Stochastic Doubly Robust Gradient

Anonymous Authors[1]

Abstract

When training a model with observational data, it is often encountered that some values are systematically missing. Learning from such incomplete data using conventional gradient descent may lead to biased estimates of model parameters and even harm the fairness of the decision outcome. In this paper, inspired by doubly robust estimator in observational studies, we propose stochastic doubly robust gradient (SDRG), which is a stochastic gradient descent (SGD) that can deal with the causal missingness in training data. Also, we identify the connection between double robustness and variance reduction in SGD by demonstrating SDRG within a unifying framework for variance reduced SGD. The performance of our approach is empirically tested by showing the convergence in training image classifiers with several examples of missing data.

1. Introduction

The missing data problem is commonly encountered when training a machine learning model with real world data: unlike in the case of clean-cut experimental data, one or more covariates are often missing in recorded observations. Learning from this incomplete data may introduce an undesirable bias, especially when the missingness mechanism is not completely at random. More specifically, if the missingness depends on some covariates (e.g., gender, age, religion, and race) involved in generating the data, the estimation based on these unequally collected observations can be significantly different from the ideal result. This does not only interfere with the consistency in the learning process, but may also have a profound effect on the fairness of learning outcome (Rotnitzky et al., 1998; Tu et al., 2019).

To mitigate this problem, one may want to infer the causal effect of covariates on the missingness mechanism when training models. *Doubly robust estimator* (Robins et al., 1994; Rotnitzky et al., 1998), first introduced in the area of observational study, has been known as an effective method to deal with such causal missingness and still remains popular (Bang & Robins, 2005; Kang et al., 2007; Rotnitzky et al., 2012; Han & Wang, 2013; Zubizarreta, 2015). It employs two well-known approaches, regression adjustment and inverse propensity score weighting, and by its interesting theoretical property, guarantees that the estimate remains unbiased as long as either of the two is specified correctly. In recent years, the double robustness has emerged in wide range of machine learning areas including covariate shift (Reddi et al., 2015b), adversarial training (Kallus, 2018), and reinforcement learning (Dudík et al., 2011; Jiang & Li, 2016; Thomas & Brunskill, 2016; Farajtabar et al., 2018).

In this paper, we introduce the concept of doubly robust estimator to stochastic gradient descent (SGD) to correct the bias induced by the causal missingness in training data while reducing the variance of SGD. Our approach, namely *stochastic doubly robust gradient (SDRG)*, consists of per-covariate control variates and weight-corrected gradients that serve as the methods for regression adjustment and inverse propensity score weighting, respectively. To the best of our knowledge, SDRG is the first SGD algorithm that provides the property, double robustness.

Recently, the use of control variate methods have been excessively studied in the literature of variance reduction of SGD for accelerating the convergence (Roux et al., 2012; Defazio et al., 2014; Johnson & Zhang, 2013; Wang et al., 2013; Reddi et al., 2016). Among these works, stochastic variance reduced gradient (SVRG) (Johnson & Zhang, 2013) and SAGA (Defazio, 2015), which is an unbiased estimate of stochastic average gradient (SAG) (Roux et al., 2012), are closely related to SDRG in that both of them and SDRG can be viewed within a generic framework for control variates of variance reduced SGDs (Reddi et al., 2015a), as SDRG involves the use of per-covariate control variates. Finding the connection between the ways control variates are used, we will show how pursuing double robustness in gradient estimation is aligned with reducing the variance.

Although SDRG and the variance reduced SGDs look similar to each other, there is a notable difference in the situation they can handle: each gradient estimate should be weighted

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

unequally to reflect the missingness in training data and the weight-corrected gradients in SDRG are devised to address this need, whereas the aforementioned variance reduction methods can only consider the equal contribution of individual gradients. The SDRG algorithm can be straightforwardly applied to practical scenarios such as class imbalance problem, as we demonstrate in this paper considering some contextual information (such as class-label or any kind of tag) of training data as the covariates.

In summary, our contributions are as follows:

- We propose the first doubly robust SGD algorithm, called SDRG, and demonstrate that SDRG can be devised in much the same way of SAGA and SVRG with comparable convergence guarantee.
- We define per-covariate momentum functions as control variates of SDRG, and show that it does not require to either periodically calculate (as SVRG) or store (as SAGA) the full gradients.
- We provide a relation between SDRG and momentum, which is a much more direct derivation than the previous relationship presented in (Roux et al., 2012).
- We experimentally show the performance of SDRG in training image classifiers with class-imbalanced MNIST and fashion-MNIST datasets since they are simple, yet commonly arising form of missing data problems.

To clarify, we remark that our work is not aligned with the approaches that employ non-uniform importance sampling for variance reduction in SGD based (Needell et al., 2014; Zhao & Zhang, 2015; Katharopoulos & Fleuret, 2018; Kern & Gyorgy, 2016; Shen et al., 2016). Rather than proposing a sampling criteria, the purpose of our work is to develop a robust learning algorithm when the weights are already determined with regards to the causal missingness.

2. Background and Related Work

A principled optimization problem in modern machine learning is that of the *finite-sum* form: minimization of an objective function $f(\boldsymbol{\theta})$ that are naturally expressed as a summation over a finite set of data $\mathcal{D} = \{x_i\}_{i=1}^n$, which is described as,

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} f(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} w_i f_i(\boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{\theta}$ is a parameter to be optimized over, each term $f_i(\boldsymbol{\theta})$ contributes with the weight w_i, and $w_i = 1$ for typical setup. Such objective in Eqn. (1) commonly appears in the empirical risk minimization framework where the objective is the average of losses computed over the data in $\mathcal{D}$, that is, $f_i(\boldsymbol{\theta}) := L(\boldsymbol{\theta}; x_i)$.

Algorithm 1 Generic Control Variate Method in SGD

Initialize: $\boldsymbol{\theta}^0 \in \mathbb{R}^d$, $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^0$, $\forall i : g_i(\tilde{\boldsymbol{\theta}}) = 0$, $\eta > 0$

1: **for** $t = 0, ..., (T-1)$ **do**
2: (Uniform-) randomly pick an $i_t \in \{1, ..., n\}$
3: Compute the surrogate estimation of $\Delta\boldsymbol{\theta}^t$:
$$\Delta\boldsymbol{\theta}^t = \nabla f_{i_t}(\boldsymbol{\theta}^t) - g_{i_t}(\tilde{\boldsymbol{\theta}}) + \frac{1}{n}\sum_{i=1}^{n} g_{i_t}(\tilde{\boldsymbol{\theta}})$$
4: Update the parameter $\boldsymbol{\theta}^{t+1}$:
$\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta\,\Delta\boldsymbol{\theta}^t$
5: Update the schedule $g_{i_t}(\cdot)$ and/or $\tilde{\boldsymbol{\theta}}$:
Option I (SVRG): update $g_i(\cdot)$, $\tilde{\boldsymbol{\theta}}$ using Eqn. (3)
Option II (SAGA): update $g_{i_t}(\tilde{\boldsymbol{\theta}})$ using Eqn. (4)
6: **end for**
7: **return** $\boldsymbol{\theta}^T$

Stochastic gradient descent (SGD) is a method of choice to deal with such optimization problem. It iteratively updates the design parameter as follows: for each training iteration $t = 1, 2, ..., T$,

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta\,\Delta\boldsymbol{\theta}^t$$
$$\Delta\boldsymbol{\theta}^t = \nabla f_{i_t}(\boldsymbol{\theta}^t)$$

where $\eta > 0$ is a learning rate, $i_t \in \{1, ...n\}$ and $f_{i_t}(\cdot)$ is the loss computed with x_{i_t} which is drawn iteratively from a training set $\mathcal{D}$.

In recent years, a class of algorithms to improve the convergence of SGD by reducing the variance of the estimates has been proposed (Roux et al., 2012; Johnson & Zhang, 2013; Wang et al., 2013; Defazio et al., 2014). Especially, Reddi et al. (Reddi et al., 2015a) provides a formal unifying framework as Alg. 1 for stochastic variance reduction methods proposed in the literature, including SVRG (Johnson & Zhang, 2013), SAGA (Defazio et al., 2014), and SAG (Roux et al., 2012). The basic idea behind the variance reduction methods is to augment the gradient with a control variate and its expectation as,

$$\Delta\boldsymbol{\theta}^t = \nabla f_{i_t}(\boldsymbol{\theta}^t) - g_{i_t}(\tilde{\boldsymbol{\theta}}) + \mathbb{E}[g_i(\tilde{\boldsymbol{\theta}})] \tag{2}$$

where $\tilde{\boldsymbol{\theta}}$ is an approximation of $\boldsymbol{\theta}$. The resulted estimate is unbiased, and has smaller variance if $g_{i_t}(\tilde{\boldsymbol{\theta}})$ has a high correlation with the target estimate $\nabla f_{i_t}(\boldsymbol{\theta}^t)$. That is, for the control variates to be effective and sound, they must satisfy that: (i) they have a high correlation with the target gradient and (ii) their expectation (with respect to random data samples) is inexpensive to compute.

As studied in (Reddi et al., 2015a), the mechanisms of updating control variates, $\{g_i(\tilde{\boldsymbol{\theta}})\}_{i=1}^n$ can be arranged within the unifying framework (see line 5 of Alg. 1) for the well-known variance reduction methods:

SVRG The control variate $g_i(\tilde{\boldsymbol{\theta}})$ is updated using the gradient $\nabla f_i(\tilde{\boldsymbol{\theta}})$ at every iteration, but the parameter $\tilde{\boldsymbol{\theta}}$ is updated after every $m > 0$ iterations as:

$$g_i(\tilde{\boldsymbol{\theta}}) = \nabla f_i(\tilde{\boldsymbol{\theta}}) \;\text{ for all } i$$
$$\tilde{\boldsymbol{\theta}} = \begin{cases} \boldsymbol{\theta}^t & \text{if } t \bmod m = 0 \\ \tilde{\boldsymbol{\theta}} & \text{otherwise.} \end{cases} \tag{3}$$

SAGA The gradients of all functions $\{\nabla f_i(\boldsymbol{\theta})\}_{i=1}^n$ are kept in memory, and one of them corresponding to the training instance is updated at every iteration as:

$$g_i(\tilde{\boldsymbol{\theta}}) = \begin{cases} \nabla f_i(\boldsymbol{\theta}^t) & \text{if } i = i_t \\ g_i(\tilde{\boldsymbol{\theta}}) & \text{otherwise.} \end{cases} \tag{4}$$

For SAG, the only difference with SAGA is that the line 3 of Alg. 1 is changed into,

$$\Delta\boldsymbol{\theta}^t = \frac{1}{n}\left[\nabla f_{i_t}(\boldsymbol{\theta}^t) - g_{i_t}(\tilde{\boldsymbol{\theta}})\right] + \frac{1}{n}\sum_{i=1}^{n} g_{i_t}(\tilde{\boldsymbol{\theta}}). \tag{5}$$

One may notice that SAG update rule does not exactly fit in the formulation of Eqn. (2) since the last term in Eqn. (5) does not become an expectation of the control variate by the scale of $1/n$. However, we categorize SAG as control variate-based variance reduction methods along with other methods, since they are all similar in the sense of incorporating an additional parameter to reduce the variance of estimates.

The aforementioned approaches are originally under strong convexity assumptions and has been extended to non-convex optimization problems (Allen-Zhu & Hazan, 2016; Reddi et al., 2016). Asynchronous (Reddi et al., 2015a; Meng et al., 2016; Huo & Huang, 2017), proximal (Xiao & Zhang, 2014; Allen-Zhu, 2017) and accelerated variants have also been proposed.

3. Stochastic Doubly Robust Gradient

Before to introduce our main algorithm, we begin by formalizing the notion of weighted finite-sum problem that we are interested in and introduce notation that we use throughout this paper.

3.1. Problem Setting

We consider the cases where the individual loss term, f_i, in Eqn. (1) contributes unequally to the optimization, and thus, should be weighted differently according to certain criteria (i.e., $w_i \neq 1$). The weight w_i is defined by the generic importance sampling literature (Weiss et al., 2013; Thomas & Brunskill, 2017; Doroudi et al., 2017): when training and testing data come from different distributions, it can be specified to correct the difference. To elaborate, we are only given a collected set of data from a distribution q (which we call the *sampling distribution*). We originally intend to compute the expected loss over some *proper* distribution p (which we refer to as the *target distribution*), that is, $\mathbb{E}_p[f(\boldsymbol{\theta}; \boldsymbol{x})]$, $\boldsymbol{x} \sim q$. Since we may not have direct access to p, however, we want to do a finite-sum approximate to the expectation over samples $\{\boldsymbol{x}_i\}_{i=1}^n$, $\boldsymbol{x}_i \in \mathbb{R}^d$ drawn from the distribution at hand:

$$\mathbb{E}_p[f(\boldsymbol{\theta})] = \mathbb{E}_q\left[\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} f(\boldsymbol{\theta}; \boldsymbol{x})\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n} \frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} f_i(\boldsymbol{\theta}), \;\; \boldsymbol{x}_i \sim q \tag{6}$$

where $p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i)$, the ratio between target distribution and sampling distribution is regarded as the weighting factor w_i.

3.2. Weight-Corrected Gradient with Variance Reduction

To solve the weighted finite-sum problem as Eqn. (6), in the standard SGD algorithms, the t-th iteration involves picking an instance from sampling distribution q over all instances and updates parameters as

$$\Delta\boldsymbol{\theta}^t = w_{i_t} \nabla f_{i_t}(\boldsymbol{\theta}^t) \tag{7}$$

where $w_{i_t} := p(\boldsymbol{x}_{i_t})/q(\boldsymbol{x}_{i_t})$ is per-sample importance weight. This weight-corrected gradient can be thought of as an inverse propensity score estimate, and we call it *importance weighted SGD*.

Within the literature of doubly robust estimator, our goal is to reduce the variance of stochastic gradient algorithm by introducing a control variate method to accelerate SGD. Given Eqn. (2) and Eqn. (7), an intuitive approach to employ a control variate to stochastic weighted gradient descent is as follows,

$$\Delta\boldsymbol{\theta}^t = w_{i_t} \nabla f_{i_t}(\boldsymbol{\theta}^t) - w_{i_t} g_{i_t}(\tilde{\boldsymbol{\theta}}) + g(\tilde{\boldsymbol{\theta}}) \tag{8}$$

where $\tilde{\boldsymbol{\theta}}$ is required to be highly correlated with $\boldsymbol{\theta}$ and $g(\tilde{\boldsymbol{\theta}}) := \mathbb{E}_p[g_i(\tilde{\boldsymbol{\theta}})]$. The resulted estimate is unbiased as the stochastic weighted gradient in Eqn. (7) with reduced variance.

The constructed gradient estimator in Eqn. (8) involves two variables: *per-sample importance weight* w_{i_t} and *control variate* g_{i_t}. In other words, the estimation accuracy of our approach relies on how correctly the two could be specified. From this perspective, we see an advantage of our formulation by observing that either one of two models needs to be correctly specified to obtain an unbiased estimator:

Seen in a broader context, such distinguishing property of our formulation in Eqn. (8) arises in many areas. Indeed, in

the area of observational study, the property called *double robustness* has been well studied (Robins et al., 1994; Bang & Robins, 2005; Kang et al., 2007; Rotnitzky et al., 2012): doubly robust (DR) estimators involves models for both the propensity score and the conditional mean of the outcome, and remain consistent even if one of those models (but not both) is misspecified. By observing that the constitution and the property of DR estimator are similar with that of our constructed gradient estimator, we see an opportunity to bring the insights of DR estimation into stochastic gradient optimization.

Theorem 1. *The weight-corrected gradient with variance reduction in Eqn. (8) satisfies the double robustness, and thus, Eqn. (8) gives a doubly robust estimation for gradients.*

Proof. We rewrite Eqn. (8) in two-ways:

$$\begin{aligned}\Delta\boldsymbol{\theta}^t &= w_{i_t}\nabla f_{i_t}(\boldsymbol{\theta}^t) - \left\{w_{i_t} g_{i_t}(\tilde{\boldsymbol{\theta}}) - g(\tilde{\boldsymbol{\theta}})\right\} \\ &= w_{i_t}\left\{\nabla f_{i_t}(\boldsymbol{\theta}^t) - g_{i_t}(\tilde{\boldsymbol{\theta}})\right\} + g(\tilde{\boldsymbol{\theta}}).\end{aligned}$$

First, if the per-sample importance weight w_{i_t} is assigned appropriately (i.e., $w_{i_t} = p_{i_t}/q_{i_t}$), then it is satisfied that $g(\tilde{\boldsymbol{\theta}}) = \mathbb{E}_q[w_{i_t} g_{i_t}(\tilde{\boldsymbol{\theta}})]$ since we defined $g(\tilde{\boldsymbol{\theta}}) := \mathbb{E}_p[g_{i_t}(\tilde{\boldsymbol{\theta}})]$. That is, we obtain that $\mathbb{E}[w_{i_t} g_{i_t}(\tilde{\boldsymbol{\theta}}) - g(\tilde{\boldsymbol{\theta}})] = 0$, and Eqn. (8) gives an unbiased estimate of $\Delta\boldsymbol{\theta}^t = w_{i_t}\nabla f_{i_t}(\boldsymbol{\theta}^t)$.

On the other hand, if the control variate g_{i_t} approximates $\nabla f_{i_t}(\boldsymbol{\theta}^{(t)})$ correctly, then it is satisfied that $\mathbb{E}[\nabla f_{i_t}(\boldsymbol{\theta}^t) - g_{i_t}(\tilde{\boldsymbol{\theta}})] = 0$. Thus, we obtain that $\Delta\boldsymbol{\theta}^t = g(\tilde{\boldsymbol{\theta}})$, and so Eqn. (8) is an unbiased estimate regardless of the accuracy of per-sample importance weights.

From the above two properties (i.e., if either w_{i_t} or g_{i_t} is specified correctly, then the estimation of the gradient $\Delta\boldsymbol{\theta}^t$ is unbiased), we can conclude that the gradient estimation in Eqn. (8) satisfies the double robustness. □

The double robustness of Thm. 1 shows the conditions under which unbiased estimation is possible even when some data are missing due to certain causal relationships, and the following result can be derived.

Corollary 1. *At timestep t, for each sample index i_t, if either per-sample importance weight w_{i_t} or per-sample control variate g_{i_t} is specified correctly, the weight-corrected gradient with variance reduction in Eqn. (8) converges to the same point where SGD converges when using complete data with no missing problem.*

3.3. Confounded Mini-Batch Gradient

In the statistics community, and particularly in causal inference settings, DR estimators provide an estimation on average causal effect from observational data, adjusting appropriately for confounders. The ability of DR estimators to taking account to the causal effect of confounders can be also useful in the general machine learning literature.

For instance, in supervised learning, the goal is to seek a function $h : X \to Y$, given n pairs of inputs and corresponding target outputs $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$. Meanwhile, there often exist contextual information associated with instances $\boldsymbol{x}_i$, and it is not directly used to compute the objective (loss) value but may indirectly influence the process of learning the relation between $\boldsymbol{x}_i$ and y_i. In that case, one may want to address causal effect of contextual information in the process of learning the relation between $\boldsymbol{x}_i$ and y_i. By regarding the contextual information as confounding factors in the approximation of the expected loss over training set, we propose a method to adjust the causal effect of contextual information that may confound the gradient estimation.

We maintain different models that estimate two key parameters of Eqn. (8) – importance weights and control variates, conditioned on each configuration of contextual information. In practice, the contextual information could represent a class-label or any tag associated with x_i (Gopal, 2016).

In this paper, we decide to use class-label as contextual information as they are almost always available. Given a finite set of class-labels as the observed contextual information, we confine our interest to class imbalance problem, one of the most common scenario we may encounter in classification tasks. By taking a mini-batch variant of our estimator, we view the class imbalance problem in the perspective of importance weighting: if the data set collected in the mini-batch is sampled from highly skewed distribution, we want to correct the difference between the skewed distribution and the target distribution which is assumed to be uniformly balanced in terms of classes.

Let I_t be the set of indices of mini-batch instances at training iteration t, $I_{t,c}$ is a disjoint subset of I, whose instances belong to class $c \in C$, $w_c(\cdot)$ is per-class model for estimating importance weights and $g_{t,c}(\cdot)$ is per-class control variates. Then, our proposed algorithm for class imbalance problem is described as:

$$\begin{aligned}\Delta\boldsymbol{\theta}^t = \frac{1}{C}\sum_{c=1}^{C}\Bigg[&\frac{1}{|I_{t,c}|}\sum_{i_{t,c}\in I_{t,c}} w_c(i_{t,c})\,\nabla f_{i_{t,c}}(\boldsymbol{\theta}^t) \\ &- \frac{1}{|I_{t,c}|}\sum_{i_{t,c}\in I_{t,c}} w_c(i_{t,c})\, g_{i_{t,c}}(\tilde{\boldsymbol{\theta}}_c) + g(\tilde{\boldsymbol{\theta}}_c)\Bigg].\end{aligned} \tag{9}$$

We call this method *stochastic doubly robust gradient (SDRG)*. In practical implementations, an intuitive way of setting importance weights in the above setting is to compute the proportion of the number of instances that belong to each class-label over the mini-batch size.

Corollary 2. *At timestep t, for each confounder class c and each of its sample index $i_{t,c}$, if either per-class importance weight w_c or per-class control variate $g_{i_{t,c}}$ is specified correctly, SDRG defined in Eqn. (9) converges to the same point where SGD converges when using complete data with no missing problem.*

In particular, if only per-class importance weight w_c is correctly described in Eqn. (9), SDRG can be reduced to the variance reduced SGDs such as SVRG or SAGA according to the per-class control variate $g_{i_{t,c}}$, thus inheriting the properties of the algorithms.

Corollary 3. *If per-class importance weight w_c is specified correctly but not per-class control variate $g_{i_{t,c}}$, SDRG in Eqn. (9) converges to SVRG in case of Eqn. (3) and SAGA in case of Eqn. (4), depending on which variance reduced SGD is selected.*

3.4. SDRG with Per-Covariate Momentum

We show that the existing variance reduction methods are related to SDRG in the sense their mechanism of updating control variates and computing the expectation of them (line 5 of Alg. 1 can be directly used in SDRG update rule). In the following, we suggest a practical way to use SDRG, which takes a per-covariate momentum to replace the per-covariate control variate.

SDRG-M Parameter $\tilde{\boldsymbol{\theta}}_c$ is updated after every m iterations as the same way as SVRG in Eqn. (3):

$$\tilde{\boldsymbol{\theta}}_c = \begin{cases} \boldsymbol{\theta}^t & \text{if } t \bmod m = 0 \\ \tilde{\boldsymbol{\theta}}_c & \text{otherwise} \end{cases}$$

where m is the parameter update frequency and $\tilde{\boldsymbol{\theta}}_c$ is initialized by $\boldsymbol{\theta}^0$ for all c, and we use the per-covariate control variate $g^t(\tilde{\boldsymbol{\theta}}_c)$ as the following per-covariate momentum:

$$g^t(\tilde{\boldsymbol{\theta}}_c) = \eta \, \frac{1}{|I_{t,c}|} \sum_{i_{t,c} \in I_{t,c}} \nabla f_{i_{t,c}}(\boldsymbol{\theta}^t) + \gamma \, g^{t-1}(\tilde{\boldsymbol{\theta}}_c)$$

where $g^0(\tilde{\boldsymbol{\theta}}_c)$ is initialized by 0 for all c.

Using momentum as a per-covariate control variate may be considered a good way to improve the practical learning performance, but unfortunately it is not known to lead to a faster convergence rate (Roux et al., 2012).

4. Relation to Momentum

In practical implementations, it is natural to take $\tilde{\boldsymbol{\theta}}$ as the average or a snapshot from the past iterations (Johnson & Zhang, 2013). However, we propose to set $\tilde{\boldsymbol{\theta}}$ as a geometric weighting of previous gradients and by doing so, show an simple analysis on the relation between the above formulation and momentum optimizer (Qian, 1999): the proposed control variate method which is described as,

$$\begin{aligned} \boldsymbol{\theta}^{t+1} &= \boldsymbol{\theta}^t - \Delta\boldsymbol{\theta}^t \\ &= \boldsymbol{\theta}^t - \Big(w_{i_t} \nabla f_{i_t}(\boldsymbol{\theta}^t) - w_{i_t} g_{i_t}(\tilde{\boldsymbol{\theta}}) + g(\tilde{\boldsymbol{\theta}}) \Big) \end{aligned}$$

reduces to the formulation of momentum under a special setting where $g_{i_t}(\tilde{\boldsymbol{\theta}}) = g(\tilde{\boldsymbol{\theta}}) = \frac{\gamma}{1-\eta} \Delta\boldsymbol{\theta}^{t-1}$ and $w_{i_t} = \eta$ $(\neq 1)$, as follows,

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \big(\eta \, \nabla f_{i_t}(\boldsymbol{\theta}^t) + \gamma \, \Delta\boldsymbol{\theta}^{t-1} \big) \tag{10}$$

where γ is a momentum coefficient. In other words, the control variate method with constant importance weight has the exact same formulation with momentum update rule.

Proposition 1. *The weight-corrected gradient with variance reduction in Eqn. (8) generalizes the momentum update rule in Eqn. (10) into the cases of importance weighting.*

From the perspective of the classic control variate schemes that involves an additional parameter and its expectation, momentum method can be regarded as a biased estimator since the expectation of weighted control variates $w_{i_t} g_{i_t}$ does not correspond to $g(\tilde{\boldsymbol{\theta}})$. The interpretation on the momentum as a biased estimator can be also find in SAG of Eqn. (5) and from this observation, we see a connection between momentum and SAG using our control variate formulation. It is noteworthy that there has been an attempt to find a relation between SAG and momentum optimizer: in the work of (Roux et al., 2012), Eqn. (5) and Eqn. (10) methods can be expressed in the following formulation,

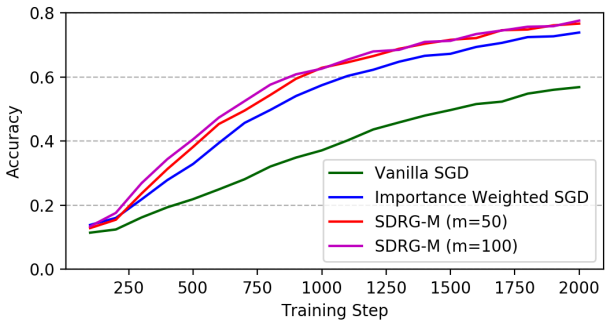
$$\text{SAG}: \quad \boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \eta \sum_{j=1}^{t} S(j, i_{1:t}) \, \nabla f_{i_t}(\boldsymbol{\theta}^j)$$

$$\text{Momentum}: \quad \boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \eta \sum_{j=1}^{t} (\gamma\eta)^{t-j} \, \nabla f_{i_t}(\boldsymbol{\theta}^j)$$

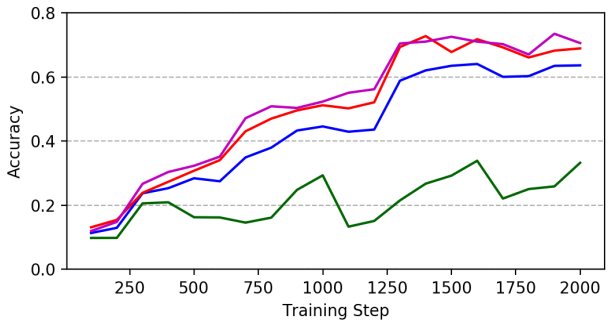
where $S(i, i_{1:t})$ is the selection function and equal to $1/n$ if j corresponds to the last iteration where $j = i_t$ and is set to 0 otherwise. Roux et al. (Roux et al., 2012) finds a connection of SAG on momentum method by showing that they can be written in a similar formation. However, we provide an simpler but stronger analysis by showing a direct connection of SAG on momentum method by proving that they reduce to the exactly same form of equation (Prop. 1).

5. Experiments

In this section, we experiment with SDRG-M in comparison with importance weighted gradient descent for classification task with MNIST (LeCun, 1998) and Fashion-MNIST (Xiao et al., 2017) datasets. Both of MNIST and Fashion-MNIST



(a) Sampling distribution is skewed consistently for a single class (i.e. class 0)



(b) Sampling distribution is skewed for the classes in turn (i.e. class 0, 1, 2, ...)

Figure 1: Accuracy vs training timestep in MNIST

are well balanced for 10 classes, but as we want to test the convergence rate in the setting of class imbalance problem, we modify the sampling distribution to make the sampled instances in mini-batches to be highly unbalanced in terms of class-label. Assuming that we want the training instances are uniformly distributed for all classes, we correct the gradient computed over the mini-batch samples drawn from the skewed distribution by employing importance weighting. In such cases where importance weighting is employed, we demonstrate that our SDRG-M which augments the importance weighted stochastic gradient with control variates whose expectation is replaced by a momentum, shows empirical improvements on the convergence rate.

MNIST MNIST is a large database of handwritten digits from 0 to 9 that is commonly used to evaluate various machine learning algorithms. The images are gray-scale with size of 28 x 28 pixel and they are uniformly balanced for all 10 classes.

We devise two different mechanisms to generate settings for the class imbalance problem: (a) first, the sampling distribution is skewed consistently for a single class during the entire training process. For instance, the probability for the instances from class 0 to be sampled is forced as 0.8, where the instances from the rest of 9 classes are sampled uniformly. (b) Otherwise, the class to be skewed with the sampling probability of 0.8 is selected in turn from class 0, 1, 2, ... to 9.

We test the performance of SDRG-M update rule which can be specifically written as the follows for 10 class classification task: Let I_t be a set of indices for mini-batch samples at t th training iteration, and $I_{t,c}$ be the disjoint subset of I_t that is the set of indices of instances that belong to class-label c.

$$\Delta\boldsymbol{\theta}^t = \frac{1}{10}\sum_{c=0}^{9}\left[\frac{1}{|I_{t,c}|}\sum_{i_{t,c}\in I_{t,c}}\nabla f_{i_{t,c}}(\boldsymbol{\theta}^t) - \frac{1}{|I_{t,c}|}\sum_{i_{t,c}\in I_{t,c}} g_{i_{t,c}}(\tilde{\boldsymbol{\theta}}_c) + g^t(\tilde{\boldsymbol{\theta}}_c)\right]$$

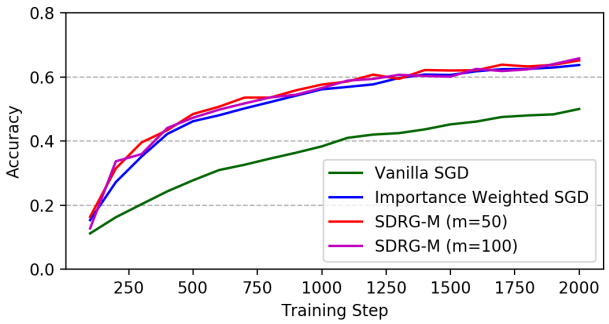
We compare the performance of SDRG-M with two algorithms: importance weighted gradient descent which is described as,

$$\Delta\boldsymbol{\theta}^t = \frac{1}{10}\sum_{c=0}^{9}\frac{1}{|I_{t,c}|}\sum_{i_{t,c}\in I_{t,c}}\nabla f_{i_{t,c}}(\boldsymbol{\theta}^t)$$

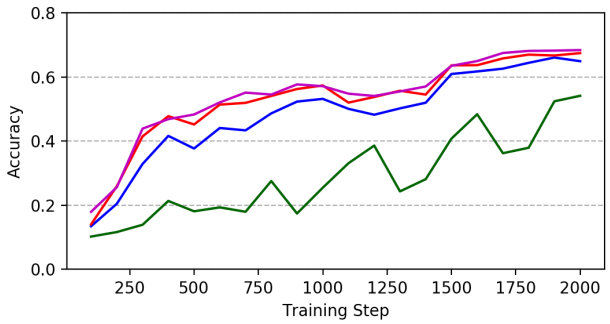
and vanilla SGD.

To compare SDRG-M and other algorithms, we train a neural network (with one fully-connected hidden layer of 100 nodes and ten softmax output nodes) using cross-entropy loss with mini-batches of size 20 and learning rate of 0.01. We evaluate the performance of SDRG-M with different frequency of updating $\tilde{\boldsymbol{\theta}}$: $m = 50, 100$. For momentum parameters, $\gamma = 0.9$ and $\eta = 0.1$ are used. And we add two extra parameters α (0.5 for (a) and 1.5 for (b)) and β (1.5 for (a) and 0.5 for (b)) for additional weighting sample gradients and control variate function, respectively. These parameters are corresponding to weights of $w_i = \eta$ and $g_{i_t}(\tilde{\boldsymbol{\theta}}) = g(\tilde{\boldsymbol{\theta}}) = \frac{\gamma}{1-\eta}\Delta\boldsymbol{\theta}^{t-1}$ in Eqn. (10). The results are all generated by taking the average from 20 runs of experiments. The confidence intervals are too insignificant to be noted and we decided not to include them in the figures.

In both Fig. 1 (a) and (b), SDRG-M empirically shows a faster convergence rate than importance weighted SGD and vanilla SGD. However, in Fig 1. (b) where the class



(a) Sampling distribution is skewed consistently for a single class (i.e. class 0)



(b) Sampling distribution is skewed for the classes in turn (i.e. class 0, 1, 2, ...)

Figure 2: Accuracy vs training timestep in Fashion-MNIST

to be skewed is changed after m iterations, which might be a harder case than (b), the overall convergance rates fluctuate more than Fig, 1 (a), but SDRG-M shows more robust convergence than the other algorithms.

Fashion-MNIST Fashion-MNIST is an MNIST-like database of clothes. The images are grayscale, with size of 28 x 28 and associated with labels from 10 classes. We evaluate the performance of SDRG-M in comparison with importance weighted SGD and vanilla SGD, under experimental settings which are exactly same with the MNIST experiment above.

In Fig 2. we can observe the same tendency in the convergence rate with Fig. 1, where SDRG-M shows less variant converging pattern for both of cases (a) and (b).

6. Conclusion and Future Work

In this paper, we proposed a SGD algorithm that addresses the caual effects of covariates on the missingness of incomplete data. Along with the previous studies that extended the use of doubly robust estimators to a variety of machine learning areas (Reddi et al., 2015b; Kallus, 2018; Dudík et al., 2011; Jiang & Li, 2016; Thomas & Brunskill, 2016; Farajtabar et al., 2018), this paper has been the first approach to apply the idea of doubly robust estimator to stochastic optimization. In SDRG, employing control variates for regression adjustment allowed us to view the proposed method in the framework for variance reduced SGDs that also utilize control variate schemes, except that ours include the additional weight correction term. For the efficiency in computing and storing control variates, we suggested to choose momentum functions as control variates in SDRG-M, and by doing so, the direct derivation from SDRG to the momentum has been found as a byproduct. In addition to these notable findings, empirical results have demonstrated that the proposed SDRG shows faster convergence than vanilla SGD and importance weighted SGD.

Future work includes further empirical studies to evaluate the performance of SDRG in various cases of missing data problem: for instance, the setting where multiple covariates are involved in the missingness mechanism. It would be essential to do deeper investigation on how the property of double robustness affects the convergence of SGD, compared to the existing variance reduced SGDs such as SAGA and SVRG.

References

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1200–1205. ACM, 2017.

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pp. 699–707, 2016.

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, 2005.

Defazio, A. New optimisation methods for machine learning. *arXiv preprint arXiv:1510.02533*, 2015.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pp. 1646–1654, 2014.

Doroudi, S., Thomas, P. S., and Brunskill, E. Importance sampling for fair policy selection. In *Proceedings of*

the Thirty-Third Conference on Uncertainty in Artificial Intelligence, 2017.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1447–1456, 2018.

Gopal, S. Adaptive sampling for SGD by exploiting side information. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 364–372, 2016.

Han, P. and Wang, L. Estimation with missing data: beyond double robustness. *Biometrika*, 100(2):417–430, 2013.

Huo, Z. and Huang, H. Asynchronous mini-batch gradient descent with variance reduction for non-convex optimization. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 652–661, 2016.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pp. 315–323, 2013.

Kallus, N. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. *arXiv preprint arXiv:1802.05664*, 2018.

Kang, J. D., Schafer, J. L., et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

Katharopoulos, A. and Fleuret, F. Not all samples are created equal: Deep learning with importance sampling. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2525–2534, 2018.

Kern, T. and Gyorgy, A. SVRG++ with non-uniform sampling. 2016.

LeCun, Y. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Meng, Q., Chen, W., Yu, J., Wang, T., Ma, Z., and Liu, T.-Y. Asynchronous accelerated stochastic gradient descent. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 1853–1859, 2016.

Needell, D., Ward, R., and Srebro, N. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. In *Advances in Neural Information Processing Systems*, pp. 1017–1025, 2014.

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

Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. J. On variance reduction in stochastic gradient descent and its asynchronous variants. In *Advances in Neural Information Processing Systems*, pp. 2647–2655, 2015a.

Reddi, S. J., Poczos, B., and Smola, A. J. Doubly robust covariate shift correction. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2949–2955, 2015b.

Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pp. 314–323, 2016.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association*, 93(444):1321–1339, 1998.

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456, 2012.

Roux, N. L., Schmidt, M., and Bach, F. R. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in neural information processing systems*, pp. 2663–2671, 2012.

Shen, Z., Qian, H., Zhou, T., and Mu, T. Adaptive variance reducing for stochastic gradient descent. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 1990–1996, 2016.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2139–2148, 2016.

Thomas, P. S. and Brunskill, E. Importance sampling with unequal support. In *Thirty-First AAAI Conference on Artificial Intelligence*, pp. 2646–2652, 2017.

Tu, R., Zhang, C., Ackermann, P., Kjellström, H., and Zhang, K. Causal discovery in the presence of missing data. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Wang, C., Chen, X., Smola, A. J., and Xing, E. P. Variance reduction for stochastic gradient optimization. In *Advances in Neural Information Processing Systems*, pp. 181–189, 2013.

Weiss, J. C., Natarajan, S., and Page, C. D. Learning when to reject an importance sample. *AAAI Late-Breaking Developments*, pp. 17, 2013.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xiao, L. and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Zhao, P. and Zhang, T. Stochastic optimization with importance sampling for regularized loss minimization. In *international conference on machine learning*, pp. 1–9, 2015.

Zubizarreta, J. R. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.