# A Technical Report about SAG,SAGA,SVRG

**Tao Hu**
Department of Computer Science
Peking University
No.5 Yiheyuan Road Haidian District, Beijing, P.R.China
taohu@pku.edu.cn

**Notice: this is a course project for <Algorithms for Big Data Analysis> conducted by Zaiwen Wen at Peking University; and <Deep Learning> conducted by Zhihua Zhang at Peking University.**

- **for <ALgorithms for Big Data Analysis>, I completed the SAG, SAGA, SVRG algorithm programming.**
- **for <Deep Learning>, I completed different gradient method ablation study.**

**I just write it as a entirety for for completeness and continuity!**

## Abstract

The project report mainly includes the ablation study about the regularization term, stochastic gradient method, etc.

## 1 Overview

There exists many gradient method, Full Gradient(FG) is proposed first. as more and more data generate, a light-weight method named Stochastic Gradient(SG) appears. Based on the SG method, there emerges several Stochastic Method: Stochastic Gradient Method(SGD)[1], Momemtum[5], AdaDelta[9], RmsProp[8], SDCA[7]. Recently, NL Roux[6] proposed a Variation Method Stochastic Average Gradient (SAG) that realizes linear convergence. lots of variety based on Variance Method such as SVRG[3] SAGA[2], appears. In this project report I will compare several aspects about some of these methods.

The problem I experiment on is regularized logistic regression.

$$min_w \quad P(w) = \frac{1}{n}\sum_{i=1}^{n} ln(1 + exp(-w^T x_i y_i)) + \lambda L(w)$$

where w is weight, L(w) means regression item which can be Ridge Regression or Lasso Regression.

as we use gradient method, we have the gradient formula:

$$\nabla P(w) = \frac{1}{n}\sum_{i=1}^{n} \frac{exp(-y_i w^T x_i)}{1+exp(-y_i w^T x_i)}(-y_i x_i) + \lambda L^{'}(w)$$

if I use *l1*-regularized logistic regression, then $L^{'}(w) = sign(w)$. otherwise, I will choose *l2*-regularized logistic regression which $L^{'}(w) = 2w$.

the dataset we use is mnist[4], so the $x_i$ I use above is a 785-by-1 vector, $y_i$ is a scalar, w is 1-by-785 vector. the mnist data size is 28-28, adding a bias totally is 785. for each mnist image, I have normalized it to 1 for convenience of training.as the logistic regression is a binary classification problem, hence I just divide the mnist data into even and odd for later study.

the final binary digit classfier is:

$$y = sign(w^T x)$$

y=1 means odd, otherwise means even.

## 2 Experiment

### 2.1 How to obtain the optimal

In the following Experiment, I will use a parameter named " Objective minus Optimum" which we need the optimum, I will choose the optimal value by SGD with a exponential decay learning rate:

$$\mu_t = \eta_0 a^{\lfloor \frac{t}{step} \rfloor}$$

$\eta_0, a$ are the grid search parameter,here we choose the step = 3, which means to decay the learning rate exponentially for every 3 epochs. 100 epochs were processed finally. the grid search result is in Table 1.

as the result shows, the minimal is 0.3707, we keep the same grid search parameter where $\eta_0 = 1, a = 0.9$ and increase the final epochs to 200, then further achieve minimal value 0.3638/0.9018, we choose the optimum 0.3638 in the following experiment.

Table 1: Optimum Grid Search Table

| a \| $\eta_0$ | 10 | 1 | 0.1 | 0.01 | 0.001 |
|---|---|---|---|---|---|
| 0.9 | 0.4200/0.9013 | **0.3707/0.9018** | 0.4173/0.8955 | 0.5830/0.8457 | 1.0758/0.7315 |
| 0.7 | 0.4118/0.9037 | 0.3826/0.9 | 0.4764/0.8705 | 0.5966/0.8097 | 1.8617/0.6378 |
| 0.5 | 0.4616/0.8997 | 0.3887/0.8975 | 0.4878/0.8615 | 0.8215/0.7896 | 3.113/0.5356 |

the element in every grid means "loss/test accuracy", the result is selected by the highest test accuracy in 100 epochs. best optimal achieves when a=0.9,$\eta_0 = 1$

### 2.2 Ablation Study

#### 2.2.1 Regulation

There exists different regulation items including L1 norm, L2 norm. here I will compare different regression factor $\lambda$ in Figure 1. different regulation type influence on test accuracy is also compared in Figure 2.

in those Figures, I choose 100 epochs to run all the methods.

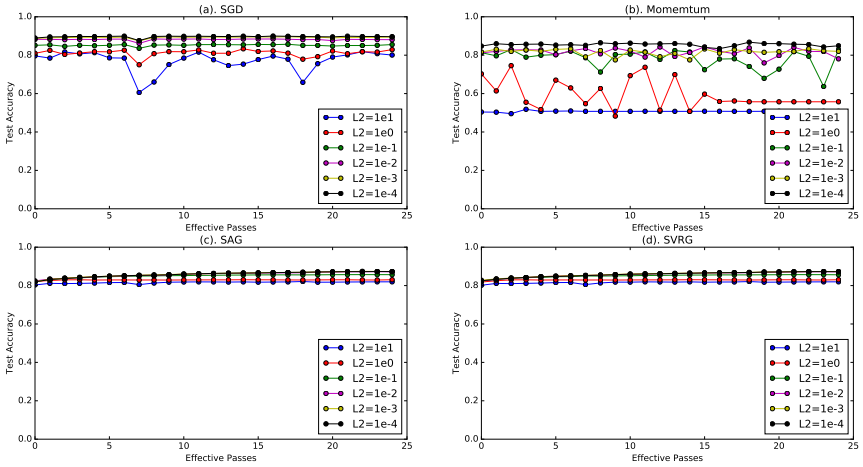


Figure 1: Regulation Value Experiment.all experiment use *l2*-regularized logistic regression as default. (a) Test Accuracy between different regulation $\lambda$ with *SGD*. (b) Test Accuracy between different regulation $\lambda$ with *SGD with Momentum*. (c) Test Accuracy between different regulation $\lambda$ with *SAG*. (d) Test Accuracy between different regulation $\lambda$ with *SVRG*.This figure is best viewed in colour.

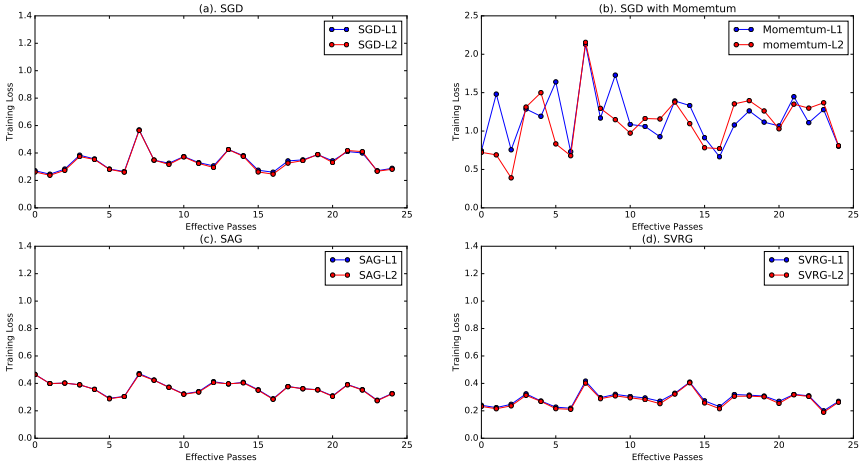


Figure 2: Regulation Type Experiment (a) Test Accuracy between *l1* and *l2* regularized logistic regression with *SGD*. (b) Test Accuracy between *l1* and *l2* regularized logistic regression *SGD with Momentum*. (c) Test Accuracy *l1* and *l2* regularized logistic regression with *SAG*. (d) Test Accuracy between *l1* and *l2* regularized logistic regression with *SVRG*.This figure is best viewed in colour.

### 2.2.2 Step Size Strategy

I mainly compared these step size strategy: fixed, exponential decay(step_size=3), backtracking. the training loss and test accuracy influenced by the step size strategy is displayed in Figure 3. 100 epochs are processed for every method.

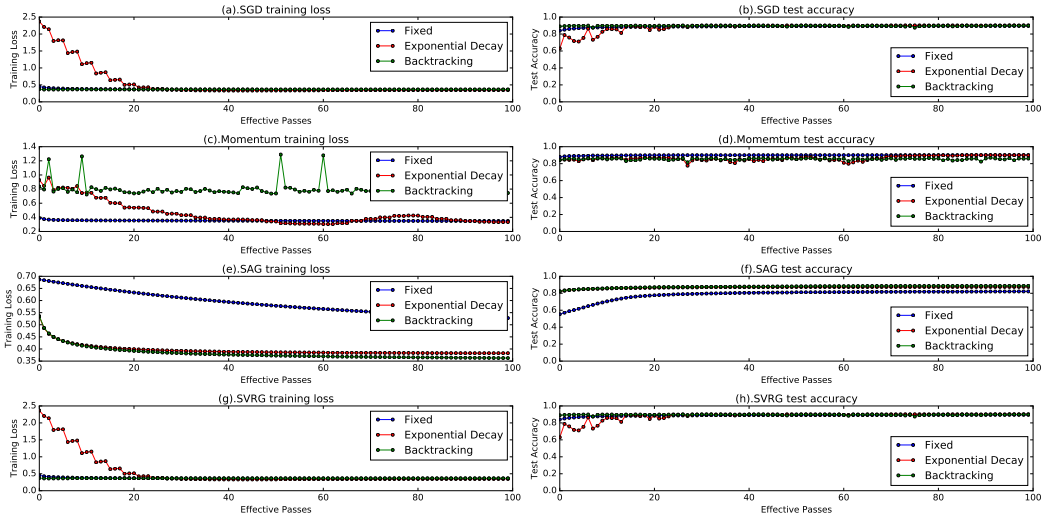


Figure 3: Step Size Strategy Experiment (a,b) training loss and test accuracy with three step size strategy with *SGD*. (c,d) training loss and test accuracy with three step size strategy *SGD with Momentum*. (e,f) training loss and test accuracy with three step size strategy with *SAG*. (g,h) training loss and test accuracy with three step size strategy with *SVRG*.This figure is best viewed in colour.

### 2.2.3 Different Method

Finally, different method including Stochastic Gradient Descend(SGD), SGD with Momentum, accelerated SGD with Momentum, SAG, SAGA, SVRG are compared in Figure 4. Training Loss, Validation Loss, Test Accuracy tendencies are listed.

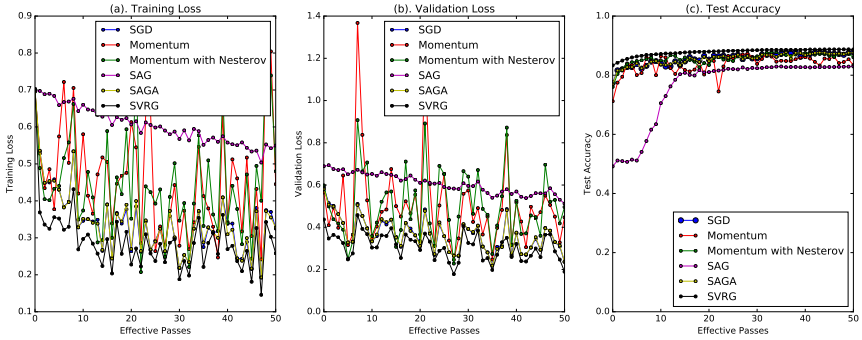


Figure 4: SG Method Experiment (a). training loss of different method. (b). validation loss of different method. (c). test accuracy of different method. This figure is best viewed in colour.

Time Consume Experiment is show in Figure **??**

### 2.2.4 Vectorization Programming

vectorization programming is very important in numerical calculation. it's a bad habit using "for loop" too frequently. here I will list two different writing style concerning the "non-vectorization" and "vectorization" programming.

vectorization is necessary especially in algorithm like SAG, SVRG, because the gradient computing is very frequently.

**Algorithm 1** Non-vectorization

**Require:** initial value w,x,y
  **for** i in 1:n **do**
    calculate gradient:$\nabla P(w) = \frac{1}{n}\sum_{i=1}^{n} \frac{exp(-y_i w^T x_i)}{1+exp(-y_i w^T x_i)}(-y_i x_i) + \lambda L^{'}(w)$
  **end for**
  Output $\nabla(w)$

**Algorithm 2** vectorization

**Require:** initial value w:d-1,x:n-d,y:n-1
  $tmp = \frac{exp(-y \odot (xw))}{1+exp(-y \odot (xw)}$
  $\nabla P(w) = x^T(tmp \odot (-y))$
  Output $\nabla(w)$

# 3 Conclusion

NIPS requires electronic submissions. The electronic submission site is

`https://cmt.research.microsoft.com/NIPS2017/`

Please read carefully the instructions below and follow them faithfully.

## 3.1 Experiment

Papers to be submitted to NIPS 2017 must be prepared according to the instructions presented here. Papers may only be up to eight pages long, including figures. This does not include acknowledgments and cited references which are allowed on subsequent pages. Papers that exceed these limits will not be reviewed, or in any other way considered for presentation at the conference.

The margins in 2017 are the same as since 2007, which allow for $\sim 15\%$ more words in the paper compared to earlier years.

Authors are required to use the NIPS LaTeX style files obtainable at the NIPS website as indicated below. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

## 3.2 Future Work

The style files for NIPS and other conference information are available on the World Wide Web at

`http://www.nips.cc/`

The file `nips_2017.pdf` contains these instructions and illustrates the various formatting requirements your NIPS paper must satisfy.

The only supported style file for NIPS 2017 is `nips_2017.sty`, rewritten for LaTeX $2_\varepsilon$. **Previous style files for LaTeX 2.09, Microsoft Word, and RTF are no longer supported!**

The new LaTeX style file contains two optional arguments: `final`, which creates a camera-ready copy, and `nonatbib`, which will not load the `natbib` package for you in case of package clash.

At submission time, please omit the `final` option. This will anonymize your submission and add line numbers to aid review. Please do *not* refer to these line numbers in your paper as they will be removed during generation of camera-ready copies.

The file `nips_2017.tex` may be used as a "shell" for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

The formatting instructions contained in these style files are summarized in Sections 4, 5, and 6 below.

# 4 Conclusion

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in Section 6 regarding figures, tables, acknowledgments, and references.

# 5 Headings: first level

All headings should be lower case (except for first word and proper nouns), flush left, and bold.

First-level headings should be in 12-point type.

## 5.1 Headings: second level

Second-level headings should be in 10-point type.

### 5.1.1 Headings: third level

Third-level headings should be in 10-point type.

**Paragraphs** There is also a `\paragraph` command available, which sets the heading in bold, flush left, and inline with the text, with the heading followed by 1 em of space.

# 6 Citations, figures, tables, references

These instructions apply to everyone.

## 6.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

Hasselmo, et al. (1995) investigated…

If you wish to load the `natbib` package with options, you may add the following before loading the `nips_2017` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{nips_2017}
```

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous."

### 6.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[2]

### 6.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

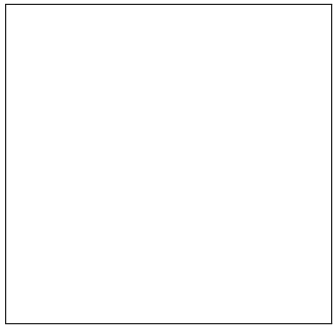

Figure 5: Sample figure caption.

### 6.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 2.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 2.

[1] Sample of the first footnote.

[2] As in this example.

Table 2: Sample table title

| | Part | |
|---|---|---|
| Name | Description | Size ($\mu$m) |
| Dendrite | Input terminal | $\sim$100 |
| Axon | Output terminal | $\sim$10 |
| Soma | Cell body | up to $10^6$ |

## 7 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

## 8 Preparing PDF files

Please prepare submission files with paper size "US Letter," and not, for example, "A4."

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NIPS. Please see `http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf`
- `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

  ```
  \usepackage{amsfonts}
  ```

  followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for $\mathbb{R}$, $\mathbb{N}$ or $\mathbb{C}$. You can also use the following workaround for reals, natural and complex:

  ```
  \newcommand{\RR}{I\!\!R} %real numbers
  \newcommand{\Nat}{I\!\!N} %natural numbers
  \newcommand{\CC}{I\!\!\!\!C} %complex numbers
  ```

  Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

### 8.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the graphics bundle documentation (`http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

**Acknowledgments**

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

## References

## References

[1] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[2] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[3] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[4] Yann LeCun, Corinna Cortes, and Christopher JC Burges. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010.

[5] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[6] Nicolas L Roux, Mark Schmidt, and Francis R Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.

[7] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

[8] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012.

[9] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.