

COMBINING RESAMPLING AND REWEIGHTING FOR FAITHFUL STOCHASTIC OPTIMIZATION*

JING AN[†] AND LEXING YING[‡]

Abstract. Many machine learning and data science tasks require solving non-convex optimization problems. When the loss function is a sum of multiple terms, a popular method is the stochastic gradient descent. Viewed as a process for sampling the loss function landscape, the stochastic gradient descent is known to prefer flat minima. Though this is desired for certain optimization problems such as in deep learning, it causes issues when the goal is to find the global minimum, especially if the global minimum resides in a sharp valley.

Illustrated with a simple motivating example, we show that the fundamental reason is that the difference in the Lipschitz constants of multiple terms in the loss function causes stochastic gradient descent to experience different gradient variances at different minima. In order to mitigate this effect and perform faithful optimization, we propose a combined resampling-reweighting scheme to balance the variance at local minima and extend to general loss functions. We explain from the numerical stability perspective how the proposed scheme is more likely to select the true global minimum, and from the local convergence analysis perspective how it converges to a minimum faster when compared with the vanilla stochastic gradient descent. Experiments from robust statistics and computational chemistry are provided to demonstrate the theoretical findings.

Keywords. Resampling; reweighting; stochastic asymptotics; non-convex optimization; stability.

AMS subject classifications. 82M60; 93E15.

1. Introduction

This paper is concerned with optimizing a non-convex smooth loss function. Identifying a global minimum is known to be computationally hard, especially in the high-dimensional setting. One possible approach, originating from early work in statistical mechanics and Monte Carlo methods, is to turn this into the task of sampling approximately the Gibbs distribution associated with the loss function at a sufficiently low temperature. The rationale is that the samples from the Gibbs distribution have a good chance of being near the global minimum.

In modern machine learning, the loss function often takes the form of an empirical sum of individual terms from finitely many sampled data points. Due to the large size of the dataset, efficient optimization methods such as the stochastic gradient descent (SGD) are commonly used. For non-convex loss function, an increasingly more popular viewpoint is to consider SGD as a sampling algorithm.

One important feature of stochastic gradient-type algorithms is that the noise drives SGD to escape from sharp minima quickly and hence SGD prefers flat minima [8, 25, 29, 31, 32]. Such a bias towards flat minima leads to better generalization properties for problems such as deep learning. However, when the ultimate goal is to identify the global minimum and the landscape around the global minimum happens to be sharper compared to the non-global local minima, this bias is often not desired as SGD often misses the global minimum in a sharp valley¹.

For many data science and physical science problems, the ultimate goal is to find the global minimum for a non-convex landscape, independent of whether it is sharp or

*Received: December 22, 2021; Accepted (in revised form): November 22, 2022. Communicated by Wotao Yin.

[†]Max-Planck Institute for Mathematics in the Sciences, Inselstr. 22, 04103 Leipzig, Germany (jing.an@mis.mpg.de).

[‡]Department of Mathematics, Stanford University, Stanford, CA 94305, US (lexing@stanford.edu).

¹The measure of sharpness adopted comes from [12]: Given a small box B_l with length l centered

flat. In data sciences, one example is the handling of contaminated data, where a simple approach is to use non-convex loss functions in robust statistics [3, 9, 18, 24]. However, as shown later in a motivating example, if we naively apply the vanilla stochastic gradient over a dataset comprised of two subgroups, where one has much larger features (more sensitive) compared to the other (less sensitive), the resulting optimal parameter will be biased towards the less sensitive subgroup. This scenario is not rare in real applications: For example, when researchers adjust a medicine’s ingredients by evaluating the tested group’s responses, inherently different hormone levels in individuals can affect the faithfulness of the evaluation. In physical sciences, examples of non-convex global minimization include finding the ground state wave function in quantum many body problems [13], geometry optimization of the potential energy surface of a molecule in computational chemistry [14], etc. Though applying vanilla stochastic gradient can reduce computational cost for these large scale problems, one also takes the risk of missing the global minimum.

1.1. Main contributions. Below we summarize the main contributions of this paper.

- (1) Starting from a motivating example, we identify the fundamental reason behind the selection bias is that the difference in the Lipschitz constants of multiple terms in the loss function causes stochastic gradient descent to experience different gradient variances at different local minima.
- (2) To mitigate this selection bias, we propose a combined resampling-reweighting strategy for faithful minimum selection. We also derive stochastic differential equation (SDE) models to shed lights on how the proposed strategy balances variances in different regions. This proposed strategy also recovers the importance sampling SGD for faster training (for example, [11, 30]) from a different perspective.
- (3) For a comparison purpose, we provide a stability analysis and a local convergence rate for the importance sampling SGD in non-convex optimization problems. The quantitative numerical results demonstrate how the combined resampling-reweighting strategy improves stability and convergence.
- (4) We show also empirically that the proposed strategy outperforms SGD with examples from robust statistics and computational chemistry.

1.2. Related work. Our proposed combined resampling-reweighting strategy can be viewed as a form of importance sampling. This line of work can be traced back to the randomized Kaczmarz method [26] that selects rows with probability proportional to their squared norms. Later, [22] connects the randomized Kaczmarz method with a SGD algorithm with importance sampling. In convex optimization, many works [4, 22, 30] show that the importance sampling among stochastic gradients can improve the convergence speed. Since importance sampling reduces the stochastic gradient’s variance, this method and its variants can also accelerate the neural networks training [1, 10, 11, 16]. However, it has not been studied yet that how the importance sampling impacts the minimum selection in learning non-convex problems.

around the parameter θ , the sharpness of loss $L(\theta)$ is defined as

$$\text{sharpness} := \frac{\max_{\theta' \in B_t} L(\theta') - L(\theta)}{1 + L(\theta)} \times 100.$$

Taking Taylor expansions, one can show that the sharpness is proportional to the local gradient $\|\nabla L(\theta)\|_2$.

Our approach to understanding the dynamics of the combined resampling-reweighting strategy is based on the numerical analysis for stochastic systems. We mention that the dynamical stability perspective has been used in, for example, [17, 28]. Studying convergence rates for stochastic gradient algorithms for non-convex loss functions is also rapidly growing in recent years [7, 20, 27]. On the other hand, taking the continuous-time limit and using SDEs to analyze stochastic algorithms have become popular especially for stochastic non-convex problems. Using the developed stochastic analysis can lead to numerous new insights of the non-convex optimization [5, 6, 15, 19]. Here we take the SDE approximation approach as it gives us a clearer picture of the global minimum selection.

2. Main idea from a motivating example

Given a dataset consisting of n samples $\{x_i\}_{i=1}^n$, we consider an optimization problem

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^n V(x_i, \theta).$$

The samples are assumed to come from m different subgroups, each representing a proportion $a_j \in (0, 1)$ of the overall population, i.e., $\sum_{j=1}^m a_j = 1$. Assuming for simplicity that the loss term $V(x_i, \theta)$ only depends on the subgroup index of x_i , i.e., $V(x_i, \theta) = V_j(\theta)$ if x_i is from subgroup j , the optimization problem can be simplified as

$$\theta^* = \operatorname{argmin}_{\theta} V(\theta), \quad V(\theta) \equiv \sum_{j=1}^m a_j V_j(\theta), \tag{2.1}$$

in the large n limit. If the terms $V_j(\theta)$ are non-convex loss functions, the overall loss function $V(\theta)$ is in general non-convex as well. From the next example, we will show that applying the vanilla SGD to solve (2.1) becomes problematic when V_j 's exhibit drastically different Lipschitz constants.

2.1. An illustrative example. Consider the case of two subgroups with the following loss functions,

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

with $\epsilon \in (0, 1)$ small, $K > 1$, and $a_2 > a_1$. The total loss function is $V(\theta) = a_1 V_1(\theta) + a_2 V_2(\theta)$ with local minima $\theta = -1$ and $\theta = 1/K$. By construction, the loss function has a sharp global minimum at $\theta = 1/K$ and a flat local minimum at $\theta = -1$ (see for example Figure 2.1 (1)).²

The following lemma states that the optimization trajectory of the vanilla SGD is biased towards one of the local minima, in the small learning rate η limit.

LEMMA 2.1. *When η is sufficiently small, the equilibrium distribution of the vanilla SGD is given by*

$$p(\theta) \sim \begin{cases} \exp\left(-\frac{2}{a_1 a_2 \eta} V(\theta)\right) & \text{for } \theta < 0, \\ \frac{1}{K^2} \exp\left(-\frac{2}{K^2 a_1 a_2 \eta} V(\theta)\right) & \text{for } \theta > 0, \end{cases}$$

²It is necessary to have $O(\epsilon)$ terms in the loss function for SGD to work. Without the $O(\epsilon)$ terms, if the SGD starts in $(-\infty, 0)$ it will stay in this region because there is no drift from $V_2(x)$. Similarly, if the SGD starts in $(0, \infty)$, it will stay in this region. That means the result of SGD only depends on the initialization when $O(\epsilon)$ term is missing.

up to a normalizing constant.

The derivation follows from approximating the SGD updates by an SDE with a numerical error of order $O(\sqrt{\eta})$ in the weak sense. We construct this example with a piecewise linear loss function in order to illustrate the point because the approximate SDEs are of Langevin dynamics form with piecewise constant noise coefficients. In particular, based on the convergence of probability distribution to the equilibrium measure (e.g., [23]), when the stochastic dynamics reaches equilibrium, the stationary distribution of the stochastic process is a Gibbs distribution. The detailed computations of Lemma 2.1 are given in the Appendix A.

With the longtime asymptotics derived in Lemma 2.1, we can explain how the dynamics prefers different minima depending on the value of K . This example can be also extended to higher dimensions under piecewise affine constraints, in a similar spirit as Lemma 7 of our previous work [2].

From Lemma 2.1, we can make the following surprising observation: *When $K \gg 1$, even though $\theta = 1/K$ is the global minimum, the SGD trajectory spends most of the time near the non-global local minimum $\theta = -1$.* This is illustrated in Figure 2.1 (2) and the fundamental reason is that the Lipschitz constant of the individual loss term affects the SGD variance at individual local minimum, thus resulting in an undesired equilibrium distribution.

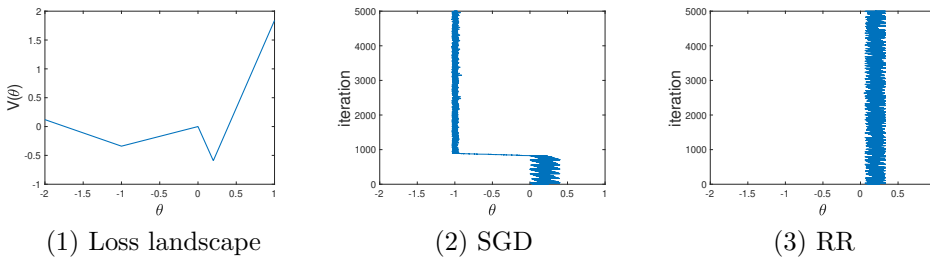


Fig. 2.1: We set $a_1 = 0.4$, $a_2 = 0.6$, $\epsilon = 0.1$, $K = 5$ so that the global minimum is at $\theta = 1/5$. For both the vanilla SGD and the resampling-reweighting (RR) scheme, we start from $\theta_0 = 0.25$ and use a fixed step size $\eta = 0.04$. We can see that the vanilla SGD jumps to the non-global local minimum $\theta = -1$ after several iterations while the RR method stays around the global minimum $\theta = 1/K$ all the time. We include more comparisons with various learning rates in the Appendix C to show that the RR scheme is stable for a wider range of η .

To fix this issue, we propose to *resample* two subgroups with proportion f_1 and f_2 , respectively (with $f_1, f_2 > 0, f_1 + f_2 = 1$). In order to maintain the same overall loss function, we also need to *reweight* each loss term with weights $w_1 := a_1/f_1, w_2 := a_2/f_2 > 0$. The values of f_1, f_2, w_1 , and w_2 are to be determined depending on a_1, a_2 and K . After reweighting and resampling, the loss function can be reformulated equivalently as

$$V(\theta) = f_1 \cdot \left(\frac{a_1}{f_1} V_1(\theta) \right) + f_2 \cdot \left(\frac{a_2}{f_2} V_2(\theta) \right). \quad (2.2)$$

In each iteration, a data point is sampled from the two subgroups following proportion f_1 and f_2 , and then either $\frac{a_1}{f_1} V_1(\theta)$ or $\frac{a_2}{f_2} V_2(\theta)$ is used for computing the stochastic gradient. In what follows, we refer to this approach as the *resampling-reweighting* (RR) scheme.

Although the expectation of the stochastic gradients in (2.2) remains the same by design, the variance experienced in different regions now can be balanced by the

parameters f_1, f_2 . A direct computation shows that in the four regions $(-\infty, -1)$, $(-1, 0)$, $(0, 1/K)$, and $(1/K, \infty)$:

- with probability f_1 , the gradients are $-a_1/f_1, a_1/f_1, \epsilon a_1/f_1, \epsilon a_1/f_1$ respectively;
- with probability f_2 , the gradients are $-\epsilon a_2/f_2, -\epsilon a_2/f_2, -K a_2/f_2, K a_2/f_2$ respectively;
- the variances of the gradients are equal to $\left(a_1\sqrt{\frac{f_2}{f_1}} - \epsilon a_2\sqrt{\frac{f_1}{f_2}}\right)^2, \left(a_1\sqrt{\frac{f_2}{f_1}} + \epsilon a_2\sqrt{\frac{f_1}{f_2}}\right)^2, \left(\epsilon a_1\sqrt{\frac{f_2}{f_1}} + K a_2\sqrt{\frac{f_1}{f_2}}\right)^2, \left(\epsilon a_1\sqrt{\frac{f_2}{f_1}} - K a_2\sqrt{\frac{f_1}{f_2}}\right)^2$, respectively.

The dynamics of RR becomes straightforward if we view it as a numerical approximation of SDEs: with a sufficiently small step size $\eta > 0$, taking $\epsilon \rightarrow 0$, it is given by

$$\begin{aligned} d\Theta_t &= -V'(\Theta_t)dt + a_1\sqrt{f_2/f_1}\sqrt{\eta}dW_t, & \text{in the region } \theta < 0, \\ d\Theta_t &= -V'(\Theta_t)dt + K a_2\sqrt{f_1/f_2}\sqrt{\eta}dW_t, & \text{in the region } \theta > 0. \end{aligned}$$

In order to balance the variance at the two local minima $\theta = -1$ and $\theta = 1/K$, we impose

$$a_1\sqrt{f_2/f_1} = K a_2\sqrt{f_1/f_2} \implies f_1 = \frac{a_1}{a_1 + K a_2}, \quad f_2 = 1 - f_1 = \frac{K a_2}{a_1 + K a_2}.$$

As a result, the assigned weights for the two subgroups are

$$w_1 := \frac{a_1}{f_1} = \frac{a_1 + K a_2}{1}, \quad w_2 := \frac{a_2}{f_2} = \frac{a_1 + K a_2}{K}.$$

The above computation suggests that in order to fix the selection bias, each subgroup should be reweighted by the reciprocal of its Lipschitz constant. We then undersample from subgroup 1 and oversample from subgroup 2 so that their sample size ratio approaches $K a_2/a_1$. In numerical tests, comparing Figure 2.1 (2) and (3), we can see that the RR scheme adjusts the dynamics of the stochastic optimization trajectory to stay around the global minimum.

REMARK 2.1. The motivating example considers different slopes 1 and K , which can reflect the feature magnitude disparities in data science. Suppose a dataset containing n samples $\{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$, and a_1 -proportion of the data features has magnitude $\|x_i\|_2 \sim K$, while the rest has magnitude $\|x_i\|_2 \sim 1$, then an optimization problem of the form

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^n (f(\theta \cdot x_i) - y_i)^2 \tag{2.3}$$

can be thought of as a more complex model of our motivating example.

3. Analysis of the general case

In this section, we consider the general empirical loss for $\theta \in \mathbb{R}^d, d \geq 1$.

$$L(\theta) = \frac{1}{n} \sum_{i=1}^n l_i(\theta). \tag{3.1}$$

3.1. The general RR scheme. Motivated by the illustrative example in Section 2, we propose the following resampling-reweighting scheme in \mathbb{R}^d , $d \geq 1$: at each k -th iteration with current parameter θ_k , for each term $i = 1, 2, \dots, n$,

- (1) reweight the i -th term $l_i(\theta_k)$ with a weight proportional to $1/\|\nabla l_i(\theta_k)\|_2$,
- (2) set the resampling probability for the (reweighted) i -th term to $\|\nabla l_i(\theta_k)\|_2$.

The resampling proportions and weights are decided by comparing the RR scheme with SGD in the continuous-time limit. More specifically, since the proposed RR scheme is designed for variance-balancing, in the limit it should share the same drift with the vanilla SGD but experience more balanced noises across different local minima. We lay out computation details in the next subsection.

3.1.1. RR scheme derivations. Recall for the vanilla SGD, the update rule with time step size $\eta > 0$ is given by

$$\theta_{k+1} = \theta_k - \eta \nabla l_j(\theta_k), \tag{3.2}$$

where the index j is chosen from 1 to n with uniform probability $1/n$. Note that

$$m(\theta_k) = \mathbb{E}_{p_1}[\nabla l_i(\theta_k)] = \frac{1}{n} \sum_{i=1}^n \nabla l_i(\theta_k) = \nabla L(\theta_k), \tag{3.3}$$

and we can rewrite (3.2) as

$$\theta_{k+1} = \theta_k - \eta m(\theta_k) + \sqrt{\eta} V_1(\theta_k), \quad \text{with } V_1(\theta_k) = \sqrt{\eta} (m(\theta_k) - \nabla l_j(\theta_k)).$$

By taking the simplifying assumption that the gradient noise is Gaussian³, the dynamics can be approximated by

$$d\Theta_t = -m(\Theta_t)dt + \sqrt{\eta} \sigma_1(\Theta_t) dW_t, \tag{3.4}$$

where $\Sigma_1(\Theta_t) := \sigma_1(\Theta_t)\sigma_1(\Theta_t)^\top$ is given by $\Sigma_1(\Theta_t) = \frac{1}{n} \sum_{i=1}^n \nabla l_i(\Theta_t)\nabla l_i(\Theta_t)^\top - m(\Theta_t) \otimes m(\Theta_t)$.

On the other hand, as indicated in the illustrative example, the RR scheme with the same time step size $\eta > 0$ should be given by

$$\theta_{k+1} = \theta_k - \eta \frac{C(\theta_k)}{\|\nabla l_j(\theta_k)\|_2} \nabla l_j(\theta_k). \tag{3.5}$$

Here we set $C(\theta_k) = \frac{1}{n} \sum_{i=1}^n \|\nabla l_i(\theta_k)\|_2$, and the index j is chosen from 1 to n with probability $\|\nabla l_i(\theta_k)\|_2/Z(\theta_k)$, with the normalizing factor $Z(\theta_k) = \sum_{i=1}^n \|\nabla l_i(\theta_k)\|_2$, so that the mean for this approach matches with the one for SGD (3.3),

$$\sum_{i=1}^n \frac{C(\theta_k)}{\|\nabla l_i(\theta_k)\|_2} \nabla l_i(\theta_k) \frac{\|\nabla l_i(\theta_k)\|_2}{Z(\theta_k)} = \frac{1}{n} \sum_{i=1}^n \nabla l_i(\theta_k) = m(\theta_k).$$

Now we rewrite (3.5) in the form

$$\theta_{k+1} = \theta_k - \eta m(\theta_k) + \sqrt{\eta} V_2(\theta_k), \quad \text{with } V_2(\theta_k) = \sqrt{\eta} \left(m(\theta_k) - \frac{C(\theta_k)}{\|\nabla l_j(\theta_k)\|_2} \nabla l_j(\theta_k) \right).$$

³We are aware that this approximation might not be valid in many situations. However, this assumption streamlines the SDE analysis.

Thus, the resulting dynamics can be approximated as

$$d\Theta_t = -m(\Theta_t)dt + \sqrt{\eta}\sigma_2(\Theta_t)dW_t, \tag{3.6}$$

where $\Sigma_2(\Theta_t) = \sigma_2(\Theta_t)\sigma_2(\Theta_t)^\top$ is given by $\Sigma_2(\Theta_t) = \frac{Z(\Theta_t)}{n^2} \sum_{i=1}^n \frac{\nabla l_i(\Theta_t)\nabla l_i(\Theta_t)^\top}{\|\nabla l_i(\Theta_t)\|_2} - m(\Theta_t)^{\otimes 2}$.

Note that (3.4) and (3.6) are derived in the same time scale and have the same drift, so that the comparison of those two reduces to comparing covariance matrices Σ_1, Σ_2 .

REMARK 3.1. The RR scheme (3.5) from the derivation turns out to be similar to [30] for faster convergence of regularized convex minimization problems, [11] for training deep learning with importance sampling, and even [26] for solving linear systems of equations earlier on.

In terms of computational complexity, recomputing the weights and sampling proportions in each iteration is obviously expensive. In practice, as discussed in [22], one can use a rejection sampling scheme to select stochastic gradients from a weighted distribution. Algorithm 1 in [11] presents a practical guidance on how to speedup the importance sampling SGD/RR scheme in deep learning, by only updating parameters by (3.5) when the variance of gradients can be reduced.

3.2. Stability analysis. Let us step back to discrete-time stochastic algorithms, and compare SGD and RR scheme from the numerical analysis perspective. In fact, especially for stiff problems, the RR scheme allows a wider range of step-sizes to keep stochastic linear stability compared to SGD. We consider minimizing the empirical loss function (3.1) of the form

$$L(\theta) = \frac{1}{n} \sum_{i=1}^n l_i(\theta) = \frac{1}{2n} \sum_{i=1}^n (f(x_i, \theta) - y_i)^2, \tag{3.7}$$

where f is the learning model and $\{(x_i, y_i)\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$ are i.i.d sampled data points. We assume that for the model f , there is an interpolation solution θ^* such that

$$y_i = f(x_i, \theta^*), \quad \forall 1 \leq i \leq n.$$

By definition, a stationary point $\tilde{\theta}$ is *stochastically stable* if there exists a uniform constant $0 < C \leq 1$ such that $\mathbb{E}[\|\theta_k - \tilde{\theta}\|_2^2] \leq C\|\theta_0 - \tilde{\theta}\|_2^2, k \geq 1$, where θ_k is the k -th iterate of the stochastic algorithm. Suppose $\{\theta_k\}$ are sufficiently close to the interpolation solution θ^* , then SGD iteration can be written as

$$\begin{aligned} \theta_{k+1} &= \theta_k - \eta(f(x_i, \theta_k) - y_i)\nabla_\theta f(x_i, \theta_k) \\ &\approx \theta_k - \eta\nabla_\theta f(x_i, \theta^*)\nabla_\theta f(x_i, \theta^*)^\top(\theta_k - \theta^*), \end{aligned} \tag{3.8}$$

where we take a Taylor expansion approximation $f(x_i, \theta_k) - y_i = f(x_i, \theta_k) - f(x_i, \theta^*) \approx \nabla_\theta f(x_i, \theta^*)^\top(\theta_k - \theta^*)$ and also take $\nabla_\theta f(x_i, \theta_k) \approx \nabla_\theta f(x_i, \theta^*)$ approximately. Let us denote

$$H_i := \nabla_\theta f(x_i, \theta^*)\nabla_\theta f(x_i, \theta^*)^\top \quad \text{and} \quad H := \frac{1}{n} \sum_{i=1}^n H_i, \tag{3.9}$$

then we have the following stability conditions for SGD and RR scheme.

LEMMA 3.1. *Suppose that the starting point θ_0 is close to the interpolation solution θ^* , that is, there exists small $\epsilon > 0$ such that $\|\theta_0 - \theta^*\|_2 \leq \epsilon$, and $\nabla_\theta f(x_i, \theta)$ is bounded and*

continuous around θ^* for all $1 \leq i \leq n$. Then the condition for the SGD to be stochastically stable around θ^* is

$$\lambda_{max} \left\{ (I - \eta H)^2 + \eta^2 \left(\frac{1}{n} \sum_{i=1}^n H_i^2 - H^2 \right) + O(\epsilon) \right\} \leq 1, \tag{3.10}$$

and the condition for the RR scheme to be stochastically stable around θ^* is, for each $k \geq 0$,

$$\lambda_{max} \left\{ (I - \eta H)^2 + \eta^2 \left(\frac{1}{n} \sum_{i=1}^n \tilde{G}_{i,k} H_i^2 - H^2 \right) + O(\epsilon) \right\} \leq 1, \tag{3.11}$$

where $\tilde{G}_{i,k} := \frac{\frac{1}{n} \sum_{j=1}^n \|\nabla_{\theta} f(x_j, \theta^*) \nabla_{\theta} f(x_j, \theta^*)^{\top} (\theta_k - \theta^*)\|_2}{\|\nabla_{\theta} f(x_i, \theta^*) \nabla_{\theta} f(x_i, \theta^*)^{\top} (\theta_k - \theta^*)\|_2}$.

Proof. The first part (3.10) closely follows from [28]. Indeed, the iteration (3.8) can be rewritten more precisely as

$$\theta_{k+1} - \theta^* = (I - \eta H_i + O(\epsilon)) (\theta_k - \theta^*). \tag{3.12}$$

Conditioned on $\{\|\theta_k - \theta^*\|_2 \leq \epsilon\}$, we have the expectation on the second moment as

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}[\|\theta_{k+1} - \theta^*\|_2^2] &= \mathbb{E}_{\mathcal{D}} \left[(\theta_k - \theta^*)^{\top} (I - \eta H_i + O(\epsilon))^2 (\theta_k - \theta^*) \right] \\ &= (\theta_k - \theta^*)^{\top} \left((I - \eta H)^2 + \eta^2 \left(\frac{1}{n} \sum_{i=1}^n H_i^2 - H^2 \right) + O(\epsilon) \right) (\theta_k - \theta^*). \end{aligned} \tag{3.13}$$

The step (3.13) can be iterated down to θ_0 , and the expectation is taken over uniform distribution \mathcal{D} . To ensure that the stochastic stability condition is satisfied, we then need (3.10). On the other hand, the iteration for the RR scheme reads

$$\begin{aligned} \theta_{k+1} &= \theta_k - \eta G(x_i, \theta_k) (f(x_i, \theta_k) - y_i) \nabla_{\theta} f(x_i, \theta_k) \\ &= \theta_k - \eta (\tilde{G}_{i,k} + O(\epsilon)) (\nabla_{\theta} f(x_i, \theta^*) \nabla_{\theta} f(x_i, \theta^*)^{\top} + O(\epsilon)) (\theta_k - \theta^*), \end{aligned} \tag{3.14}$$

where each data sample $x_i \sim \mathcal{D}^{(w)}$ is selected with probability

$$\begin{aligned} p(x_i, \theta_k) &= \frac{\|(f(x_i, \theta_k) - y_i) \nabla_{\theta} f(x_i, \theta_k)\|_2}{\sum_{j=1}^n \|(f(x_j, \theta_k) - y_j) \nabla_{\theta} f(x_j, \theta_k)\|_2} \\ &= \frac{\|\nabla_{\theta} f(x_i, \theta^*) \nabla_{\theta} f(x_i, \theta^*)^{\top} (\theta_k - \theta^*)\|_2}{\sum_{j=1}^n \|\nabla_{\theta} f(x_j, \theta^*) \nabla_{\theta} f(x_j, \theta^*)^{\top} (\theta_k - \theta^*)\|_2} + O(\epsilon) \end{aligned}$$

given $\{\|\theta_k - \theta^*\|_2 \leq \epsilon\}$, and $G(x_i, \theta_k) = C(\theta_k) / \|\nabla_{\theta} f(x_i, \theta_k)\|_2$, which can be rewritten here as

$$\begin{aligned} G(x_i, \theta_k) &= \frac{\frac{1}{n} \sum_{j=1}^n \|(f(x_j, \theta_k) - y_j) \nabla_{\theta} f(x_j, \theta_k)\|_2}{\|(f(x_i, \theta_k) - y_i) \nabla_{\theta} f(x_i, \theta_k)\|_2} \\ &= \frac{\frac{1}{n} \sum_{j=1}^n \|\nabla_{\theta} f(x_j, \theta^*) \nabla_{\theta} f(x_j, \theta^*)^{\top} (\theta_k - \theta^*)\|_2}{\|\nabla_{\theta} f(x_i, \theta^*) \nabla_{\theta} f(x_i, \theta^*)^{\top} (\theta_k - \theta^*)\|_2} + O(\epsilon) = \tilde{G}_{i,k} + O(\epsilon). \end{aligned}$$

Therefore, the RR iteration (3.14) can be rewritten as

$$\theta_{k+1} - \theta^* = (I - \eta \tilde{G}_{i,k} H_i + O(\epsilon)) (\theta_k - \theta^*). \tag{3.15}$$

Conditioned on $\{\|\theta_k - \theta^*\|_2 \leq \epsilon\}$, we compute the expectation on the second moment

$$\begin{aligned} \mathbb{E}_{\mathcal{D}^{(w)}}[\|\theta_{k+1} - \theta^*\|_2^2] &= \mathbb{E}_{\mathcal{D}^{(w)}}\left[(\theta_k - \theta^*)^\top \left(I - \eta \tilde{G}_i H_i + O(\epsilon)\right)^2 (\theta_k - \theta^*)\right] \\ &= (\theta_k - \theta^*)^\top \left((I - \eta H)^2 + \eta^2 \left(\frac{1}{n} \sum_{i=1}^n \tilde{G}_{i,k} H_i^2 - H^2 \right) + O(\epsilon) \right) (\theta_k - \theta^*). \end{aligned} \tag{3.16}$$

Similar to what we analyzed for SGD, the stability condition for RR scheme is deduced to be (3.11). \square

REMARK 3.2. When $d = 1$, all matrices in Lemma 3.1 are scalar. It is straightforward to see that, for the second component in RR scheme, by Cauchy-Schwarz inequality, we get

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \tilde{G}_{i,k} H_i^2 - H^2 &= \frac{1}{n} \sum_{i=1}^n \frac{\frac{1}{n} \sum_{j=1}^n |\nabla_\theta f(x_j, \theta^*)|^2}{|\nabla_\theta f(x_i, \theta^*)|^2} |\nabla_\theta f(x_i, \theta^*)|^4 - H^2 \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n |\nabla_\theta f(x_i, \theta^*)|^2 \right)^2 - H^2 \leq \frac{1}{n} \sum_{i=1}^n |\nabla_\theta f(x_i, \theta^*)|^4 - H^2 = \frac{1}{n} \sum_{i=1}^n H_i^2 - H^2. \end{aligned}$$

When the model f has drastically different gradient magnitude in different data samples, compared to SGD, the RR scheme fundamentally acts as gradient magnitude averaging to allow a broader range of learning rates η for stochastic stability. The RR scheme exhibits its particular strength of maintaining stability for stiff problems.

3.3. Local convergence analysis. In the non-convex scenario, it is relatively easier to obtain convergence results locally. In order to make a direct comparison with SGD in terms of local convergence rates, we adapt Theorem 4 and its proof strategy from [20] to establish convergence analysis of the RR scheme. We use the same setup as in [20], which takes a step-size schedule of the form $\eta_k = \frac{\gamma}{(k+m)^p}$ for some $p \in (1/2, 1]$ with sufficiently large $\gamma, m > 0$. The key ingredient is the following decomposition for stochastic gradients

$$V_{k,j} := \nabla l_j(\theta_k) = \nabla L(\theta_k) + Z_j(\theta_k), \quad j \sim \mathcal{D} \quad \text{for SGD}, \tag{3.17}$$

$$\tilde{V}_{k,j} := G_j(\theta_k) \nabla l_j(\theta_k) = \nabla L(\theta_k) + \tilde{Z}_j(\theta_k), \quad j \sim \mathcal{D}^{(w)} \quad \text{for RR}, \tag{3.18}$$

where \mathcal{D} denotes the uniform distribution over samples, and $\mathcal{D}^{(w)}$ is the weighted distribution used in the RR scheme. $G_j(\theta_k) = \frac{1}{n} \sum_{i=1}^n \|\nabla l_i(\theta_k)\|_2 / \|\nabla l_j(\theta_k)\|_2$ as before. Note that for both cases $\mathbb{E}_{\mathcal{D}}[Z_j(\theta_k)] = \mathbb{E}_{\mathcal{D}^{(w)}}[\tilde{Z}_j(\theta_k)] = 0$, and we further assume that given a neighborhood \mathcal{U} of θ_* , for each $1 \leq j \leq n$, there exists $\sigma_j > 0$ such that

$$\sup_{\theta \in \mathcal{U}} \|\nabla l_j(\theta)\|_2^2 \leq \sigma_j^2. \tag{3.19}$$

THEOREM 3.1. *Suppose in a convex compact neighborhood of θ_* , there exists $\alpha > 0$ so that $\nabla^2 L(\theta) \succcurlyeq \alpha I$. Given the assumptions and the step-size schedule introduced above, for a fixed $\delta \in (0, 1)$, there exist neighborhoods $\mathcal{U}, \mathcal{U}_1$ containing θ_* such that*

$$\mathbb{P}(E_\infty = \{\theta_k \in \mathcal{U} \text{ for all } k \geq 1\} | \theta_1 \in \mathcal{U}_1) \geq 1 - \delta. \tag{3.20}$$

Furthermore, we have the local convergence rate for SGD

$$\mathbb{E}[\|\theta_k - \theta_*\|_2^2 | E_\infty] \leq \frac{\gamma^2}{(1-\delta)(2\alpha\gamma-1)k} \sup_j \sigma_j^2 + o(1/k), \quad \text{if } p=1, \tag{3.21}$$

$$\mathbb{E}[\|\theta_k - \theta_*\|_2^2 | E_\infty] \leq \frac{\gamma}{2\alpha(1-\delta)k^p} \sup_j \sigma_j^2 + o(1/k^p), \quad \text{if } p < 1. \tag{3.22}$$

And for the RR scheme, we have

$$\mathbb{E}[\|\theta_k - \theta_*\|_2^2 | E_\infty] \leq \frac{\gamma^2}{(1-\delta)(2\alpha\gamma-1)k} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i^2 \right) + o(1/k), \quad \text{if } p=1, \tag{3.23}$$

$$\mathbb{E}[\|\theta_k - \theta_*\|_2^2 | E_\infty] \leq \frac{\gamma}{2\alpha(1-\delta)k^p} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i^2 \right) + o(1/k^p), \quad \text{if } p < 1. \tag{3.24}$$

When $p=1$, we choose γ large so that $2\alpha\gamma > 1$.

Proof. First of all, due to the local convexity of $L(\theta)$ near θ_* , we have

$$\langle \theta - \theta_*, \nabla L(\theta) \rangle \geq \alpha \|\theta - \theta_*\|_2^2,$$

for all $\theta \in \mathcal{K}$, where \mathcal{K} is a convex compact neighborhood of θ_* . We have the stochastic gradient updates as $\theta_{k+1} = \theta^k - \eta_k V_{k,j}$. Let $D_k := \|\theta_k - \theta_*\|_2^2$, then

$$\begin{aligned} D_{k+1} &= \|\theta_k - \theta_* - \eta_k V_{k,j}\|_2^2 \\ &= \|\theta_k - \theta_*\|_2^2 - 2\eta_k \langle \theta_k - \theta_*, V_{k,j} \rangle + \eta_k^2 \|V_{k,j}\|_2^2 \\ &= \|\theta_k - \theta_*\|_2^2 - 2\eta_k \langle \theta_k - \theta_*, \nabla L(\theta_k) \rangle - 2\eta_k \langle \theta_k - \theta_*, Z_j(\theta_k) \rangle + \eta_k^2 \|V_{k,j}\|_2^2 \\ &\leq (1 - 2\alpha\eta_k) D_k + 2\eta_k \xi_{k,j} + \eta_k^2 \|V_{k,j}\|_2^2, \end{aligned} \tag{3.25}$$

where $\xi_{k,j} = -\langle \theta_k - \theta_*, Z_j(\theta_k) \rangle$ and $\mathbb{E}[\xi_{k,j} | \mathcal{F}_k] = 0$. Same computations hold for RR scheme as well by replacing notations accordingly. The key idea of the proof is to control the error aggregation in $2\eta_k \xi_{k,j} + \eta_k^2 \|V_{k,j}\|_2^2$. We will only sketch the main steps and highlight where the difference between SGD and RR emerges, since the proof details can be found in [20]. From now on, unless it is necessary, we omit the second subscript in ξ, V for simplicity as they change in each iteration. For the error terms, one can define

$$M_k = 2 \sum_{l=1}^k \eta_l \xi_l \quad \text{and} \quad S_k = \sum_{l=1}^k \eta_l^2 \|V_l\|_2^2. \tag{3.26}$$

Define the cumulative mean square error $R_k = M_k^2 + S_k$, then we have

$$\begin{aligned} R_k &= (M_{k-1} + 2\eta_k \xi_k)^2 + S_{k-1} + \eta_k^2 \|V_k\|_2^2 \\ &= R_{k-1} + 4M_{k-1}\eta_k \xi_k + 4\eta_k^2 \xi_k^2 + \eta_k^2 \|V_k\|_2^2. \end{aligned} \tag{3.27}$$

It is easy to check that R_k is sub-martingale, $\mathbb{E}[R_k | \mathcal{F}_k] \geq R_{k-1}$. The proof uses a finer condition by introducing the following events, let \mathcal{U} be a neighborhood of θ^* and $\epsilon > 0$,

$$E_k = \{\theta_l \in \mathcal{U} \text{ for all } 1 \leq l \leq k\} \quad \text{and} \quad H_k = \{R_l \leq \epsilon \text{ for all } 1 \leq l \leq k\}. \tag{3.28}$$

The property analysis of E_k and H_k is the same as in Lemma D.2 in [20], except for (D.19). Notice that (D.26) can be estimated as

$$\mathbb{E}_{\mathcal{D}}[\|V_{k,j}\|_2^2 \mathbb{1}_{H_{k-1}}] \leq \sup_j \|\nabla l_j(\theta_k)\|_2^2, \quad \text{for SGD}, \tag{3.29}$$

$$\mathbb{E}_{\mathcal{D}^{(w)}}[\|\tilde{V}_{k,j}\|_2^2 \mathbb{1}_{H_{k-1}}] \leq \mathbb{E}_{\mathcal{D}^{(w)}}[G_j(\theta_k)^2 \|\nabla l_j(\theta_k)\|_2^2] = \frac{1}{n^2} \left(\sum_{j=1}^n \|\nabla l_j(\theta_k)\|_2 \right)^2, \quad \text{for RR}, \tag{3.30}$$

and moreover,

$$\mathbb{E}_{\mathcal{D}}[\xi_{k,j}^2 \mathbb{1}_{H_{k-1}}] \leq 2\|\theta_k - \theta_*\|_2^2 (\sup_j \|\nabla l_j(\theta_k)\|_2^2 + \|\nabla L(\theta_k)\|_2^2), \quad \text{for SGD}, \tag{3.31}$$

$$\mathbb{E}_{\mathcal{D}^{(w)}}[\xi_{k,j}^2 \mathbb{1}_{H_{k-1}}] \leq 2\|\theta_k - \theta_*\|_2^2 \left(\frac{1}{n^2} \left(\sum_{j=1}^n \|\nabla l_j(\theta_k)\|_2 \right)^2 + \|\nabla L(\theta_k)\|_2^2 \right), \quad \text{for RR}. \tag{3.32}$$

Putting together with other terms, we have (D.12) modified to be

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}[R_k \mathbb{1}_{H_{k-1}}] &\leq \mathbb{E}_{\mathcal{D}}[R_{k-1} \mathbb{1}_{H_{k-2}}] - \epsilon \mathbb{P}(H_{k-2} \setminus H_{k-1}) \\ &\quad + ((8\|\theta_k - \theta_*\|_2^2 + 1) \sup_j \|\nabla l_j(\theta_k)\|_2^2 + 8\|\theta_k - \theta_*\|_2^2 \|\nabla L(\theta_k)\|_2^2) \eta_k^2 \end{aligned} \tag{3.33}$$

for SGD, and

$$\begin{aligned} \mathbb{E}_{\mathcal{D}^{(w)}}[R_k \mathbb{1}_{H_{k-1}}] &\leq \mathbb{E}_{\mathcal{D}^{(w)}}[R_{k-1} \mathbb{1}_{H_{k-2}}] - \epsilon \mathbb{P}(H_{k-2} \setminus H_{k-1}) \\ &\quad + ((8\|\theta_k - \theta_*\|_2^2 + 1) \frac{1}{n^2} \left(\sum_{j=1}^n \|\nabla l_j(\theta_k)\|_2 \right)^2 + 8\|\theta_k - \theta_*\|_2^2 \|\nabla L(\theta_k)\|_2^2) \eta_k^2 \end{aligned} \tag{3.34}$$

for the RR scheme.

One should control the probability of escaping the neighborhood of θ_* . Denote $r_{\mathcal{U}} = \sup_{\theta \in \mathcal{U}} \|\theta - \theta_*\|_2$, by taking the telescoping sums of (3.33) and (3.34), we get that

$$\mathbb{E}_{\mathcal{D}}[R_k \mathbb{1}_{H_{k-1}}] \leq ((8r_{\mathcal{U}}^2 + 1) \sup_j \sigma_j^2 + 8r_{\mathcal{U}}^2 \sum_{i=1}^n \sigma_i^2) \sum_{l=1}^k \eta_l^2 - \epsilon \sum_{l=1}^k \mathbb{P}(H_{l-2} \setminus H_{l-1}) \tag{3.35}$$

for SGD, and

$$\mathbb{E}_{\mathcal{D}^{(w)}}[R_k \mathbb{1}_{H_{k-1}}] \leq ((8r_{\mathcal{U}}^2 + 1) \frac{1}{n} \sum_{i=1}^n \sigma_i^2 + 8r_{\mathcal{U}}^2 \sum_{i=1}^n \sigma_i^2) \sum_{l=1}^k \eta_l^2 - \epsilon \sum_{l=1}^k \mathbb{P}(H_{l-2} \setminus H_{l-1}) \tag{3.36}$$

for the RR scheme. Since the left-hand sides are non-negative, we get that

$$\sum_{l=1}^k \mathbb{P}(H_{l-1} \setminus H_l) \leq \frac{R_*}{\epsilon} \sum_{l=1}^{\infty} \eta_l^2 = \frac{R_* \gamma^2}{\epsilon} \sum_{l=1}^{\infty} \frac{1}{(l+m)^{2p}}, \tag{3.37}$$

with $R_* = (8r_{\mathcal{U}}^2 + 1) \sup_j \sigma_j^2 + 8r_{\mathcal{U}}^2 \sum_{i=1}^n \sigma_i^2$ for SGD and $R_* = (8r_{\mathcal{U}}^2 + 1) \frac{1}{n} \sum_{i=1}^n \sigma_i^2 + 8r_{\mathcal{U}}^2 \sum_{i=1}^n \sigma_i^2$ for RR scheme, respectively. Because $p > 1/2$, the sum on the right-hand side is finite. For any $\delta \in (0, 1)$, we can choose m sufficiently large, so that $\frac{R_* \gamma^2}{\epsilon} \sum_{l=1}^{\infty} \frac{1}{(l+m)^{2p}} < \delta$. With that, for any k , we get

$$\mathbb{P}(H_k) = 1 - \sum_{l=1}^k \mathbb{P}(H_{l-1} \setminus H_l) \geq 1 - \delta. \tag{3.38}$$

As a consequence, since $E_\infty = \bigcap_{k=1}^\infty E_k$, we will obtain

$$\mathbb{P}(E_\infty) = \inf_k \mathbb{P}(E_k) \geq \inf_k \mathbb{P}(H_{k-1}) \geq 1 - \delta. \tag{3.39}$$

For the convergence rate, from (3.25) we have

$$\mathbb{E}[D_{k+1} \mathbb{1}_{E_{k+1}}] \leq \mathbb{E}[D_{k+1} \mathbb{1}_{E_k}] \leq (1 - 2\alpha\eta_k)\mathbb{E}[D_k \mathbb{1}_{E_k}] + \eta_k^2 \mathbb{E}[\|V_k\|_2^2 \mathbb{1}_{E_k}]. \tag{3.40}$$

The estimate of the second term above is similar to (3.29) and (3.30). Now we insert the expression of η_k and only consider $p=1$ for shortness, since the result can be derived for $p < 1$ verbatim, we get that

$$\mathbb{E}_{\mathcal{D}}[D_{k+1} \mathbb{1}_{E_{k+1}}] \leq \left(1 - \frac{2\alpha\gamma}{k+m}\right)\mathbb{E}_{\mathcal{D}}[D_k \mathbb{1}_{E_k}] + \frac{\gamma^2}{(k+m)^2} \sup_j \sigma_j^2, \quad \text{for SGD}, \tag{3.41}$$

$$\mathbb{E}_{\mathcal{D}^{(w)}}[D_{k+1} \mathbb{1}_{E_{k+1}}] \leq \left(1 - \frac{2\alpha\gamma}{k+m}\right)\mathbb{E}_{\mathcal{D}^{(w)}}[D_k \mathbb{1}_{E_k}] + \frac{\gamma^2}{n(k+m)^2} \sum_{i=1}^n \sigma_i^2, \quad \text{for RR}. \tag{3.42}$$

Therefore, we eventually get that, with γ large so that $2\alpha\gamma > 1$,

$$\mathbb{E}_{\mathcal{D}}[\|\theta_k - \theta_*\|_2^2 | E_\infty] \leq \frac{\mathbb{E}_{\mathcal{D}}[D_k \mathbb{1}_{E_\infty}]}{\mathbb{P}(E_\infty)} \leq \frac{\gamma^2}{(1-\delta)(2\alpha\gamma-1)k} \sup_j \sigma_j^2 + o(1/k) \tag{3.43}$$

for SGD, and

$$\mathbb{E}_{\mathcal{D}^{(w)}}[\|\theta_k - \theta_*\|_2^2 | E_\infty] \leq \frac{\gamma^2}{(1-\delta)(2\alpha\gamma-1)k} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i^2\right) + o(1/k) \tag{3.44}$$

for the RR scheme. □

REMARK 3.3. The convergence rate comparison result is similar to [22], in the spirit that for loss functions with drastically different subfunction slopes, the RR scheme performs as averaging to speed up the convergence rate. We should point it out that the analysis in [22] only considers strong convex objectives, and the weighted SGD they investigate has weights proportional to Lipschitz constants of ∇l_i rather than $\|\nabla l_i\|_2$.

3.4. Comments from the asymptotic viewpoint. With the derivations of (3.4) and (3.6), one would hope to leverage stochastic calculus tools to give a short and illustrative picture for stochastic algorithm behavior comparisons. We make following remarks which fully use the trace information of covariance matrices, which may indicate the faster local convergence and less oscillation behaviors of the RR scheme.

REMARK 3.4 (Local convergence rate in the continuous-time limit). Suppose θ_* is a local minimum, and there exists $r > 0$ such that the stochastic trajectories Θ_t with the starting point $\Theta_0 = \theta_0 \in B(\theta_*, r)$ stay inside the ball $B(\theta_*, r)$ for $t \geq 0$. Also the local strong convexity holds: There exists $c_0 > 0$ such that $(x-y)^\top (\nabla L(x) - \nabla L(y)) \geq c_0 \|x-y\|_2^2$ for $x, y \in B(\theta_*, r)$, then the trajectory driven by (3.6) converges faster to θ_* compared to the trajectory driven by (3.4) in the $L^2(0, t; L^2(\mathbb{R}^d))$ -sense.

Let us elaborate on the Remark 3.4. We denote the solutions to (3.4) and (3.6) as Θ^1 and Θ^2 respectively and compute their convergence rate to the local minimum in $L^2(0, t; L^2(\mathbb{R}^d))$. Consider the function $f(x) = \frac{1}{2}(x - \theta_*)^\top (x - \theta_*)$. By Ito's lemma,

$$df(\Theta_t^i) = \left(-(\Theta_t^i - \theta_*)^\top m(\Theta_t^i) + \frac{\eta}{2} \text{Tr}(\Sigma_i(\Theta_t^i))\right) dt + (\Theta_t^i - \theta_*)^\top \sigma_i(\Theta_t^i) dW_t$$

for $i = 1, 2$. Therefore, for $i = 1, 2$,

$$\begin{aligned} \mathbb{E}[\|\Theta_t^i - \theta_*\|_2^2] &= \|\theta_0 - \theta_*\|_2^2 + \mathbb{E}\left[\int_0^t \left(-(\Theta_s^i - \theta_*)^\top \nabla L(\Theta_s^i) + \frac{\eta}{2} \text{Tr}(\Sigma_i(\Theta_s^i))\right) ds\right] \\ &\leq \|\theta_0 - \theta_*\|_2^2 + \frac{\eta}{2} \int_0^t \mathbb{E}[\text{Tr}(\Sigma_i(\Theta_s^i))] ds - c_0 \int_0^t \mathbb{E}[\|\Theta_s^i - \theta_*\|_2^2] ds, \end{aligned}$$

where the last line uses the local strong convexity and $\nabla L(\theta_*) = 0$. Then we deduce that

$$\int_0^t \mathbb{E}[\|\Theta_s^i - \theta_*\|_2^2] ds \leq \int_0^t e^{-c_0(t-s)} \left(\|\theta_0 - \theta_*\|_2^2 + \frac{\eta}{2} \int_0^s \mathbb{E}[\text{Tr}(\Sigma^i(\Theta_\tau^i))] d\tau \right) ds.$$

Unfortunately, a direct comparison between two independent stochastic processes is unclear and out of our reach, but we speculate that in order to obtain a faster convergence to the global minimum, it should have a smaller trace of the covariance matrix. Indeed, by Cauchy-Schwarz, the covariance matrix from (3.6) has a smaller trace at every small neighborhood of points coinciding with the trajectory from (3.4) since

$$\begin{aligned} \text{Tr}(\Sigma^2(\Theta_t)) &= \frac{Z(\Theta_t)}{n^2} \sum_{i=1}^n \|\nabla l_i(\Theta_t)\|_2 - \|m(\Theta_t)\|_2^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n \|\nabla l_i(\Theta_t)\|_2^2 - \|m(\Theta_t)\|_2^2 = \text{Tr}(\Sigma^1(\Theta_t)). \end{aligned} \tag{3.45}$$

Back to the discrete algorithms, the above computations also indicate that the RR scheme (3.5) improves the convergence rate to a local minimum.

REMARK 3.5 (Deviation from the deterministic gradient descent). Suppose that ∇L is uniformly Lipschitz continuous, i.e., there exists $B > 0$ such that $\|\nabla L(x) - \nabla L(y)\|_2 \leq B\|x - y\|_2$ for all $x, y \in \mathbb{R}^d$. Let ϕ_t denote the solution to

$$\dot{\phi}_t = -\nabla L(\phi_t)$$

with $\phi_0 = \theta$, and Θ_t be the solution to

$$d\Theta_t = -\nabla L(\Theta_t) + \sqrt{\eta}\sigma(\Theta_t)dW_t$$

with $\Theta_0 = \theta$, then we have an upper bound on the probability of deviation.

LEMMA 3.2. For any $\delta > 0$ and $0 < T < \infty$, we have the inequality

$$\mathbb{P}_\theta \left(\sup_{t \in [0, T]} \|\Theta_t - \phi_t\|_2 > \delta \right) \leq \eta c' \mathbb{E}_\theta \left[\int_0^T \text{Tr}(\sigma(\Theta_s)\sigma(\Theta_s)^\top) ds \right], \tag{3.46}$$

where c' only depends on δ, T and B .

The proof of inequality (3.46) can be found in the Appendix B. Now if we consider $\Theta_{1,t}$ in (3.4) for SGD along with $\Theta_{2,t}$ in (3.6) for the RR scheme, the trace bound in Lemma 3.2 implies that (3.6) is closer to ϕ_t with a higher probability, simply by taking (3.45). In terms of discrete stochastic algorithms, it also indicates that the RR scheme is more deterministic compared to SGD.

4. Experiments

The analysis above suggests that, for non-convex optimization problems, the RR scheme is more likely to find the global minimum compared to the vanilla SGD, especially when the global minimum lies in the sharp valley. We empirically verify this in several examples from both data sciences and physical sciences. In particular, we study (1) robust statistics problems under data feature disparities, and (2) geometric optimization problems in computational chemistry. For real applications in deep learning, we refer to empirical works in [11, 30].

4.1. Robust classification/regression. In this example, we consider to use the Welsch loss (see [3]) from robust statistics, where the corresponding loss for each data sample i is

$$l_i(\theta) = 1 - \exp(-(y_i - \theta^\top x_i)^2/2). \quad (4.1)$$

This loss function has found many applications in regression problems dealing with outliers. Suppose that the goal is to find a global optimizer for a mixed population of multiple subgroups: part of them are quite sensitive in a certain trait, while the rest are much less sensitive in the same trait. In the following two examples, the RR scheme randomly selects the sub-population j with the probability proportional to $a_j \|\nabla l_j(\theta_k)\|_2$ with replacement in each iteration, where a_j denotes the sub-population proportion.

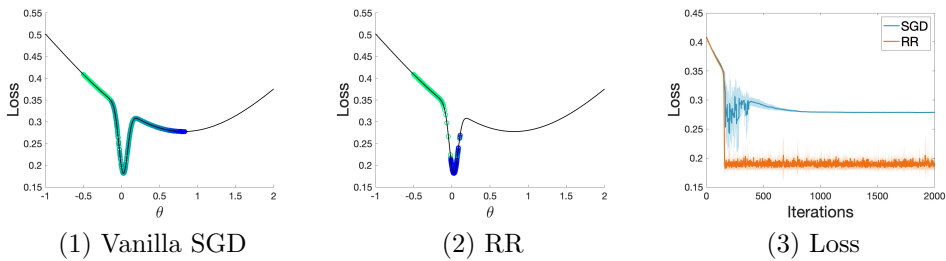


Fig. 4.1: We set the starting point $\theta_0 = -0.5$ with a constant learning rate $\eta = 0.015$ in the robust classification problem. The color gradient of circles from green to blue shows how iterations proceed. Plots (1) and (2) show the trajectories for one trial under vanilla SGD and RR, respectively. Plot (3) show the loss comparisons over 10 trials.

4.1.1. Classification. The dataset $\{(x_i, y_i)\}_{1 \leq i \leq N}$, $x_i \in \mathbb{R}$, $y_i \in \{0, 1\}$ consists of x from two subgroups

- Subgroup 1: $x_i = 20 + \mathcal{N}(0, 1)$; total number $N_1 = 800$.
- Subgroup 2: $x_i = 0.5 + \mathcal{N}(0, 1/4)$; total number $N_2 = 4000$.

Here $a_1/a_2 = 1/5$. The class $y_i \sim \text{Ber}(1/2)$ is preassigned for each data point, as we assume that each subgroup contains individuals belonging to different classes. The goal is to find the global minimum rather than a local minimum, even though it is flat. From Figure 4.1 we can see that in contrast to the vanilla SGD trajectory escaping to the nearby flat local minimum, the RR scheme trajectory stays inside the sharp valley to reach the global minimum.

4.1.2. Regression. The dataset $\{(\mathbf{x}_i, y_i)\}_{1 \leq i \leq N}$, $\mathbf{x}_i \in \mathbb{R}^d$, $d = 10$ is composed of samples from two subgroups

- Subgroup 1: $\mathbf{x}_i = 20\mathbf{e} + \mathcal{N}(0, I_d)$; total number $N_1 = 2000$.

- Subgroup 2: $\mathbf{x}_i = \frac{1}{4}\mathbf{e} + \frac{1}{2}\mathcal{N}(0, I_d)$; total number $N_2 = 800$.

Here $a_1/a_2 = 5/2$. The exact regression coefficient $\beta^* \in \mathbb{R}^d$ is picked by $\beta^* \sim \mathcal{N}(0, I_d)$ and the uncorrupted response variables are $y_i^* = \mathbf{x}_i^\top \beta^*$. The corrupted response variables are generated by $y_i = y_i^* + u_i + \epsilon_i$, where $u_i \sim \text{Unif}([-3\|y^*\|_\infty, 3\|y^*\|_\infty])$, $\epsilon_i \sim \frac{1}{10}\mathcal{N}(0, 1)$. Figure 4.2 shows that for a wide range of learning rate choices, the RR scheme selects a better minimum at a faster speed compared to SGD.

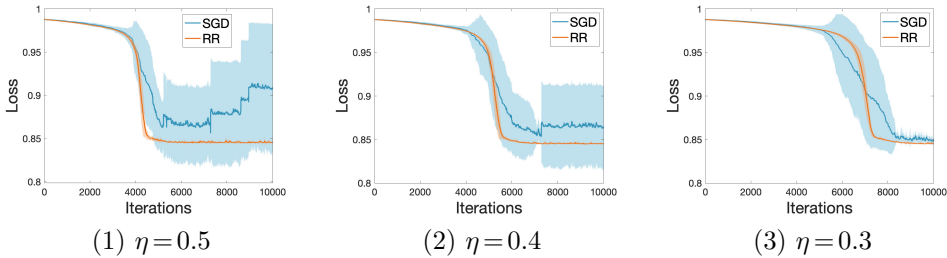


Fig. 4.2: Comparisons of the vanilla SGD and the RR scheme for various learning rates over 10 trials in robust regression. For each plot, we start from the same $\beta_0 = \beta^* + 5\mathbf{e} + \epsilon$, $\epsilon \sim \mathcal{N}(0, I_d)$. We observe with a relatively large learning rate, SGD can deviate while the RR scheme is stable around the local minimum. Even when the learning rate decreases and SGD converges to the same minimum that the RR scheme arrives, the convergence speed of SGD is slower.

4.2. Computational chemistry. The problem here is to find the global minimum of a potential energy surface (PES) that typically gives a mathematical description of the molecular structure and its energy. In general, assuming that there are n atoms to form a molecule, we consider minimizing the particle interacting energy of the form

$$E(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m) = \sum_{i < j}^m V(\mathbf{z}_i, \mathbf{z}_j),$$

where V is a bi-atom potential function, and $\mathbf{z}_k \in \mathbb{R}^d, 1 \leq k \leq m, d \geq 1$ denote the atom’s position. In particular, the global minimum represents the most stable conformation with respect to location arrangements of atoms. Though using SGD for a large system is computationally efficient, trying to find the global minimum with SGD can be difficult, especially when the global minimum lies in a sharp valley. The RR scheme outperforms SGD in terms of the likelihood of identifying the global minimum. This is demonstrated in Figure 4.3 by looking at two examples. The first one is the Müller-Brown potential [21]. The second one is an artificial large system with the interacting function of Gaussian type

$$V_k(\mathbf{z}_i, \mathbf{z}_k) = \exp(-(\mathbf{z}_i - \mathbf{z}_k)^\top M_k (\mathbf{z}_i - \mathbf{z}_k)), \quad M_k = \begin{bmatrix} a_k & b_k/2 \\ b_k/2 & c_k \end{bmatrix},$$

with $\mathbf{z} = (x, y) \in \mathbb{R}^2$. We assume that except for the atom i , the rest of atoms’ positions are fixed, then it is to consider minimizing the potential energy $\min_{\mathbf{z}_i} \frac{1}{m-1} \sum_{k \neq i}^m V_k(\mathbf{z}_i, \mathbf{z}_k)$ and finding the optimal \mathbf{z}_i . Detailed parameters setups are provided in the Appendix C.

5. Conclusion

To deal with data feature disparities in non-convex optimization problems, we propose in this paper a combined resampling-reweighting (RR) scheme to balance variances

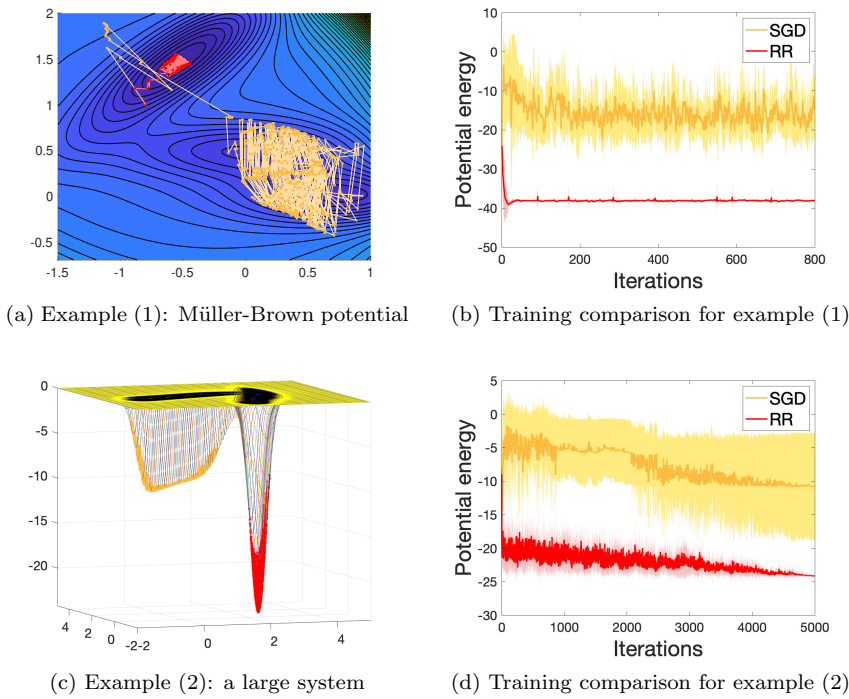


Fig. 4.3: *Example (1): the global optimization problem for the 1/4-rescaled Müller-Brown potential with a learning rate $\eta=0.002$. Both start at $(x_0, y_0)=(-0.8, 1.0)$. Example (2): a large system with $m=1000$ atoms in 5000 iterations. Because this problem is rather stiff, we use a monotonically decreasing learning rate starting from $\eta=0.002$ and ending at $\eta=1e-5$ for the optimization. Both trajectories start at $(x_0, y_0)=(3.0, 1.0)$. For both examples (1) and (2), the left plot shows the trajectories for one trial, and the right plot shows the potential energy over 10 trials.*

experienced in different regions. The RR scheme connects with the importance sampling SGD that was previously proposed and analyzed for convex optimization problems. We extend the analysis of the importance sampling SGD to non-convex problems from the viewpoints of stochastic stability and local convergence speed. Numerical experiments verify that the RR scheme outperforms the SGD in capturing the sharp global minimum, making it more reliable and faithful for optimization purposes.

Acknowledgments. The work of L.Y. is partially supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC) program and also by the National Science Foundation under award DMS-1818449. J.A. was supported by Joe Olinger Fellowship from Stanford University.

Appendix A. Proof of Lemma 2.1.

Proof. Due to the construction of the loss, the stochastic iterates experience different gradients depending on the location. In different regions, the associated variances are different:

- In regions $(-\infty, -1)$, $(-1, 0)$, $(0, 1/K)$, $(1/K, \infty)$.
- With probability a_1 , the gradients are -1 , 1 , ϵ , ϵ respectively.

- With probability a_2 , the gradients are $-\epsilon, -\epsilon, -K, K$ respectively.
- Corresponding variances are $a_1a_2(1-\epsilon)^2, a_1a_2(1+\epsilon)^2, a_1a_2(K+\epsilon)^2, a_1a_2(K-\epsilon)^2$ in each region.

With the above computed, we then derive the SGD, and use the first order Euler-Maruyama approximation (see [15]) to obtain stochastic modified equations. Indeed, we take $\epsilon \rightarrow 0$. Let \mathcal{N} denote a standard normal random variable. For $x < 0$, the SGD is approximately

$$\theta \leftarrow \theta - V'(\theta)\eta + \sqrt{a_1a_2}\eta\mathcal{N}, \tag{A.1}$$

and the corresponding SDE is

$$d\Theta_t = -V'(\Theta_t)dt + \sqrt{a_1a_2}\eta dW_t. \tag{A.2}$$

For $x > 0$, the SGD is approximately

$$\theta \leftarrow \theta - V'(\theta)\eta + K\sqrt{a_1a_2}\eta\mathcal{N}, \tag{A.3}$$

and the corresponding SDE is

$$d\Theta_t = -V'(\Theta_t)dt + K\sqrt{a_1a_2}\eta dW_t. \tag{A.4}$$

The resulted equilibrium measures on two sides are

$$p(\theta) = \frac{1}{Z_1} \exp\left(-\frac{2}{a_1a_2\eta}V(\theta)\right) \text{ for } \theta < 0, \quad p(\theta) = \frac{1}{Z_2} \exp\left(-\frac{2}{K^2a_1a_2\eta}V(\theta)\right) \text{ for } \theta > 0. \tag{A.5}$$

To determine the normalization constants Z_1 and Z_2 , consider a SDE with non-smooth diffusion

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dB_t,$$

the corresponding Kolmogorov forward equation is

$$\partial_s p(x, s) = -\partial_x[\mu(x, s)p(x, s)] + \frac{1}{2}\partial_x^2[\sigma^2(x, s)p(x, s)],$$

for $s \geq t$. The equilibrium measure should satisfy $\partial_s p(\theta) = 0$, which suggests that $\sigma^2 p$ is continuous at $\theta = 0$. Because $V(0) = 0$, the continuity implies that

$$\frac{1}{Z_1}a_1a_2\eta = \frac{1}{Z_2}K^2a_1a_2\eta \implies Z_2 = K^2Z_1.$$

□

Appendix B. Proof of Lemma 3.2.

Proof. Since there exists $B > 0$ such that $\|\nabla L(x) - \nabla L(y)\|_2 \leq B\|x - y\|_2$ for all $x, y \in \mathbb{R}^d$. Then we have

$$\|\Theta_t - \phi_t\|_2 \leq B \int_0^T \|\Theta_s - \phi_s\|_2 ds + \sqrt{\eta} \left\| \int_0^T \sigma(\Theta_s) dW_s \right\|_2.$$

By Gronwall’s inequality, we get

$$\sup_{t \in [0, T]} \|\Theta_t - \phi_t\|_{\otimes} \leq \sqrt{\eta} e^{BT} \sup_{t \in [0, T]} \left\| \int_0^T \sigma(\Theta_s) dW_s \right\|_2.$$

Therefore,

$$\begin{aligned} \mathbb{P}_\theta \left(\sup_{t \in [0, T]} \|\Theta_t - \phi_t\|_2 > \delta \right) &\leq \mathbb{P}_\theta \left(\sup_{t \in [0, T]} \left\| \int_0^T \sigma(\Theta_s) dW_s \right\|_2 > \frac{\delta}{\sqrt{\eta}} e^{-BT} \right) \\ &\leq \frac{\eta}{\delta^2 e^{-2BT}} \mathbb{E}_\theta \left[\left(\sup_{t \in [0, T]} \left\| \int_0^T \sigma(\Theta_s) dW_s \right\|_2 \right)^2 \right] \\ &\leq \eta c' \mathbb{E}_\theta \left[\int_0^T \text{Tr}(\sigma(\Theta_s) \sigma(\Theta_s)^\top) ds \right], \end{aligned}$$

where we use Chebyshev’s inequality for the second last inequality and Burkholder-Davis-Gundy maximal inequality for the last inequality. \square

Appendix C. Extended numerical results and parameters setup.

C.1. Numerical comparisons with different learning rates. Here we show more numerical comparisons between SGD and the RR scheme with various learning rates. See Figure C.1.

C.2. Parameters setup in Section 4.2.

Müller-Brown potential

$$\begin{aligned} V(x, y) &= \sum_{i=1}^4 A_i \exp(a_i(x - x_i)^2 + b_i(x - x_i)(y - y_i) + c_i(y - y_i)^2), \\ A &= (-150, -100, -170, 15), \quad a = (-1, -1, -6.5, 0.7), \\ b &= (0, 0, 11, 0.6), \quad c = (-10, -10, -6.5, 0.7), \\ x &= (1, 0, -0.5, -1), \quad y = (0, 0.5, 1.5, 1). \end{aligned} \tag{C.1}$$

A large system

$$\begin{aligned} x_k &= 2.0 - 0.006k, \quad \text{for } 1 \leq k \leq 500; \quad x_k = 1.8 + 0.0024k, \quad \text{for } 500 \leq k \leq 1000. \\ y_k &= 2.0 - 0.006k, \quad \text{for } 1 \leq k \leq 500; \quad y_k = -1.0 + 0.006k, \quad \text{for } 500 \leq k \leq 1000. \\ A_k &= -50.0 - 0.15k, \quad a_k = -2.0 - 0.018k, \\ b_k &= -0.1 + 0.0002k, \quad c_k = -10 + 0.009k, \quad \text{for } 1 \leq k \leq 1000. \end{aligned} \tag{C.2}$$

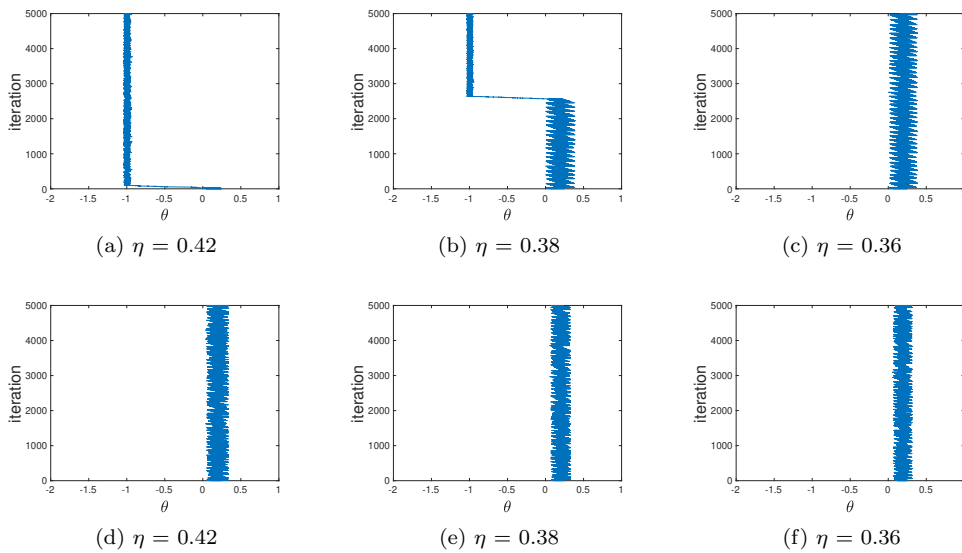


Fig. C.1: *Extended comparisons of SGD (upper row) and RR (lower row) with $a_1=0.4, a_2=0.6, \epsilon=0.1, K=5$ at various learning rates η . All experiments start at $\theta_0=0.25$. We can see that unless the learning rate $\eta \leq 0.36$, RR is more reliable in the sense that its trajectory stays around the desired minimum. Even when SGD and RR both stay around the sharp global minimum, the oscillation in RR is smaller.*

REFERENCES

- [1] G. Alain, A. Lamb, C. Sankar, A. Courville, and Y. Bengio, *Variance reduction in SGD by distributed importance sampling*, International Conference on Learning Representations, 2016. 1.2
- [2] J. An, L. Ying, and Y. Zhu, *Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients*, International Conference on Learning Representations, 2021. 2.1
- [3] J.T. Barron, *A general and adaptive robust loss function*, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 4331–4339, 2019. 1, 4.1
- [4] O. Canévet, C. Jose, and F. Fleuret, *Importance sampling tree for large-scale empirical expectation*, Proc. Mach. Learn. Res., 48:1454–1462, 2016. 1.2
- [5] P. Chaudhari and S. Soatto, *Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks*, 2018 Information Theory and Applications Workshop (ITA), IEEE, 1–10, 2018. 1.2
- [6] X. Cheng, D. Yin, P. Bartlett, and M. Jordan, *Stochastic gradient and Langevin processes*, Proc. Mach. Learn. Res., 119:1810–1819, 2020. 1.2
- [7] B. Fehrman, B. Gess, and A. Jentzen, *Convergence rates for the stochastic gradient descent method for non-convex objective functions*, J. Mach. Learn. Res., 21(1):5354–5401, 2020. 1.2
- [8] E. Hoffer, I. Hubara, and D. Soudry, *Train longer, generalize better: Closing the generalization gap in large batch training of neural networks*, Proc. Int. Conf. Neur. Inf. Process. Syst., 1729–1739, 2017. 1
- [9] P. Jain and P. Kar, *Non-convex optimization for machine learning*, Found. Trends[®] Mach. Learn., 10(3-4):142–336, 2017. 1
- [10] T.B. Johnson and C. Guestrin, *Training deep models faster with robust, approximate importance sampling*, Adv. Neur. Inf. Process. Syst., 31:7265–7275, 2018. 1.2
- [11] A. Katharopoulos and F. Fleuret, *Not all samples are created equal: Deep learning with importance sampling*, Proc. Int. Conf. Mach. Learn., 2525–2534, 2018. 2, 1.2, 3.1, 4
- [12] N.S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P.T.P. Tang, *On large-batch training for deep learning: Generalization gap and sharp minima*, International Conference on

- Learning Representations, 2017. 1
- [13] D. Kochkov and B.K. Clark, *Variational optimization in the AI era: Computational graph states and supervised wave-function optimization*, arXiv preprint, [arXiv:1811.12423](https://arxiv.org/abs/1811.12423), 2018. 1
- [14] A.R. Leach, *Molecular Modelling: Principles and Applications*, Pearson Education, 2001. 1
- [15] Q. Li, C. Tai, and W. E, *Stochastic modified equations and adaptive stochastic gradient algorithms*, Proc. Mach. Learn. Res., **70:2101–2110**, 2017. 1.2, A
- [16] I. Loshchilov and F. Hutter, *Online batch selection for faster training of neural networks*, arXiv preprint, [arXiv:1511.06343](https://arxiv.org/abs/1511.06343), 2015. 1.2
- [17] C. Ma and L. Ying, *The Sobolev regularization effect of stochastic gradient descent*, arXiv preprint, [arXiv:2105.13462](https://arxiv.org/abs/2105.13462), 2021. 1.2
- [18] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey, *Normalized loss functions for deep learning with noisy labels*, Proc. Int. Conf. Mach. Learn., **6543–6553**, 2020. 1
- [19] S. Mandt, M. Hoffman, and D. Blei, *A variational analysis of stochastic gradient algorithms*, Proc. Int. Conf. Mach. Learn., **354–363**, 2016. 1.2
- [20] P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher, *On the almost sure convergence of stochastic gradient descent in non-convex problems*, NeurIPS’20: The 34th International Conference on Neural Information Processing Systems, 2020. 1.2, 3.3, 3.3, 3.3
- [21] K. Müller and L.D. Brown, *Location of saddle points and minimum energy paths by a constrained simplex optimization procedure*, Theor. Chim. Acta, **53(1):75–93**, 1979. 4.2
- [22] D. Needell, R. Ward, and N. Srebro, *Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm*, Adv. Neur. Inf. Process. Syst., **27:1017–1025**, 2014. 1.2, 3.1, 3.3
- [23] G.A. Pavliotis, *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, Springer, **60**, 2014. 2.1
- [24] P.J. Rousseeuw, *Least median of squares regression*, J. Amer. Stat. Assoc., **79(388):871–880**, 1984. 1
- [25] U. Simsekli, L. Sagun, and M. Gurbuzbalaban, *A tail-index analysis of stochastic gradient noise in deep neural networks*, International Conference on Machine Learning, **5827–5837**, 2019. 1
- [26] T. Strohmer and R. Vershynin, *A randomized Kaczmarz algorithm with exponential convergence*, J. Four. Anal. Appl., **15(2):262–278**, 2009. 1.2, 3.1
- [27] S. Wojtowysch, *Stochastic gradient descent with noise of machine learning type. Part I: Discrete time analysis*, J. Nonlinear Sci., **33:45**, 2023. 1.2
- [28] L. Wu, C. Ma, and W. E, *How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective*, Adv. Neur. Inf. Process. Syst., **8279–8288**, 2018. 1.2, 3.2
- [29] Z. Xie, I. Sato, and M. Sugiyama, *A diffusion theory for deep learning dynamics: Stochastic gradient descent escapes from sharp minima exponentially fast*, International Conference on Learning Representations, 2021. 1
- [30] P. Zhao and T. Zhang, *Stochastic optimization with importance sampling for regularized loss minimization*, Proc. Int. Conf. Mach. Learn., **1–9**, 2015. 2, 1.2, 3.1, 4
- [31] P. Zhou, J. Feng, C. Ma, C. Xiong, S. Hoi, and W. E, *Towards theoretically understanding why SGD generalizes better than ADAM in deep learning*, Adv. Neur. Inf. Process. Syst., **33**, 2020. 1
- [32] Z. Zhu, J. Wu, B. Yu, L. Wu, and J. Ma, *The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects*, Proc. Mach. Learn. Res., **97:7654–7663**, 2019. 1