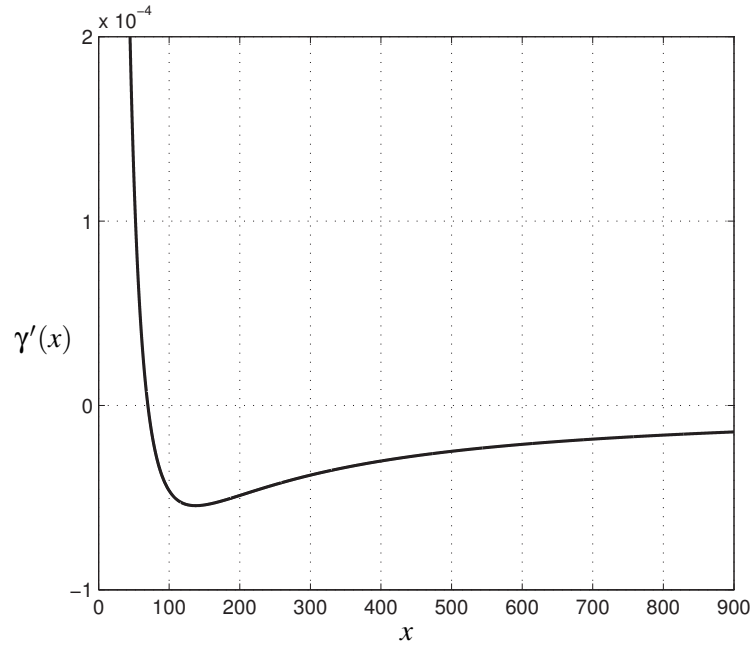

Figure 2: The Function Curve of $\gamma'(x)$

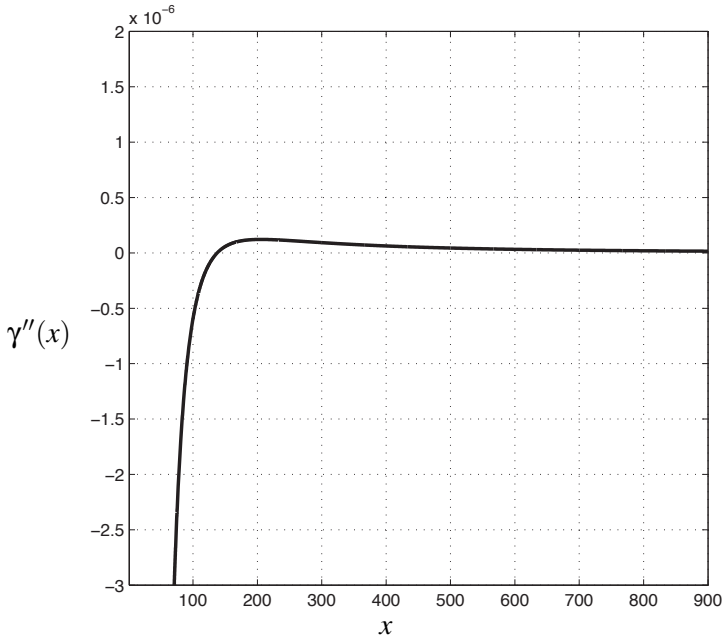

Figure 3: The Function Curve of $\gamma''(x)$

and draw the function curve of $\gamma''(x)$ in Figure 3. This figure shows that there is a solution to the equation $\gamma''(x) = 0$ and its value approximately equals to 137.67. Moreover, according to (36), we arrive at

$$\lim_{x \to +\infty} \gamma''(x) = 0. \tag{37}$$

Therefore, by combining (34), (35), (36) and (37), we obtain that $\gamma(x)$ has only one global maximum point when $x > 1.719$ and thus the solution $\widehat{x}$ to the equation $\gamma'(x) = 0$ also achieves

$$\widehat{x} = \arg \max_{x > 1.719} \gamma(x).$$

Our further numerical experiment shows that the value of $\widehat{x}$ approximately equals to 69.85 and the maximum of $\gamma(x)$ $(x > 1.719)$ is not larger than 1.3 (see Figure 4). This completes the proof. ■

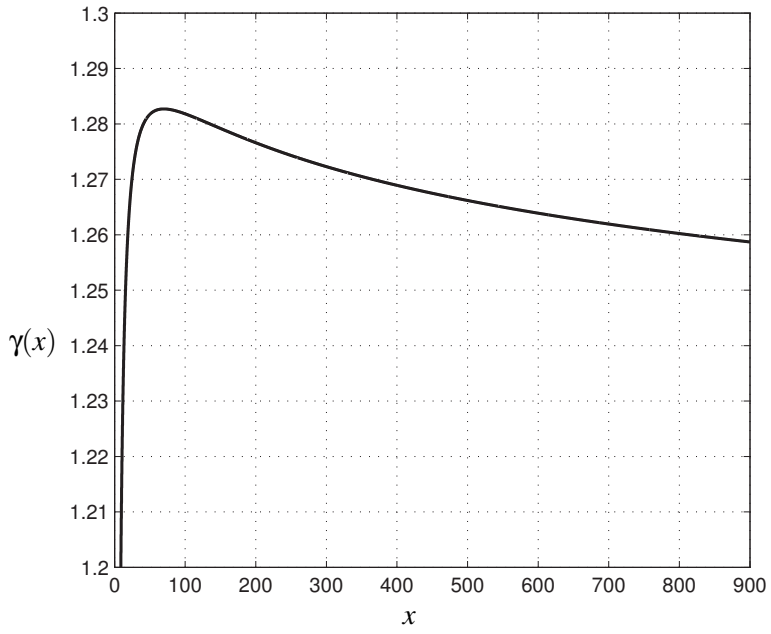

Figure 4: The Function Curve of $\gamma(x)$

References

D. Applebaum. *Lévy Processes and Stochastic Calculus*. Cambridge: Cambridge Press, 2004a.

D. Applebaum. Lévy processes-from probability to finance and quantum groups. *Notices of the American Mathematical Society*, 51:1336–1347, 2004b.

O.E. Barndorff-Nielsen, T. Mikosch, and S.I. Resnick. *Lévy Processes: Theory and Applications*. Birkhauser, 2001.

P.L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

M. Biguesh and A.B. Gershman. Training-based mimo channel estimation: a study of estimator tradeoffs and optimal training signals. *IEEE Transactions on Signal Processing*, 54(3):884–893, 2006.

A. Bose, A. Dasgupta, and H. Rubin. A contemporary review and bibliography of infinitely divisible distributions and processes. *The Indian Journal of Statistics, Series A*, 64:763–819, 2002.

O. Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. *Advanced Lectures on Machine Learning*, pages 169–207, 2004.

N. Cesa-Bianchi and C. Gentile. Improved risk tail bounds for on-line algorithms. *IEEE Transactions on Information Theory*, 54(1):386–390, 2008.

R. Cont and P. Tankov. Retrieving lévy processes from option prices: Regularization of an ill-posed inverse problem. *SIAM Journal on Control and Optimization*, 45(1):1–25, 2006.

T.E. Duncan. Mutual information for stochastic signals and lévy processes. *IEEE Transactions on Information Theory*, 56(1):18–24, 2009.

J.E. Figueroa-López and C. Houdré. Risk bounds for the non-parametric estimation of lévy processes. *Lecture Notes-Monograph Series*, pages 96–116, 2006.

C. Houdré. Remarks on deviation inequalities for functions of infinitely divisible random vectors. *Annals of probability*, pages 1223–1237, 2002.

C. Houdré and P. Marchal. Median, concentration and fluctuations for lévy processes. *Stochastic Processes and their Applications*, 118(5):852–863, 2008.

C. Houdré, V. Pérez-Abreu, and D. Surgailis. Interpolation, correlation identities, and inequalities for infinitely divisible variables. *Journal of Fourier Analysis and Applications*, 4(6):651–668, 1998.

M. Jacobsen. *Point Process Theory and Applications: Marked Point and Piecewise Deterministic Processes*. Birkhäuser Boston, 2005.

W. Jiang. On the uniform deviations of general empirical risks with unboundedness, dependence, and high dimensionality. *Journal of Machine Learning Research*, 10:977–996, 2009.

O. Kallenberg. Canonical representations and convergence criteria for processes with interchangeable increments. *Probability Theory and Related Fields*, 27(1):23–36, 1973.

O. Kallenberg et al. On symmetrically distributed random measures. *Trans. Amer. Math. Soc*, 202: 105–121, 1975.

K. Kim. Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1): 307–319, 2003.

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

A. Kyprianou. *Introductory Lectures on Fluctuations of Lévy Processes with Applications (Universitext)*. Springer, 2006.

D.J. Love, R.W. Heath, V.K.N. Lau, D. Gesbert, B.D. Rao, and M. Andrews. An overview of limited feedback in wireless communication systems. *IEEE Journal on Selected Areas Communications*, 26(8):1341–1365, 2008.

S. Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE Transactions on Information Theory*, 48(1):251–263, 2002.

S. Mendelson. A few notes on statistical learning theory. *Advanced Lectures on Machine Learning*, pages 1–40, 2003.

S. Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Transactions on Information Theory*, 54(8):3797–3803, 2008.

M. Mohri and A. Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11:798–814, 2010.

S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using support vector machines. In *IEEE Workshop on Neural Networks for Signal Processing VII*, pages 511–520, 1997.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

A.B. Nobel and A. Dembo. A note on uniform laws of averages for dependent processes. *Statistics & Probability Letters*, 17(3):169–172, 1993.

K.S. Pedersen, R. Duits, and M. Nielsen. On α kernels, lévy processes, and natural image statistics. In Kimmel, Sochen, and Weickert, editors, *Scale Space and PDE Methods in Computer Vision*, pages 468–479, 2005.

S.T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. New York: Wiley, 1991.

M. Reid and B. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

M. Sanchez-Fernandez, M. de Prado-Cumplido, J. Arenas-Garcia, and F. Perez-Cruz. Svm multiregression for nonlinear channel estimation in multiple-input multiple-output systems. *IEEE Transactions on Signal Processing*, 52(8):2298–2307, 2004.

K. Sato. *Lévy Processes and Infinite Divisible Distributions (Cambridge Studies in Advanced Mathematics)*. USA: Cambridge University Press, 2004.

B.K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G.R.G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

A. Sutivong, M. Chiang, T.M. Cover, and Y.H. Kim. Channel capacity and state estimation for state-dependent gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1486–1495, 2005.

A.M. Tulino, A. Lozano, and S. Verdú. Impact of antenna correlation on the capacity of multiantenna channels. *IEEE Transactions on Information Theory*, 51(7):2491–2509, 2005.

A. Van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

V.N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, 22(1):94–116, 1994.

C. Zhang and D. Tao. Risk bounds for lévy processes in the pac-learning framework. *Journal of Machine Learning Research-Proceedings Track*, 9:948–955, 2010.

C. Zhang and D. Tao. Generalization bound for infinitely divisible empirical process. *J. Mach. Learn. Res.-Proc. Track*, 15:864–872, 2011a.

C. Zhang and D. Tao. Risk bounds for infinitely divisible distribution. In *The 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 2011b.

V.M. Zolotarev. Probability metrics. *Theory of Probability and its Application*, 28(1):278–302, 1984.