WHY RESAMPLING OUTPERFORMS REWEIGHTING FOR CORRECTING SAMPLING BIAS

Anonymous authors
Paper under double-blind review

ABSTRACT

A data set sampled from a certain population is biased if the subgroups of the population are sampled at proportions that are significantly different from their underlying proportions. Training machine learning models on biased data sets requires correction techniques to compensate for the bias. We consider two commonly-used techniques, resampling and reweighting, that rebalance the proportions of the subgroups to maintain the desired objective function. Though statistically equivalent, it has been observed that reweighting outperforms resampling when combined with stochastic gradient algorithms. By analyzing illustrative examples, we explain the reason behind this phenomenon using tools from dynamical stability and stochastic asymptotics. We also present experiments from regression, classification, and off-policy prediction to demonstrate that this is a general phenomenon. We argue that it is imperative to consider the objective function design and the optimization algorithm together while addressing the sampling bias.

1 INTRODUCTION

A data set sampled from a certain population is called *biased* if the subgroups of the population are sampled at proportions that are significantly different from their underlying population proportions. Applying machine learning algorithms naively to biased training data can raise serious concerns and lead to controversial results (Sweeney, 2013; Kay et al., 2015; Menon et al., 2020). In many domains such as demographic surveys, fraud detection, identification of rare diseases, and natural disasters prediction, a model trained from biased data tends to favor oversampled subgroups by achieving high accuracy there while sacrificing the performance on undersampled subgroups. Although one can improve by diversifying and balancing during the data collection process, it is often hard or impossible to eliminate the sampling bias due to historical and operational issues.

In order to mitigate the biases and discriminations against the undersampled subgroups, a common technique is to preprocess the data set by compensating the mismatch between *population proportion* and the *sampling proportion*. Among various approaches, two commonly-used choices are *reweighting* and *resampling*. In reweighting, one multiplies each sample with a ratio equal to its population proportion over its sampling proportion. In resampling, on the other hand, one corrects the proportion mismatch by either generates new samples for the undersampled subgroups or selecting a subset of samples for the oversampled subgroups. Both methods result in statistically equivalent models in terms of the loss function (see details in Section 2). However, it has been observed in practice that resampling often outperforms reweighting significantly, such as boosting algorithms in classification (Galar et al., 2011; Seiffert et al., 2008), off-policy prediction in reinforcement learning (Schlegel et al., 2019) and so on. The obvious question is why.

Main contributions. Our main contribution is to provide an answer to this question: resampling outperforms reweighting because of the stochastic gradient-type algorithms used for training. To the best of our knowledge, our explanation is the first theoretical quantitative analysis for such this phenomenon. With stochastic gradient descent (SGD) being the dominant method for model training, our analysis is based on some recent developments for understanding SGD. We show via simple and explicitly analyzable examples why resampling generates expected results while reweighting performs undesirably. Our theoretical analysis is based on two points of view, one from the dynamical stability perspective and the other from stochastic asymptotics.

In addition to the theoretical analysis, we present experimental examples from three distinct categories (classification, regression, and off-policy prediction) to demonstrate that resampling outperforms reweighting in practice. This empirical study illustrates that this is a quite general phenomenon when models are trained using stochastic gradient type algorithms.

Our theoretical analysis and experiments show clearly that adjusting only the loss functions is not sufficient for fixing the biased data problem. The output can be disastrous if one overlooks the optimization algorithm used in the training. In fact, recent understanding has shown that objective function design and optimization algorithm are closely related, for example optimization algorithms such as SGD play a key role in the generalizability of deep neural networks. Therefore in order to address the biased data issue, we advocate for *considering data, model, and optimization as an integrated system.*

Related work. In a broader scope, resampling and reweighting can be considered as instances of preprocessing the training data to tackle biases of machine learning algorithms. Though there are many well-developed resampling (Mani & Zhang, 2003; He & Garcia, 2009; Maciejewski & Stefanowski, 2011) and reweighting (Kumar et al., 2010; Malisiewicz et al., 2011; Chang et al., 2017) techniques, we only focus on the reweighting approaches that do not change the optimization problem. It has been well-known that training algorithms using disparate data can lead to algorithmic discrimination (Bolukbasi et al., 2016; Caliskan et al., 2017), and over the years there have been growing efforts to mitigate such biases, for example see (Amini et al., 2019; Kamiran & Calders, 2012; Calmon et al., 2017; Zhao et al., 2019; López et al., 2013). We also refer to (Haixiang et al., 2017; He & Ma, 2013; Krawczyk, 2016) for a comprehensive review of this growing research field.

Our approaches for understanding the dynamics of resampling and reweighting under SGD are based on tools from numerical analysis for stochastic systems. Connections between numerical analysis and stochastic algorithms have been rapidly developing in recent years. The dynamical stability perspective has been used in (Wu et al., 2018) to show the impact of learning rate and batch size in minima selection. The stochastic differential equations (SDE) approach for approximating stochastic optimization methods can be traced in the line of work (Li et al., 2017; 2019; Rotskoff & Vanden-Eijnden, 2018; Shi et al., 2019), just to mention a few.

2 Problem setup

Let us consider a population that is comprised of two different groups, where a proportion a_1 of the population belongs to the first group, and the rest with the proportion $a_2 = 1 - a_1$ belongs to the second (i.e., $a_1, a_2 > 0$ and $a_1 + a_2 = 1$). In what follows, we shall call a_1 and a_2 the *population proportions*. Consider an optimization problem for this population over a parameter θ. For simplicity, we assume that each individual from the first group experiences a loss function $V_1(\theta)$, while each individual from the second group has a loss function of type $V_2(\theta)$. Here the loss function $V_1(\theta)$ is assumed to be identical across all members of the first group and the same for $V_2(\theta)$ across the second group, however it is possible to extend the formulation to allow for loss function variation within each group. Based on this setup, a minimization problem over the whole population is to find

$$\theta^* = \arg\min_{\theta} V(\theta), \quad \text{where } V(\theta) \equiv a_1 V_1(\theta) + a_2 V_2(\theta). \tag{1}$$

For a given set Ω of N individuals sampled uniformly from the population, the empirical minimization problem is

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{r \in \Omega} V_{i_r}(\theta), \tag{2}$$

where $i_r \in \{1, 2\}$ denotes which group an individual r belongs to. When N grows, the empirical loss in (2) is consistent with the population loss in (1) as there are approximately a_1 fraction of samples from the first group and a_2 fraction of samples from the second.

However, the sampling can be far from uniformly random in reality. Let n_1 and n_2 with $n_1 + n_2 = N$ denote the number of samples from the first and the second group, respectively. It is convenient to define $f_i, i = 1, 2$ as the sampling proportions for each group, i.e., $f_1 = n_1/N$ and $f_2 = n_2/N$ with $f_1 + f_2 = 1$. The data set is biased when the sampling proportions f_1 and f_2 are different from the

population proportions a_1 and a_2. In such a case, the empirical loss is $f_1 V_1(\theta) + f_2 V_2(\theta)$, which is clearly wrong when compared with (1).

Let us consider two basic strategies to adjust the model: reweighting and resampling. In reweighting, one assigns to each sample $r \in \Omega$ a weight a_{i_r}/f_{i_r} and the reweighting loss function is

$$V_w(\theta) \equiv \frac{1}{N} \sum_{r \in \Omega} \frac{a_{i_r}}{f_{i_r}} V_{i_r}(\theta) \approx a_1 V_1(\theta) + a_2 V_2(\theta). \tag{3}$$

In resampling, one either adds samples to the minority group (i.e., oversampling) or removing samples from the majority group (i.e., undersampling). Although the actual implementation of oversampling and undersampling could be quite sophisticated in order to avoid overfitting or loss of information, mathematically we interpret the resampling as constructing a new set of $\{s\}$ samples of size M, among which $a_1 M$ samples are of the first group and $a_2 M$ samples of the second. The resampling loss function is

$$V_s(\theta) \equiv \frac{1}{M} \sum_{s} V_{i_s}(\theta) \approx a_1 V_1(\theta) + a_2 V_2(\theta). \tag{4}$$

Notice that both $V_w(\theta)$ and $V_s(\theta)$ are consistent with the population loss function $V(\theta)$. This means that, under mild conditions on $V_1(\theta)$ and $V_2(\theta)$, a deterministic gradient descent algorithm from a generic initial condition converges to similar solutions for $V_w(\theta)$ and $V_s(\theta)$. However, as we shall explain below, the behavior can be drastically different when a stochastic gradient algorithm is used. The key reason is that the variances experienced can be very different.

3 STABILITY ANALYSIS

Let us use a simple example to illustrate why resampling outperforms reweighting under SGD, from the viewpoint of stability. Consider two loss functions V_1 and V_2 with disjoint supports,

$$V_1(\theta) = \begin{cases} \frac{1}{2}(\theta+1)^2 - \frac{1}{2}, & \theta \le 0 \\ 0, & \theta > 0, \end{cases} \quad V_2(\theta) = \begin{cases} 0, & \theta \le 0 \\ \frac{1}{2}(\theta-1)^2 - \frac{1}{2}, & \theta > 0, \end{cases} \tag{5}$$

each of which is quadratic on its support. The population loss function is $V(\theta) = a_1 V_1(\theta) + a_2 V_2(\theta)$, with two local minima at $\theta = -1$ and $\theta = 1$. The gradients for V_1 and V_2 are

$$\nabla V_1(\theta) = \begin{cases} \theta + 1, & \theta \le 0 \\ 0, & \theta > 0. \end{cases}, \quad \nabla V_2(\theta) = \begin{cases} 0, & \theta \le 0 \\ \theta - 1, & \theta > 0. \end{cases}$$

Suppose that the population proportions satisfy $a_2 > a_1$, then $\theta = 1$ is the global minimum and it is desired that SGD should be stable near it. However, as shown in Figure 1, when the sampling proportion f_2 is significantly less than the population proportion a_2, for reweighting $\theta = 1$ can easily become unstable: even if one starts near the global minimum $\theta = 1$, the trajectories for reweighting always gear towards $\theta = -1$ after a few steps (see Figure 1(1)). On the other hand, for resampling $\theta = 1$ is quite stable (see Figure 1(2)).

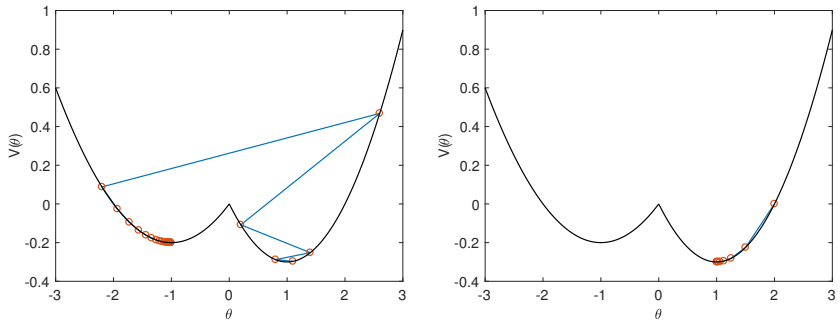
The expectations of the stochastic gradient are the same for both methods. It is the difference in the second moment that explains why trajectories near the two minima exhibit different behaviors. Our explanation is based on the stability analysis framework used in (Wu et al., 2018). By definition, a stationary point θ^* is *stochastically stable* if there exists a uniform constant $C > 0$ such that $\mathbb{E}[\|\theta_k - \theta^*\|^2] \le C\|\theta_0 - \theta^*\|^2$, where θ_k is the k-th iterate of SGD. The stability conditions for resampling and reweighting are stated in the following two lemmas.

Lemma 1. *For resampling, the conditions for the SGD to be stochastically stable around $\theta = -1$ and $\theta = 1$ are respectively*

$$(1 - \eta a_1)^2 + \eta^2 a_1 a_2 \le 1, \quad (1 - \eta a_2)^2 + \eta^2 a_1 a_2 \le 1.$$

Lemma 2. *For reweighting, the condition for the SGD to be stochastically stable around $\theta = -1$ and $\theta = 1$ are respectively*

$$(1 - \eta a_1)^2 + \eta^2 f_1 f_2 \left(\frac{a_1}{f_1}\right)^2 \le 1, \quad (1 - \eta a_2)^2 + \eta^2 f_1 f_2 \left(\frac{a_2}{f_2}\right)^2 \le 1.$$



(1) Reweighting (2) Resampling

Figure 1: Comparison of reweighting and resampling with $a_1/a_2 = 0.4/0.6$ and $f_1/f_2 = 0.9/0.1$ at the learning rate $\eta = 0.5$. Both experiments start near $\theta = 1$. (1) For reweighting, the trajectory can end up at $\theta = -1$ after a few iterations, but $\theta = -1$ is not the global minimizer. (2) For resampling, the trajectory stays close to the desired minimizer $\theta = 1$. Hence resampling is more reliable than reweighting.

Note that the stability conditions for resampling are independent of the sampling proportions (f_1, f_2), while the ones for reweighting clearly depend on (f_1, f_2). We refer the detailed computations to Appendix A.

Lemma 2 shows that reweighting makes problems stiffer in terms of the stability criterion. Let us consider the case $a_1 = \frac{1}{2} - \epsilon, a_2 = \frac{1}{2} + \epsilon$ with a small constant $\epsilon > 0$ and $f_2/f_1 \ll 1$. For reweighting, the global minimum $\theta = 1$ is stochastically stable only if $\eta(1 + f_1/f_2) \leq 4 + O(\epsilon)$. This condition becomes rather stringent in terms of the learning rate η since $f_1/f_2 \gg 1$. On the other hand, the local minimizer $\theta = -1$ is stable if $\eta(1 + f_2/f_1) \leq 4 + O(\epsilon)$, which could be satisfied for a broader range of η because $f_2/f_1 \ll 1$. In other words, for a fixed learning rate η, when the ratio f_2/f_1 between the sampling proportions is sufficiently small, the desired minimizer $\theta = 1$ is no longer statistically stable with respect to SGD.

4 SDE ANALYSIS

The stability analysis can only be carried for a learning rate η of a finite size. However, even for a small learning rate η, one can show that the reweighting method is still unreliable from a different perspective. This section applies stochastic differential equation analysis to demonstrate it.

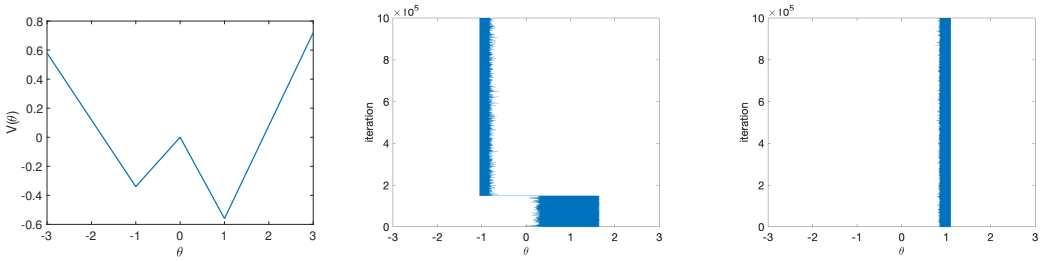
Let us again use a simple example to illustrate the main idea. Consider the following two loss functions,

$$V_1(\theta) = \begin{cases} |\theta + 1| - 1, & \theta \leq 0 \\ \epsilon\theta, & \theta > 0 \end{cases}, \quad V_2(\theta) = \begin{cases} -\epsilon\theta, & \theta \leq 0 \\ |\theta - 1| - 1, & \theta > 0 \end{cases},$$

with $0 < \epsilon \ll 1$. The population loss function is $V(\theta) = a_1 V_1(\theta) + a_2 V_2(\theta)$ with local minimizers $\theta = -1$ and $\theta = 1$. Note that the $O(\epsilon)$ terms are necessary. Without it, if the SGD starts in $(-\infty, 0)$, all iterates will stay in this region because there is no drift from $V_2(\theta)$. Similarly, if the SGD starts in $(0, \infty)$, no iterates will move to $(-\infty, 0)$. That means the result of SGD only depends on the initialization when $O(\epsilon)$ term is absent.

In Figure 2, we present numerical simulations of the resampling and reweighting methods for the designed loss function $V(\theta)$. If $a_2 > a_1$, then the global minimizer of $V(\theta)$ is $\theta = 1$ (see the Figure 2(1)). Consider a setup with population proportions $a_1/a_2 = 0.4/0.6$ along sampling proportions $f_1/f_2 = 0.9/0.1$, which are quite different. Figures 2(2) and (3) show the dynamics under the reweighting and resampling methods, respectively. The plots show that, while the trajectory for resampling is stable across all the time, the trajectory for reweighting quickly escapes to the (non-global) local minimizer $\theta = -1$ even when it starts near the global minimizer $\theta = 1$.

When the learning rate is sufficiently small, one can approximate the SGD by an SDE, which in this piece-wise linear loss example is approximately a Langevin dynamics with a piecewise constant mobility. In particular when the dynamics reaches equilibrium, the stationary distribution of the



(1) Loss function $V(\theta)$ (2) Reweighting (3) Resampling

Figure 2: Comparison of reweighting and resampling with learning rate $\eta = 0.12$. We set $a_1/a_2 = 0.4/0.6$, $f_1/f_2 = 0.9/0.1$ and $\epsilon = 0.1$. Both experiments start near $\theta = 1$. In (2) where reweighting is used, the trajectory skips to the local minimizer $\theta = -1$ later. In (3) where resampling is used, it stabilizes at the global minimizer $\theta = 1$ all the time.

stochastic process is approximated by a Gibbs distribution, which gives the probability densities at the stationary points. Let us denote $p_s(\theta)$ and $p_w(\theta)$ as the stationary distribution over θ under resampling and reweighting, respectively. Suppose that $a_2 > a_1$ and hence $V(1) < V(-1)$. The following lemmas quantitatively summarize the results.

Lemma 3. *When $a_2 > a_1$, $V(1) < V(-1)$. The stationary distribution for resampling satisfies the relationship*

$$\frac{p_s(1)}{p_s(-1)} = \exp\left(-\frac{2}{a_1 a_2 \eta}(V(1) - V(-1))\right) + O(\epsilon) > 1.$$

Lemma 4. *With $a_2 > a_1$, $V(1) < V(-1) < 0$. Under the condition $\frac{f_2}{f_1} \leq \frac{a_2}{a_1}\sqrt{\frac{V(-1)}{V(1)}}$ for the sampling proportions, the stationary distribution for reweighting satisfies the relationship*

$$\frac{p_w(1)}{p_w(-1)} = \frac{a_1^2/f_1^2}{a_2^2/f_2^2} \exp\left(-\frac{2f_2/f_1}{a_2^2\eta}V(1) + \frac{2f_1/f_2}{a_1^2\eta}V(-1)\right) + O(\epsilon) < 1.$$

The proofs of the above two lemmas can be found in Appendix B. Lemma 3 shows that for resampling it is always more likely to find θ at the global minimizer 1 than at the local minimizer -1. Lemma 4 states that for reweighting it is more likely to find θ at the local minimizer -1 when $\frac{f_2}{f_1} \leq \frac{a_2}{a_1}\sqrt{\frac{V(-1)}{V(1)}}$. Together, they explain the phenomenon shown in Figure 2.

To better understand the condition in Lemma 4, let us consider the case $a_1 = \frac{1}{2} - \epsilon, a_2 = \frac{1}{2} + \epsilon$ with a small constant $\epsilon > 0$. Under this setup, $V(-1)/V(1) \approx 1$. Whenever the ratio of the sampling proportions f_2/f_1 is significantly less than the ratio of the population proportions $a_2/a_1 \approx 1$, reweighting will lead to the undesired behavior. The smaller the ratio f_2/f_1 is, the less likely the global minimizer will be visited.

General results. Let us now consider the minimization of $V(\theta) = a_1 V_1(\theta) + a_2 V_2(\theta)$ for more general V_1, V_2 and also θ in high dimensions. It is in fact not clear how to extend the above stochastic analysis to more general functions $V(\theta)$. Instead we focus on the *transition time* from one stationary point to another in order to understand the behavior of resampling and reweighting. For this purpose, we again resort to the SDE approximation of the SGD in the continuous time limit.

Such a SDE approximation, first introduced in (Li et al., 2017), involves a data-dependent covariance coefficient for the diffusion term and is justified in the weak sense with an error of order $O(\sqrt{\eta})$. More specifically, the dynamics can be approximated by

$$d\Theta = \nabla V(\Theta)dt + \sqrt{\eta}\Sigma(\Theta)^{1/2}dB, \tag{6}$$

where $\Theta(t = k\eta) \approx \theta_k$ for the step k parameter θ_k, η is the learning rate, and $\Sigma(\Theta)$ is the covariance of the stochastic gradient at location Θ. In the SDE theory, the drift term $\nabla V(\cdot)$ is usually assumed to be Lipschitz. However, in machine learning (for example neural network training with non-smooth activation functions), it is common to encounter not-Lipschitz gradients of loss functions (as

in the example presented in Section 3). To fill this gap, we provide in Appendix C a justification of SDE approximation for the drift with jump discontinuities, based on the proof presented in (Müller-Gronbach et al., 2020). The following two lemmas summarize the transition times between the two local minimizers.

Lemma 5. *Assume that there are only two local minimizers θ_1^*, θ_2^* for the objective function $V(\theta)$. Let $\tau_{\theta_1^* \to \theta_2^*}$ be the transition time for $\Theta(t)$ in (6) from the ϵ-neighborhood of θ_1^* (a closed ball of radius ϵ centered at θ_1^*) to the ϵ-neighborhood of θ_2^* and $\tau_{\theta_2^* \to \theta_1^*}$ be the transition time in the opposite direction. Then*

$$\frac{\mathbb{E}[\tau_{\theta_1^* \to \theta_2^*}]}{\mathbb{E}[\tau_{\theta_2^* \to \theta_1^*}]} = \sqrt{\frac{\det(\nabla^2 L(\theta_2^*))}{\det(\nabla^2 L(\theta_1^*))}} \exp\left(\frac{2}{\eta}\left(\frac{\delta V(\theta_1^*)}{\Sigma(\theta_1^*)} - \frac{\delta V(\theta_2^*)}{\Sigma(\theta_2^*)}\right)\right) + O(\sqrt{\epsilon}).$$

Here $\det(\nabla^2 L(\theta_1^*))$ and $\det(\nabla^2 L(\theta_2^*))$ are the determinants of the Hessians at θ_1^* and θ_2^*, respectively. $\delta V(\theta_k^*) \equiv V(\theta^*) - V(\theta_k^*)$ for $k = 1, 2$ where θ° is the saddle point between θ_1^*, θ_2^*.[1]

This lemma is known in the diffusion process literature as the Eyring-Kramers formula; see, e.g., (Berglund, 2011; Bovier et al., 2004; 2005). Using the above lemma, we obtain the following result for the transition times for resampling and reweighting.

Lemma 6. *Assume that there are only two local minimizers θ_1^*, θ_2^* for the objective function $V(\theta)$. Also assume that the loss function $V_1(\cdot)$ for the first group is $O(\epsilon)$ in the ϵ-neighborhood of θ_2^* and the loss function $V_2(\cdot)$ for the second group is $O(\epsilon)$ in the ϵ-neighborhood of θ_1^*. In addition, assume that the determinants of the Hessian at two local minimizers are the same. Then the ratio of the transition times between the two local minimizers for resampling is*

$$\frac{\mathbb{E}[\tau^s_{\theta_1^* \to \theta_2^*}]}{\mathbb{E}[\tau^s_{\theta_2^* \to \theta_1^*}]} = \exp\left(\frac{2}{\eta}\left(\frac{\delta V(\theta_1^*)}{a_1 \nabla V_1(\theta_1^*) \nabla V_1(\theta_1^*)^\top} - \frac{\delta V(\theta_2^*)}{a_2 \nabla V_2(\theta_2^*) \nabla V_2(\theta_2^*)^\top}\right)\right) + O(\sqrt{\epsilon})$$

and the ratio for reweighting is

$$\frac{\mathbb{E}[\tau^w_{\theta_1^* \to \theta_2^*}]}{\mathbb{E}[\tau^w_{\theta_2^* \to \theta_1^*}]} = \exp\left(\frac{2}{\eta}\left(\frac{f_1 \delta V(\theta_1^*)}{a_1^2 \nabla V_1(\theta_1^*) \nabla V_1(\theta_1^*)^\top} - \frac{f_2 \delta V(\theta_2^*)}{a_2^2 \nabla V_2(\theta_2^*) \nabla V_2(\theta_2^*)^\top}\right)\right) + O(\sqrt{\epsilon}).$$

See Appendix B for the proof. When the ratio is larger than 1, it means that θ_1^* is more stable than θ_2^*. This result shows that for reweighting the relative stability of the two minimizers highly depends on the sampling proportions (f_1, f_2). On the other hand, for resampling it is independent of (f_1, f_2). To see how the sampling proportions affect the behavior of reweighting, let us consider a simple case where θ_1^* is the global minimizer, $\nabla V_1(\theta_1^*) \nabla V_1(\theta_1^*)^\top = \nabla V_2(\theta_2^*) \nabla V_2(\theta_2^*)^\top$, $a_1 = \frac{1}{2} + \epsilon, a_2 = \frac{1}{2} - \epsilon$, and $f_1 \ll f_2$. This ensures that $\delta V(\theta_1^*) > \delta V(\theta_2^*)$ and the above ratio for resampling is larger than 1, which is the desired result. However, $f_1 \ll f_2$ implies that $\frac{f_1}{a_1^2} \ll 1$, $\frac{f_2}{a_2^2} \gg 1$, and the above ratio for reweighting is much smaller than 1, which means that the local minimizer θ_2^* is more stable than the global minimizer θ_1^*.

5 Experiments

This section examines the empirical performance of resampling and reweighting for problems from classification, regression, and reinforcement learning.

Classification. This experiment uses the Porto Seguros safe driver prediction data set[2] to predict the probability of a driver to file a claim. One issue of the provided training data set after preprocessing is that the data distribution under the column "target", which indicates whether a claim is filed, is already highly skewed: the ratio of unfiled claims and filed claims is $a_1/a_2 = 26.4/1$. Assuming this as the underlying population distribution since filing a claim can be considered as a rare event,

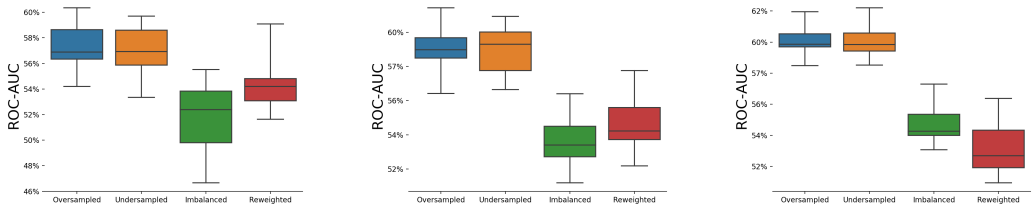
[1] The formal definition of θ°: Let $\theta(t)$ be a path with $\theta(0) = \theta_1^*, \theta(1) = \theta_2^*$, then $\hat{\theta}(t) = \arg\min_{\theta(t)} \sup_{t \in [0,1]} V(\theta(t))$ is the path with minimal saddle point height among all continuous paths. $\theta^\circ = \sup_{t \in (0,1)} \hat{\theta}(t)$ is the saddle point of this path.

[2] `https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data`

we create biased data sets with different sampling proportions f_1/f_2 (see Figure 3) by extracting samples uniformly randomly.

The code from the imbalanced-learn package[3] is adapted to compare the performance between resampling (including both oversampling and undersampling methods) and reweighting. The neural network used to train the data is implemented in Keras with 4 layers. As we discussed in Section 2, the resampling methods adjust sample size to make the population proportions close to a_1/a_2, and reweighting multiplies each sample by a_i/f_i along with a normalizing factor.

To estimate the performance, rather than using the classification accuracy that can be misleading for biased data, we use the metric that computes the area under the receiver operating characteristic curve (ROC-AUC) from the prediction scores. The ROC curves plots the true positive rate on the y-axis versus the false positive rate on the x-axis. As a result, a larger area under the curve indicates a better performance of a classifier. As shown in Figure 3, resampling outperforms reweighting in terms of the ROC-AUC scores for different sampling proportions f_1/f_2.



(1) $f_1/f_2 = 1/5.8$ (2) $f_1/f_2 = 1/2.9$ (3) $f_1/f_2 = 1.8/1$

Figure 3: Box plots for ROC-AUC scores among random oversampling (ROS), random undersampling (RUS), reweighting, and the benchmark training without balancing.

Nonlinear regression. Using the Labeled Faces in the Wild (LFW) dataset, we compare the performance of resampling and reweighting for the facial recognition problem. Two celebrities are picked from the LFW dataset: the majority group among selected samples consists of 530 photos of President George W. Bush, while the minority group contains 29 photos of President Bill Clinton. The sampling proportions are $f_1/f_2 = 18.3/1$ and the underlying population proportions are $a_1/a_2 = 1/1$.

We again adapt the code from the imbalanced-learn package and experiment on several oversampling methods (ADASYN, ROS, SMOTE) and undersampling methods (RUS, NearMiss), as well as reweighting, by conducting SGDClassifier using logistic regression. Figure 4 plots the ROC curves for all these methods mentioned above. We emphasize that this is not a claim that SGDClassifier gives the best prediction score. Instead, our motivation is to illustrate that the performance of SGD deteriorates when reweighting is used.

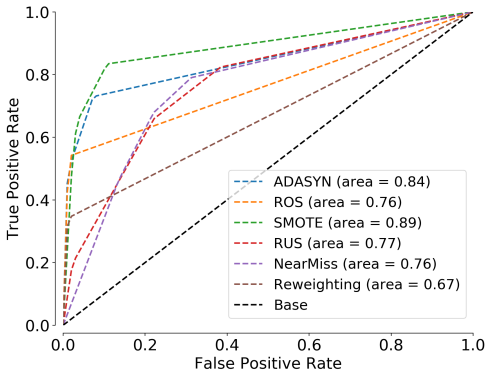


Figure 4: The ROC curve comparisons among various resampling methods and reweighting. We see that reweighting outputs the poorest AUC score (0.67) under SGDClassifier.

Off-policy prediction. In the off-policy prediction problem in reinforcement learning, the objective is to find the value function of police π using the trajectory $\{(a_t, s_t, s_{t+1})\}_{t=1}^T$ generated by a behavior policy μ. To achieve this, the standard approach is to update the value function based on

[3] https://imbalanced-learn.org/stable/index.html

the behavior policy's temporal difference (TD) error $\delta_\mu(s_t) = R(s_t) + \gamma V(s_{t+1}) - V(s_t)$ with an importance weight $\mathbb{E}[\delta_\pi|s_t = s] = \sum_{a\in\mathbb{A}} \frac{\pi(a|s)}{\mu(a|s)} \mathbb{E}[\delta_\pi|s_t = s, a_t = a]\mu(a|s)$, where the summation is taken over the action space $\mathbb{A}$. The resulting reweighting TD learning for policy π is

$$V_{t+1}(s_t) = V_t(s_t) - \eta \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} (R(s_t) + \gamma V_t(s_{t+1}) - V_t(s_t)),$$

where η is the learning rate. This update rule is an example of reweighting. On the other hand, the expected TD error can also be written in the resampling form, $\mathbb{E}[\delta_\pi|s_t = s] = \sum_{a\in\mathbb{A}} \mathbb{E}[\delta_\pi|s_t = s, a_t = a]\pi(a|s) = \sum_{a\in\mathbb{A}} \sum_{j=1}^{\pi(a|s)N} \mathbb{E}[\delta_\pi^j|s_t = s, a_t = a]$, where N is the total number of samples for $s_t = s$. This results to a resampling TD learning algorithm: at step t,

$$V_{t+1}(s_t) = V_t(s_t) - \eta(R(s_k) + \gamma V_t(s_{k+1}) - V_t(s_k)),$$

where (a_k, s_k, s_{k+1}) is randomly chosen from the data set $\{(a_j, s_j, s_{j+1})\}_{s_j=s_t}$ with probability $\pi(a_k|s_t)$.

Consider a simple example with discrete state space $\mathbb{S} = \{i\}_{i=0}^{n-1}$, action space $\mathbb{A} = \{\pm 1\}$, and transition dynamics $s_{t+1} = \text{mod}(s_t + a_t, n)$, where the operator $\text{mod}\ (m, n)$ gives the remainder of m divided by n. Figure 5 shows the results of the off-policy TD learning by these two approaches, with the choice of $n = 32$ and $r(s) = 1 + \sin(2\pi s/n)$. The target policy is $\pi(a_i|s) = \frac{1}{2}$ while the behavior policy is $\mu(a_i|s) = \frac{1}{2} + ca_i$. The difference between the two policies becomes larger as the constant $c \in [0, 1/2]$ increases. From the previous analysis, if one group has much fewer samples as it should have, then the minimizer of the reweighting method is highly affected by the sampling bias. This is verified in the plots: as c becomes larger, the performance of reweighting deteriorates, while resampling is rather stable and almost experiences no difference with the on-policy prediction in this example.

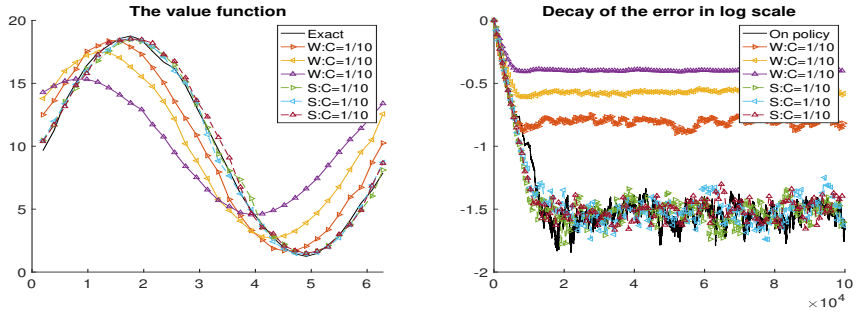


Figure 5: The left plot shows the approximate value function obtained by the two methods. The right plot is the evolution of the relative error $\log(\frac{e_t}{e_0})$, where the absolute error $e_t = \|V_t(s) - V^*(s)\|_2^2$. The value function obtained from resampling is fairly close to the exact value function, while the results of reweighting gets worse as the behavior policy gets further from the target policy.

6 Discussions

This paper examines the different behaviors of reweighting and resampling for training on biasedly sampled data with the stochastic gradient descent. From both the dynamical stability and stochastic asymptotics viewpoints, we explain why resampling is numerically more stable and robust than reweighting. Based on this theoretical understanding, we advocate for considering data, model, and optimization as an integrated system, while addressing the bias.

An immediate direction for future work is to apply the analysis to more sophisticated stochastic training algorithms and understand their impact on resampling and reweighting. Another direction is to extend our analysis to unsupervised learning problems.

REFERENCES

Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 289–295, 2019.

Nils Berglund. Kramers' law: Validity, derivations and generalisations. *Markov Processes Relat. Fields*, 19(3):459–490, 2011.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pp. 4349–4357, 2016.

Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes i. sharp asymptotics for capcities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004.

Anton Bovier, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes ii: Precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7 (1):69–99, 2005.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3992–4001, 2017.

Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *Advances in Neural Information Processing Systems*, pp. 1002–1012, 2017.

Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2011.

Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Haibo He and Yunqian Ma. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3819–3828, 2015.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in neural information processing systems*, pp. 1189–1197, 2010.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pp. 2101–2110, 2017.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *J. Mach. Learn. Res.*, 20:40–1, 2019.

Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences*, 250:113–141, 2013.

Tomasz Maciejewski and Jerzy Stefanowski. Local neighbourhood extension of smote for mining imbalanced data. In *2011 IEEE symposium on computational intelligence and data mining (CIDM)*, pp. 104–111. IEEE, 2011.

Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Ensemble of exemplar-svms for object detection and beyond. In *2011 International conference on computer vision*, pp. 89–96. IEEE, 2011.

Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126, 2003.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2437–2445, 2020.

Thomas Müller-Gronbach, Larisa Yaroslavtseva, et al. On the performance of the euler–maruyama scheme for sdes with discontinuous drift coefficient. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 56, pp. 1162–1178. Institut Henri Poincaré, 2020.

Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in neural information processing systems*, pp. 7146–7155, 2018.

Matthew Schlegel, Wesley Chung, Daniel Graves, Jian Qian, and Martha White. Importance resampling for off-policy prediction. In *Advances in Neural Information Processing Systems*, pp. 1799–1809, 2019.

Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Resampling or reweighting: A comparison of boosting implementations. In *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, volume 1, pp. 445–451. IEEE, 2008.

Bin Shi, Simon S Du, Weijie Su, and Michael I Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *Advances in Neural Information Processing Systems*, pp. 5744–5752, 2019.

Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10–29, 2013.

Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, pp. 8279–8288, 2018.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. Conditional learning of fair representations. In *International Conference on Learning Representations*, 2019.

A Proofs in section 3

A.1 Proof of Lemma 1

Proof. In resampling, near $\theta = -1$ the gradient is $\theta + 1$ with probability a_1 and 0 with probability a_2. Let us denote the random gradient at each step by $W(\theta + 1)$, where W is a Bernoulli random variable with mean $\mathbb{E}(W) = a_1$ and variance $\mathbb{V}(W) = a_1 a_2$. At the learning rate η, the iteration can be written as

$$(\theta_{k+1} + 1) = (1 - \eta W)(\theta_k + 1).$$

The first and second moments of the iterates are

$$\begin{aligned} \mathbb{E}[\theta_k + 1] &= (1 - \eta a_1)^k (\theta_0 + 1), \\ \mathbb{E}[(\theta_k + 1)^2] &= ((1 - \eta a_1)^2 + \eta^2 a_1 a_2)^k (\theta_0 + 1)^2. \end{aligned} \tag{7}$$

According to the definition of the stochastic stability, SGD is stable around $\theta = -1$ if the multiplicative factor of the second equation is bounded by 1, i.e.

$$(1 - \eta a_1)^2 + \eta^2 a_1 a_2 \leq 1. \tag{8}$$

Consider now the stability around $\theta = 1$, the iteration can be written as

$$(\theta_{k+1} - 1) = (1 - \eta W)(\theta_k - 1),$$

where W is again a Bernoulli random variable with $\mathbb{E}(W) = a_2$ and $\mathbb{V}(W) = a_1 a_2$. The same computation shows that the second moment follows

$$\mathbb{E}[(\theta_k - 1)^2] = ((1 - \eta a_2)^2 + \eta^2 a_1 a_2)^k (\theta_0 - 1)^2.$$

Therefore, the condition for the SGD to be stable around $\theta = 1$ is

$$(1 - \eta a_2)^2 + \eta^2 a_1 a_2 \leq 1. \tag{9}$$

□

A.2 Proof of Lemma 2

Proof. In reweighting, near $\theta = -1$ the gradient is $\frac{a_1}{f_1}(\theta + 1)$ with probability f_1 and 0 with probability f_2. Let us denote the random gradient at each step by $W(\theta + 1)$, where W is a Bernoulli random variable with $\mathbb{E}(W) = a_1$ and $\mathbb{V}(W) = f_1 f_2 \left(\frac{a_1}{f_1}\right)^2$. At the learning rate η, the iteration can be written as

$$(\theta_{k+1} + 1) \leftarrow (1 - \eta W)(\theta_k + 1).$$

Hence the second moments of the iterates are given by

$$\mathbb{E}[(\theta_k + 1)^2] = ((1 - \eta a_1)^2 + \eta^2 f_1 f_2 (a_1/f_1)^2)^k (\theta_0 + 1)^2.$$

Therefore, the condition for the SGD to be stable around $\theta = -1$ is

$$(1 - \eta a_1)^2 + \eta^2 f_1 f_2 \left(\frac{a_1}{f_1}\right)^2 \leq 1.$$

Consider now the stability around $\theta = 1$, the gradient is 0 with probability f_1 and $\frac{a_2}{f_2}(\theta - 1)$ with probability f_2. An analysis similar to the case $\theta = -1$ shows that the condition for the SGD to be stable around $\theta = 1$ is

$$(1 - \eta a_2)^2 + \eta^2 f_1 f_2 \left(\frac{a_2}{f_2}\right)^2 \leq 1.$$

□

B Proofs in section 4

B.1 Proof of Lemma 3

Proof. In resampling, with probability a_1 the gradients over the four intervals $(-\infty, -1)$, $(-1, 0)$, $(0, 1)$, and $(1, \infty)$ are -1, 1, ϵ, and ϵ. With probability a_2, they are $-\epsilon$, $-\epsilon$, -1, and 1 across these four intervals. The variances of the gradients are $a_1 a_2 (1 - \epsilon)^2$, $a_1 a_2 (1 + \epsilon)^2$, $a_1 a_2 (1 + \epsilon)^2$, $a_1 a_2 (1 - \epsilon)^2$, respectively, across the same intervals.

Since $\epsilon \ll 1$, the variance can be written as $a_1 a_2 + O(\epsilon)$ across all intervals. Then the SGD dynamics with learning rate η can be approximated by

$$\theta_{k+1} \leftarrow \theta_k - \eta \left(V'(\theta_k) + \sqrt{a_1 a_2 + O(\epsilon)} Z \right),$$

where $Z \sim \mathcal{N}(0, 1)$ is a normal random variable. When η is small, one can approximate the dynamics by a stochastic differential equation of form

$$d\Theta = -V'(\Theta) dt + \sqrt{\eta} \sqrt{a_1 a_2 + O(\epsilon)} dB$$

by identifying $\theta_k \approx \Theta(t = k\eta)$ (see Appendix C for details). The stationary distribution of this stochastic process is

$$p_s(\theta) = \frac{1}{Z} \exp\left(-\frac{2}{(a_1 a_2 + O(\epsilon))\eta} V(\theta)\right).$$

Plugging in $\theta = -1, 1$ results in

$$\begin{aligned}\frac{p_s(1)}{p_s(-1)} &= \exp\left(-\frac{2}{(a_1 a_2 + O(\epsilon))\eta}\left(V(1) - V(-1)\right)\right) = \exp\left(-\frac{2}{a_1 a_2 \eta}(V(1) - V(-1)) + O(\epsilon)\right) \\ &= \exp\left(-\frac{2}{a_1 a_2 \eta}(V(1) - V(-1))\right) + O(\epsilon).\end{aligned}$$

Under the assumption that $\epsilon \ll 1$, the last term is negligible. When $a_2 > a_1$, $V(\theta)$ is minimized at $\theta = 1$, which implies $-(V(1) - V(-1)) > 0$. Hence, this ratio is larger than 1. □

B.2 Proof of Lemma 4

Proof. In reweighting, with probability f_1 the gradients are $-\frac{a_1}{f_1}$, $\frac{a_1}{f_1}$, $\frac{a_1}{f_1}\epsilon$, and $\frac{a_1}{f_1}\epsilon$ over the four intervals $(-\infty, -1)$, $(-1, 0)$, $(0, 1)$, and $(1, \infty)$, respectively. With probability f_2, they are $-\frac{a_2}{f_2}\epsilon$, $-\frac{a_2}{f_2}\epsilon$, $-\frac{a_2}{f_2}$, and $\frac{a_2}{f_2}$. The variances of the gradients are $f_1 f_2(\frac{a_1}{f_1} - \frac{a_2}{f_2}\epsilon)^2$, $f_1 f_2(\frac{a_1}{f_1} + \frac{a_2}{f_2}\epsilon)^2$, $f_1 f_2(\frac{a_1}{f_1}\epsilon + \frac{a_2}{f_2})^2$, and $f_1 f_2(\frac{a_1}{f_1}\epsilon - \frac{a_2}{f_2})^2$, respectively, across the same intervals.

Since $\epsilon \ll 1$, the variance can be written as $f_1 f_2 \frac{a_1^2}{f_1^2} + O(\epsilon)$ for $\theta < 0$ and $f_1 f_2 \frac{a_2^2}{f_2^2} + O(\epsilon)$ for $\theta > 0$.

With $\theta_k \approx \Theta(k\eta)$, the approximate SDE for $\theta < 0$ is given by

$$d\Theta = -V'(\Theta)dt + \sqrt{\eta}\sqrt{f_1 f_2 \frac{a_1^2}{f_1^2} + O(\epsilon)}dB$$

while the one for $\theta > 0$ is

$$d\Theta = -V'(\Theta)dt + \sqrt{\eta}\sqrt{f_1 f_2 \frac{a_2^2}{f_2^2} + O(\epsilon)}dB$$

(see Appendix C for the SDE derivations). The stationary distributions for $\theta < 0$ and $\theta > 0$ are, respectively,

$$\frac{1}{Z_1} \exp\left(-\frac{2}{\left(f_1 f_2 \frac{a_1^2}{f_1^2} + O(\epsilon)\right)\eta} V(\theta)\right), \quad \frac{1}{Z_2} \exp\left(-\frac{2}{\left(f_1 f_2 \frac{a_2^2}{f_2^2} + O(\epsilon)\right)\eta} V(\theta)\right).$$

Plugging in $\theta = -1, 1$ results in

$$\begin{aligned}\frac{p_w(1)}{p_w(-1)} &= \frac{Z_1}{Z_2} \exp\left(-\frac{2}{\left(f_1 f_2 \frac{a_2^2}{f_2^2} + O(\epsilon)\right)\eta} V(1) + \frac{2}{\left(f_1 f_2 \frac{a_1^2}{f_1^2} + O(\epsilon)\right)\eta} V(-1)\right) \\ &= \frac{Z_1}{Z_2} \exp\left(-\frac{2 f_2/f_1}{a_2^2 \eta} V(1) + \frac{2 f_1/f_2}{a_1^2 \eta} V(-1) + O(\epsilon)\right).\end{aligned} \tag{10}$$

The next step is to figure out the relationship between Z_1 and Z_2. Consider an SDE with non-smooth diffusion $d\Theta = -V'(\Theta)dt + \sigma dB$. The Kolmogorov equation for the stationary distribution is

$$0 = p_t = \left(V'(\theta)p + \left(\frac{\sigma^2}{2}p\right)_\theta\right)_\theta. \tag{11}$$

This suggests that $\sigma^2 p$ is continuous at the discontinuity $\theta = 0$. In our setting, since $V(0) = 0$, this simplifies to

$$\left(f_1 f_2 \frac{a_1^2}{f_1^2} + O(\epsilon)\right)\eta \cdot \frac{1}{Z_1} = \left(f_1 f_2 \frac{a_2^2}{f_2^2} + O(\epsilon)\right)\eta \cdot \frac{1}{Z_2}.$$

This simplifies to

$$\frac{Z_1}{Z_2} = \frac{f_1 f_2 \frac{a_1^2}{f_1^2} + O(\epsilon)}{f_1 f_2 \frac{a_2^2}{f_2^2} + O(\epsilon)} = \frac{a_1^2/f_1^2}{a_2^2/f_2^2} + O(\epsilon).$$

Inserting this into (10) results in

$$\begin{aligned}\frac{p_w(1)}{p_w(-1)} =& \left(\frac{a_1^2/f_1^2}{a_2^2/f_2^2} + O(\epsilon)\right) \exp\left(-\frac{2f_2/f_1}{a_2^2\eta} V(1) + \frac{2f_1/f_2}{a_1^2\eta} V(-1) + O(\epsilon)\right) \\ =& \frac{a_1^2/f_1^2}{a_2^2/f_2^2} \exp\left(-\frac{2f_2/f_1}{a_2^2\eta} V(1) + \frac{2f_1/f_2}{a_1^2\eta} V(-1)\right) + O(\epsilon).\end{aligned}$$

By the assumption $\frac{f_2}{f_1} \leq \frac{a_2}{a_1}\sqrt{\frac{V(-1)}{V(1)}}$ and $V(1) < V(-1) < 0$, one has $\left(\frac{a_1}{a_2}\right)^2 \left(\frac{f_2}{f_1}\right)^2 \leq \frac{V(-1)}{V(1)} < 1$ and $-\frac{f_2/f_1}{a_2^2} V(1) \leq -\frac{f_1/f_2}{a_1^2} V(-1)$. Hence the above ratio is less than 1. □

B.3 Proof of Lemma 6

Proof. The variances of gradients for resampling and reweighting are respectively,

$$\begin{aligned}\mathbb{V}\left[\nabla \hat{V}_s(\theta)\right] &= a_1 \nabla V_1(\theta) \nabla V_1(\theta)^T + a_2 \nabla V_2(\theta) \nabla V_2(\theta)^T - (\mathbb{E}[\nabla \hat{V}_s(\theta)])^2, \\ \mathbb{V}\left[\nabla \hat{V}_w(\theta)\right] &= \frac{a_1^2}{f_1} \nabla V_1(\theta) \nabla V_1(\theta)^T + \frac{a_2^2}{f_2} \nabla V_2(\theta) \nabla V_2(\theta)^T - (\mathbb{E}[\nabla \hat{V}_w(\theta)])^2.\end{aligned}$$

Given that at the stationary point $\mathbb{E}[\nabla V(\theta_1^*)] = \mathbb{E}[\nabla V(\theta_2^*)] = 0$, one can omit the last term in the variance. In addition, since $\nabla V_1(\theta_2^*), \nabla V_2(\theta_1^*) = O(\epsilon) \ll \nabla V_1(\theta_1^*), \nabla V_2(\theta_2^*)$ by assumption, all the higher order terms are included in an $O(\sqrt{\epsilon})$ term. One can then derive Lemma 6 from Lemma 5. □

C A Justification of the SDE approximation

The stochastic differential equation approximation of SGD involving data-dependent covariance coefficient Gaussian noise was first introduced in (Li et al., 2017) and justified in the weak sense. Consider the SDE

$$d\Theta = b(\Theta)dt + \sigma(\Theta)dB. \tag{12}$$

The Euler-Maruyama discretization with time step η results in

$$\Theta_{k+1} = \Theta_k + \eta b(\Theta_k) + \sqrt{\eta}\sigma(\Theta_k) Z_k, \;\; Z_k \sim \mathcal{N}(0,1), \;\; \Theta_0 = \theta_0. \tag{13}$$

In our case, $b(\cdot) = -V'(\cdot)$. When b satisfies Lipschitz continuity and some technical smoothness conditions, according to (Li et al., 2017) for any function g from a smooth class $\mathcal{M}$, there exists $C > 0$ and $\alpha > 0$ such that for all $k = 0, 1, 2, \cdots, N$,

$$|E[g(\Theta_{k\eta})] - E[g(\theta_k)]| \leq C\eta^\alpha.$$

However, as the loss function considered in this paper has jump discontinuous in the first derivative, the classical approximation error results for SDE do not apply. In fact, the problem $V \notin C^1(\mathbb{R}^n)$ is a common issue in machine learning and deep neural networks, as many loss functions involves non-smooth activation functions such as ReLU and leaky ReLU. In our case, we need to justify the SDE approximation adopted in Section 3. It turns out that strong approximation error can be obtained if

- the noise coefficient σ is Lipschitz continuous and non-degenerate, and
- the drift coefficient b is piece-wise Lipschitz continuous, in the sense that b has finitely many discontinuity points $-\infty = \xi_0 < \xi_1 < \cdots < \xi_m < \xi_{m+1} = \infty$ and in each interval (ξ_{i-1}, ξ_i), b is Lipschitz continuous.

Under these conditions, the following approximation result holds: for all $k = 0, 1, 2, \cdots, N$, there exists $C > 0$ such that

$$E[|\Theta_{k\eta} - \theta_k|] \leq C\sqrt{\eta}. \tag{14}$$

Here $\Theta_{k\eta}$ is the solution to SDE at time $k\eta$. The proof strategy closely follows from (Müller-Gronbach et al., 2020). The key is to construct a bijective mapping $G : \mathbb{R} \to \mathbb{R}$ that transforms (12) to SDE with Lipschitz continuous coefficients. With such a bijection G, one can define a stochastic process $Z : [0, T] \times \Omega \to \mathbb{R}$ by $Z_t = G(\Theta_t)$ and the transformed SDE is

$$dZ_t = \tilde{b}(Z_t)dt + \tilde{\sigma}dB_t, \;\; t \in [0, T], \;\; Z_0 = G(\Theta_0), \tag{15}$$

$$\text{with } \tilde{b} = (G' \cdot b + \frac{1}{2}G'' \cdot \sigma^2) \circ G^{-1} \text{ and } \tilde{\sigma} = (G' \cdot \sigma) \circ G^{-1}. \tag{16}$$

As the SGD updates can essentially be viewed as data from the Euler-Maruyama scheme, considering Z_k as updates from Euler-Maruyama scheme leads to

$$\begin{aligned}\mathbb{E}[|\Theta_{k\eta} - \theta_k|] &\leq c_1\mathbb{E}[|Z_{k\eta} - G \circ \theta_k|] = c_1\mathbb{E}[|Z_{k\eta} - Z_k + Z_k - G \circ \theta_k|] \\ &\leq c_2\sqrt{\eta} + c_1\mathbb{E}[|Z_k - G \circ \theta_k|].\end{aligned}$$

To control the second item, we introduce

$$\theta_t := \theta_k + b(\theta_k)(t - k\eta) + \sqrt{t - k\eta}\sigma(\theta_k)Z_k,$$

where $t \in [0, k\eta]$. Then as shown in (Müller-Gronbach et al., 2020),

$$\mathbb{E}[|Z_k - G \circ \theta_k|] \leq c\sqrt{\eta} + c\mathbb{E}\left[\left|\int_0^{k\eta} 1_B(\theta_t, \theta_k)dt\right|\right],$$

with B being the set of pairs $(y_1, y_2) \in \mathbb{R}^2$ where the joint Lipschitz estimate $|b(y_1) - b(y_2)|$ does not apply due to at least one discontinuity. In (Müller-Gronbach et al., 2020), it is estimated by

$$\mathbb{E}\left[\left|\int_0^{k\eta} 1_B(\theta_t, \theta_k)dt\right|\right] \leq c\sqrt{\eta},$$

which leads us to (14).