Proceedings of the 36th Chinese Control Conference
July 26-28, 2017, Dalian, China

A New Convergent Algorithm for Online Empirical Risk Minimization

Xinqiang Wang[1], Lei Guo[2]

1. Key Lab. of Systems and Control, Institute of Systems Science, AMSS, CAS, Beijing 100190, P. R. China
E-mail: wangxq@amss.ac.cn

2. Key Lab. of Systems and Control, Institute of Systems Science, AMSS, CAS, Beijing 100190, P. R. China
E-mail: lguo@amss.ac.cn

Abstract: The generalization ability of learning algorithms is the focus of machine learning research, where the empirical risk minimization (ERM) plays an important role when the population distribution of observations is unknown. Most of the previous results are mainly based on computational learning theory, which is interested in how many samples are needed to make sure the estimated expected risk satisfies a given accuracy with high probability. In this paper, we will propose a new algorithm by combining the advantages of both random search and gradient descent algorithms, and show that given an accuracy level of the estimated expected risk, we can generate a hypothesis by our algorithm to guarantee the accuracy with probability 1, and our algorithm will converge in finite steps. In addition, we will relax the conventional independently and identically distributed(i.i.d.) assumption on the observations to a kind of weakly dependent condition. We will also provide some simulations to demonstrate our algorithm's advantages over either random search or gradient descent algorithms.

Key Words: ERM, random search, convergent, gradient descent, machine learning

1 Introduction

There has been considerable attention paid to the analysis of the machine learning algorithms recently. With the advances of the computing and network technology, more and more valuable information can be discovered by the machine learning algorithms, especially when combined with the deep neural networks. Traditionally, it is assumed that the observations are obtained from an unknown i.i.d.distribution and that the empirical risk minimization(ERM) is a main method to approximate the expected risk commonly known as generalization ability of the learning algorithms. A key problem is whether the expected risk is minimized when the empirical risk is minimized. Vapnik [1] first obtained the bounds on both the rate of uniform convergence to the expected risk and the rate of relative uniform convergence for i.i.d. observations. Cucker and Smale [2] studied the least squares loss case and established the bound on the sample error and the approximation error respectively for i.i.d. observations on a compact hypothesis space.

Much efforts has been made in relaxing the i.i.d. requirements on the data in both machine learning and statistical inference. For examples, Yu [4] analysed the rate of uniform convergence for stationary mixing sequences. Vidyasagar [5] established the bound on the rate of uniform convergence of the empirical means to their expectation for mixing sequences. Zhou and Li [6] obtained the bound on the rate of uniform convergence for learning algorithms by using Bernstein's inequality. Zhou, Li and Xu [7] established the exponentially bound on the rate of relative uniform convergence of the ERM algorithm with exponentially strongly mixing observations.

However, all the above-mentioned works did not provide any concrete algorithms for minimizing the empirical risk. In this paper, we will propose a leaning algorithm by combining the advantages of the well-known random search and gradient descent algorithms for iterative optimization of the empirical risk with possibly non-convex loss function. It is conceivable and provable that our optimization algorithm will avoid the possibility of sticking at a local minima (or saddle point) as the pure gradient descent method, and at the same time has a much faster convergence rate than the pure random search.

The rest of this paper is organized as follows: In Sect.2, we introduce our algorithm with some notations. In Sect.3, we present the main results of this paper together with some lemmas. In Sect.4, we provide some simulations to demonstrate the advantages of our algorithm. Finally, we conclude the paper with some remarks.

2 Preliminaries

In this section, we introduce some definitions and notations used throughout the paper.

Let $\mathbf{Z} = \{z_i = (x_i, y_i)\}_{i=1}^{\infty}$ be a stationary real-valued sequence on a probability space $(\Omega, \mathcal{B}, P)$. For $0 \le t < \infty$, let $\mathcal{F}_t^{\infty} := \sigma\{z_i, i > t\}$, and $\mathcal{F}_0^t := \sigma\{z_i, 0 \le i \le t\}$. With these notations, we give the definition of the *-mixing process in [3].

Definition 2.1 ([3]): $\{z_i, i \ge 1\}$ is called *-mixing if there exists an integer M, and a function $\phi(m)$ for which $\phi(m) \to 0$ as $m \to \infty$, and for any $A \in \mathcal{F}_i^n, B \in \mathcal{F}_{m+n}^{m+n}$ implies

$$|P(A \cap B) - P(A)P(B)| \le \phi(m)P(A)P(B),\ \forall m \ge M, n \ge 1,$$

where $\phi(m)$ is called the *-mixing coefficient.

Definition 2.2 ([3]): $\{z_i, i \ge 1\}$ is called strong ϕ-mixing if there there exist an integer M, and a function $\phi(m)$ for which $\phi(m) \to 0$ as $m \to \infty$, and for any $A \in \mathcal{F}_i^n, B \in \mathcal{F}_{m+n}^{\infty}$ implies

$$|P(A \cap B) - P(A)P(B)| \le \phi(m)P(A)P(B),\ \forall m \ge M, n \ge 1.$$

Next we give the relevant notations on ERM. Denote by $\mathbf{z}_t$ the sample set of size t observations $\mathbf{z}_t = \{z_1, z_2, \cdots, z_t\}$ drawn from the *-mixing sequence $\mathbf{Z}$.

Let us consider the following stochastic optimization problem

$$\min L(w) := E[\ell(w,z)] = \int_Z \ell(w,z)P(dz) \qquad (1)$$

where w takes values in a compact set $D \subset \mathbb{R}^d$ and z is a random element of $\mathbf{Z}$ with an unknown probability law P, $\ell(w,z)$ is the loss function which is used to measure the loss for predictions and classifications, and $L(w)$ is called the expected risk. For instance, the loss function of the least square estimation is $\ell(w,z) = (y - w^T x)^2$ and the one of logistic regression is $\ell(w,z) = y\log(1+\exp(-w^T x)) + (1-y)\log(1+\exp(w^T x))$. Here we simply describe the loss function in a parametric form. The goal of machining learning is to generate a hypothesis $\hat{w} \in D$ with small expected excess risk

$$E[L(\hat{w})] - L^* \qquad (2)$$

where $L^* := \inf_{w\in D} L(w)$ which is the minimum value of the expected risk on hypothesis space D, and the expectation is with respect to the observations $\mathbf{z}_t$ and any additional randomness used by the algorithm for generating $\hat{w}$.

Since the distribution P is unknown, according to the principle of Empirical Risk Minimization, we attempt to (approximately) minimize

$$L_t(w) := \frac{1}{t}\sum_{i=1}^{t} \ell(w,z_i) \qquad (3)$$

called the empirical risk of a hypothesis $w \in D$ on an observation set $\mathbf{z}_t$. So we can consider the minimization of $L_t(w)$ as an approximation to the minimum of the expected risk $L(w)$. The algorithm to be proposed in this paper is denoted as RSAGD, shorted for random search and gradient descent, which is defined by following recursion for any given excess risk $\epsilon > 0$,

$$w_t = \begin{cases} \eta_t & \text{if } L_t(\eta_t) \le L_t(w_{t-1}) - \epsilon \text{ and } L_t(\eta_t) \le L_t(\tilde{w}_t) \\ \tilde{w}_t & \text{if } L_t(\tilde{w}_t) \le L_t(w_{t-1}) - \epsilon \text{ and } L_t(\tilde{w}_t) < L_t(\eta_t) \\ w_{t-1} & \text{otherwise} \end{cases} \qquad (4)$$

where $\tilde{w}_t = \Pi_D\{w_{t-1} - \mu\nabla L_t(w_{t-1})\}$, $\nabla L_t(w_{t-1})$ denotes the gradient of $L_t(w)$ at $w = w_{t-1}$, $\Pi_D\{\cdot\}$ denotes the projection on D, η_t is an i.i.d. random sample with uniform distribution on D, which is independent of $\mathcal{F}_t$, where $\mathcal{F}_t := \sigma\{z_i, \eta_{i-1}, i \le t\}$.

Next we give an assumption on the loss function.

Assumption 2.1: Let $\ell(w,z)$ be continuous in w and z, and M and L defined by the following are finite,

$$M = \sup_{w\in D}\sup_{z\in Z} |\ell(w,z)|$$

$$L = \sup_{w,u\in D}\sup_{z\in Z} \frac{|\ell(w,z) - \ell(u,z)|}{|w-u|}.$$

Assumption 2.1 shows that the loss function $\ell(w,z)$ is Lipschitz in w on compact set D.

3 Main results

In this section, we will give our main result of this paper. First we need a lemma about strong law of large numbers of the *-mixing sequence.

Lemma 3.1 ([3] Theorem 3.3.2): Let $\{\xi_i, i \ge 1\}$ be *-mixing with $E[\xi_i] = 0$ and $E[|\xi_i|] \le K < \infty$ for each $i \ge 1$, and $\sum_{i=1}^{\infty} \frac{E[\xi_i^2]}{i^2} < \infty$, then $\sum_{i=1}^{n} \xi_i/n \to 0, a.s.$

Owing to Lemma 3.1, it is easy to obtain the following lemma.

Lemma 3.2: For $\forall w \in D$,$\lim_{t\to\infty} L_t(w) = L(w), a.s.$

Proof. We let $\xi_i = \ell(w,z_i) - E[\ell(w,z_i)]$. Since $\{z_i, i \ge 1\}$ is a stationary sequence,we have

$$E[\ell(w,z_i)] = E[\ell(w,z)] = L(w),\ \forall i \ge 1.$$

From the definition of $L_t(w)$, we get

$$L_t(w) - L(w) = \frac{1}{t}\sum_{i=1}^{t} \xi_i.$$

Since $\{z_i, i \ge 1\}$ is a *-mixing sequence, $\{\xi_i, i \ge 1\}$ is also a *-mixing sequence. It is easy to know that

$$E[\xi_i] = E[\ell(w,z_i)] - E[\ell(w,z_i)] = 0.$$

From Assumption 2.1, we have $E[|\xi_i|] \le 2M$ for each $i \ge 1$ and

$$\sum_{i=1}^{\infty} \frac{E[\xi_i^2]}{i^2} \le 4M^2 \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty.$$

By Lemma 3.1, we have

$$\sum_{i=1}^{t} \xi_i/t \to 0, a.s.$$

Thus we have $L_t \to L(w),\ a.s.$. Hence the proof is complete. □

With Lemma 3.2, we can obtain our main result.

Theorem 3.1: Let observations $\{z_i, i \ge 1\}$ be a stationary *-mixing sequence and Assumption 2.1 be satisfied,take the RSAGD algorithm (4), then we have

$$\lim_{t\to\infty} w_t = w_\infty,\ a.s. \qquad (5)$$

and $w_\infty \in S(\epsilon)$, where $S(\epsilon) := \{w \in D : L(w) \le L^* + \epsilon\}$, which is the ϵ-neighbourhood of the global minimum value of $L(w)$ on D.

Proof. The proof idea is similar to that of Guo [8].

Step 1: we prove the uniform convergence of $L_t(w)$ on D. From the condition of the theorem and by Lemma 3.2, we have

$$\lim_{t\to\infty} L_t(w) = L(w),\ a.s.,\ \forall w \in D \qquad (6)$$

$\forall \varepsilon > 0$, for D is a compact set in $\mathbb{R}^d$, there exist n_0 δ-neighbourhoods $\{u_1, u_2, \cdots, u_{n_0}\}$ that cover D totally, and

the δ-neighbourhood subjects to that if $|w-u_i| < \delta$ for some $i \in \{1, 2, \cdots, n_0\}$, then

$$|L_t(w) - L_t(u_i)| < \varepsilon/3, \text{ for } \forall t \geq 1,$$

and

$$|L(w) - L(u_i)| < \varepsilon/3.$$

From the definition of $L_t(w)$ and Assumption 2.1, we know δ exists. By equation (6), we can find a common T_0 for $\{u_1, \cdots, u_{n_0}\}$ such that when $t > T_0$,

$$|L_t(u_i) - L(u_i)| < \varepsilon, \text{ for all } \quad i = 1, 2, \cdots, n_0.$$

For $\forall w \in D$, there exists some u_i such that $|w - u_i| < \delta$, therefore

$$|L_t(w) - L_t(u_i)| < \varepsilon/3, \ \forall t \geq 1,$$

and

$$|L(w) - L(u_i)| < \varepsilon/3.$$

When $t > T_0$, we have

$$|L_t(u_i) - L(u_i)| < \varepsilon/3.$$

Therefore, when $t > T_0$,

$$\begin{aligned} |L_t(w) - L(w)| &\leq |L_t(w) - L_t(u_i)| + |L_t(u_i) - L(u_i)| \\ &\quad + |L(u_i) - L(w)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \\ &= \varepsilon. \end{aligned}$$

By the arbitrariness of w, we know for $\forall w \in D$, when $t > T_0$, we have $|L_t(w) - L(w)| < \varepsilon$. This shows the uniform convergence of $L_t(w)$ on D.
Step 2:we prove that there exists a positive random variable $\delta_\infty > 0$ such that

$$\liminf_{t\to\infty} L_t(\eta_t) \leq \delta_\infty, \ a.s. \tag{7}$$

Define: $\delta_t := \min_{w\in D} L_t(w)$, $D_t := \{w \in D : L_t(w) - \delta_t \leq \frac{\epsilon}{3}\}$ and $\mathcal{F}_t := \sigma\{z_i, \eta_{i-1}, i \leq t\}$. Notice that η_t and $\mathcal{F}_t$ is independent, from the property of conditional expectation, we have

$$\begin{aligned} P(L_t(\eta_t) \leq \delta_t + \frac{\epsilon}{3}|\mathcal{F}_t) &= \int_{w\in D} I(L_t(w) \leq \delta_t + \frac{\epsilon}{3})\mu(dw) \\ &= \int_{w\in D_t} \mu(dw) \\ &= \mu(D_t) \end{aligned} \tag{8}$$

where $\mu(\cdot)$ is uniform probability measure on D. Because $L_t(w)$ uniformly converges to $L(w)$ on D, we have $\delta_t \to \delta_\infty$, where $\delta_\infty = L^*$.
Define:

$$D_\infty = \{w \in D : L(w) \leq L^* + \frac{\epsilon}{3}\}.$$

By the continuation of $L(w)$ in w, we have $\mu(D_\infty) > 0$, Therefore $[L_t(w), \delta_t]$ converges to $[L(w), \delta_\infty]$ and for large enough t, $\mu(D_t) > \frac{\mu(D_\infty)}{2}$, and this shows $\mu(D_t) \nrightarrow 0$. By equation (8), we have

$$\sum_{t=1}^{\infty} P(L_t(\eta_t) \leq \delta_t + \frac{\epsilon}{3}|\mathcal{F}_t) = \infty, \ a.s.$$

and by Borel-Contelli-Lévy lemma, we have

$$\sum_{t=1}^{\infty} I(L_t(\eta_t) \leq \delta_t + \frac{\epsilon}{3}) = \infty, \ a.s..$$

This shows

$$\liminf_{t\to\infty} L_t(\eta_t) \leq \lim_{t\to\infty} \delta_t + \frac{\epsilon}{3} = \delta_\infty + \frac{\epsilon}{3}, \ a.s..$$

Step 3:we will prove there exists t_0 such that

$$L(w_t) \leq \delta_\infty + \epsilon, \ a.s., \ \forall t \geq t_0 \tag{9}$$

From algorithm (4), we may take $\frac{\epsilon}{3}$ instead of ϵ, then we have

$$L_t(w_t) \leq L_t(\eta_t) + \frac{\epsilon}{3}, \ a.s.$$

and then

$$\liminf_{t\to\infty} L_t(w_t) \leq \delta_\infty + \frac{2\epsilon}{3}, \ a.s.$$

Because of the uniform convergence of $L_t(w)$, we know there exists positive integer t_0 such that

$$L_{t_0}(w_{t_0}) \leq \delta_\infty + \frac{2\epsilon}{3} + \frac{\epsilon}{6} = \delta_\infty + \frac{5\epsilon}{6} \tag{10}$$

and

$$|L_t(w_s) - L(w_s)| \leq \frac{\epsilon}{6}, \ \forall t \geq t_0, \ \forall s \geq 0 \tag{11}$$

Next we will prove formula (9) by induction.
First, when $t = t_0$, from (10) and (11), we know

$$L(w_{t_0}) \leq L_{t_0}(w_{t_0}) + \frac{\epsilon}{6} \leq \delta_\infty + \epsilon.$$

Second, suppose formula (9) holds when $t = k \geq t_0$, we consider the situation when $t = k + 1$. If $w_{k+1} = w_k$, we know formula (9) holds by the assumption of the induction. Otherwise, from algorithm (4), we have

$$L_{k+1}(w_{k+1}) \leq L_{k+1}(w_k) - \frac{\epsilon}{3} \tag{12}$$

Considering formula (11) and $L(w_k) \leq \delta_\infty + \epsilon$, we obtain

$$\begin{aligned} L(w_{k+1}) &\leq L_{k+1}(w_{k+1}) + \frac{\epsilon}{6} \\ &\leq L_{k+1}(w_k) - \frac{\epsilon}{3} + \frac{\epsilon}{6} \\ &\leq L(w_k) + \frac{\epsilon}{6} - \frac{\epsilon}{3} + \frac{\epsilon}{6} \\ &= L(w_k) \\ &\leq \delta_\infty + \epsilon \end{aligned} \tag{13}$$

Therefore formula (9) holds.
Step 4:Finally we will prove the convergence of our algorithm (4). First, we will prove $\lim_{t\to\infty} L(w_t) = L_0$ exists and $L_0 \leq \delta_\infty + \epsilon$. It suffices to prove that $L(w_t)$ for $t \geq t_0$ is decreasing. And we already have that in formula (13). In

addition, because of the uniform convergence of $L_t(w)$ on D, we have

$$\lim_{t\to\infty} L_{t+1}(w_t) = \lim_{t\to\infty} L_t(w_t) = L_0, a.s..$$

Next we will prove the convergence of our algorithm by contradiction. Suppose that the algorithm (4) does not converge, then inequality (12) will hold for infinite t_k. Take $k \to \infty$, we have

$$L_0 \le L_0 - \frac{\epsilon}{3},$$

but this is impossible. Therefore we have formula (5) is correct. Hence the proof is complete. □

Remark 1: When the algorithm converges, we have $w_\infty \in S(\epsilon)$ and $E[L(w_\infty] = L(w_\infty)$. Thus $E[L(w_\infty)] - L^* \le \epsilon$.

Remark 2: M-dependent sequence and strong ϕ-mixing sequence in [3] are also *-mixing sequences, therefore our result also applies to the M-dependent observations and the strong ϕ-mixing observations.

In particular, if $\{z_i, i \ge 1\}$ is an i.i.d. sequence, by Theorem 3.1, we directly have the following corollary.

Corollary 3.1: Let $\{z_i, i \ge 1\}$ be an i.i.d. sequence and Assumption 2.1 be satisfied, taking the algorithm (4), then we have

$$\lim_{t\to\infty} w_t = w_\infty, a.s. \tag{14}$$

and $w_\infty \in S(\epsilon)$, where $S(\epsilon) := \{w \in D : L(w) \le L^* + \epsilon\}$, which is the ϵ-neighbourhood of the global minimum value of $L(w)$ on D.

Remark 3: When the observations are i.i.d., we can remove the Lipschitz constrain on the loss function in Assumption 2.1. By the uniform strong law of large numbers (Theorem 16(a)) in [9], we can directly obtain the uniform convergence of $L_t(w)$ on D.

4 Simulations

In this section, we will provide some simulations to demonstrate the advantages of our algorithm over either random search or gradient descent algorithms.

First, let us analyze the RSAGD algorithm (4) and it is easy to know that we can take the advantage of parallel calculations in program. We can calculate three loss function $L_t(w_{t-1}), L_t(\eta_t), L_t(\tilde{w}_t)$ simultaneously. To decrease the calculating cost, we can take stochastic gradient descent method when calculating $\tilde{w}_t$, which means selecting fewer samples k ($k \ll t$) to get the gradient at w_{t-1}.

We take the loss function of logistic regression to compare the performance of the RSAGD algorithm with the gradient descent algorithm. We take $y \in \{0, 1\}$ and

$$\ell(w, z) = y\log(1+\exp(-w^T x)) + (1-y)\log(1+\exp(w^T x)) \tag{15}$$

Gradient descent algorithm is as following:

$$w_t = w_{t-1} - \mu \nabla L_n(w_{t-1})$$

where $\nabla L_n(w_{t-1})$ denotes the gradient of $L_n(w)$ at w_{t-1}, μ is the step size, n is the sample size.

Claim 4.1: The convergence rate of the RSAGD is faster than the gradient descent algorithm.

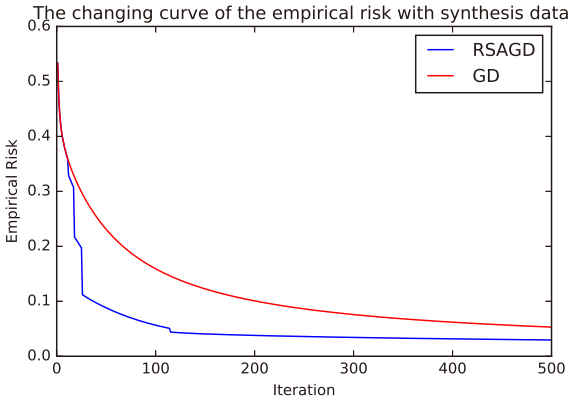

Fig. 1: The curve of empirical loss with the sample size $n = 2000, w_0 = [0, 0, 0]^T, \epsilon = 0.0003, \mu = 0.1, 0.001$ and D being a ball whose radius is 5.0 when the loss function is (15), GD denotes the gradient descent algorithm.

In the simulation generating Fig.1, we generate a data set with the sample size 2000 which contains 1000 positive samples and 1000 negative samples obeying normal distributions with means $(0, 0)^T$ and $(4, 4)^T$ respectively and the same variance. Fig.1 displays the changing curve of the empirical risk. For both the RSAGD and gradient descent algorithms, we take the same initial parameter $w_0 = [0, 0, 0]^T$ and step size $\mu = 0.1$ for the first 200 iterations and $\mu = 0.001$ for the following 300 iterations. Analyzing Fig.1, we can see that the steep descent of the blue curve shows that w_t takes the random search η_t by the RSAGD algorithm. It is precisely such updates that can increase the rate of convergence to the optimal value and escape the saddle points or some local minima. After we get a hypothesis, we test the performance of the hypothesis on a test data set with 2000 samples half of which are positive samples, and the accuracy of the RSAGD and gradient descent algorithms are 99.70% and 99.45% respectively.

Then we do simulations on a regression problem and get another property of the RSAGD. We generate a data set contains 1000 samples each of which consists of 5 attributes and a output, and the output is generated by the function $f(x) = sigmoid(w^T x)$, where $sigmoid(t) = \frac{1}{1+\exp(-t)}$ and $w = [1, 2, 1, 1, 1]^T$. We define the loss function as

$$\ell(w, z) = (y - sigmoid(w^T x))^2 \tag{16}$$

It is easy to know that this loss function is non-convex in w. When we take this loss function, we can find the following property of the RSAGD.

Claim 4.2: The RSAGD can generate a hypothesis satisfying the given accuracy level even if the loss function is non-convex.

Fig.2 and Fig.3 show that the gradient descent algorithm converges to different local minima when the initial parameters are different while the RSAGD algorithm can generate a hypothesis satisfying the given accuracy. This shows

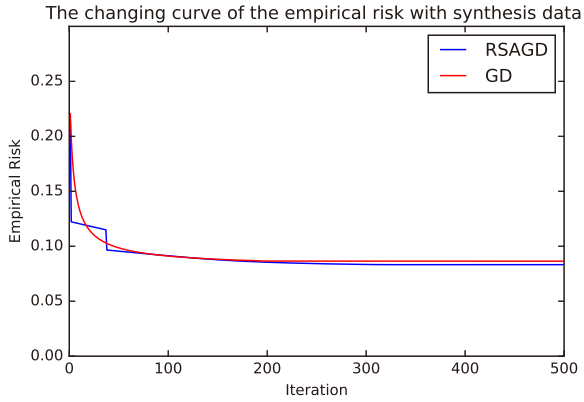

Fig. 2: The curve of empirical loss with the initial parameter $w_0 = [0, 0, 0, 0, 0]^T$, $\epsilon = 0.0003$, $\mu = 0.1$ and D being a ball whose radius is 5.0 when the loss function is (16), GD denotes the gradient descent algorithm.

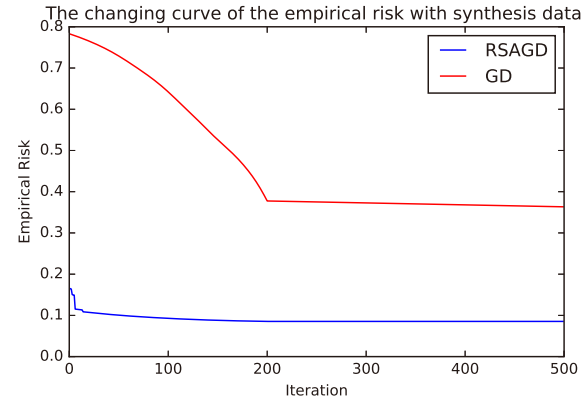

Fig. 3: The curve of empirical loss with the initial parameter $w_0 = [-1, -1, -1, -1, -1]^T$, $\epsilon = 0.0003$, $\mu = 0.1$ and D being a ball whose radius is 5.0 when the loss function is (16), GD denotes the gradient descent algorithm.

that the RSAGD can work well even if the loss function is non-convex. This property is very helpful in high dimension optimization problems.

The following property shows the advantage of the R-SAGD over the random search algorithm which is defined as following:

$$w_t = \begin{cases} \eta_t & \text{if } L_t(\eta_t) \le L_t(w_{t-1}) - \epsilon \\ w_{t-1} & \text{otherwise} \end{cases}$$

where η_t is an i.i.d. sample with uniform distribution on D.

Claim 4.3: The convergence rate of the RSAGD is faster than the random search algorithm especially when the loss function is non-convex.

From Fig.4, we find that the RSAGD can get a solution satisfying the given accuracy faster, while the random search algorithm may also get a good solution but needs more iterations. Because of the use of randomness in the RSAGD and random search algorithm, the figures generated by programs may be different, but their convergent values are basically the same in the same number of iterations.

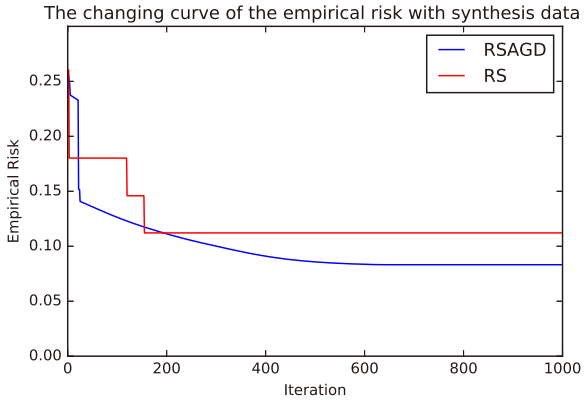

Fig. 4: The curve of empirical loss with the initial parameter $w_0 = [-1, -1, -1, -1, -1]^T$, $\epsilon = 0.0003$, $\mu = 0.1$ and D being a ball whose radius is 5.0 when the loss function is (16), RS denotes the random search algorithm.

5 Conclusion

In this paper,we have proposed a new convergent algorithm called RSAGD for ERM with *-mixing observations. The RSAGD can generate a sequence of hypothesis at which the expected risk is in the ϵ-neighbourhood of the optimal expected risk with probability 1 after finite steps. Our result is applicable to both weakly dependent data and non-convex loss functions. We also demonstrated the advantages of the new algorithm over either gradient descent algorithms or random search algorithms by simulations. For further investigation, it is desirable to reduce the cost of calculation at each iteration and to further relax the dependent assumptions on the observations.

References

[1] V.N. Vapnic, *Statistical Learning Theory*, New York:Wiley, 1998

[2] F. Cucker, S. Smale, On the mathematical foundations of learning, *Bulletin of the American Mathematical Society*, 39(1):1-49, 2002.

[3] W.F. Stout, *Almost sure convergence*, New York:Academic Press, 1974.

[4] B. Yu, Rates of convergence for empirical processes of stationary mixing sequences, *Annals of Probability*, 22(1):94-114, 1994.

[5] M. Vidyasagar, *Learning and generalization with applications to neural networks(2nd ed.)*, Berlin:Springer, 2002.

[6] B. Zhou, L. Li, The performance bounds of learning machines based on exponentially strong mixing sequences, *Computer & Mathematics with Applications*, 53(7):1050-1058, 2007.

[7] B. Zhou, L.Q. Li, Z.B. Xu, The generalization performance of ERM algorithm with strong mixing observations, *Mach Learn*, 75:275-295, 2009.

[8] L Guo, Self-convergence of weighted least-squares with applications to stochastic adaptive control, *IEEE Transactions on Automatic Control*, 40(1):79-89, 1996.

[9] T.S. Furguson, *A course in large sample theory*, Chapman & Hall, 1996.