SGD model and its applications

Chenze Li
Department of Statistics, University of California, Davis, CA 95616
czeli@ucdavis.edu

Abstract

I first repeat the results from a proposed method which tracks the dynamics of stochastic gradient descent (SGD) model when data is isotropic. I also try a numerical simulation method to solve a specific kind of Volterra equations. Besides of that, I investigate the effect of batch size on SGD model. Finally I show that SGD model with streaming data (training data is updated in each iteration) and the application of SGD mdoel on the least squares problem with penalty converge with the increase of iterations.

1 Introduction

Stochastic Gradient Descent (SGD) model proposed by Robbins and Monro (1951) is an important learning machine method to minimize the specific kind of problems which have the form $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$. And because of its efficiency, many researchers commit a lot into it. Courtney Paquette, Lee,Pedregosa and Elliot Paquette (2021) provided an alternative model to accurately track the convergence trace of SGD model under specific assumptions. They also compare their model with other techniques which analyses the dynamics of SGD model such as the stochastic differential equations (SDE) paradigm (Li et al., 2017) and the stochastic modified equation (SME) (Li et al., 2017).

In this report, besides repeating the results shown by Paquette et al.(2021), I also try another numerical method to solve the Volterra equation model(Paquette et al.,2021). At the same time, the effect of batch size on SGD model, online SGD model(regenerating data at each update) and the application of SGD model on the ridge regression are also investigated. I hope these premature research could give a hint to the future investigation.

2 Asymptotic analysis of SGD with a Volterra integral equation

2.1 Stochastic gradient descent

We consider to get the minimum of a high dimensional polynomial function $f(x)$, that is,

$$\arg\min_{\boldsymbol{x}\in R^d} f(\boldsymbol{x}) = \frac{1}{2n}\sum_{i=1}^{n}(\boldsymbol{a_i}\boldsymbol{x} - b_i)^2 \tag{1}$$

where $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$ and b_i is the ith element in $\boldsymbol{b}$, $\boldsymbol{A} \in R^{n\times d}$ is a random matrix in which $\boldsymbol{a_i}$ is ith row, and $\boldsymbol{x} \in R^{d\times 1}$ is the signal vector. And we set the dimension of the samples as d and the number of samples as n. In the report the ratio of d and n is $r = \frac{d}{n}$.

To solve the problem, I will use SGD method. In kth iteration, I generate a randomly selected subset B_k of size β from $\{1, 2, \cdots, n\}$. Then the kth update will be:

$$\boldsymbol{x_{k+1}} = \boldsymbol{x_k} - \frac{\gamma}{n}\boldsymbol{A}^\top \boldsymbol{P_k}(\boldsymbol{A}\boldsymbol{x_k} - \boldsymbol{b}),\ with\ \boldsymbol{P_k} = \sum_{i\in B_k} \boldsymbol{e_i}\boldsymbol{e_i}^\top \tag{2}$$

This report is a summary of summer research under the supervision of Professor Xiucai Ding

where $\boldsymbol{e_i}$ is the ith standard vector and γ is the step size.

2.2 Volterra integral equation

For SGD algorithm, Courtney Paquette, Lee,Pedregosa and Elliot Paquette (2021) proposed that under specific conditions(c.f.[Paquette et al.,2021, Assumption 1.1 and Assumption 1.2]), the values of $f(x)$ under SGD model converge to the solution of a Volterra integral equation. The theorem could be stated as following:

Theorem 1 *(*Paquette et al.,2021, Theorem 1.1) *Suppose the stepsize satisfies* $\gamma < \frac{2}{r}\left(\int_0^\infty x d\mu(x)\right)^{-1}$ *and the batch size satisfies* $\beta(n) \leq n^{1/5-\delta}$ *for some* $\delta > 0$*. Under specific assumptions (c.f.[Paquette et al.,2021, Assumption 1.1 and Assumption 1.2]),*

$$\sup_{0\leq t\leq T} |f\left(\boldsymbol{x}_{\lfloor \frac{n}{\beta}t \rfloor}\right) - \psi_0(t)| \xrightarrow[n\to\infty]{Pr} 0 \tag{3}$$

where the function $\psi_0(t)$ *is the solution to the Volterra equation*

$$\begin{aligned} \psi_0(t) &= \frac{R}{2}h_1(t) + \frac{\widetilde{R}}{2}\left(rh_0(t) + (1-r)\right) + \int_0^t \gamma^2 r h_2(t-s)\psi_0(s)ds, \\ and \quad h_k(t) &= \int_0^\infty x^k e^{-2\gamma t x} d\mu(x) \quad for\ all \quad k \geq 0. \end{aligned} \tag{4}$$

In the Theorem 1, $R = \|\boldsymbol{x_0} - \boldsymbol{x}\|$ and $\widetilde{R}$ is a constant value related to the noise term. To be specific, if $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} + \sqrt{n}\boldsymbol{\eta}$, then $\mathrm{E}[\|\boldsymbol{\eta}\|_2^2] = \widetilde{\mathrm{R}}$ but we could set $\widetilde{R} = 0$ since in the report $\boldsymbol{\eta} = \boldsymbol{0}$. μ is the limit measure of the eigenvalue distribution of $\boldsymbol{H} = \frac{1}{n}\boldsymbol{A}^\top\boldsymbol{A}$ which means $\mu_{\boldsymbol{H}} = \frac{1}{d}\sum_{i=1}^d \delta_{\lambda_i} \to \mu$ with δ_{λ_i} denoting the Dirac delta with mass at λ_i and λ_i is the ith smallest eigenvalue of $\boldsymbol{H}$.

At the same time, Paquette et al.(2021) also gives analytical solution to the Volterra equation 4 for isotropic models. And we will introduce isotropic model first.

Isotropic model The model requires entries of $\boldsymbol{A}$ generated from standard Gaussian distribution, that is, $A_{ij} \sim N(0,1)\ \forall i,j$. Research by Marčenko and Pastur (1967) showed that when sample size n and feature dimensions d tends to infinity proportionally, that is, $\frac{d}{n} \to r \in (0,\infty)$, the distribution of eigenvalues of $\boldsymbol{H}$ converges to a deterministic measure μ whose distribution is

$$d\mu(x) = \left(1 - \frac{1}{r}\right)_+ \delta_0(x) + \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi r x}\mathbf{1}_{x\in[\lambda_-,\lambda_+]}dx$$

where

$$\lambda_- = (1-\sqrt{r})^2 \quad \lambda_+ = (1+\sqrt{r})^2$$

Then the following theorem shows analytical solutions to the equation 4 for the isotropic model when $\widetilde{R} = 0$.

Theorem 2 *(*Paquette et al.,2021, Theorem E.4) *Suppose* $\widetilde{R} = 0$ *and the batchsize satisfies* $\beta(n) \leq n^{1/5-\delta}$ *for some* $\delta > 0$*, and the stepsize is* $0 < \gamma < \frac{2}{r}$*. Define the critical stepsize* $\gamma_* \in \mathbb{R}$ *and constants* $\theta > 0$ *and* $\omega \in \mathbb{C}$*,*

$$\gamma_* = \frac{2}{\sqrt{r}(r-\sqrt{r}+1)},\ \theta = \frac{1+r}{2}\left(1 - \frac{r\gamma}{2}\right),\ and\ \omega = \frac{1}{4}\left(1-\frac{r\gamma}{2}\right)^2\left(\frac{8}{\gamma} - (1+r)^2\right).$$

The iterates of SGD satisfy if $\gamma \leq \gamma_*$

$$f\left(\boldsymbol{x}_{\lfloor \frac{n}{\beta}t \rfloor}\right) \xrightarrow[n\to\infty]{Pr} \frac{R}{\gamma}\left(1 - \frac{r\gamma}{2}\right)\int_0^\infty \frac{xe^{-2\gamma x t}}{(x-\theta)^2 + \omega} d\mu(x) \tag{5}$$

and if $\gamma > \gamma_*$ *the iterates of SGD satisfy*

$$f\left(\boldsymbol{x}_{\lfloor \frac{n}{\beta} t \rfloor}\right) \xrightarrow[n\to\infty]{Pr} \frac{R}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \frac{xe^{-2\gamma x t}}{(x-\theta)^2+\omega} d\mu(x) + \frac{R}{4\sqrt{|\omega|}}\left[\theta+\sqrt{|\omega|}-\left(\frac{2}{\gamma}\right)^2\left(1-\frac{r\gamma}{2}\right)^2\frac{\theta-\sqrt{|\omega|}}{r(\theta^2-|\omega|)}\right] \quad (6)$$

2.3 Numerical method for solving Volterra equation

A major challenge of approximating SGD with the Volterra equation is how to solve the equation. In this section, I will introduce a numerical method developed by Costarelli and Spigler(Summer 2013).

For a linear Volterra integral equations with the convolutional form,

$$y(t) = f(t) + \int_a^t K(t-s)y(s)ds, \quad t \in [a,b] \quad (7)$$

Costarelli and Spigler(Summer 2013) constructed an unit step function to approximate the solution of a Volterra equation in 7, which has the form

$$(G_N y)(t) = \sum_{k=1}^{N} y_k H(t-t_k) + y_0 H(t-t_{-1}), \quad t \in [a,b] \quad (8)$$

where $t_k = a + hk, k = -1, 0, \cdots, N$, $h = (b-a)/N$ and $y_0, y_1, \cdots, y_N, N \in \mathbb{N}^+$ are unknown variables. N is set in advance and in the simulation $N = 1000$. For the Volterra equation,

$$y(t) = \psi_0(t), f(t) = \frac{R}{2}h_1(t), K(t-s) = \gamma^2 r h_2(t-s)$$

Besides, $H(t)$ is a Heaviside function which means $H(t) = 1$ for $t \geq 0$ and $H(t) = 0$ for $t < 0$. Then all we need to do is to get the values of $y_0, y_1, \cdots, y_N, N \in \mathbb{N}^+$. Then Costarelli and Spigler(Summer 2013) proved the following theorem:

Theorem 3 *(*Costarelli & Spigler, 2013, Corollary 3.3) *Let Y_N equal to $(y_0, y_1, \cdots, y_N)^\top$ and F_N be $(f(x_0), f(x_1), \cdots, f(x_N))^\top$. $x_0, x_1, \cdots, x_N$ are 'N equal points' which means $x_k = x_0 + kh,\ k = 0, 1, \cdots, N$ and $x_0 = a$ and $x_N = b$. The collocation method for solving 7, based on unit step functions, admits a unique solution. Moreover, the real matrix M_N is a lower triangular Toeplitz matrix, for every $N \in \mathbb{N}^+$. M_N could be represented as*

$$M_N = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ m_1 & 1 & 0 & \cdots & 0 \\ m_2 & m_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ m_N & m_{N-1} & \cdots & m_1 & 1 \end{bmatrix} \quad (9)$$

where

$$m_i = 1 - \int_a^{t_i} K(t_i - s)ds \quad i = 1, 2, \cdots, N$$

Then Y_N satisfies the following equation:

$$M_N Y_N = F_N, N \in \mathbb{N}^+ \quad (10)$$

Therefore, we could get the values of Y_N and $Y_N = M_N^{-1} F_N$. After we get the values of $y_0, y_1, \cdots, y_N$, I just plug these values into the equation 8 and in the tth epoch of SGD model the value of the target function denoted as $f\left(\boldsymbol{x}_{\lfloor \frac{n}{\beta} t \rfloor}\right)$ is approximated by $(G_N y)\left(\boldsymbol{x}_{\lfloor \frac{n}{\beta} t \rfloor}\right)$.

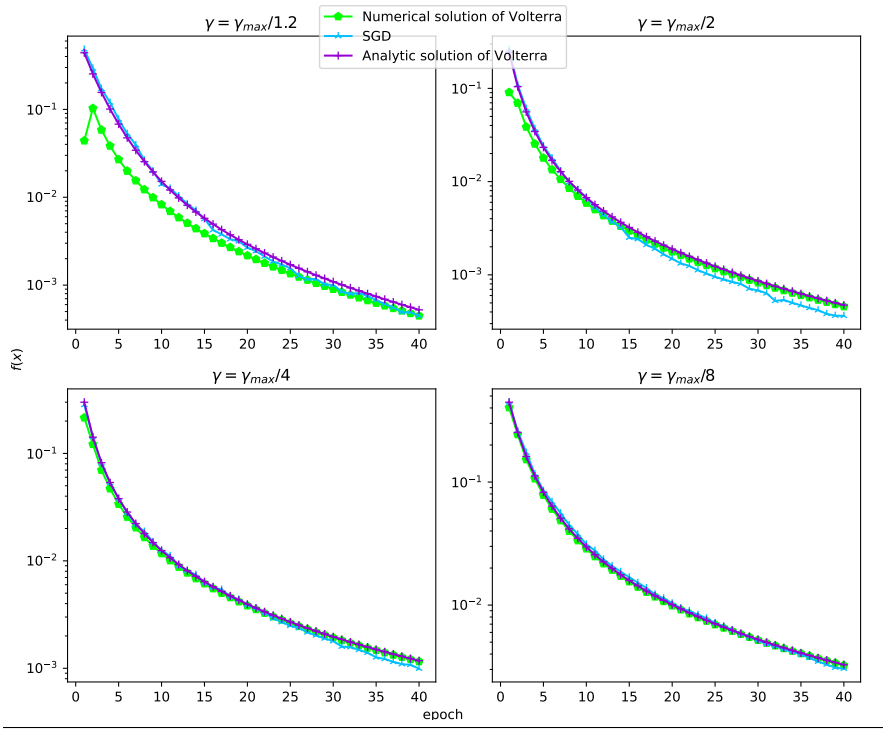

Figure 1: **Comparison between SGD model and Volterra equation**. $r = 1.2$, $n = 1000$ and $\beta = 3$. Data matrix $\boldsymbol{A}$ is generated according to the isotropic model. γ is step size and $\gamma_{max} = \frac{2}{r}\left(\frac{1}{d}tr(H)\right)^{-1}$. The initial vector $\boldsymbol{x_0}$ and the signal vector $\boldsymbol{x}$ are i.i.d. generated from $\boldsymbol{N}(\boldsymbol{0}, \frac{\boldsymbol{1}}{\boldsymbol{d}}\boldsymbol{I})$. For the variables of numerical method to solve Volterra equation, $N = 1000$, $a = 1$ and $b = 40$.

3 Numerical simulation

3.1 SGD model and Volterra equation

In this section, apart from using SGD model on the problem 1, I also apply the analytical solution and numerical method to solve the Volterra equation. The results are shown in Figure 1. In figure 1, we could see that with step size γ taking different values, the analytic solution of Volterra equation $\psi_0(x)$ approximately converges to the results of SGD model with the increase of epochs. Besides, we could find that $(G_{Ny})(t)$ from the numerical method is a good approximation to $\psi_0(t)$.

3.2 SGD model on ridge regression

For the above SGD model, I only consider the situation without penalty terms and now I will consider applying SGD model on the ridge regression containing ℓ_2 penalty, and the problem turns into minimizing the following function, that is,

$$\arg\min_{\boldsymbol{x}\in R^d} f_{ridge}(\boldsymbol{x}) = \frac{1}{2n}\sum_{i=1}^{n}(\boldsymbol{a_i}\boldsymbol{x} - b_i)^2 + \lambda\|\boldsymbol{x}\|_2^2 \tag{11}$$

where λ is to control the effect of penalty.

Figure 2 shows the effects of SGD model on the ridge regression with different λ. Since some large λ may cause $f(x_k)$ converge to infinity with the increase of k, then we ignore and drop those λs. Therefore, it could be seen in figure 2 that with different step size, curves truncate at different λ. And I refer the smallest values of λ at which $f(x_k)$ diverges as the **threshold of penalty**. And in the

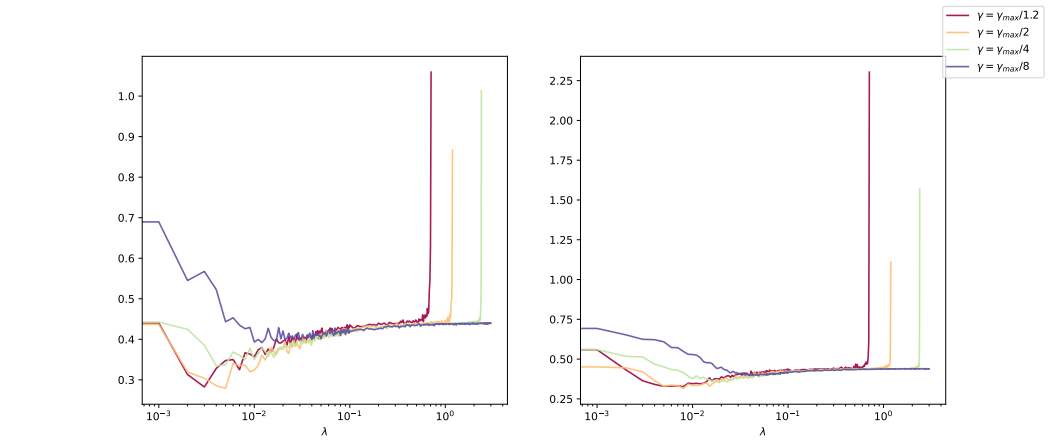

Figure 2: **Ridge regression with λ varying**. $r = 1.2$, $n = 1000$ and batch size $\beta = 3$. Data matrix $\boldsymbol{A}$ is generated according to the isotropic model. γ is step size and $\gamma_{max} = \frac{2}{r}\left(\frac{1}{d}tr(H)\right)^{-1}$. The initial vector $\boldsymbol{x_0}$ and the signal vector $\boldsymbol{x}$ are i.i.d. generated from $\boldsymbol{N}(\boldsymbol{0}, \frac{\boldsymbol{1}}{\boldsymbol{d}}\boldsymbol{I})$. The left figure shows the value of $f(x_k)$ with the largest k in the set $\{k : \|f(\boldsymbol{x_k}) - f(\boldsymbol{x_{k-4}})\| < 4 \times 10^{-3}\}$ when $|\{k : \|f(\boldsymbol{x_k}) - f(\boldsymbol{x_{k-4}})\| < 4 \times 10^{-3}\}| = 5$ for each λ. The plot on the right shows the value of $f(x_k)$ with $k = 100$ for each λ. And I drop those λ when for the specific λ, the value of $f(x_k)$ increases with the increase of k.

figure 3, y axis means the value of $f(x_k)$ of some k with some λ. k satisfies different conditions for the two plots. x axis means different values of λ.

3.3 The effect of batch size on SGD model

In this section, I will investigate how batch size affects the converge rate of SGD models. The target function is the same as Equation 1. The batch sizes are from set $\{1, 2, \cdots, 40\}$ with different step size.The results are shown in figure 4. The plot shows the function between the values of $f(x)$ at the 400th iteration and the batch size β.

3.4 Online SGD

For this part, I will regenerate data matrix $\boldsymbol{A}$ in each update and the setting of the problem is mentioned in Section 2. Figure 5 displays how $f(x_k)$ converges under SGD model with regenerating data matrix $\boldsymbol{A}$ for each update. The discussion will be shown later. Besides I also give an example of the convergence trace of SGD model with step size $\gamma = 0.001$ for streaming data which means only five samples are given for each update.

4 Conclusions and Discussions

SGD model and Volterra equation of Isotropic model From figure 1, it is shown that with different step size Volterra equation shows great performance on the approximation of dynamics of SGD model when epoch increases (each epoch means $\lfloor \frac{n}{\beta} \rfloor$ iterations). And we could see in the situation of $\gamma = \gamma_{max}/8$ curves from three models shows nearly the same pattern and when γ turns larger, the numerical method works worse.

SGD model on ridge regression Figure 2 shows that with different stop criteria, $f(x_k)$ shows the "U" pattern with the increase of λ and the larger the step size is, the smaller the threshold penalty

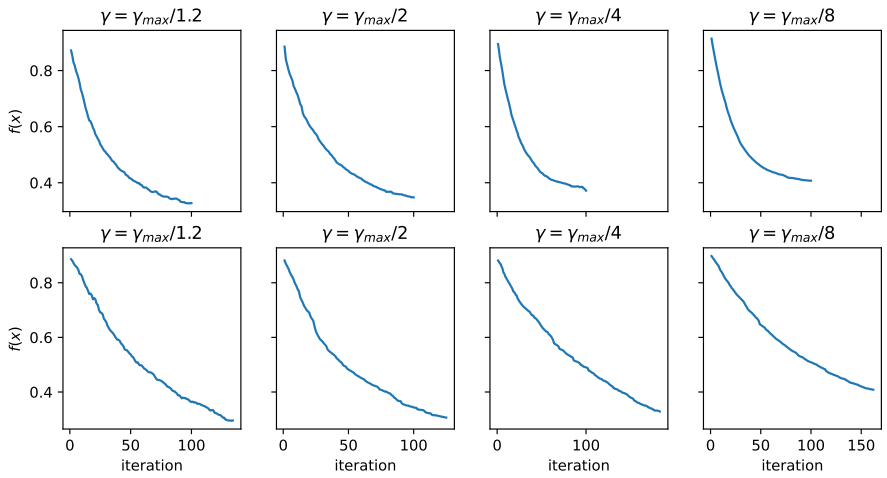

Figure 3: **The convergence trace of SGD model on the ridge regression**.The setting is the same as mentioned in the note of figure 2. With two types of convergence criteria, I select λ corresponding to the smallest values of $f(x_k)$ in figure 2. Figure 3 shows the convergence track of SGD model with this smallest λ with different step sizes. The first row matches the criteria that iterations stop when $k = 100$. The bottom row matches the criteria that iterations stop when k is the largest k in the set $\{k : \|f(\boldsymbol{x_k}) - f(\boldsymbol{x_{k-4}})\| < 4 \times 10^{-3}\}$ when $|\{k : \|f(\boldsymbol{x_k}) - f(\boldsymbol{x_{k-4}})\| < 4 \times 10^{-3}\}| = 5$

is. And the similarity of values of $f(x_k)$ from two figures indicates that SGD model converges very slowly after small amounts of updates.

The effect of batch size on SGD model Figure 4 shows the value of $f(x_k)$ at the 400-th iteration of different batch size.It displays batch size has great effects on SGD model. The greater the batch size is, at the same iteration the smaller of $f(x_k)$ is. Besides, larger step size helps to increase the rate of convergence.

Online SGD Figure 5 shows the results of SGD model with regenerating data matrix $\boldsymbol{A}$ at each update. It could be seen that SGD model also makes a difference on the streaming data but the rate of convergence seems to slow. From figure 6, when we only train five samples for each update, the convergence rate becomes slower and there shows large fluctuation for each iteration.

5 Acknowledgements

Thanks for the instructions from Professor Xiucai Ding. I feel really appreciated that he could give me advise during the process of completing the report. Also, I really thank Professor Courtney Paquette for answering my questions about her paper.

6 Appendix

[1].Paquette, C., Lee, K., Pedregosa, F., & Paquette, E. (2021). SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality. arXiv preprint arXiv:2102.04396.

[2].Costarelli, D., & Spigler, R. (2013). Solving Volterra integral equations of the second kind by sigmoidal functions approximation. Journal of Integral Equations and Applications, 25(2), 193-222.

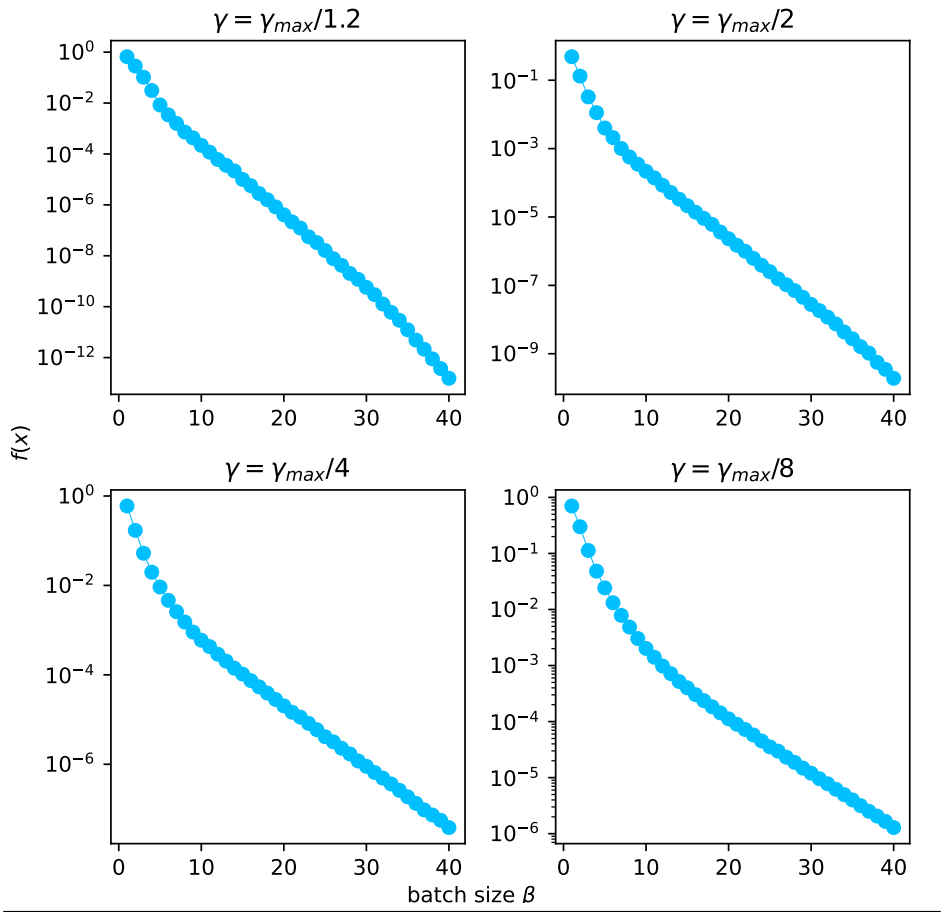

Figure 4: **The effects of batch size β on SGD model**. $r = 1.2$ and $n = 1000$.The figure shows the value of $f(x_k)$ with $k = 400$ for each batch size. Each element of data matrix $\boldsymbol{A}$ is generated from standard Gaussian distribution. γ is step size and $\gamma_{max} = \frac{2}{r}\left(\frac{1}{d}tr(H)\right)^{-1}$. The initial vector $\boldsymbol{x_0}$ and the signal vector $\boldsymbol{x}$ are i.i.d. generated from $\boldsymbol{N}(\boldsymbol{0}, \frac{\boldsymbol{1}}{\boldsymbol{d}}\boldsymbol{I})$.

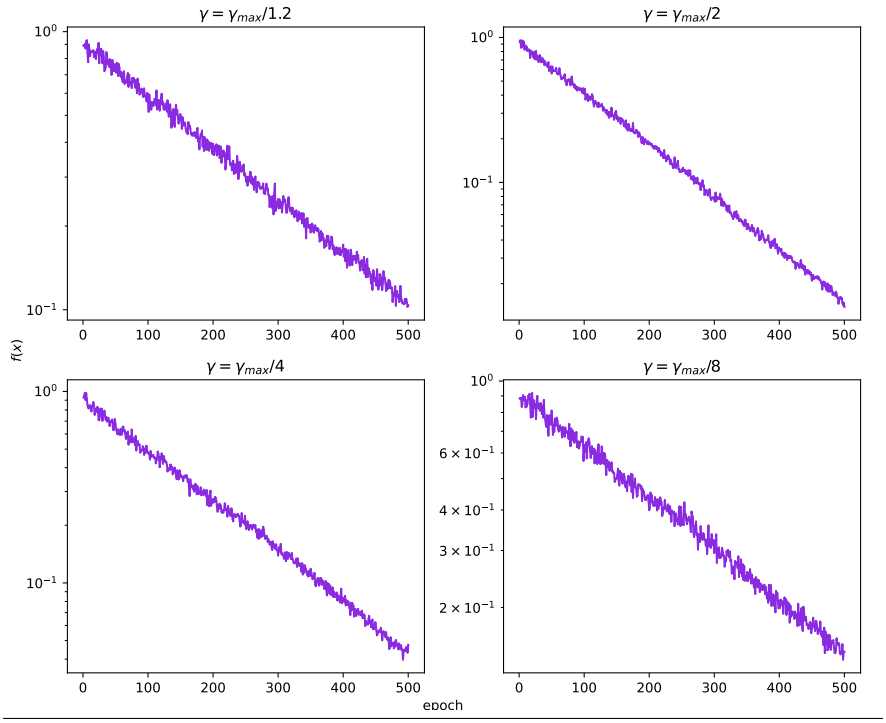

Figure 5: **Online SGD model(regenerate $\boldsymbol{A}$ for each update)**. $r = 1.2$, $n = 1000$ and $\beta = 10$. Each element of data matrix $\boldsymbol{A}$ is generated from standard Gaussian distribution. γ is step size and $\gamma_{max} = \frac{2}{r}\left(\frac{1}{d}tr(H)\right)^{-1}$. The initial vector $\boldsymbol{x_0}$ and the signal vector $\boldsymbol{x}$ are i.i.d. generated from $\boldsymbol{N}(\boldsymbol{0}, \frac{\boldsymbol{1}}{\boldsymbol{d}}\boldsymbol{I})$.

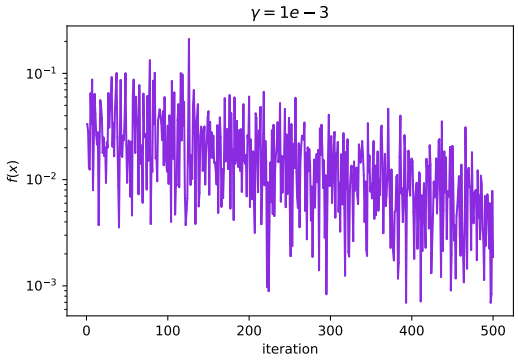

Figure 6: **Online SGD model**. This plot shows the convergence trace of SGD model when I generate five samples for each update, which means data matrix for each update is $\boldsymbol{A}_{3\times d}$ with $d = 1200$. Each element of data matrix $\boldsymbol{A}$ is generated from standard Gaussian distribution. γ is step size. The initial vector $\boldsymbol{x_0}$ and the signal vector $\boldsymbol{x}$ are i.i.d. generated from $\boldsymbol{N}(\boldsymbol{0}, \frac{\boldsymbol{1}}{\boldsymbol{d}}\boldsymbol{I})$.