Accelerating Stratified Sampling SGD by Reconstructing Strata

Weijie Liu[1,2] , **Hui Qian**[2] , **Chao Zhang**[2] , **Zebang Shen**[3*] , **Jiahao Xie**[2] and **Nenggan Zheng**[1]

[1]Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, Zhejiang, China
[2]College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang, China
[3]University of Pennsylvania, Philadelphia, Pennsylvania
{westonhunter, qianhui, zczju, xiejh, zng}@zju.edu.cn, zebang@seas.upenn.edu

Abstract

In this paper, a novel stratified sampling strategy is designed to accelerate the mini-batch SGD. We derive a new iteration-dependent surrogate which bound the stochastic variance from above. To keep the strata minimizing this surrogate with high probability, a stochastic stratifying algorithm is adopted in an adaptive manner, that is, in each iteration, strata are reconstructed only if an easily verifiable condition is met. Based on this novel sampling strategy, we propose an accelerated mini-batch SGD algorithm named SGD-RS. Our theoretical analysis shows that the convergence rate of SGD-RS is superior to the state-of-the-art. Numerical experiments corroborate our theory and demonstrate that SGD-RS achieves at least 3.48-times speed-ups compared to vanilla mini-batch SGD.

1 Introduction

Over the past decades, mini-batch Stochastic Gradient Descent (SGD) has become the main workhorse in classical machine learning tasks due to its balance between efficacy and scalability [Finkel *et al.*, 2008; Krizhevsky *et al.*, 2012; Sutskever, 2013]. To form a gradient estimate, the vanilla implementation of mini-batch SGD samples a small and fixed number of training instances *uniformly* in each iteration. Although uniform sampling is easy to implement, the resulting estimate may have a rather high variance since the component gradient of each sample can vary considerably, which results in the slow convergence of the underlying optimization procedure [Zhao and Zhang, 2015; Borsos *et al.*, 2018].

A strategy that can significantly ameliorate the high variance issue is *importance sampling*, which constructs a *non-uniform* sampling distribution to ensure that important training instances are sampled more frequently. Obtaining the sampling distribution that minimizes the variance requires calculating all component gradients and is hence prohibitive. Zhao and Zhang [2015] and Needell *et al.* [2016] use training instances to calculate a fixed sampling distribution. Recently, there has been abundant research in finding iteration-dependent sampling distributions. Katharopoulos and Fleuret [2018] propose an upper bound of the gradient norm that can be calculated in the forward pass of the neural network and use this upper bound to calculate the sampling distribution. Johnson and Guestrin [2018] compute the sampling distribution using robust optimization and find the sampling distribution that is minimax optimal with respect to an uncertainty set. Salehi *et al.* [2017] and Borsos *et al.* [2018] formulate the sampling distribution learning as an online optimization problem with the bandit feedback.

Another line of research that also aims to effectively reduce the variance of gradient estimates is *stratified sampling* [Botev and Ridder, 2014]. Stratified sampling involves two phases: first dividing a population into subpopulations and then applying uniform sampling methods to each subpopulation to form the estimates. When applying to mini-batch SGD, calculating and stratifying all component gradients at every iteration are computationally impractical. To reduce this overhead, Zhao and Zhang [2014] propose a practical algorithm SGD-ss forming fixed strata by clustering training instances instead of component gradients, which essentially determines strata by minimizing a surrogate bounding the stochastic variance from above. More recently, methods based on *repulsive point process* generalize the idea of SGD-ss by adopting a measure of similarity between data points, and allowing strata to overlap [Zhang *et al.*, 2017; Zhang *et al.*, 2019]. Although these methods substantially alleviate the overhead of stratification, their strata are predetermined before the main gradient descent iterations and cannot partition the time-variant component gradients well.

In this paper, a novel stratified sampling strategy is designed to accelerate the mini-batch SGD. Specifically, we derive a new iteration-dependent surrogate and propose a stochastic stratifying algorithm to find the strata that minimize it. Strata are reconstructed in an adaptive manner, that is, in each iteration, strata are reconstructed only if an easily verifiable condition is met. Based on this novel sampling strategy, we propose an improved stratified mini-batch SGD algorithm named SGD-RS. Our contributions can be summarized as follows.

- We derive a new surrogate bounding the stochastic variance from above. Compared with the counterpart used in SGD-ss, ours is closer to the stochastic variance and does not require the component gradient to be Lipschitz continuous with respect to the feature vector. As a re-

*Corresponding Author

sult, our theoretical analysis shows better convergence rates than SGD-ss for both convex and non-convex objective functions.

- We design a strategy to substantially alleviate the overhead of reconstructing strata. Specifically, a reconstructing condition is devised for our method to avoid per-iteration computational load of stratification. We prove that, when current strata do not minimize the surrogate, the reconstruction is invoked with high probability. Therefore, strata maintained by this strategy attain the minimum of the surrogate with high probability for all SGD iterations, with only sporadical stratification.

Besides, we prove that our stratifying algorithm can find the optimal strata with high probability in polynomial time. Numerical experiments are conducted on popular applications including logistic regression and neural network based image classification. On the rcv1 binary classification dataset, SGD-RS achieves a more than 4-times speed-up compared to SGD-ss, with only 1.20% iterations reconstructing strata. On the CIFAR10 image classification dataset, a more than 3.5-times speed-up is obtained by our method compared to SGD-ss, with merely 0.33% iterations reconstructing strata. On average, strata reconstructing occurs at a rate of 0.61%.

Notation. We use bold lowercase symbols to denote vectors (e.g., $\mathbf{x}$). We denote the ℓ_2 norm of vector $\mathbf{a}$ by $\|\mathbf{a}\|$ and the cardinality of set A by $|A|$. For a random variable x, $\mathbb{E}\,x$ is its expectation.

2 Preliminaries

Throughout this paper, we use the *empirical risk minimization* setting to demonstrate the advantages of our method, from which the nature of stratification can be gleaned. Given training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_s$ is the feature vector and y_s is the label for all $s \in \{1, \ldots, n\}$, we solve the empirical risk minimization problem

$$\min_{\mathbf{w}\in W} f(\mathbf{w}) = \frac{1}{n}\sum_{s=1}^{n} f_s(\mathbf{w}), \tag{1}$$

where $W \subseteq \mathbb{R}^d$ is a convex set, and the component function $f_s(\mathbf{w}) = \ell(\mathbf{w}; \mathbf{x}_s, y_s)$ designates the loss of model $\mathbf{w}$ encountered on datum $(\mathbf{x}_s, y_s)$. Let $\mathbf{g}$ be an unbiased estimate of the true gradient ∇f. The variance of $\mathbf{g}$ is defined as

$$V(\mathbf{g}) := \mathbb{E}\,\|\mathbf{g} - \nabla f\|^2. \tag{2}$$

2.1 Stratified Sampling for Mini-batch SGD

We now introduce stratified sampling for mini-batch SGD. At the t-th iteration, we partition the training set into k disjoint subsets $A^t = \{A_1^t, \ldots, A_k^t\}$. Each subset has cardinality $n_i^t = |A_i^t|$. We sample subsets B_i^t from A_i^t, i.e., $B_i^t \subseteq A_i^t$. The mini-batch is the union of these sampled subsets $B^t = B_1^t \cup \cdots \cup B_k^t$. Let $b_i^t = |B_i^t|$. The pre-determined mini-batch size is $b = \sum_{i=1}^k b_i^t$.

SGD-ss assumes that $\frac{\partial \ell(\mathbf{w}^t; \mathbf{x}_s, y_s)}{\partial \mathbf{w}}$ is L-Lipschitz continuous with respect to $\mathbf{x}_s$ [Zhao and Zhang, 2014] and uses iteration-independent strata A throughout training. b_i^t is set as an iteration-independent constant $b_i = \frac{bn_i\sqrt{u_i}}{\sum_{j=1}^n n_j\sqrt{u_j}}$ where $u_i = \frac{1}{n_i}\sum_{s\in A_i}\left\|\mathbf{x}_s - \frac{1}{|A_i|}\sum_{r\in A_i}\mathbf{x}_r\right\|^2$. SGD-ss then has gradient estimate

$$\hat{\nabla} f_{\text{ss}}(\mathbf{w}^t) = \frac{1}{n}\sum_{i=1}^{k}\frac{|A_i|}{b_i}\sum_{s\in B_i^t}\frac{\partial \ell(\mathbf{w}^t; \mathbf{x}_s, y_s)}{\partial \mathbf{w}},$$

and the upper bound of the variance of $\hat{\nabla} f_{\text{ss}}(\mathbf{w}^t)$, i.e.,

$$V\big(\hat{\nabla} f_{\text{ss}}(\mathbf{w}^t)\big) \le \frac{L^2}{nb}\sum_{i=1}^{k}\sum_{s\in A_i}\left\|\mathbf{x}_s - \frac{1}{|A_i|}\sum_{r\in A_i}\mathbf{x}_r\right\|^2. \tag{3}$$

The fixed strata in SGD-ss are obtained by minimizing the right hand side of (3).

2.2 Importance Sampling for Mini-batch SGD

At the t-th iteration, importance sampling methods assign each $s \in \{1, \ldots, n\}$ a probability $q_s^t \ge 0$ such that $\sum_{s=1}^n q_s^t = 1$, and select training instances according to the sampling distribution $\mathbf{q}^t = (q_1^t, \ldots, q_n^t)$. The gradient estimate of importance sampling is then

$$\hat{\nabla} f_{\text{is}}(\mathbf{w}^t) = \frac{1}{|B_{\text{is}}^t|}\sum_{s\in B_{\text{is}}^t}\frac{\nabla f_s(\mathbf{w}^t)}{nq_s^t}, \tag{4}$$

where B_{is}^t is the sampled mini-batch. According to [Johnson and Guestrin, 2018], its variance achieves minimum when $q_s^t \propto \|\nabla f_s(\mathbf{w}^t)\|$ for all s , i.e,

$$q_s^t = \frac{\|\nabla f_s(\mathbf{w}^t)\|}{\sum_{j=1}^n \|\nabla f_j(\mathbf{w}^t)\|}. \tag{5}$$

3 Method

In the section, we discuss two critical details of SGD-RS, strata construction and reconstructing condition.

3.1 Constructing Strata

The gradient estimate adopting both stratified sampling and importance sampling is

$$\hat{\nabla} f_{\text{ss-is}}(\mathbf{w}^t) = \frac{1}{n}\sum_{i=1}^{k}\frac{n_i^t}{b_i^t}\sum_{s\in B_i^t}\frac{1}{n_i^t p_s^t}\frac{\partial \ell(\mathbf{w}^t; \mathbf{x}_s, y_s)}{\partial \mathbf{w}}, \tag{6}$$

where p_s^t is the probability assigned to s in A_i^t.

Proposition 1. *(6) is an unbiased gradient estimate, and its variance can be bounded from above by*

$$V_0^t(A^t, \mathbf{p}^t) = \frac{1}{nb}\sum_{i=1}^{k}\sum_{s\in A_i^t} n_i^t p_s^t\left\|\frac{\nabla f_s(\mathbf{w}^t)}{n_i^t p_s^t} - \frac{\sum_{s\in A_i^t}\nabla f_s(\mathbf{w}^t)}{n_i^t}\right\|^2. \tag{7}$$

Finding the partition A^t and the probability $\mathbf{p}^t$ minimizing (7) can be solved in an alternating manner, that is, we first find the best A^t with fixed $\mathbf{p}^t$, and then vice versa. As we shall see in Section 4.3, if we choose $\mathbf{p}^t$ as a uniform distribution, we can obtain a good partition that already guarantees

Algorithm 1 Stochastic Stratifying

Input: $\mathbf{w}^t$, k, A^0, mini-batch size β and β_0, learning rate sequence λ_i^τ, termination_criterion.
Sample S_0 of size β_0 uniformly with replacement
$G_{S_0} = \{\nabla f_s(\mathbf{w}^t), \forall s \in S_0\}$
$\mathbf{c}^0 \leftarrow$ run Single-Linkage on G_{S_0}
for $\tau = 0, \ldots$ **do**
  Sample indices S of size β uniformly with replacement; Set counter m_i as 0 and S_i as an empty set for all $i = 1, \ldots, k$
  $G_S = \{\nabla f_s(\mathbf{w}^t), \forall s \in S\}$
  for $\nabla f_s(\mathbf{w}^t) \in G_S$ **do**
    Find $i = \arg\min_l \|\nabla f_s(\mathbf{w}^t) - \mathbf{c}_l^\tau\|^2$ and I the current stratum s belongs to
    $S_i = S_i \cup \{s\}$; $m_i = m_i + 1$
    if $i \neq I$ **then** $A_i^\tau = A_i^\tau \cup \{s\}$; $A_I^\tau = A_I^\tau - \{s\}$
    end if
  end for
  for $i = 1, 2, \ldots, k$ **do**
    if $m_i > 0$ **then**
      $\hat{\mathbf{c}}_i^{\tau+1} = \frac{\sum_{s \in S_i} \nabla f_s(\mathbf{w}^t)}{m_i}$
      $\mathbf{c}_i^{\tau+1} = (1 - \lambda_i^\tau)\mathbf{c}_i^\tau + \lambda_i^\tau \hat{\mathbf{c}}_i^{\tau+1}$
    else
      $\mathbf{c}_i^{\tau+1} = \mathbf{c}_i^\tau$
    end if
  end for
  if termination_criterion is satisfied **then** return A^τ
  end if
end for

a better performance than SGD-ss. With such $\mathbf{p}^t$, $V_0^t(A^t, \mathbf{p}^t)$ is reduced to the following form:

$$\hat{V}_0^t(A^t) = \frac{1}{n b} \sum_{i=1}^k \sum_{s \in A_i^t} \left\| \nabla f_s(\mathbf{w}^t) - \frac{\sum_{s \in A_i^t} \nabla f_s(\mathbf{w}^t)}{n_i^t} \right\|^2, \quad (8)$$

which is also the objective function of the standard k-means. We can find optimal A^t by algorithm 1 which is a modified version of the stochastic k-means [Tang and Monteleoni, 2016]. The difference is that stochastic k-means only updates the centroids and does not maintain the strata. As we shall see in Section 4.1, we prove that Algorithm 1 can find the strata minimizing (8) in polynomial time by extending Tang and Monteleoni's result.

Note that in constrast to SGD-ss, we partition the training set by Euclidian distances between gradient vectors. We use Single-Linkage to initialize the centroids [Wikipedia contributors, 2019]. The stratification results for a 2-D logistic regression task are demonstrated in Figure 1 which shows that SGD-RS forms better strata than what SGD-ss does, and the gradient estimate is also closer to the true gradient.

3.2 Reconstructing Condition

To avoid per-iteration reconstruction of strata, we design a reconstructing condition that exploits a subset of the sampled mini-batch $D^t = D_1^t \cup \cdots \cup D_k^t$, where $D_1^t \subseteq B_1^t, \ldots, D_k^t \subseteq B_k^t$. It is defined as

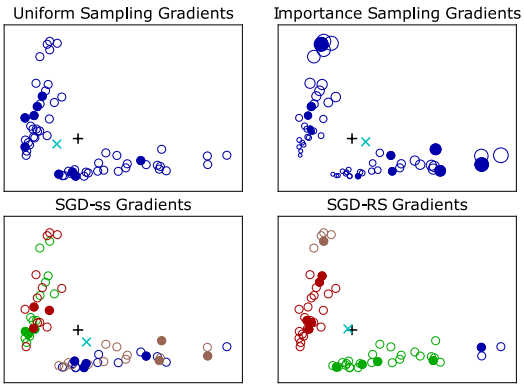

Figure 1: Comparision of gradient estimates in a toy 2-D logistic regression task. Each circle is a gradient vector. If a gradient vector is used to form the gradient estimate, it is denoted by a filled circle. Otherwise it is an unfilled circle. In the picture of importance sampling, the size of each circle is proportional to the possiblity that this gradient vector is used. In the second row, the color of each point indicates the clustering assignment. The black $+$ and the cyan $\times$ in each subplot denote the true gradient and the gradient estimate respectively. As is shown, SGD-ss that only forms strata by clustering training instances cannot partition gradient vectors well. The gradient estimate of SGD-RS is the closest to the true gradient.

$$\zeta(D^t) = \begin{cases} true, & \text{if } \exists v_i^t(s_1, s_2) \geq \epsilon^t, \\ false, & \text{else,} \end{cases} \quad (9)$$

where $v_i^t(s_1, s_2)$ is the distance between two component gradients within the same stratum i, i.e.,

$$v_i^t(s_1, s_2) = \|\nabla f_{s_1}(\mathbf{w}^t) - \nabla f_{s_2}(\mathbf{w}^t)\|, \text{ for } s_1, s_2 \in D_i^t. \quad (10)$$

The intuition is that we reconstruct the strata only when component gradients from the same stratum are not homogeneous.

We summarize the proposed SGD-RS in Algorithm 2. SGD-RS reconstructs strata when the reconstructing condition $\zeta(D^t)$ is true. After calculating $\hat{\nabla} f(\mathbf{w}^t)$, we perform stochastic gradient descent and project $\mathbf{w}^{t+1}$ into W.

4 Theoretical Analysis

In this section, we conduct theoretical analysis about SGD-RS. All proofs are referred to the long version of this paper.

4.1 Algorithm 1 Terminates in Finite Steps

We prove that Algorithm 1 can find A^t that minimizes (8) in finite steps. We first introduce the *proximity condition* and illustrate it in Figure 2.

Definition 1 (Proximity condition (cf. [Kumar and Kannan, 2010])). *Assume point set X admits ground truth non-degenerate*[1] *k-clustering (strata) $T = \{T_1, \ldots, T_k\}$ with ground truth centroids $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$. Let $\bar{\mathbf{g}}$ be the projection of $\mathbf{g}$ onto the $\boldsymbol{\mu}_i$ to $\boldsymbol{\mu}_j$ line. We say a point $\mathbf{g} \in T_i$*

[1] k-clustering is degenerate if any of its k clusters is empty.

Algorithm 2 SGD-RS

Input: initialization $\mathbf{w}^0$, learning rate sequence η^t
for $t = 0, 1, \ldots, T-1$ **do**
  For all i, sample B_i^t from A_i^t
  Calculate gradient vectors $\{\nabla f_s(\mathbf{w}^t), s \in B^t\}$
  if $\zeta(D^t)$ is true **then**
    Run Algorithm 1 to reconstruct A^t
    For all i, sample B_i^t from A_i^t
    Calculate gradient vectors $\{\nabla f_s(\mathbf{w}^t), s \in B^t\}$
  end if
  $\hat{\nabla} f(\mathbf{w}^t) = \frac{1}{n}\sum_{i=1}^k \frac{n_i^t}{b_i^t}\sum_{s\in B_i^t}\nabla f_s(\mathbf{w}^t)$
  $\mathbf{w}^{t+1} = \Pi_W\big(\mathbf{w}^t - \eta^t\hat{\nabla} f(\mathbf{w}^t)\big)$; $A^{t+1} = A^t$
end for
Output: Select $\tilde{\mathbf{w}}^T$ from $\{\mathbf{w}^0, \ldots, \mathbf{w}^{T-1}\}$ according to probability $P(\tilde{\mathbf{w}}^T = \mathbf{w}^t) = \frac{\eta^t}{\sum_{t=0}^{T-1}\eta^t}$

satisfies the proximity condition if for any $i \neq j$, $\bar{\mathbf{g}}$ is at least Δ_{ij} closer to $\boldsymbol{\mu}_i$ than to $\boldsymbol{\mu}_j$, i.e.,

$$\Delta_{ij} \leq \Big|\|\bar{\mathbf{g}} - \boldsymbol{\mu}_j\| - \|\bar{\mathbf{g}} - \boldsymbol{\mu}_i\|\Big|, \forall j \neq i, \mathbf{g} \in T_i. \tag{11}$$

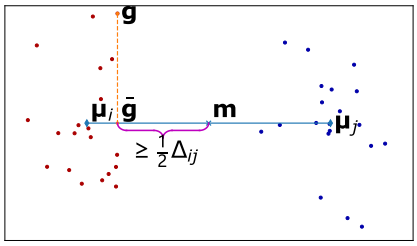

Figure 2: Visualization of Definition 1. $\mathbf{m}$ is the middle point between $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$. Because $\Big|\|\bar{\mathbf{g}} - \boldsymbol{\mu}_j\| - \|\bar{\mathbf{g}} - \boldsymbol{\mu}_i\|\Big| = 2\|\bar{\mathbf{g}} - \mathbf{m}\|$, if $\|\bar{\mathbf{g}} - \mathbf{m}\| \geq \frac{1}{2}\Delta_{ij}$, $\mathbf{g}$ satisfies the proximity condition.

We assume that at iteration t the set of gradient vectors G^t admits ground truth non-degenerate k-clustering (strata) $T^t = \{T_1^t, \ldots, T_k^t\}$ with ground truthcentroids $\boldsymbol{\mu}^t = \{\boldsymbol{\mu}_1^t, \ldots, \boldsymbol{\mu}_k^t\}$. The ground truth strata obtain the minimum of the surrogate. We further make following assumptions about G^t. These assumptions are commonly used in clustering literature (see e.g. [Tang and Monteleoni, 2016]).

Assumption 1 (δ^t-general dataset)**.** *G^t is a general dataset with δ^t-*margin *if all points satisfy the proximity condition and*

$$\delta^t = \min_{i,j\neq i}\Delta_{ij}^t, \delta^t > 0. \tag{12}$$

Assumption 2 ($F(a)$-clusterability)**.** *We say that G^t is $F(a)$-clusterable if G^t has δ^t-margin such that for all $i \neq j$,*

$$\delta^t \geq F(a)\sqrt{\phi^t}\Big(\frac{1}{\sqrt{|T_i^t|}} + \frac{1}{\sqrt{|T_j^t|}}\Big), \tag{13}$$

where $\phi^t = \sum_{i=1}^k\sum_{\mathbf{g}\in T_i^t}\|\mathbf{g} - \boldsymbol{\mu}_i^t\|^2$ and $F(a) \geq \max\big\{\alpha^t, 64, \frac{5a+5}{256a}\big\}$ with $0 < a < 1$ and $\alpha^t = \max_{i,j}\frac{|T_i^t|}{|T_j^t|}$.

Our proof is based on Tang and Monteleoni [2016]'s result which we include here for completeness.

Lemma 2 (Theorem 3 of [Tang and Monteleoni, 2016])**.** *Suppose G^t satisfies Assumptions 1 and 2. Fix any $0 < \theta \leq \frac{1}{e}$. If we set parameters $\beta > \frac{\ln(1-\sqrt{a})}{\ln\left(1-\min_i\frac{4|T_i^t|}{5n}\right)}$, $\lambda_i^\tau = \frac{c'}{\tau_0+\tau}$, where $c' > \frac{1}{1-\sqrt{a}-\left(1-\min_i\frac{4|T_i^t|}{5n}\right)^\beta}$ and $\tau_0 \geq 768(c')^2\Big(1 + \frac{16}{\big(F(a)\big)^2}\Big)^2 n^2\ln^2\big(\frac{1}{\theta}\big)$, and β_0 satisfying $\frac{n\ln\frac{2k}{\xi}}{\min_i|T_i^t|} < \beta_0 < \frac{\xi}{2}\exp\Big(2\big(\frac{F(a)}{4}-1\big)^2(\omega^t)^2\Big)$, then with probability at least $(1-\theta)(1-\xi)$, Algorithm 1 guarantees that*

$$\mathbb{E}\sum_{i=1}^k|T_i^t|\|\mathbf{c}_i^\tau - \boldsymbol{\mu}_i^t\|^2 = O(\frac{1}{\tau}) \tag{14}$$

Algorithm 1 can assign all points to ground truth clusters in polynomial time when Assumptions 1 and 2 hold. The main idea is that $\sum_{i=1}^k\sum_{j\neq i}|T_i^t \cap A_j^\tau|$ can be bounded using $\sum_{i=1}^k|T_i^t|\|\mathbf{c}_i^\tau - \boldsymbol{\mu}_i^t\|^2$. Extending Lemma 2, we have the following result.

Theorem 3. *When Assumptions 1 and 2 hold, if we use parameters in Lemma 2, Algorithm 1 guarantees that*

$$\mathbb{E}\sum_{i=1}^k\sum_{j\neq i}|T_i^t \cap A_j^\tau| = O(\frac{1}{\tau}), \tag{15}$$

with probability at least $(1-\xi)(1-\theta)(1-\nu^\tau)$, where $\nu^\tau = O\big(\frac{k}{\tau}\big)$.

4.2 Reconstructing Condition

In this subsection, we prove that A^t maintained by SGD-RS minimizes (8) for all t with high probability.

The intuition that we can reconstruct strata sporadically is simple: the gradient vectors that are far away from each other at iteration t tend to be still far away at iteration $t+1$, while the gradient vectors that are close to each other at iteration t tend to be still close at iteration $t+1$. More specifically,

$$\Big|\|\nabla f_s(\mathbf{w}^{t+1}) - \nabla f_r(\mathbf{w}^{t+1})\| - \|\nabla f_s(\mathbf{w}^t) - \nabla f_r(\mathbf{w}^t)\|\Big|$$
$$\leq\|\nabla f_s(\mathbf{w}^{t+1}) - \nabla f_s(\mathbf{w}^t)\| + \|\nabla f_r(\mathbf{w}^{t+1}) - \nabla f_r(\mathbf{w}^t)\|.$$

When $\mathbf{w}^t$ and $\mathbf{w}^{t+1}$ are located within a locally smooth region, $\|\nabla f_s(\mathbf{w}^{t+1}) - \nabla f_s(\mathbf{w}^t)\|$ and $\|\nabla f_r(\mathbf{w}^{t+1}) - \nabla f_r(\mathbf{w}^t)\|$ are of order $O\big(\eta^t\|\hat{\nabla} f(\mathbf{w}^t)\|\big)$ which is small if we set small learning rates. SGD-RS requires ϵ^t to satisfy $\epsilon^t \leq \delta^t$, where δ^t is the margin in Assumption 1. When s_2 and s_1 are from different ground truth clusters, $v_i^t(s_1, s_2) \geq \delta^t$, $\zeta(D^t)$ is true. The following lemma states that w.h.p. we reconstruct strata when there exists i such that $A_i^t \neq T_i^t$.

Lemma 4. *Assume that G^t is a δ^t-general dataset. When there exists i, such that $A_i^t \neq T_i^t$, $\zeta(D^t)$ is false with probability at most*

$$P\Big(\zeta(D^t) = false\Big|\exists i, \text{ s.t. } A_i^t \neq T_i^t\Big) \leq \prod_{i=1}^k\Big(\sum_{j=1}^k(p_{ij}^t)^{|D_i^t|}\Big), \tag{16}$$

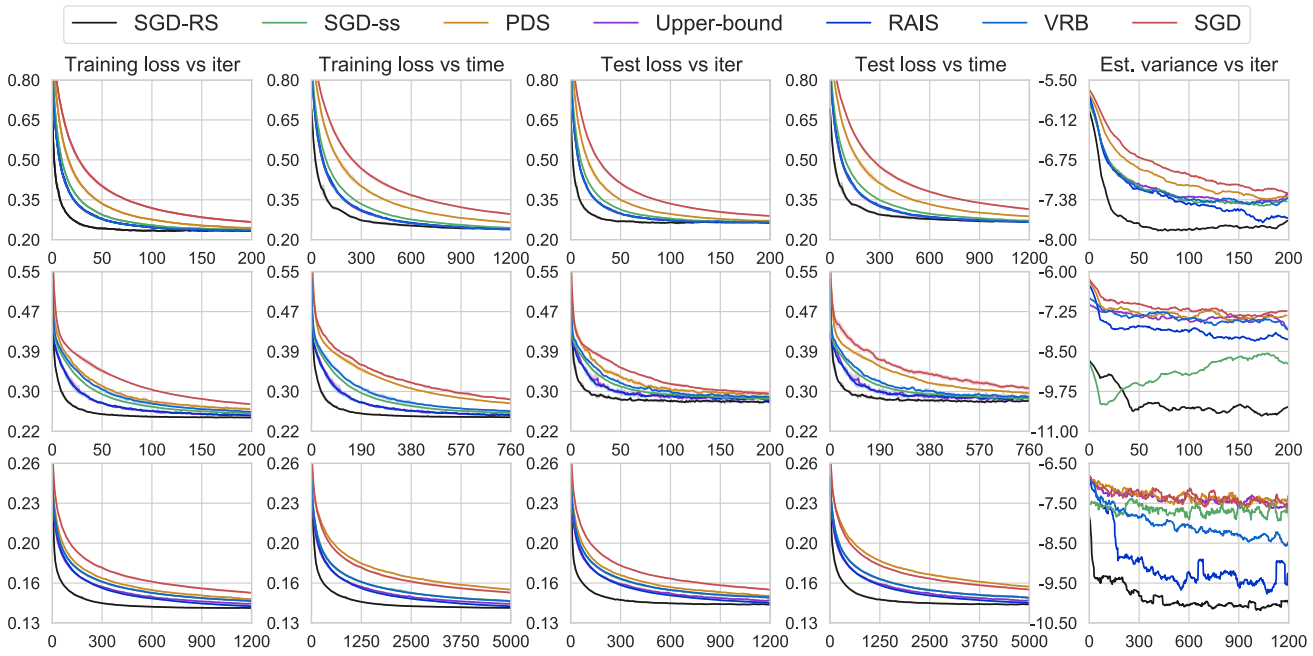

Figure 3: Logistic regression results on rcv1 (top row), ijcnn1 (middle row), and w8a (bottom row). From left to right, we report training loss vs iterations, training loss vs time (second), testing loss vs iterations, testing loss vs time (second), and estimated variance vs iterations. Estimated variance is in log-scale. Filled areas signify ± 1.5 times standard error of the mean.

where $p_{ij}^t = \frac{|A_i^t \cap T_j^t|}{|A_i^t|}$, *for all* $i, j = 1, \ldots, k$.

Remark. As Lemma 4 states, if we choose an appropriate k, $\prod_{i=1}^k \sum_{j=1}^k (p_{ij}^t)^{|D_i^t|}$ is close to 0 even if some $\sum_{j=1}^k (p_{ij}^t)^{|D_i^t|}$'s approach 1 that implies $\zeta(D^t)$ holds with high probability when there exists i, such that $A_i^t \neq T_i^t$.

4.3 Faster Convergence

We now prove that SGD-RS improves the convergence rates for both convex and non-convex objectives by obtaining tighter variance upper bounds. We first give the following lemma.

Lemma 5. *The minimum of* V_0^t *is less than or equal to the minimal upper bound of SGD-ss, i.e.,*

$$\min_{A^t, \mathbf{p}^t} V_0^t(A^t, \mathbf{p}^t) \leq \min_A \frac{L^2}{nb} \sum_{i=1}^k \sum_{s \in A_i} \left\| \mathbf{x}_s - \frac{1}{|A_i|} \sum_{r \in A_i} \mathbf{x}_r \right\|^2. \quad (17)$$

Furthermore, SGD-RS has a tighter variance bound than SGD-ss, i.e.,

$$\min_{A^t} \hat{V}_0^t(A^t) \leq \min_A \frac{L^2}{nb} \sum_{i=1}^k \sum_{s \in A_i} \left\| \mathbf{x}_s - \frac{1}{|A_i|} \sum_{r \in A_i} \mathbf{x}_r \right\|^2. \quad (18)$$

Now we formally give the convergence rates. Let $\mathbf{w}^* = \arg\min_{\mathbf{w} \in W} f(\mathbf{w})$ and $\sigma^2 \geq \max_t \min_{A^t} \hat{V}_0^t(A^t)$ where A^t is maintained by SGD-RS.

Theorem 6. *When* $\|\nabla f(\mathbf{w})\|^2 \leq M^2$, *and* $f_s(\mathbf{w})$ *is convex for all* $s = 1, \ldots, n$, *then the objective can be bounded as follows*

$$\mathbb{E} f(\tilde{\mathbf{w}}^t) - f(\mathbf{w}^*) \leq \frac{\|\mathbf{w}^0 - \mathbf{w}^*\| \sqrt{\sigma^2 + M^2}}{\sqrt{T}}. \quad (19)$$

Theorem 7. *If* $f(\mathbf{w})$ *is non-convex and* γ*-Lipschitz smooth, then the number of iterations for* $\mathbb{E} \|\nabla f(\tilde{\mathbf{w}}^T)\|^2$ *to be smaller than* ϵ^2 *is*

$$T = \Omega\left(\left(\frac{\gamma}{\epsilon^2} + \frac{\gamma \sigma^2}{\epsilon^4} \right) (f(\mathbf{w}_0) - f(\mathbf{w}^*)) \right), \quad (20)$$

where $\Omega(\cdot)$ *is the asymptotic lower bound.*

Combining Lemma 4 and Lemma 5, SGD-RS has a tighter variance bound than SGD-ss for every iteration with probability at least $1 - \prod_{i=1}^k \left(\sum_{j=1}^k (p_{ij}^t)^{|D_i^t|} \right)$, which implies SGD-RS converges faster than SGD-ss for both convex and non-convex cases.

5 Experiments

In this section, we compare SGD-RS with state-of-the-art algorithms, including SGD-ss [Zhao and Zhang, 2014], PDS [Zhang *et al.*, 2019], Upper-bound [Katharopoulos and Fleuret, 2018], RAIS [Johnson and Guestrin, 2018], VRB [Borsos *et al.*, 2018], and vanilla mini-batch SGD. SGD-ss pre-stratifies the dataset. PDS generalizes the idea of SGD-ss by adopting a measure of similarity between data points and also has iteration-independent sampling behaviours. Upper-bound, RAIS, and VRB are importance sampling algorithms.

We run SGD-RS and baseline methods 5 times and report the averaged objective values, averaged testing losses, and the *estimated variances* of gradient estimates. The overhead of stratification and updating importance sampling distributions is considered in the comparison. To estimate the variances, we calculate the distances between gradient estimates

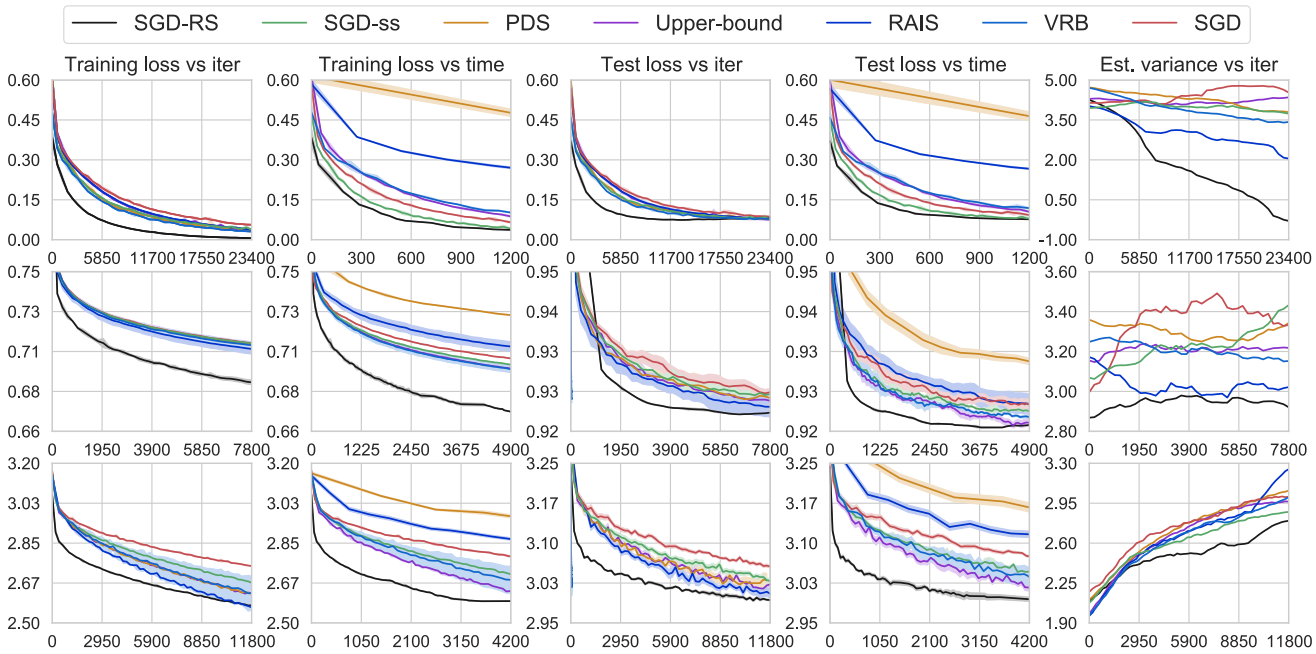

Figure 4: Image classification results on MNIST (top row), CIFAR10 (middle row), and CIFAR100 (bottom row). From left to right, we report training loss vs iterations, training loss vs time (second), testing loss vs iterations, testing loss vs time (second), and estimated variance vs iterations. Estimated variance is in log-scale. Filled areas signify ± 1.5 times standard error of the mean.

and true gradients each time we run the algorithms and average the distances over the 5 rounds. The datasets and the corresponding parameter setup are summarized in Table 1.

Dataset	rcv1	ijcnn1	w8a	MNIST	CIFAR10	CIFAR100
b	100	200	300	100	100	100
k	50	100	50	50	50	50

Table 1: A summary of the datasets and parameter choices in the logistic regression with ℓ_2 norm experiment and the neural network based image classification experiment.

5.1 Logistic Regression

We conduct logistic regression experiments on three real-world benchmark datasets: rcv1, ijcnn1, and w8a[2].

We can see from Figure 3 that both stratified sampling and importance sampling can reduce the variance of gradient estimates and accelerate the convergence. The gradient estimates in SGD-RS are more accurate than baseline methods and hence SGD-RS converges fastest. The variance of SGD-RS can be as small as 8.2% of the variance of SGD-ss. We observe that strata reconstructing happens at rates of 1.20%, 2.90% and 2.07% on rcv1, ijcnn1, and w8a, respectively.

5.2 Image Classification

We evaluate the empirical performance of SGD-RS in image classification benchmark datasets: MNIST, CIFAR10, and CIFAR100. On MNIST, we train a simple network that has three fully-connected layers and two ReLU layers. We train VGG-11 [Simonyan and Zisserman, 2014] on CIFAR10 and ResNet-18 [He *et al.*, 2016] on CIFAR100 respectively and the networks are initialized by running vanilla mini-batch SGD for 50 epochs.

As is shown in Figure 4, SGD-RS consistently outperforms baseline methods both in terms of the objective value and the generalization performance. Strata reconstructing happens at rates of 0.27%, 0.33% and 1.26% on MNIST, CIFAR10, and CIFAR100 respectively. SGD-RS is also the most stable.

[2]These datasets are downloaded from libsvm websites https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

6 Conclusion

We propose a novel stratified sampling strategy for mini-batch SGD. It maintains strata by minimizing an iteration-dependent surrogate majorizing the variance of gradient estimates and accelerates the convergence. To alleviate the overhead of stratifying, we devise a reconstructing condition and reconstruct strata when the condition is met. We demonstrate its superiority in extensive experiments. This strategy is complementary to other variance reduction methods like SVRG [Johnson and Zhang, 2013] and can be used simultaneously.

Acknowledgments

This work is supported by Zhejiang Provincial Natural Science Foundation of China (Grant No: LZ18F020002, LR19F020005), and National Natural Science Foundation of China (Grant No: 61672376, 61751209, 61472347, 61572433, 61972347).

References

[Borsos *et al.*, 2018] Zalán Borsos, Andreas Krause, and Kfir Y Levy. Online variance reduction for stochastic optimization. *arXiv preprint arXiv:1802.04715*, 2018.

[Botev and Ridder, 2014] Zdravko Botev and Ad Ridder. Variance reduction. *Wiley StatsRef: Statistics Reference Online*, pages 1–6, 2014.

[Finkel *et al.*, 2008] Jenny Rose Finkel, Alex Kleeman, and Christopher D Manning. Efficient, feature-based, conditional random field parsing. In *Proceedings of ACL-08: HLT*, pages 959–967, 2008.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Johnson and Guestrin, 2018] Tyler B Johnson and Carlos Guestrin. Training deep models faster with robust, approximate importance sampling. In *Advances in Neural Information Processing Systems*, pages 7265–7275, 2018.

[Johnson and Zhang, 2013] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[Katharopoulos and Fleuret, 2018] Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. *arXiv preprint arXiv:1803.00942*, 2018.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[Kumar and Kannan, 2010] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 299–308. IEEE, 2010.

[Needell *et al.*, 2016] Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Mathematical Programming*, 155(1-2):549–573, 2016.

[Salehi *et al.*, 2017] Farnood Salehi, L Elisa Celis, and Patrick Thiran. Stochastic optimization with bandit sampling. *arXiv preprint arXiv:1708.02544*, 2017.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Sutskever, 2013] Ilya Sutskever. *Training recurrent neural networks*. University of Toronto Toronto, Ontario, Canada, 2013.

[Tang and Monteleoni, 2016] Cheng Tang and Claire Monteleoni. convergence rate of stochastic k-means. *arXiv preprint arXiv:1610.04900*, 2016.

[Wikipedia contributors, 2019] Wikipedia contributors. Single-linkage clustering — Wikipedia, the free encyclopedia, 2019. [Online; accessed 21-January-2020].

[Zhang *et al.*, 2017] Cheng Zhang, Hedvig Kjellstrom, and Stephan Mandt. Determinantal point processes for mini-batch diversification. *arXiv preprint arXiv:1705.00607*, 2017.

[Zhang *et al.*, 2019] Cheng Zhang, Cengiz Öztireli, Stephan Mandt, and Giampiero Salvi. Active mini-batch sampling using repulsive point processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5741–5748, 2019.

[Zhao and Zhang, 2014] Peilin Zhao and Tong Zhang. Accelerating minibatch stochastic gradient descent using stratified sampling. *arXiv preprint arXiv:1405.3080*, 2014.

[Zhao and Zhang, 2015] Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *international conference on machine learning*, pages 1–9, 2015.