Uniform Stability and Generalization Error of GD and SGD on Fixed-Point Parameters

Jonghyun Shin
Department of Artificial Intelligence
Korea University
Seoul, Korea
uenjgieonj5448@korea.ac.kr

Sejun Park*
Department of Artificial Intelligence
Korea University
Seoul, Korea
sejun.park000@gmail.com

Abstract

We analyze generalization error, uniform stability, and uniform argument stability of gradient descent (GD) and stochastic gradient descent (SGD) over discrete parameter spaces, where each update involves deterministic or stochastic rounding. We show that deterministic rounding degrades the generalization error of GD on convex, Lipschitz, and smooth loss functions, increasing the rate from $O(T/n)$ to $O(T/\sqrt{n})$, and establish matching lower bounds. We further prove that uniform stability of GD becomes $\Omega(T)$, showing that stability-based generalization bounds are vacuous in this setting. In contrast, for the same losses, stochastic gradient descent with deterministic rounding admits nontrivial uniform stability guarantees, which differ qualitatively from the real-valued case and exhibit distinct dependencies on the number of iterations and the dimension: we prove tight bounds $O(T/n)$ for one dimension and $O(T^2/n)$ for higher dimensions. We also show that stochastic rounding can introduce generalization error that increases with the dimension; such a phenomenon is absent in standard real-valued optimization and in the deterministic rounding case. Finally, we provide upper bounds on uniform argument stability for stochastic rounding schemes and show that these bounds are tight when the loss can be represented as a sum of coordinate-wise functions.

1 Introduction

Gradient-based optimization algorithms are the backbone of modern machine learning and are widely used to train models ranging from linear predictors to deep neural networks. Starting from classical methods such as gradient descent (GD) and stochastic gradient descent (SGD), numerous variants have been developed to improve convergence speed and robustness in large-scale and non-convex settings [Duchi et al., 2011, Tieleman and Hinton, 2012, Kingma and Ba, 2015, Liu et al., 2025]. These algorithms have been extensively studied, with a rich theory characterizing their convergence behavior under various assumptions [Bottou et al., 2018, Reddi et al., 2018, Zhou et al., 2024], as well as their generalization performance in statistical learning frameworks [Hardt et al., 2016, Bassily et al., 2020, Zhou et al., 2020, Nguyen et al., 2022]. However, most existing analyses rely on an idealized setting in which the optimization variables lie in a continuous space and all arithmetic operations are performed exactly.

In practice, optimization algorithms are implemented on digital hardware using discrete parameters, such as fixed-point or floating-point representations [Goldberg, 1991, Gupta et al., 2015]. In this setting, the parameter space is inherently discrete, and every arithmetic operation incurs round-off errors. As a result, the optimization dynamics can differ fundamentally from their real-valued counterparts: updates are quantized and numerical errors accumulate over time. Recent works have

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

begun to study gradient-based methods in discrete or quantized settings by establishing convergence guarantees [Markov et al., 2023, Xin et al., 2025, Xia et al., 2025]. Nevertheless, a theoretical understanding of how discrete parameters affect generalization remains limited.

Understanding the generalization properties of optimization algorithms is a central question in statistical learning theory. Classical approaches based on uniform convergence provide bounds on the gap between empirical and population risks uniformly over a hypothesis class. However, such bounds typically depend on the ambient dimension [Shalev-Shwartz and Ben-David, 2014, Feldman, 2016] and are agnostic to the specific optimization algorithm used, often leading to overly pessimistic bounds in modern high-dimensional settings. This limitation has motivated *algorithm-dependent* analyses, which take into account the dynamics of the optimization process.

A canonical framework in this direction is *algorithmic stability* [Bousquet and Elisseeff, 2002, Shalev-Shwartz et al., 2010]. In this framework, a learning algorithm A maps a dataset $S = (z_1, \ldots, z_n)$ to a parameter vector $A(S)$, and the generalization error is bounded by the sensitivity of $A(S)$ to small perturbations of the dataset. In particular, *uniform stability* measures how much the loss changes when a single training example is replaced, while *uniform argument stability* measures how much the learned parameters themselves change under such perturbations. When the loss function is Lipschitz, these two notions are directly related, allowing one to bound the generalization error by tracking how perturbations propagate through the optimization dynamics. This framework has been particularly successful for gradient-based methods, yielding dimension-independent generalization guarantees for GD and SGD [Hardt et al., 2016, Chen et al., 2018, Bassily et al., 2020].

1.1 Our contributions

In this paper, we revisit this framework with fixed-point parameters. Specifically, we consider gradient-based optimization over a discrete parameter space $\{\Delta x : x \in \mathbb{Z}\}^d$, where each update involves deterministic or stochastic rounding as in training under quantized weights with high-precision gradients. Within this setting, we analyze how finite-precision effects influence generalization error, uniform stability, and uniform argument stability[2] of gradient-based optimization algorithms when the loss functions are convex, Lipschitz, and smooth. We show that the discrete and inexact nature of machine arithmetic fundamentally alters the behavior of these stability measures, leading to qualitatively different generalization guarantees compared to the classical real-valued setting. In particular, our contributions can be summarized as below (see also Table 1). Here, DR-GD, SR-GD, DR-SGD, and SR-SGD denote (stochastic) gradient descent on $\{\Delta x : x \in \mathbb{Z}\}^d$ with deterministic/stochastic rounding, and GD and SGD denote standard (stochastic) gradient descent on $\mathbb{R}^d$ without rounding operations. In addition, we hide the Lipschitz constant and Δ for concise representation; we present the precise bounds including them in Sections 3 and 4.

- We first prove that the generalization error of T-step DR-GD is upper bounded by $O(\eta T/\sqrt{n})$[3] on a constant-size finite sample space (Theorem 2). We also show the tightness of $1/\sqrt{n}$ rate for an arbitrary finite sample space (Theorem 3). Since the generalization error of GD is $O(\eta T/n)$ for convex, Lipschitz, and smooth losses, these results imply that deterministic rounding to the fixed-point parameter space degrades the sample complexity of GD.
- However, a stability-based approach cannot achieve this bound: the uniform stability of DR-GD can be lower bounded by $\Omega(T)$ (Theorem 4). Hence, the uniform stability of DR-GD can only produce vacuous generalization bounds in this case.
- On the other hand, this is not the case for DR-SGD. We prove that the uniform stability of DR-SGD is bounded by $O(\eta T/n)$ for $d = 1$ (Theorem 5), and it is bounded by $O(\eta T^2/n)$ for $d > 1$ (Theorem 8). We further show the tightness of these bounds (Theorems 6 and 7), which implies that depending on the ambient dimension d, the uniform stability can have a different dependence on T, while such a dependency was not observed in SGD.
- We next consider stochastic rounding and show that the generalization errors of SR-GD and SR-SGD can increase with d. Specifically, Theorem 9 states that for a finite sample space, there exists a convex, Lipschitz, and smooth loss and a data distribution for which the generalization errors of SR-GD and SR-SGD are $\Omega(\sqrt{\eta d^{1/2}}/n)$.[4] We note that such a dimension dependency is from the stochasticity in rounding, and does not arise in GD, SGD, DR-GD, and DR-SGD.

[2] generalization error $\leq$ uniform stability $\leq$ Lipschitz constant $\times$ uniform argument stability (see Section 2.4).

[3] η denotes the learning rate, n denotes the number of samples.

[4] We assume $\eta d^{1/2} = \Omega(1)$ here. See Section 4 for bounds without this assumption.

Table 1: A summary of generalization and stability bounds for iterative algorithms. All results are for convex, L-Lipschitz, and smooth losses except for [Bassily et al., 2020], which considers convex, Lipschitz but non-smooth losses. $\varepsilon_g, \varepsilon_{us}, \varepsilon_{uas}$ denote the generalization error, uniform stability, and uniform argument stability, respectively. For simplicity, we consider $L, \Delta = \Theta(1)$, $n > T$, $\eta = O(1)$, and $\eta d^{1/2} = \Omega(1)$ here; the precise bounds are in Sections 3 and 4.

Algorithm	Reference	Bound
GD	Nikolakakis et al. [2025]	$\varepsilon_g \gtrsim \eta T/n$
	Chen et al. [2018]	$\varepsilon_{us} \asymp \eta T/n$
	Bassily et al. [2020]	$\varepsilon_{uas} \asymp \eta\sqrt{T} + \eta T/n$ [(1)]
SGD	Hardt et al. [2016]	$\varepsilon_{us} \lesssim \eta T/n$
	Zhang et al. [2022]	$\varepsilon_{us} \gtrsim \eta T/n$
	Bassily et al. [2020]	$\varepsilon_{uas} \asymp \eta T^{3/2}/n$ [(1)]
DR-GD	Theorem 2	$\varepsilon_g \lesssim \eta T/\sqrt{n}$ [(2)]
	Theorem 3	$\varepsilon_g \gtrsim T/\sqrt{n}$
	Theorem 4	$\varepsilon_{us} \gtrsim T$
DR-SGD	Theorem 5	$\varepsilon_{us} \lesssim T/n$ [(3)]
	Theorem 6	$\varepsilon_g \gtrsim T/n$
	Theorem 7	$\varepsilon_{us} \gtrsim T^2/n$ [(4)]
	Theorem 8	$\varepsilon_{us} \lesssim \eta T^2/n$
SR-GD	Theorem 9	$\varepsilon_g \gtrsim \sqrt{\eta d^{1/2}}/n$ [(5)]
	Theorem 11	$\varepsilon_{uas} \lesssim \eta T/n + \min\{1, \eta d^{1/2} T/n\}\sqrt{\eta d^{1/2} T}$
	Theorem 12	$\varepsilon_{uas} \lesssim \sqrt{\eta d^{1/2} T/n}$ [(6)]
	Theorem 13	$\varepsilon_{uas} \gtrsim \eta T/n + \sqrt{\min\{1, \eta d^{1/2} T/n\}\eta d^{1/2} T/n}$ [(6)]
SR-SGD	Theorem 9	$\varepsilon_g \gtrsim \sqrt{\eta d^{1/2}}/n$ [(5)]
	Theorem 10	$\varepsilon_{uas} \lesssim \sqrt{\eta d^{1/2} T^3}/n$
	Theorem 12	$\varepsilon_{uas} \lesssim T\sqrt{\eta d^{1/2}}/n$ [(6)]
	Theorem 13	$\varepsilon_{uas} \gtrsim T\sqrt{\eta d^{1/2}}/n$ [(6)]

n, η, T, and d denote the number of samples, learning rate, total number of iterations, and parameter dimension, respectively. [(1)] requires $d \geq \min\{T, 1/\eta^2\}$ for the lower bounds, [(2)] requires the sample space to have constant size, [(3)] requires $d = 1$, [(4)] requires $d \geq 2$, [(5)] requires $T = 1$, and [(6)] requires that the loss function is a sum of coordinate-wise functions.

- We also prove upper bounds on the uniform argument stability: $O(\sqrt{\eta d^{1/2} T^3}/n)$ for SR-SGD and $O(\eta T/n + \min\{1, \eta d^{1/2} T/n\}\sqrt{\eta d^{1/2} T})$ for SR-GD (Theorems 10 and 11).[4] We further reduce the $T^{3/2}$ dependency here to T for losses that can be represented as a sum of coordinate-wise functions (Theorem 12), and show the tightness of these reduced bounds (Theorem 13).

2 Preliminaries

2.1 Notations and problem setup

For $n \in \mathbb{N}$, we use $[n]$ to denote $\{1, \cdots, n\}$. For $d \in \mathbb{N}$, we use $\mathbf{0}_d$ to denote $(0, \cdots, 0) \in \mathbb{R}^d$ and $\mathbf{1}_d$ to denote $(1, \cdots, 1) \in \mathbb{R}^d$. For $n \in \mathbb{N}$ and $S, S' \in \mathbb{R}^n$, we say S and S' are neighboring, denoted by $S \simeq S'$, if they differ at a single entry. For $\Delta > 0$, we use $\mathbb{Z}_\Delta$ to denote the set $\{\Delta n : n \in \mathbb{Z}\}$. For any $x \in \mathbb{R}$, we use $\lfloor x \rfloor_\Delta$ to denote $\max\{y \in \mathbb{Z}_\Delta : y \leq x\}$. We say a function $f : \mathbb{R}^d \to \mathbb{R}$ is separable if there exist functions $f_1, \cdots, f_d : \mathbb{R} \to \mathbb{R}$ such that $f(x_1, \cdots, x_d) = \sum_{i=1}^d f_i(x_i)$. For a logical proposition P, we define $\mathbb{1}[P] = 1$ if P is true and $\mathbb{1}[P] = 0$ if P is false.

2.2 Supervised learning and generalization error

We consider a supervised learning problem. Let $\mathcal{D}$ be an underlying distribution over a sample space $\mathcal{Z}$ with at least two elements, and let $S = (z_1, \cdots, z_n)$ be a dataset of n samples drawn i.i.d. from $\mathcal{D}$.

For a parameter space $\mathbb{R}^d$, a loss function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$, and a learning algorithm $A_f : \mathcal{Z}^n \to \mathbb{R}^d$, we define the *generalization error* of A_f as follows:

$$\varepsilon_g(A_f; \mathcal{D}) \triangleq |\mathbb{E}_{A_f, S=(z_1,\ldots,z_n)}[f(A_f(S); S) - f(A_f(S); \mathcal{D})]|$$

where $f(w; \mathcal{D}) \triangleq \mathbb{E}_{z\sim\mathcal{D}}[f(w; z)]$ and $f(w; S) \triangleq \frac{1}{n}\sum_{i=1}^n f(w; z_i)$ denote the population loss and the empirical loss, respectively. Throughout this paper, we primarily focus on loss functions such that for any $z \in \mathcal{Z}$, $f(\cdot, z)$ is differentiable, convex, Lipschitz, and smooth.

Definition 1. $f : \mathbb{R}^d \to \mathbb{R}$ is convex if for any $u, v \in \mathbb{R}^d$, $f(u) \geq f(v) + \langle \nabla f(v), u - v\rangle$.

Definition 2. $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz if for any $u, v \in \mathbb{R}^d$, $|f(u) - f(v)| \leq L\|u - v\|_2$.

Definition 3. $f : \mathbb{R}^d \to \mathbb{R}$ is M-smooth if for any $u, v \in \mathbb{R}^d$, $\|\nabla f(u) - \nabla f(v)\|_2 \leq M\|u - v\|_2$.

2.3 Gradient descent and stochastic gradient descent with rounding operations

Gradient descent (GD) and stochastic gradient descent (SGD) are iterative learning algorithms defined as follows: given a learning rate $\eta > 0$, initial parameter $w_0 \in \mathbb{R}^d$, total number of iterations T,

$$A^{\text{GD}}_{f,\theta}(S) = G^{\text{GD}}_{f,\eta,S,T} \circ \cdots \circ G^{\text{GD}}_{f,\eta,S,1}(w_0) \text{ and } A^{\text{SGD}}_{f,\theta}(S) = G^{\text{SGD}}_{f,\eta,S,T} \circ \cdots \circ G^{\text{SGD}}_{f,\eta,S,1}(w_0) \quad (1)$$

where $\theta = (\eta, T, w_0)$,

$$G^{\text{GD}}_{f,\eta,S,t}(w) = w - \eta \cdot \nabla f(w; S), \text{ and } G^{\text{SGD}}_{f,\eta,S,t}(w) = w - \eta \cdot \nabla f(w; z_{i_t}) \quad (2)$$

for all $t \in [T]$. Here, i_t is a random variable following the uniform distribution over $[n]$.

Given a discrete grid $\mathbb{Z}^d_\Delta$, we define deterministic rounding DR and stochastic rounding SR that maps $\mathbb{R}^d$ to $\mathbb{Z}^d_\Delta$ as follows:

$$\mathsf{DR}(x)_j \in \arg\min_{y\in\mathbb{Z}_\Delta} |y - x_j|, \quad \mathsf{SR}(x)_j = \lfloor x_j \rfloor_\Delta + \Delta \cdot \mathbb{1}\left[u_j \leq \frac{x_j - \lfloor x_j \rfloor_\Delta}{\Delta}\right] \quad (3)$$

where $\mathsf{DR}(x)_j$ is the even multiple of Δ if $\min_{y\in\mathbb{Z}_\Delta} |y - x_j| = 0.5\Delta$ and u_j is a random variable following the uniform distribution over $[0, 1]$. We note that choosing an arbitrary $\mathsf{DR}(x)_j$ in $\arg\min_{y\in\mathbb{Z}_\Delta} |y - x_j|$ does not affect our main results as long as it is deterministic.

Given $\theta = (\eta, T, w_0 \in \mathbb{Z}^d_\Delta)$, we define GD with deterministic rounding onto $\mathbb{Z}^d_\Delta$ (DR-GD) as

$$A^{\text{DR-GD}}_{f,\theta}(S) = G^{\text{DR-GD}}_{f,\eta,S,T} \circ \cdots \circ G^{\text{DR-GD}}_{f,\eta,S,1}(w_0) \text{ where } G^{\text{DR-GD}}_{f,\eta,S,t}(w) = w - \mathsf{DR}(\eta \cdot \nabla f(w; S)).$$

We similarly define GD with stochastic rounding (SR-GD) and SGD with deterministic and stochastic rounding (DR-SGD and SR-SGD, respectively). For SR-GD and SR-SGD, SR at the t-th iteration uses fresh independent random variables $u_{t,1}, \cdots, u_{t,d}$ in Eq. (3).

2.4 Algorithmic stability

Algorithmic stability is a framework for bounding the generalization error of a learning algorithm based on its sensitivity to dataset perturbations. Specifically, if a learning algorithm A is *symmetric* (i.e., the distribution of $A(S)$ does not depend on the order of the samples in S), then we can represent the generalization error as stated in the following lemma.

Lemma 1 (Lemma 7 in [Bousquet and Elisseeff, 2002])**.** *Let $\mathcal{D}$ be a distribution over $\mathcal{Z}$, $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ be a loss function, $S = (z_1, \cdots, z_n)$ and $S' = (z'_1, z_2, \cdots, z_n)$ be neighboring datasets where $z'_1, z_1, \ldots, z_n$ are drawn from $\mathcal{D}$ independently. If A_f is symmetric, then*

$$\varepsilon_g(A_f; \mathcal{D}) = \left|\mathbb{E}_{A_f, S', z_1}\left[f(A_f(S); z_1) - f(A_f(S'); z_1)\right]\right|.$$

From Lemma 1, one can observe that the *uniform stability* of A_f defined as

$$\varepsilon_{us}(A_f) = \sup_{S\simeq S', z} \mathbb{E}_{A_f}[f(A_f(S); z) - f(A_f(S'); z)]$$

upper bounds the generalization error. In particular, if the loss function is L-Lipschitz, a typical process to bound the uniform stability is to bound the *uniform argument stability* defined as follows:

$$\varepsilon_{uas}(A_f) = \sup_{S\simeq S'} \mathbb{E}_{A_f}[\|A_f(S) - A_f(S')\|_2].$$

For randomized algorithms, the expectation here is taken under *coupled* randomness: for alg $\in \{\mathsf{SGD}, \mathsf{DR\text{-}SGD}, \mathsf{SR\text{-}SGD}\}$, $A^{\text{alg}}_{f,\theta}(S)$ and $A^{\text{alg}}_{f,\theta}(S')$ share the same random indices $i_1, \cdots, i_T$ in Eq. (2), and for alg $\in \{\mathsf{SR\text{-}GD}, \mathsf{SR\text{-}SGD}\}$, stochastic rounding operations use the same random variable $u_{t,j}$ for each $t \in [T]$ and $j \in [d]$ in Eq. (3). We note that this coupling is used only for our upper bounds for $\varepsilon_{uas}(A_f)$; our lower bounds hold for any joint distribution of $A_f(S)$ and $A_f(S')$.

Then, the following relation between the generalization error, uniform stability, and uniform argument stability holds:

$$\varepsilon_g(A_f; \mathcal{D}) \le \varepsilon_{us}(A_f) \le L \cdot \varepsilon_{uas}(A_f).$$

We say $A^{\text{alg}}_{f,\theta}$ is *non-expansive* if for any $u, v \in \mathbb{Z}^d_\Delta$ and $t \in [T]$, there exists a joint distribution p of $G^{\text{alg}}_{f,\eta,S,t}(u) \triangleq u'$ and $G^{\text{alg}}_{f,\eta,S,t}(v) \triangleq v'$ such that $\mathbb{E}_p[\|u' - v'\|_2] \le \|u - v\|_2$.

3 Generalization error and stability of DR-GD and DR-SGD

In this section, we formally present our bounds on generalization error, uniform stability, and uniform argument stability of gradient methods over $\mathbb{Z}^d_\Delta$ with deterministic rounding schemes. In our theoretical results, we use asymptotic notation to hide universal constants. Specifically, for nonnegative functions f and g of n, T, d, η, Δ, M, and L, we write $f \lesssim g$ if there exists a universal constant $C > 0$ such that $f \le C \cdot g$. We write $f \gtrsim g$ if $g \lesssim f$, and $f \asymp g$ if both $f \lesssim g$ and $f \gtrsim g$ hold. We use $f = O(g)$, $f = \Omega(g)$, and $f = \Theta(g)$ to denote $f \lesssim g$, $f \gtrsim g$, and $f \asymp g$, respectively.

We first provide upper and lower bounds on the generalization error of DR-GD. The proofs of Theorems 2 and 3 are in Appendices A.1 and A.2, respectively.

Theorem 2. *Let $m, n, d, T \in \mathbb{N}$, $\eta, \Delta, L > 0$, $w_0 \in \mathbb{Z}^d_\Delta$, $\mathcal{D}$ be a distribution over a sample space $\mathcal{Z}$ of size m, and $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ be an L-Lipschitz loss function. Then, for $\theta = (\eta, T, w_0)$,*

$$\varepsilon_g(A^{\mathsf{DR\text{-}GD}}_{f,\theta}; \mathcal{D}) \le 2\eta L^2 T \sqrt{\frac{m-1}{n}}.$$

Theorem 3. *Let $c > 0.5$ be a fixed constant, $\eta, \Delta, L > 0$ with $\eta L > c\Delta$, $n, T \in \mathbb{N}$, $\mathcal{Z}$ be a finite sample space, and $\mathcal{D}$ be the uniform distribution on $\mathcal{Z}$. There exists an L-Lipschitz, affine loss $f : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$ such that for $\theta = (\eta, T, w_0 = 0)$,*

$$\varepsilon_g(A^{\mathsf{DR\text{-}GD}}_{f,\theta}; \mathcal{D}) \gtrsim \max\left\{\frac{\Delta L T}{\sqrt{n}}, \frac{\eta L^2 T}{n}\right\}.$$

Theorem 2 shows that the generalization error of DR-GD is $O(\eta L^2 T\sqrt{(m-1)/n})$ when the size of the sample space is m; here, the $n^{-1/2}$ rate cannot be improved by Theorem 3. These results indicate that DR-GD and GD have fundamentally different generalization properties since a stability-based approach can achieve $\varepsilon_g(A^{\mathsf{GD}}_{f,\theta}; \mathcal{D}) = O(\eta L^2 T/n)$ [Hardt et al., 2016, Chen et al., 2018], i.e., the deterministic rounding onto $\mathbb{Z}_\Delta$ induces worse sample complexity. We note that our lower-bound construction requires $\eta L > \Delta/2$, since L-Lipschitzness forces $\mathsf{DR}(\eta \nabla f(w; z)) = 0$ when $\eta L < \Delta/2$. The $\Delta L T/\sqrt{n}$ term dominates when $\Delta \gg \eta L/\sqrt{n}$, and the $\eta L^2 T/n$ term dominates when $\Delta \ll \eta L/\sqrt{n}$, which matches to the existing uniform stability bound $O(\eta L^2 T/n)$ [Hardt et al., 2016, Chen et al., 2018].

We further note that the $n^{-1/2}$ dependence on the sample size in Theorem 3 also extends to the continuous sample space $\mathcal{Z} = [0, 1]$; see Appendix C.1. Moreover, our one-dimensional lower-bound constructions also extend to any dimension $d > 1$ by defining $g(w_1, \cdots, w_d; z) = f(w_1; z)$, so that the loss depends only on the first coordinate.

We now sketch the proof of Theorem 3. For simplicity, we consider a data space $\mathcal{Z} = \{0, 1\}$ and a loss function $f(w, z) = -zLw$ for $z \in \mathcal{Z}$; the full proof in Appendix A.2 extends this construction to the uniform distribution on a finite $\mathcal{Z}$. In our proof, we show that for some neighboring datasets $S \simeq S'$ where S' contains more ones compared to S, the DR-GD iterate for S' moves strictly faster than that for S. That is, we have $\mathsf{DR}(\eta \nabla f(w_t; S)) > \mathsf{DR}(\eta \nabla f(w'_t; S'))$, which implies $|w'_T - w_T| \ge \Delta T$, and hence, $|f(A^{\mathsf{DR\text{-}GD}}_{f,\theta}(S); 1) - f(A^{\mathsf{DR\text{-}GD}}_{f,\theta}(S'); 1)| \ge L\Delta T$. Here, we choose the underlying data distribution over $\mathcal{Z}$ so that such an event occurs with probability $\Omega(1/\sqrt{n})$,

which results in the $\Omega(\Delta LT/\sqrt{n})$ lower bound. We note that such an observation only holds for GD with deterministic rounding: for GD, SGD, DR-SGD, SR-GD, and SR-SGD, $\mathbb{E}[|w'_T - w_T|] \to 0$ as $n \to \infty$ (see Section 4 for the SR-GD and SR-SGD cases). We present the full proofs of Theorems 2 and 3 in Appendices A.1 and A.2, respectively.

We next show that the bound in Theorem 2 cannot be achieved by using the uniform stability. While the bound in Theorem 2 converges to zero as $n \to \infty$, the following theorem shows that it is not the case for the uniform stability of DR-GD. In particular, the uniform stability can be independent of n but can increase with T at the same time. The proof of Theorem 4 is presented in Appendix A.3.

Theorem 4. *Let $\mathcal{Z}$ be a sample space, $n, T \in \mathbb{N}$, $\eta, L, \Delta > 0$ with $\eta L > 0.5\Delta$. Then, there exists an L-Lipschitz, affine loss function $f : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$ such that for $\theta = (\eta, T, w_0 = 0)$,*

$$\varepsilon_{us}(A^{\text{DR-GD}}_{f,\theta}) \gtrsim \max\left\{\Delta LT, LT\left(\frac{\eta L}{n} - \Delta\right)\right\}.$$

Theorem 4 shows the existence of an affine function f with $|f'(x; z)| \leq L$ such that the uniform stability-based approach can only produce a vacuous bound $\Omega(\Delta LT)$. As in Theorem 3, the bound in Theorem 4 naturally extends to a general input dimension $d > 1$. In this lower bound, the $LT(\eta L/n - \Delta)$ term dominates when $\Delta \ll \eta L/n$, whereas the ΔLT term dominates when $\Delta \gg \eta L/n$, subject to $\eta L > 0.5\Delta$.

The main proof idea behind Theorem 4 is almost identical to that of Theorem 3: we also use the counterexample loss function $f = -zLw$ and two datasets $S \simeq S'$ such that $\mathsf{DR}(\eta \nabla f(w_t; S)) > \mathsf{DR}(\eta \nabla f(w'_t; S'))$. However, unlike the $\Omega(\Delta LT/\sqrt{n})$ bound in Theorem 3, the bound $\Omega(\Delta LT)$ in Theorem 4 does not contain the $1/\sqrt{n}$ factor. This is because the definition of the uniform stability ε_{us} does not depend on the underlying data distribution, which incurs the $1/\sqrt{n}$ factor in Theorem 3 (see our discussions after Theorem 3 and Appendix A.2). See Appendix A.3 for the full proof.

Bassily et al. [2020] also derived the lower bound $\Omega(\eta LT/n + \eta L\sqrt{T})$ on the uniform argument stability of GD for high-dimensional non-smooth and convex losses, by exploiting the expansiveness of GD for such losses. Specifically, they designed loss functions so that at the t-th iteration, one GD iterate using S moves $\Theta(\eta)$ along the t-th coordinate, while the other GD iterate using $S' \simeq S$ stays at the origin. Hence, the distance between two iterates scales with $\sqrt{T}$. We note that DR-GD is non-expansive for convex and smooth losses if $d = 1$ (see Lemma 16 in Appendix A.4), but the rounding operation incurs a similar effect so that the bound in Theorem 4 follows. In addition, while there is a clear gap between the generalization error and the uniform stability of DR-GD (Theorems 2 and 4), to our knowledge, such an observation has not been made for GD with convex, Lipschitz, but non-smooth losses due to the lack of tight bounds on the generalization error.

On the other hand, the uniform stability of DR-SGD converges to zero as $n \to \infty$. We present the proof of Theorem 5 in Appendix A.4.

Theorem 5. *Let $\mathcal{Z}$ be a sample space, $n, T \in \mathbb{N}$, $\Delta, L, \eta, M > 0$ with $\eta < 2/M$, $w_0 \in \mathbb{Z}_\Delta$, and $f : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$ be a convex, L-Lipschitz, and M-smooth loss function. Then, for $\theta = (\eta, T, w_0)$,*

$$\varepsilon_{us}(A^{\text{DR-SGD}}_{f,\theta}) \leq \frac{(2\eta L + \Delta)LT}{n}.$$

Theorem 5 states that when $d = 1$, the uniform stability of DR-SGD for convex, Lipschitz, and smooth losses is upper bounded by $(2\eta L + \Delta)LT/n$. As $\Delta \to 0$, this bound converges to the existing bound $O(\eta L^2 T/n)$ for SGD [Hardt et al., 2016].

Unlike generalization bounds $\Theta(\eta L^2 T/n)$ for GD and SGD with a convex, Lipschitz, and smooth loss [Hardt et al., 2016, Chen et al., 2018], these results show that there is a fundamental difference between DR-GD and DR-SGD in terms of the sample complexity: for convex, Lipschitz, and smooth loss function, we have $\varepsilon_g(A^{\text{DR-GD}}_{f,\theta}) \geq \Omega(\Delta LT/\sqrt{n})$ in general but $\varepsilon_g(A^{\text{DR-SGD}}_{f,\theta}) \leq O((\eta L + \Delta)LT/n)$ when $d = 1$. Such a difference is due to the stochasticity in DR-SGD. Specifically, for coupled DR-SGD iterates for neighboring datasets $S \simeq S'$, at each iteration, they choose the same loss function with probability $(n-1)/n$. Hence, the coupled SGD iterates are non-expansive with probability $(n-1)/n$ before rounding, which also implies that they are non-expansive after rounding since $d = 1$. Namely, the distance between them can increase with probability $1/n$.

We also show the tightness of the bound in Theorem 5. The proof of Theorem 6 is in Appendix A.5.

Theorem 6. *Let $\mathcal{Z}$ be a finite sample space, $n, T \in \mathbb{N}$ and $\Delta, L, \eta > 0$ with $\eta L > 0.5\Delta$, let $\mathcal{D}$ be a uniform distribution on $\mathcal{Z}$. Then there exists an L-Lipschitz affine loss function $f : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$ such that for $\theta = (\eta, T, w_0 = 0)$,*

$$\varepsilon_g(A_{f,\theta}^{\text{DR-SGD}}; \mathcal{D}) \gtrsim \frac{(2\eta L - \Delta)LT}{n}.$$

However, the uniform stability bound linear in T in Theorem 5 does not extend to general $d > 1$. Our next result shows that the uniform stability of DR-SGD can increase with T^2 when $T/n \leq 1$. We present the proof of Theorem 7 in Appendix A.6.

Theorem 7. *Let $\mathcal{Z}$ be a sample space, $n, d, T \in \mathbb{N}$ with $n, d, T \geq 2$, and $\Delta > 0$. Then, there exists a convex, Δ-Lipschitz and $\left(\frac{9}{10T^2} + \frac{8}{5}\right)$-smooth loss function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ such that for $\theta = (\eta = 1, T, w_0 = \mathbf{0}_d)$,*

$$\varepsilon_{us}(A_{f,\theta}^{\text{DR-SGD}}) \gtrsim \min\left\{1, \frac{T}{n}\right\} \cdot \Delta^2 T.$$

Theorem 7 also implies the difference between the generalization properties of SGD and DR-SGD when $d > 1$ since $\varepsilon_{us}(A_{f,\theta}^{\text{SGD}}) = O(\eta L^2 T/n)$ [Hardt et al., 2016]. Here, the main difference between SGD and DR-SGD is that when $d > 1$, DR-SGD can be expansive due to the rounding scheme, even if a loss function is convex and smooth. Using this property, we design a counterexample loss function $f : \mathbb{R}^d \times \{0, 1\} \to \mathbb{R}$ that satisfies $\mathsf{DR}(\eta \nabla f(\mathbf{0}_d; 0)) = \mathbf{0}_d$ and $\mathsf{DR}(\eta \nabla f(\mathbf{0}_d; 1)) = u \neq 0$, i.e., DR-SGD iterate never moves with dataset $S = (0, \ldots, 0)$ but it can move with $S' = (1, 0, \ldots, 0) \simeq S$, with probability $1/n$. Here, we design f so that once the DR-SGD iterate touches u, then for each iteration, the distance between the iterate and the origin increases by $\Theta(\Delta)$. Here, the $\Theta(T/n)$ term of the bound in Theorem 7 can be considered as the probability that one is sampled from S' in the first $T/2$ DR-SGD iterations. Once one is sampled, the distance between the T-th DR-SGD iterate for S' and the origin will be $\Theta(\Delta T)$, which results in a $\Theta(\Delta^2 T)$ difference in the loss function values $f(A_{f,\theta}^{\text{DR-SGD}}(S); 0)$ and $f(A_{f,\theta}^{\text{DR-SGD}}(S'); 0)$. For the full proof, see Appendix A.6.

We next show that the bound in Theorem 7 is tight in general, by providing a matching upper bound in the following theorem. The proof of Theorem 8 is presented in Appendix A.7.

Theorem 8. *Let $\mathcal{Z}$ be a sample space, $n, d, T \in \mathbb{N}$, $\Delta, L, \eta > 0$, $w_0 \in \mathbb{Z}_\Delta^d$, and $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ be an L-Lipschitz loss function. Then, for $\theta = (\eta, T, w_0)$,*

$$\varepsilon_{us}(A_{f,\theta}^{\text{DR-SGD}}) \leq \min\left\{1, \frac{T}{n}\right\} \cdot 8\eta L^2 T.$$

We remark that for SGD with non-smooth and convex losses, the uniform argument stability is $\Theta(\eta LT/n + \min\{1, T/n\}\eta L\sqrt{T})$ [Bassily et al., 2020]. This bound is also derived using the expansiveness of counterexample losses. While the distance between the coupled DR-SGD iterates can increase by Δ per iteration, the squared distance between the coupled SGD iterates can increase by $\Theta(\eta^2)$, yielding the $T^{3/2}$ factor in the bound.

4 Generalization error and stability of SR-GD and SR-SGD

In this section, we present our bounds for SR-GD and SR-SGD. We first provide the lower bound on the generalization error of SR-GD and SR-SGD. The proof of Theorem 9 is in Appendix B.1.

Theorem 9. *Let $\eta, L, \Delta > 0$ and $n, d \in \mathbb{N}$ with $n \geq 2$ and $\eta L < \Delta\sqrt{d}$. Let $\mathcal{Z}$ be a finite sample space, and let $\mathcal{D}$ be uniform on $\mathcal{Z}$. There exists a convex, L-Lipschitz, and $(4L/\Delta)$-smooth loss function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ such that for $\theta = (\eta, T = 1, w_0 = \mathbf{0}_d)$,*

$$\varepsilon_g(A_{f,\theta}^{\text{SR-GD}}, \mathcal{D}), \varepsilon_g(A_{f,\theta}^{\text{SR-SGD}}, \mathcal{D}) \gtrsim \frac{\sqrt{\min\{\Delta, \eta L d^{1/2}\}\eta L^3 d^{1/2}}}{n}.$$

Theorem 9 shows that the generalization error of SR-GD and SR-SGD is inherently dependent on the ambient dimension: there exist a data distribution $\mathcal{D}$ and a convex, L-Lipschitz, and $(4L/\Delta)$-smooth loss function f such that the generalization error of SR-GD and SR-SGD with $T = 1$ scale with (at

least) $d^{1/4}$. This result shows that the generalization properties of SR-GD and SR-SGD fundamentally differ from those of GD and SGD, which are independent of d [Hardt et al., 2016]. We note that both lower bounds also hold for the uniform distribution on $\mathcal{Z} = [0, 1]$; see Appendix C.1.

Here, the dimension dependency in Theorem 9 is from the stochastic rounding. For GD and SGD, the distance between coupled iterates for neighboring datasets increases at most $O(\eta L/n)$. Then, after T iterations, the distance can be bounded by $O(\eta LT/n)$, and hence, the difference in the function value can be bounded by $O(\eta L^2 T/n)$. However, if we apply the stochastic rounding to two points with distance ℓ, the resulting distance can be $\Omega(\ell\sqrt{d})$ in expectation (e.g., consider one point is $\mathbf{0}_d$ and the other point is $(\ell/\sqrt{d}, \dots, \ell/\sqrt{d})$). Motivated by this observation, we design the loss function and a data distribution so that the generalization error scales with d. See Appendix B.1 for the full proof.

We now shift our focus to the stability bounds of SR-GD and SR-SGD for general $T > 1$. To this end, we present an upper bound of the uniform argument stability of SR-SGD. The proof of Theorem 10 is presented in Appendix B.2.

Theorem 10. *Let $\mathcal{Z}$ be a sample space, $n, d, T \in \mathbb{N}$, $\Delta, \eta, L, M > 0$ with $\eta < 2/M$, $w_0 \in \mathbb{Z}_\Delta^d$, and $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ be convex, L-Lipschitz, and M-smooth. Then, for $\theta = (\eta, T, w_0)$,*

$$\varepsilon_{uas}(A_{f,\theta}^{\text{SR-SGD}}) \le \frac{8\eta LT}{n} + \min\left\{1, \frac{T}{n}\right\} \cdot \sqrt{2\Delta\eta L d^{1/2} T}.$$

Theorem 10 provides an upper bound on $\varepsilon_{uas}(A_{f,\theta}^{\text{SR-SGD}})$ that scales with $d^{1/4}$; such dimension dependence is unavoidable in general as shown by Theorem 9. Unlike the bound for SGD, which scales with T, our bound scales with $T^{3/2}$ as in the uniform argument stability bound for SGD on a non-smooth convex loss: $\varepsilon_{uas}(A_{f,\theta}^{\text{SGD}}) = O(\min\{1, T/n\}\eta L\sqrt{T} + \eta LT/n)$ [Bassily et al., 2020]. This follows from the potential expansiveness of SR-SGD even for convex and smooth losses. Here, the $\min\{1, T/n\}$ term corresponds to the probability that different examples (i.e. z_1 and z_1') are selected from neighboring datasets during the coupled SGD iterates, and the $\Theta(\sqrt{\Delta\eta L d^{1/2} T})$ term is the accumulated distance between two coupled SR-SGD iterates after the differing samples are selected. We also note that this bound recovers the uniform argument stability bound for SGD as $\Delta \to 0$. For the full proof of Theorem 10, see Appendix B.2.

We also provide an upper bound on the uniform argument stability of SR-GD in the following theorem, whose proof is presented in Appendix B.3.

Theorem 11. *Let $\mathcal{Z}$ be a sample space, $n, d, T \in \mathbb{N}$, $\Delta, \eta, L, M > 0$ with $\eta < 2/M$, $w_0 \in \mathbb{Z}_\Delta^d$, and $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ be a convex, L-Lipschitz and M-smooth loss function. Then, for $\theta = (\eta, T, w_0)$,*

$$\varepsilon_{uas}(A_{f,\theta}^{\text{SR-GD}}) \le \frac{6\eta LT}{n} + \min\left\{1, \frac{2\eta L\sqrt{d}T}{n\Delta}\right\} \cdot \sqrt{\Delta^2 + 2\Delta\eta L d^{1/2}\left(\frac{1}{n} + T - 1\right)}.$$

Unlike the lower bound on the uniform stability of DR-GD (Theorem 4), the bound in Theorem 11 converges to zero as $n \to \infty$ when Δ is independent of n. In DR-GD, even an $O(1/n)$ change in the gradient of empirical loss can deterministically move the update across a rounding threshold, producing a separation of order Δ between two iterates on neighboring datasets even after the first iterate (see our discussion after Theorem 4). In contrast, under stochastic rounding, a small change in the gradient of empirical loss causes different rounded values only with probability proportional to the magnitude of that change relative to Δ. Therefore, the probability that two SR-GD iterates first separate over T iterations is of order $\min\{1, \eta L\sqrt{d}T/(n\Delta)\}$, yielding the sample-dependent bound in Theorem 11. Here, the following term $\Theta(\sqrt{\Delta^2 + \Delta\eta L d^{1/2} T})$ corresponds to the accumulated distance between two SR-GD iterates after the first separation. Specifically, the squared distance between the two SR-GD iterates is at most $O(\Delta^2 + \Delta\eta L\sqrt{d}/n)$ at the first separation, and then increases at most $O(\Delta\eta L\sqrt{d})$ per iteration. See Appendix B.3 for more details.

The bound in Theorem 11 can be interpreted differently depending on the value of Δ. For $\Delta \lesssim \Delta_1 = \Theta(\eta L\sqrt{d}T/n)$, the bound becomes $\varepsilon_{uas}(A_{f,\theta}^{\text{SR-GD}}) = O(\eta LT/n + \sqrt{\Delta\eta L d^{1/2} T})$. For $\Delta_1 \lesssim \Delta \lesssim \Delta_2 = \Theta(\eta L\sqrt{d}T)$, the bound becomes $O(\eta^{3/2} L^{3/2} d^{3/4} T^{3/2}/(n\Delta^{1/2}))$, which can scale with $T^{3/2}$, e.g., for constant Δ when $\Delta_1 = O(1)$. If $\Delta \gtrsim \Delta_2$, the bound becomes $O(\eta L\sqrt{d}T/n)$. As in the SR-SGD case, we recover the uniform argument stability bound for GD as $\Delta \to 0$.

While the bounds in Theorems 10 and 11 can scale with $T^{3/2}/n$, our next results show that the bounds can be relaxed to T/n (or $\sqrt{T/n}$) if losses are separable (i.e., sum of coordinate-wise functions) as in the following theorem. The proof of Theorem 12 is presented in Appendix B.4.

Theorem 12. *Let $\mathcal{Z}$ be a sample space, $d, n, T \in \mathbb{N}$, $\Delta, \eta, M, L > 0$ such that $\eta < 1/M$, $w_0 \in \mathbb{Z}_\Delta^d$, and $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ be a convex, L-Lipschitz, M-smooth, and separable loss function. Then, for $\theta = (\eta, T, w_0)$,*

$$\varepsilon_{uas}(A_{f,\theta}^{\text{SR-GD}}) \le \frac{6\eta LT}{n} + \sqrt{\frac{2\Delta\eta L d^{1/2} T}{n}}, \quad \varepsilon_{uas}(A_{f,\theta}^{\text{SR-SGD}}) \le \frac{2\eta LT}{n} + \frac{T\sqrt{2\Delta\eta L d^{1/2}}}{n}.$$

Unlike the potential expansiveness of SR-GD and SR-SGD for convex and smooth losses, they are non-expansive if the losses are additionally separable. By exploiting this property, we can better bound the accumulated distance between coupled iterates, and derive the bounds in Theorem 12.

We lastly show that the bounds in Theorem 12 are tight in general, as stated in the following theorem. The proof of Theorem 13 is in Appendix B.5.

Theorem 13. *Let $\mathcal{Z}$ be a sample space, $\eta, \Delta, L > 0$ and $n, T, d \in \mathbb{N}$. Then, there exists an L-Lipschitz affine loss function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ such that for $\theta = (\eta, T, w_0 = \mathbf{0}_d)$, if $\eta < n\Delta\sqrt{d}/L$, then*

$$\varepsilon_{uas}(A_{f,\theta}^{\text{SR-GD}}) \gtrsim \frac{\eta LT}{n} + \frac{\sqrt{\min\{n\Delta, \eta L d^{1/2} T\}\eta L d^{1/2} T}}{n}.$$

In addition, if $\eta < \Delta\sqrt{d}/L$, then

$$\varepsilon_{uas}(A_{f,\theta}^{\text{SR-SGD}}) \gtrsim \min\left\{1, \frac{T}{n}\right\}\left(\eta L + \sqrt{\min\{\Delta, \eta L d^{1/2}\}\eta L d^{1/2}}\right).$$

Together with Theorem 12, Theorem 13 can tightly characterize the uniform argument stability of SR-GD and SR-SGD when the loss function is separable. Specifically, for $\eta LT/(n\sqrt{d}) < \Delta \lesssim \eta L\sqrt{d}T/n$, we have $\varepsilon_{uas}(A_{f,\theta}^{\text{SR-GD}}) = \Theta(\sqrt{\Delta\eta L d^{1/2} T/n})$. Likewise, for $T < n$ and $\eta L/\sqrt{d} < \Delta \lesssim \eta L\sqrt{d}$, it holds that $\varepsilon_{uas}(A_{f,\theta}^{\text{SR-SGD}}) = \Theta(\sqrt{\Delta\eta L d^{1/2}}T/n)$. This result shows a qualitative difference between SR-GD and SR-SGD: since SR-SGD contains the extra $\sqrt{T/n}$ compared to SR-GD, SR-SGD is more favorable in sample complexity, but less favorable in its dependence on T.

5 Discussions

Varying learning rates. Our results can be extended to a non-constant learning rate schedule $(\eta_t)_{t=1}^T$. Writing $H = \sum_{t=1}^T \eta_t$, the bounds in Theorems 2, 5 and 8 become $2L^2 H\sqrt{(m-1)/n}$, $(L/n)\sum_{t=1}^T(2\eta_t L + \Delta)$, and $8L^2 H\min\{1, T/n\}$, respectively. For stochastic rounding, we give explicit upper bounds in terms of H, $\sum_t \eta_t^2$, and $\max_t \eta_t$. Formal statements and proofs are given in Appendix D.

Extension to non-convex loss functions. Although our main results are stated for convex, Lipschitz, and smooth losses, our stability upper bound results can be extended beyond this class. In particular, the proofs of Theorems 5, 8 and 10–12 only use the loss values and gradients at the points in $\mathbb{Z}_\Delta^d$ that can be visited by the rounded iteration. Hence, these results may also apply to non-convex losses whose gradients on $\mathbb{Z}_\Delta^d$ can be interpolated by those of a convex, Lipschitz, and smooth function. This interpolation condition can be checked using known characterizations of smooth convex interpolation. For example, Taylor et al. [2017] show that for any finite collection $\{(x_i, g_i, f_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$, where each x_i is an interpolation point with gradient g_i and value f_i, there exists a convex and M-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ satisfying $f(x_i) = f_i$ and $\nabla f(x_i) = g_i$ for all $i \in [n]$ if and only if

$$f_i \ge f_j + g_j^\top(x_i - x_j) + \frac{1}{2M}\|g_i - g_j\|_2^2 \text{ for all } i \text{ and } j.$$

Extension to floating-point numbers. Unlike fixed-point numbers, floating-point numbers form a non-uniform grid: the spacing between representable numbers depends on the exponent. Thus,

the magnitude of rounding errors depends on the scale of the iterates themselves. Nevertheless, our analysis may extend to floating-point arithmetic in regimes where the iterates remain within a fixed exponent range. In such a local regime, the set of representable numbers behaves similarly to fixed-point numbers. However, extending our results to the full floating-point setting requires controlling changes in the exponent along the optimization trajectory, as well as the resulting scale-dependent rounding errors. Developing a stability and generalization theory that captures these effects is an important direction for future work.

Limitations.

Since the uniform stability only gives an upper bound on generalization error, large stability does not imply large generalization error. For example, the generalization upper bound for DR-GD in Theorem 2 vanishes as $n \to \infty$ with the other parameters fixed, even though its uniform stability need not vanish (See Theorem 4). Therefore, our stability lower bounds in Theorems 4, 7 and 13 do not imply generalization lower bounds, but they show the limitation of stability-based analyses. In contrast, Theorems 3 and 9 show that rounding can worsen generalization for particular losses and data distributions in the worst case. Nevertheless, these results do not imply that rounding worsens generalization in every learning problem.

In addition, our theoretical analysis isolates parameter rounding in idealized GD/SGD updates (see Section 2.3). Accordingly, it may not cover several ingredients in real implementations. First, our model assumes that the gradient of the real-valued loss is computed exactly. However, automatic differentiation over machine-representable numbers can exhibit incorrect behavior even when the loss is differentiable due to the round-off error in the intermediate computations. Second, modern low-precision training often relies on quantization-aware training [Jacob et al., 2018], which maintains high-precision parameters while training a model with fake-quantization operators in the forward pass. Thus, unlike our setup, it does not directly optimize over a discrete parameter space.

6 Conclusion

In this work, we analyze the generalization error and uniform stability of GD and SGD with deterministic and stochastic rounding onto fixed-point spaces, on convex, Lipschitz, and smooth losses. We first show that deterministic rounding can degrade the generalization error to $\Theta(T/\sqrt{n})$ for a constant-size sample space, with a vacuous $\Omega(T)$ uniform stability bound. In contrast, SGD with deterministic rounding yields tight, dimension-dependent bounds: $O(T/n)$ for one dimension and $O(T^2/n)$ for higher dimensions, which is qualitatively different from SGD without rounding. Furthermore, we show that both for GD and SGD with stochastic rounding, the generalization error can scale with the ambient dimension. We also provide the upper and lower bounds on uniform argument stability, which are tight when the loss function can be represented as a sum of coordinate-wise functions. We believe our results contribute to bridging the gap between theoretical optimization and the real implementation.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant (No. RS-2024-00415812, RS-2025-24534003), Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No. RS-2019-II190079) funded by the Korea government (MSIT).

References

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

Vitaly Feldman. Generalization of ERM in stochastic convex optimization: the dimension strikes back. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

David Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, 23(1):5–48, 1991.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *International Conference on Machine Learning (ICML)*, 2015.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, 2016.

Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Ilia Markov, Adrian Vladu, Qi Guo, and Dan Alistarh. Quantized distributed training of large models with convergence guarantees. In *International Conference on Machine Learning (ICML)*, 2023.

Han V. Nguyen, Hai Pham, Sashank J. Reddi, and Barnabás Póczos. On the algorithmic stability and generalization of adaptive optimization methods. *arXiv preprint arXiv:2211.03970*, 2022.

Konstantinos E. Nikolakakis, Amin Karbasi, and Dionysis Kalogerias. Select without fear: Almost all minibatch schedules generalize optimally. *SIAM Journal on Mathematics of Data Science*, 7(3): 965–992, 2025.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations (ICLR)*, 2018.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010. ISSN 1532-4435.

Adrien B Taylor, Julien M Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161: 307–345, 2017.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSProp: Divide the gradient by a running average of its recent magnitude, 2012. COURSERA: Neural networks for machine learning, Lecture notes.

Lu Xia, Stefano Massei, and Michiel E. Hochstenbach. On the convergence of the gradient descent method with stochastic fixed-point rounding errors under the polyak-Łojasiewicz inequality. *Computational Optimization and Applications*, 90(3):753–799, 2025.

Jihao Xin, Marco Canini, Peter Richtárik, and Samuel Horváth. Global-qsgd: Allreduce-compatible quantization for distributed learning with theoretical guarantees. In *Proceedings of the 5th Workshop on Machine Learning and Systems*. Association for Computing Machinery, 2025.

Yikai Zhang, Wenjia Zhang, Sammy Bald, Vamsi Pingali, Chao Chen, and Mayank Goswami. Stability of SGD: Tightness analysis and improved bounds. In James Cussens and Kun Zhang, editors, *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022.

Dongruo Zhou, Jinghui Chen, Yuan Cao, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *Transactions on Machine Learning Research*, 2024.

Yingxue Zhou, Belhal Karimi, Jinxing Yu, Zhiqiang Xu, and Ping Li. Towards better generalization of adaptive gradient methods. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

A Proofs for DR-GD and DR-SGD

In this section, we prove the generalization and stability bounds for DR-GD and DR-SGD. We first present the proofs of Theorems 2 and 3, and Theorem 4 for DR-GD in Appendices A.1 and A.2, and Appendix A.3, respectively. We then present the proofs of Theorem 5, Theorem 6, Theorem 7, and Theorem 8 for DR-SGD in Appendix A.4, Appendix A.5, Appendix A.6, Appendix A.7, respectively.

A.1 Proof of Theorem 2

In this section, we prove Theorem 2. Let $p_i = \mathbb{P}_{X\sim\mathcal{D}}(X = i)$ and $\widehat{p}_i = \sum_{j=1}^{n} \mathbb{1}[z_j = i]/n$. To this end, we use the following lemma, whose proof is in Appendix A.1.1.

Lemma 14. *Let $d \in \mathbb{N}$ and $\Delta, R > 0$ and $a \in \mathbb{R}^d$ such that $\|a\|_2 \le R$. Then, it holds that*

$$\|\mathsf{DR}(a)\|_2 \le 2R.$$

We now prove Theorem 2. Let $\widehat{p} = (\widehat{p}_1, \cdots, \widehat{p}_m)$ and set $g(w; z) = f(w; z) - f(w_0; z)$. Subtracting $f(w_0; z)$ changes neither the iterates nor the expected generalization gap. By Lemma 14, $\|w_T - w_0\|_2 \le 2\eta LT$, so $|g(w_T; z)| \le 2\eta L^2 T$. Consequently,

$$\begin{aligned}\varepsilon_g(A^{\text{DR-GD}}_{f,\theta}; \mathcal{D}) &= \left| \mathbb{E}\left[\sum_{j=1}^{m} (\widehat{p}_j - p_j) g(w_T; j)\right]\right| \\ &\le 2\eta L^2 T\, \mathbb{E}[\|\widehat{p} - p\|_1] \le \frac{2\eta L^2 T}{\sqrt{n}} \sum_{j=1}^{m} \sqrt{p_j(1-p_j)}.\end{aligned}$$

The last inequality uses $\mathrm{Var}(\widehat{p}_j) = p_j(1-p_j)/n$ and Jensen's inequality. By the Cauchy–Schwarz inequality, we have

$$\sum_{j=1}^{m} \sqrt{p_j(1-p_j)} \le \sqrt{\left(\sum_j p_j\right)\left(\sum_j (1-p_j)\right)} = \sqrt{m-1},$$

which completes the proof.

A.1.1 Proof of Lemma 14

In this section, we prove Lemma 14. Let $a = (a_1, \cdots, a_d)$. It suffices to show that any $x \in \mathbb{R}$, $|\mathsf{DR}(x)| \le 2|x|$. Then, we have

$$\|\mathsf{DR}(a)\|_2^2 = \sum_{i=1}^{d} |\mathsf{DR}(a_i)|^2 \le \sum_{i=1}^{d} 4|a_i|^2 = (2\|a\|_2)^2 \le (2R)^2$$

which implies $\|\mathsf{DR}(a)\|_2 \le 2R$ and this completes the proof.

Since $|x| < \Delta/2$ is trivial ($\mathsf{DR}(x) = 0$), we assume that $|x| \ge \Delta/2$. Since $|x - \mathsf{DR}(x)| \le \Delta/2$, by the triangle inequality, we have

$$|\mathsf{DR}(x)| \le |x| + \frac{\Delta}{2} \le |x| + |x| = 2|x|$$

which completes the proof.

A.2 Proof of Theorem 3

In this section, we prove Theorem 3. Without loss of generality, we assume that $\mathcal{Z} = [m]$ for $m \geq 2$. We first introduce the following lemma. The proof of Lemma 15 is in Appendix A.2.1.

Lemma 15. *Let $n, T \in \mathbb{N}$, $\mathcal{D}$ be a distribution over a sample space $\mathcal{Z}$, let $\eta, \Delta, L > 0$, and let $A : \mathcal{Z} \to [0, L]$ be a function. For $Z \sim \mathcal{D}$, define $\mu = \mathbb{E}[A(Z)]$. Suppose that there exists $b > 0$ such that*

$$Y = \frac{A(Z) - \mu}{b}, \qquad |Y| \leq 1 \text{ almost surely}, \qquad \mathbb{E}[Y^2] \geq \frac{1}{3}.$$

Suppose further that $\eta\mu \in \Delta(\mathbb{Z} + 1/2)$. Let $f(w; z) = -A(z)w$ for $(w, z) \in \mathbb{R} \times \mathcal{Z}$. Then, for $\theta = (\eta, T, w_0 = 0)$, it holds that

$$\varepsilon_g(A^{\text{DR-GD}}_{f,\theta}; \mathcal{D}) \geq \max\left\{\frac{\Delta b T}{18\sqrt{n}}, \frac{\eta b^2 T}{6n}\right\}.$$

Lemma 15 gives sufficient conditions on the slopes of the linear losses $f(\cdot; z)$ under which the generalization error of DR-GD is $\Omega(\max\{\Delta T/\sqrt{n}, T/n\})$. We now prove Theorem 3 by constructing a loss on $[m]$. Let $R = \eta L$. If $R < 3\Delta/2$, set $q = \Delta/2$. Otherwise, choose

$$q \in \operatorname*{arg\,min}_{u \in \Delta(\mathbb{Z}+1/2)} \left|u - \frac{R}{2}\right|,$$

so that $|q - R/2| \leq \Delta/2$. In the first case, $1/3 < q/R < 1/(2c)$; in the second, $1/6 \leq q/R \leq 5/6$. Thus, defining

$$a = \min\left\{\frac{1}{6}, 1 - \frac{1}{2c}\right\} > 0, \qquad \mu = \frac{q}{\eta}, \qquad b = \min\{\mu, L - \mu\}, \tag{4}$$

we have $aL \leq \mu \leq (1 - a)L$ and $b \geq aL$.

For Z uniform on $[m]$, take

$$A(i) = \mu + b\left(\frac{2(i-1)}{m-1} - 1\right), \qquad f(w; i) = -A(i)w.$$

Then $A(i) \in [0, L]$, and $Y = 2(Z - 1)/(m - 1) - 1$ has mean zero, satisfies $|Y| \leq 1$, and has variance $(m + 1)/(3(m - 1)) \geq 1/3$. Applying Lemma 15 and $b \geq aL$ completes the proof of Theorem 3.

A.2.1 Proof of Lemma 15

In this section, we prove Lemma 15. Let $Z_1, Z_1', Z_2, \cdots, Z_n$ be independent samples from $\mathcal{D}$, and set

$$S = (Z_1, Z_2, \cdots, Z_n), \qquad S' = (Z_1', Z_2, \cdots, Z_n).$$

Write $A_i = A(Z_i)$, $A_1' = A(Z_1')$, and define

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n} A_i, \qquad \overline{A}' = \overline{A} + \frac{A_1' - A_1}{n}, \qquad q = \eta\mu, \qquad Q(u) = -\mathsf{DR}(-\eta u).$$

Since $f(w; z) = -A(z)w$, the empirical gradients are constant in w. Thus, the two trajectories satisfy

$$w_t = tQ(\overline{A}), \qquad w_t' = tQ(\overline{A}'), \qquad t \in \{0, \cdots, T\}.$$

The algorithm is symmetric in the samples, so Lemma 1 gives

$$\begin{aligned}
\varepsilon_g(A^{\text{DR-GD}}_{f,\theta}; \mathcal{D}) &= T\left|\mathbb{E}\left[A_1\big(Q(\overline{A}') - Q(\overline{A})\big)\right]\right| \\
&= \frac{T}{2}\mathbb{E}\left[(A_1 - A_1')\big(Q(\overline{A}) - Q(\overline{A}')\big)\right] \\
&= \frac{T}{2}\mathbb{E}\left[|A_1 - A_1'|\,|Q(\overline{A}) - Q(\overline{A}')|\right].
\end{aligned}$$

Here, the second equality follows by exchanging Z_1 and Z_1', and the last two equalities use the monotonicity of Q and $\overline{A} - \overline{A}' = (A_1 - A_1')/n$.

Since $q \in \Delta(\mathbb{Z} + 1/2)$, we have $Q(\mu) \in \{q - \Delta/2, q + \Delta/2\}$. Define

$$\kappa = \frac{Q(\mu) - (q - \Delta/2)}{\Delta} \in \{0, 1\}, \qquad H(u) = \mathbb{1}[u > \mu] + \kappa \mathbb{1}[u = \mu].$$

This definition includes either deterministic choice at the rounding tie. If $H(u) = 0$, then $Q(u) \leq q - \Delta/2$; if $H(u) = 1$, then $Q(u) \geq q + \Delta/2$. Consequently, on the event

$$\mathcal{E} = \{H(\overline{A}) \neq H(\overline{A}')\},$$

the rounded gradients differ by at least Δ, and hence $|w_T - w_T'| \geq \Delta T$. Substituting this into the average-stability identity yields

$$\varepsilon_g(A_{f,\theta}^{\text{DR-GD}}; \mathcal{D}) \geq \frac{\Delta T}{2} \mathbb{E}\left[|A_1 - A_1'| \mathbb{1}[\mathcal{E}]\right].$$

Since H is nondecreasing and takes values in $\{0, 1\}$,

$$|A_1 - A_1'| \mathbb{1}[\mathcal{E}] = (A_1 - A_1')\big(H(\overline{A}) - H(\overline{A}')\big).$$

By exchanging A_1 and A_1', using the independence of A_1' and $\overline{A}$, and then averaging over the exchangeable variables $A_1, \cdots, A_n$, we obtain

$$\begin{aligned}\mathbb{E}\left[|A_1 - A_1'| \mathbb{1}[\mathcal{E}]\right] &= 2\mathbb{E}\left[(A_1 - \mu) H(\overline{A})\right] \\ &= 2\mathbb{E}\left[(\overline{A} - \mu) H(\overline{A})\right] = \mathbb{E}[|\overline{A} - \mu|].\end{aligned}$$

The last equality uses $\mathbb{E}[\overline{A} - \mu] = 0$ and $(u - \mu)H(u) = (u - \mu)_+$, where $v_+ = \max\{v, 0\}$.

Set $Y_i = (A_i - \mu)/b$ and $\overline{Y} = n^{-1} \sum_{i=1}^n Y_i$, so that $\overline{A} - \mu = b\overline{Y}$. By independence, $\mathbb{E}[Y] = 0$, and $|Y| \leq 1$, we have

$$\begin{aligned}\mathbb{E}[\overline{Y}^2] &= \frac{\mathbb{E}[Y^2]}{n} \geq \frac{1}{3n}, \\ \mathbb{E}[\overline{Y}^4] &= \frac{n\mathbb{E}[Y^4] + 3n(n-1)(\mathbb{E}[Y^2])^2}{n^4} \leq \frac{3}{n^2}.\end{aligned}$$

Hölder's inequality therefore gives

$$\mathbb{E}[|\overline{Y}|] \geq \frac{(\mathbb{E}[\overline{Y}^2])^{3/2}}{(\mathbb{E}[\overline{Y}^4])^{1/2}} \geq \frac{1}{9\sqrt{n}}.$$

Since $|A_1 - A_1'| \leq 2b$ almost surely, it follows that

$$\mathbb{P}(\mathcal{E}) \geq \frac{\mathbb{E}[|\overline{A} - \mu|]}{2b} \geq \frac{1}{18\sqrt{n}}, \qquad \varepsilon_g(A_{f,\theta}^{\text{DR-GD}}; \mathcal{D}) \geq \frac{\Delta T}{2} \mathbb{E}[|\overline{A} - \mu|] \geq \frac{\Delta b T}{18\sqrt{n}}.$$

Thus, the neighboring iterates separate by at least ΔT with probability $\Omega(n^{-1/2})$, and the weighted crossing contribution gives the claimed generalization lower bound.

Let $x = \overline{A} - \mu$. The same symmetrized identity, together with exchangeability and $\mathbb{E}[x] = 0$, gives

$$\varepsilon_g(A_{f,\theta}^{\text{DR-GD}}; \mathcal{D}) = T\mathbb{E}\left[x\big(Q(\mu + x) - q\big)\right].$$

Since $|Q(u) - \eta u| \leq \Delta/2$, we also have

$$\varepsilon_g(A_{f,\theta}^{\text{DR-GD}}; \mathcal{D}) \geq \eta T \mathbb{E}[x^2] - \frac{\Delta T}{2} \mathbb{E}[|x|].$$

Adding this inequality to the crossing bound $\varepsilon_g(A_{f,\theta}^{\text{DR-GD}}; \mathcal{D}) \geq (\Delta T/2)\mathbb{E}[|x|]$ yields

$$\varepsilon_g(A_{f,\theta}^{\text{DR-GD}}; \mathcal{D}) \geq \frac{\eta T}{2} \mathbb{E}[x^2] = \frac{\eta b^2 T}{2} \mathbb{E}[\overline{Y}^2] \geq \frac{\eta b^2 T}{6n}.$$

Combining the two lower bounds completes the proof.

A.3 Proof of Theorem 4

In this section, we prove Theorem 4. Throughout the proof, we fix two distinct samples in $\mathcal{Z}$ and denote them by 0 and 1 for notational simplicity. Let $S = (0, \cdots, 0)$ and $S' = (1, 0, \cdots, 0)$ be neighboring datasets. Since $\eta L > \Delta/2$, we can choose

$$a = \frac{\Delta}{2\eta} + \frac{1}{2}\min\left\{L - \frac{\Delta}{2\eta}, \frac{L}{n}\right\} \in \left(\frac{\Delta}{2\eta}, L\right).$$

We define an L-Lipschitz affine loss by $f(w; 0) = aw$ and $f(w; 1) = -Lw$, with its values at the other samples defined as above. The empirical gradients on S and S' are a and $a' = a - (a + L)/n$, respectively. By the choice of a, we have

$$a' - \frac{\Delta}{2\eta} = a - \frac{\Delta}{2\eta} - \frac{a + L}{n} \leq \frac{L}{2n} - \frac{a + L}{n} < 0.$$

Thus, $\eta a > \Delta/2$ and $\eta a' < \Delta/2$, and hence $\mathsf{DR}(\eta a) - \mathsf{DR}(\eta a') \geq \Delta$. Since the gradients do not depend on w, the two iterates satisfy

$$w'_T - w_T = T\big(\mathsf{DR}(\eta a) - \mathsf{DR}(\eta a')\big) \geq \Delta T.$$

Furthermore, $|\mathsf{DR}(u) - u| \leq \Delta/2$ implies

$$w'_T - w_T \geq T\left(\frac{\eta(a + L)}{n} - \Delta\right) \geq T\left(\frac{\eta L}{n} - \Delta\right).$$

By evaluating the loss at sample 1, we obtain

$$\varepsilon_{us}(A^{\mathsf{DR\text{-}GD}}_{f,\theta}) \geq f(w_T; 1) - f(w'_T; 1) = L(w'_T - w_T) \geq LT\max\left\{\Delta, \frac{\eta L}{n} - \Delta\right\},$$

which completes the proof.

A.4 Proof of Theorem 5

Let $S = (z_1, \cdots, z_n)$ and $S' = (z'_1, \cdots, z'_n) \in \mathcal{Z}^n$ be the neighboring datasets, $\delta_t = |w_t - w'_t|$ where $(w_t)_{t=1}^T$ and $(w'_t)_{t=1}^T$ be the coupled iterates from $A_{f,\theta}^{\text{DR-SGD}}(S)$ and $A_{f,\theta}^{\text{DR-SGD}}(S')$, respectively. Since $\varepsilon_{us}(A_{f,\theta}^{\text{DR-SGD}}) \leq L \cdot \varepsilon_{uas}(A_{f,\theta}^{\text{DR-SGD}})$, it suffices to show that $\mathbb{E}[\delta_T] \leq (2\eta L + \Delta)T/n$. By the law of total expectation, we have

$$\mathbb{E}[\delta_t] = \frac{n-1}{n}\mathbb{E}[\delta_t | z_{i_t} = z'_{i_t}] + \frac{1}{n}\mathbb{E}[\delta_t | z_{i_t} \neq z'_{i_t}].$$

We first consider the upper bound of $\mathbb{E}[\delta_t | z_{i_t} = z'_{i_t}]$. To this end, we use the following lemma:

Lemma 16. *Let $f : \mathbb{R} \to \mathbb{R}$ be a convex, M-smooth function. Let $x, y \in \mathbb{Z}_\Delta$ and $\eta > 0$ with $\eta < 2/M$. Then, it holds that*

$$|x - y - (\mathsf{DR}(\eta \nabla f(x)) - \mathsf{DR}(\eta \nabla f(y)))| \leq |x - y|.$$

Lemma 16 implies that non-expansiveness holds for 1-dimensional DR-SGD. See Appendix A.4.1 for the proof of Lemma 16. By applying Lemma 16, we have $\mathbb{E}[\delta_t | z_{i_t} = z'_{i_t}] \leq \mathbb{E}[\delta_{t-1}]$.

We now show the upper bound of $\mathbb{E}[\delta_t | z_{i_t} \neq z'_{i_t}]$. Conditioned on $z_{i_t} \neq z'_{i_t}$, we have

$$\begin{aligned}\delta_t &\leq |w_{t-1} - w'_{t-1}| + |\mathsf{DR}(\eta \nabla f(w_{t-1}; z_{i_t})) - \mathsf{DR}(\eta \nabla f(w'_{t-1}; z'_{i_t}))| \\ &\leq \delta_{t-1} + 2\eta L + \Delta\end{aligned}$$

where the first inequality follows from the triangle inequality, and the second inequality holds since $\mathsf{DR}(x) \in [x - \Delta/2, x + \Delta/2]$ for all $x \in \mathbb{R}$, and $|\eta \nabla f(x; z_{i_t}) - \eta \nabla f(y; z'_{i_t})| \leq 2\eta L$ for all $x, y \in \mathbb{R}$.

Conclusively, we have

$$\mathbb{E}[\delta_T] \leq \mathbb{E}[\delta_{T-1}] + \frac{2\eta L + \Delta}{n} \leq \frac{(2\eta L + \Delta)T}{n}$$

which completes the proof of the upper bound.

A.4.1 Proof of Lemma 16

For notational simplicity, we set $u = \eta \nabla f(x)$ and $v = \eta \nabla f(y)$. Without loss of generality, we assume that $x > y$; $x = y$ is trivial. Since $x, y \in \mathbb{Z}_\Delta$, there exists $k \in \mathbb{N}$ such that $x - y = k\Delta$. Then, $|x - y - (\mathsf{DR}(u) - \mathsf{DR}(v))| \leq |x - y|$ is equivalent to $\mathsf{DR}(u) - \mathsf{DR}(v) \in [0, 2k\Delta]$.

Since $u \geq v$ and DR is monotonically increasing, $\mathsf{DR}(u) - \mathsf{DR}(v) \geq 0$. Furthermore, since $\mathsf{DR}(x) \in [x - 0.5\Delta, x + 0.5\Delta]$ for any $x \in \mathbb{R}$, we have

$$\mathsf{DR}(u) - \mathsf{DR}(v) \leq u + 0.5\Delta - (v - 0.5\Delta) = u - v + \Delta.$$

Here, since f is convex, f' is monotonically increasing. Furthermore, since f is M-smooth, and $\eta < 2/M$, we have

$$0 \leq u - v \leq \eta M(x - y) < 2k\Delta$$

which implies $0 \leq \mathsf{DR}(u) - \mathsf{DR}(v) < (2k + 1)\Delta$. Since $\mathsf{DR}(u), \mathsf{DR}(v) \in \mathbb{Z}_\Delta$, we have $\mathsf{DR}(u) - \mathsf{DR}(v) \in [0, 2k\Delta]$. This completes the proof.

A.5 Proof of Theorem 6

Without loss of generality, we assume that $\mathcal{Z} = [m]$ for some $m \geq 2$. Let $\mathcal{D}$ be uniform on $\mathcal{Z}$, and define

$$a_z = \frac{L(2z - m - 1)}{m - 1}, \qquad f(w; z) = -a_z w.$$

The coefficients a_z are equally spaced on $[-L, L]$. Thus, the losses are affine, L-Lipschitz, and smooth, with a different slope for every sample.

Let $A = a_Z$ for $Z \sim \mathcal{D}$ and $R(a) = -\mathsf{DR}(-\eta a)$. By symmetry, $\mathbb{E}[A] = 0$. For $S = (z_1, \ldots, z_n) \sim \mathcal{D}^n$, write $A_i = a_{z_i}$, $\overline{A} = n^{-1} \sum_{i=1}^n A_i$, and $\overline{R} = n^{-1} \sum_{i=1}^n R(A_i)$. Starting from $w_0 = 0$, each DR-SGD update adds $R(A_{i_t})$, so

$$\mathbb{E}[w_T \mid S] = T\overline{R}, \qquad f(w; S) - f(w; \mathcal{D}) = -\overline{A}w.$$

By taking expectations over the sampled indices and then the dataset, we have

$$\mathbb{E}[f(w_T; S) - f(w_T; \mathcal{D})] = -T\mathbb{E}[\overline{A}\,\overline{R}] = -\frac{T}{n}\mathbb{E}[AR(A)],$$

where the cross terms vanish by independence and $\mathbb{E}[A] = 0$. Since $aR(a) \geq 0$, we obtain the exact identity $\varepsilon_g(A_{f,\theta}^{\text{DR-SGD}}; \mathcal{D}) = (T/n)\mathbb{E}[AR(A)]$.

Set $h = \Delta/(2\eta) < L$ and $s = 2L/(m-1)$. For $|a| > h$, the rounded update satisfies $|R(a)| \geq \Delta$ and

$$\eta|a| \leq |R(a)| + \frac{\Delta}{2} \leq \frac{3}{2}|R(a)|.$$

Hence $aR(a) \geq (2\eta/3)a^2$ whenever $|a| > h$. The positive coefficients above h form right endpoints of intervals of length s covering $(h, L]$. Since x^2 is increasing for $x \geq 0$, comparison with the corresponding integral gives

$$s \sum_{z: a_z > h} a_z^2 \geq \int_h^L x^2\, dx = \frac{L^3 - h^3}{3}.$$

Using the symmetry of the coefficient set, we conclude that

$$\begin{aligned}
\mathbb{E}[AR(A)] &\geq \frac{2\eta}{3m} \sum_{z: |a_z| > h} a_z^2 = \frac{4\eta}{3m} \sum_{z: a_z > h} a_z^2 \\
&\geq \frac{2\eta(m-1)}{9mL}(L^3 - h^3) \\
&\geq \frac{\eta L}{9}(L - h) = \frac{(2\eta L - \Delta)L}{18},
\end{aligned}$$

where the last inequality uses $m \geq 2$ and $L^3 - h^3 \geq L^2(L - h)$. Therefore,

$$\varepsilon_g(A_{f,\theta}^{\text{DR-SGD}}; \mathcal{D}) \geq \frac{(2\eta L - \Delta)LT}{18n},$$

which completes the proof.

A.6 Proof of Theorem 7

In this section, we prove Theorem 7. Here, we construct a counterexample function $f : \mathbb{R}^2 \times \mathcal{Z} \to \mathbb{R}$ that achieves a desired uniform stability bound in Theorem 7. We note that it can easily be extended to general $d > 2$ by considering a function $g(x_1, \cdots, x_d; z) = f(x_1, x_2; z)$. In addition, we fix two distinct samples in $\mathcal{Z}$ and denote them by 0 and 1 for notational simplicity.

To this end, we construct a loss function $f : \mathbb{R}^2 \times \mathcal{Z} \to \mathbb{R}$ by

$$f(x, y; 0) = g(x, y), \qquad f(x, y; 1) = \Delta y$$

where g is obtained from the following lemma. Both losses are convex, Δ-Lipschitz, and $\left(\frac{9}{10T^2} + \frac{8}{5}\right)$-smooth:

Lemma 17. *Let $\Delta > 0$ and $T \in \mathbb{N}$. Then, there exists $\alpha \in (0, \Delta/2)$ and a convex, Δ-Lipschitz, and $\left(\frac{9}{10T^2} + \frac{8}{5}\right)$-smooth function $g : \mathbb{R}^2 \to \mathbb{R}$ such that*

$$g(0,0) = 0,\ \mathsf{DR}(\nabla g(0,0)) = (0,0),\ g(x,y) = -\left(\frac{\Delta}{2} + \alpha\right) x - \frac{\Delta^2}{20},\ \mathsf{DR}(\nabla g(x,y)) = (-\Delta, 0)$$

for all x, y such that $x \in \{0, \Delta, 2\Delta, \cdots, T\Delta\}$ and $y \le -\Delta$.

Lemma 17 implies that there exists a convex and smooth function $g : \mathbb{R}^2 \to \mathbb{R}$ whose rounded gradient is $(0, 0)$ at the origin but consistently induces $(-\Delta, 0)$ gradient on the region $\{(x, y) : x \in \{0, \cdots, T\Delta\}, y \le -\Delta\}$. The proof of Lemma 17 is in Appendix A.6.1.

Let $S = (0, \cdots, 0)$ and $S' = (1, 0, \cdots, 0)$ be the neighboring datasets, $\{w_t\}_{t=1}^T$ and $\{w'_t\}_{t=1}^T$ be trajectories from $A_{f,\theta}^{\text{DR-SGD}}(S)$ and $A_{f,\theta}^{\text{DR-SGD}}(S')$, respectively, where $w_t = (w_{t,1}, w_{t,2})$ and $w'_t = (w'_{t,1}, w'_{t,2})$ for each $t \in [T]$. Then since $S = (0, \cdots, 0)$ and $\mathsf{DR}(\nabla f(w_0; 0)) = (0, 0)$, we have $w_t = (0, 0)$ for all $t \in [T]$.

Then, we have

$$\mathbb{E}[f(w_T; 0) - f(w'_T; 0)] = \mathbb{E}[g(w_T) - g(w'_T)] = -\mathbb{E}[g(w'_T)].$$

Here, since $\mathsf{DR}(\nabla f(w_0; 0)) = (0, 0)$, we have $w'_t = (0, 0)$ for all $t < \tau$ where $\tau := \min\{t : i_t = 1\}$. When $t = \tau$, i.e., $i_t = 1$ is observed, we have $w'_\tau = (0, -\Delta)$ since $\mathsf{DR}(\eta \nabla f(x, y; 1)) = (0, \Delta)$. After τ, whenever $i_t \neq 1$, the update uses g, and hence $w'_t = w'_{t-1} + (\Delta, 0)$. Whenever $i_t = 1$, the update only decreases the second coordinate by Δ, so the iterate remains in the region $y \le -\Delta$. Thus, we have

$$\mathbb{E}[f(w_T; 0) - f(w'_T; 0)] = -\mathbb{E}[g(w'_T)] \ge \left(\frac{\Delta}{2} + \alpha\right) \mathbb{E}[w'_{T,1}] = \Omega(\Delta \mathbb{E}[w'_{T,1}]).$$

Let I_t be an indicator variable such that $I_t = 1$ if $i_t \neq 1$ and $t > \tau$, i.e., the event that $w'_t = w'_{t-1} + (\Delta, 0)$ occurs. Then, we have

$$\mathbb{P}(I_t = 1) = \left(1 - \left(1 - \frac{1}{n}\right)^{t-1}\right)\left(1 - \frac{1}{n}\right).$$

Let $q = 1 - 1/n$. Then, we have

$$\begin{aligned}
\mathbb{E}[w'_{T,1}] &= \Delta \sum_{t=1}^{T} q(1 - q^{t-1}) \\
&= \Delta q \left(T - \frac{1 - q^T}{1 - q}\right) \\
&= \Delta \left(1 - \frac{1}{n}\right) \left(T - n\left(1 - \left(1 - \frac{1}{n}\right)^T\right)\right).
\end{aligned}$$

We complete the proof using the following technical lemma. The proof of Lemma 18 is in Appendix A.6.2.

Lemma 18. *Let $n \in \mathbb{N}$ and $T \in \mathbb{N}$ with $n \geq 2, T \geq 2$. Then, it holds that*

$$\left(1-\frac{1}{n}\right)\left(T-n\left(1-\left(1-\frac{1}{n}\right)^{T}\right)\right)=\Omega\left(\min\left\{1,\frac{T}{n}\right\}T\right).$$

By applying Lemma 18, we have

$$\mathbb{E}[w'_{T,1}]=\Omega\left(\min\left\{1,\frac{T}{n}\right\}\Delta T\right).$$

Conclusively, we have

$$\mathbb{E}[f(w_T;0)-f(w'_T;0)]=-\mathbb{E}[g(w'_T)]=\Omega(\Delta\mathbb{E}[w'_{T,1}])=\Omega\left(\min\left\{1,\frac{T}{n}\right\}\Delta^2 T\right)$$

which completes the proof.

A.6.1 Proof of Lemma 17

In this section, we prove Lemma 17. To this end, we first consider a smoothed version of a piecewise-linear function $\max\{L_1, L_2\}$ that satisfies the gradient conditions in Lemma 17.

We first consider a smoothing function: for $\varepsilon > 0$, $H_\varepsilon : \mathbb{R} \to \mathbb{R}$ is defined by

$$H_\varepsilon(x)=\begin{cases}0 & \text{if } x \leq -\varepsilon,\\ \frac{(x+\varepsilon)^2}{4\varepsilon} & \text{if } -\varepsilon < x < \varepsilon,\\ x & \text{if } x \geq \varepsilon.\end{cases}$$

One can observe that $H_\varepsilon(x)$ can be regarded as a smoothed version of $\max\{0, x\}$ and H_ε is convex.

We define a loss function $g : \mathbb{R}^2 \to \mathbb{R}$ by

$$g(x,y)=L_1(x,y)+H_\varepsilon\left(L_2(x,y)-L_1(x,y)\right)$$

where

$$L_1(x,y)=-\left(\frac{\Delta}{2}-\frac{\Delta}{10T}\right)x+\frac{2\Delta}{5}y, \quad L_2(x,y)=-\left(\frac{\Delta}{2}+\frac{\Delta}{5T}\right)x-\frac{\Delta^2}{20}$$

and $\varepsilon > 0$. We note that since $L_2 - L_1$ is affine and H_ε is convex, $H_\varepsilon \circ (L_2 - L_1)$ is convex; adding the affine function L_1 preserves convexity. We now assign an explicit value to ε to satisfy the gradient conditions of Lemma 17. To this end, we determine $\varepsilon > 0$ so that $\nabla g(0,0) = \nabla L_1(0,0)$, and $\nabla g(x,y) = \nabla L_2(x,y)$ for all $x \in \{0, \Delta, 2\Delta, \cdots, T\Delta\}$ and $y \leq -\Delta$. Then, we have

$$\nabla g(0,0)=\nabla L_1(0,0)=\left(-\frac{\Delta}{2}+\frac{\Delta}{10T},\frac{2\Delta}{5}\right), \quad \nabla g(x,y)=\nabla L_2(x,y)=\left(-\frac{\Delta}{2}-\frac{\Delta}{5T},0\right)$$

for each $x \in \{0, \Delta, \cdots, T\Delta\}$ and $y \leq -\Delta$.

We first derive the condition for $\varepsilon > 0$ such that $\nabla g(0,0) = \nabla L_1(0,0)$. Since $g(x,y) = L_1(x,y)$ if $H_\varepsilon(L_2(x,y) - L_1(x,y)) = 0$, it suffices to consider $\varepsilon > 0$ such that $L_2(0,0) - L_1(0,0) \leq -\varepsilon$. Thus, we obtain the first requirement of ε:

$$L_2(0,0)-L_1(0,0)=-\frac{\Delta^2}{20}\leq-\varepsilon \iff \varepsilon\leq\frac{\Delta^2}{20}.$$

We now consider $\nabla g(x,y) = \nabla L_2(x,y)$ such that $x \in \{0, \Delta, \cdots, T\Delta\}$ and $y \leq -\Delta$. Since $g(x,y) = L_2(x,y)$ if $H_\varepsilon(L_2(x,y) - L_1(x,y)) = L_2(x,y) - L_1(x,y)$, we consider $\varepsilon > 0$ satisfying $L_2(x,y) - L_1(x,y) \geq \varepsilon$. Here, since

$$L_2(x,y)-L_1(x,y)=-\frac{3\Delta}{10T}x-\frac{2\Delta}{5}y-\frac{\Delta^2}{20}$$

is minimized at $(T\Delta, -\Delta)$, it suffices to derive the condition for ε such that $L_2(T\Delta, -\Delta) - L_1(T\Delta, -\Delta) \geq \varepsilon$ for $x \in \{0, \cdots, T\Delta\}$. We have

$$\begin{aligned} L_2(x, -\Delta) - L_1(x, -\Delta) &= \left(-\left(\frac{\Delta}{2} + \frac{\Delta}{5T}\right)x - \frac{\Delta^2}{20}\right) - \left(-\left(\frac{\Delta}{2} - \frac{\Delta}{10T}\right)x - \frac{2\Delta^2}{5}\right) \\ &= \frac{2\Delta^2}{5} - \frac{3\Delta}{10T}x - \frac{\Delta^2}{20} \\ &= \frac{7\Delta^2}{20} - \frac{3\Delta x}{10T}. \end{aligned}$$

The above expression has a minimum value at $x = T\Delta$:

$$L_2(x, -\Delta) - L_1(x, -\Delta) \geq \frac{7\Delta^2}{20} - \frac{3\Delta^2}{10} = \frac{\Delta^2}{20}$$

Thus, we need $\varepsilon \leq \Delta^2/20$. Thus, if we choose $\varepsilon = \Delta^2/20$ then our g satisfies the gradient conditions at the origin, and at $(w, -\Delta)$ for each $w \in \{0, \cdots, T\Delta\}$.

We now evaluate the Lipschitzness and smoothness constant of g. Let $\varepsilon_0 = \Delta^2/20$ and $L_0(x, y) = L_2(x, y) - L_1(x, y)$. Then, we have

$$\nabla g(x, y) = \nabla L_1(x, y) + H'_\varepsilon(L_0(x, y))\nabla L_0(x, y).$$

Since $0 \leq H'_\varepsilon(x) \leq 1$, one can observe that ∇g is a convex combination of ∇L_1 and ∇L_2. Here, since $\|\nabla L_1\|_2 = \|(\Delta/2 - \Delta/(10T), 2\Delta/5)\|_2 \leq \Delta$ and $\|\nabla L_2\|_2 = \|(\Delta/2 + \Delta/(5T), 0)\|_2 \leq \Delta$ for all $T \geq 1$, g is Δ-Lipschitz.

Furthermore, since H'_{ε_0} is $1/(2\varepsilon_0)$-Lipschitz, $\nabla g(x, y)$ is Lipschitz with constant at most

$$\frac{1}{2\varepsilon_0}\|\nabla L_0(x, y)\|_2^2 = \frac{1}{2\varepsilon_0}\left(\frac{9\Delta^2}{100T^2} + \frac{4\Delta^2}{25}\right) = \frac{9}{10T^2} + \frac{8}{5}$$

which implies that our g is $(9/(10T^2) + 8/5)$-smooth. This completes the proof with the choice of $\alpha = \Delta/(5T)$.

A.6.2 Proof of Lemma 18

In this section, we prove Lemma 18. To this end, we show the lower bound for the following two cases: (1) $T \geq n$, and (2) $T < n$. For notational simplicity, we denote $E(n, T)$ by the given expression in Lemma 18: $E(n, T) = \left(1 - \frac{1}{n}\right)\left(T - n\left(1 - \left(1 - \frac{1}{n}\right)^T\right)\right)$. It suffices to show that

$$E(n, T) = \Omega\left(\min\left\{1, \frac{T}{n}\right\}T\right).$$

For $n \geq 2$, $1 - 1/n \geq 1/2$ holds. Thus, we have

$$E(n, T) \geq \frac{1}{2}\left(T - n\left(1 - \left(1 - \frac{1}{n}\right)^T\right)\right) \tag{5}$$

We first show that if $T \geq n$, then $E(n, T) = \Omega(T)$. Here, we consider the following two sub-cases: $T \geq 2n$ and $n \leq T \leq 2n$.

The case where $T \geq 2n$ is easily obtained since

$$T - n\left(1 - \left(1 - \frac{1}{n}\right)^T\right) \geq T - n \geq \frac{T}{2}.$$

We now suppose that $n \leq T \leq 2n$. Then, we have

$$\left(1 - \frac{1}{n}\right)^T \geq \left(1 - \frac{1}{n}\right)^{2n} \geq \frac{1}{16}$$

since $(1-1/n)^n$ is strictly increasing for $n \geq 2$. Thus, we have

$$T - n\left(1 - \left(1 - \frac{1}{n}\right)^T\right) \geq T - \frac{15n}{16} \geq \frac{T}{16}.$$

By combining these results with Eq. (5), we have $E(n,T) = \Omega(T)$ if $T \geq n$.

We now show that if $T < n$, then $E(n,T) = \Omega(T^2/n)$. To this end, we consider the binomial expansion of $(1-1/n)^T$:

$$\left(1 - \frac{1}{n}\right)^T = \sum_{i=0}^{T} \binom{T}{i} \left(-\frac{1}{n}\right)^i$$

Here, since $T < n$, the magnitude of the terms $\binom{T}{i}\left(-\frac{1}{n}\right)^i$ is monotonically decreasing:

$$\frac{\binom{T}{i+1} n^{-(i+1)}}{\binom{T}{i} n^{-i}} = \frac{T-i}{(i+1)n} \leq \frac{T}{n} < 1$$

Then, truncating the series at the negative term gives a lower bound of the binomial expansion:

$$\left(1 - \frac{1}{n}\right)^T \geq 1 - \frac{T}{n} + \frac{T(T-1)}{2n^2} - \frac{T(T-1)(T-2)}{6n^3}.$$

Then, we have

$$\begin{aligned} T - n\left(1 - \left(1 - \frac{1}{n}\right)^T\right) &\geq T - n\left(1 - \left(1 - \frac{T}{n} + \frac{T(T-1)}{2n^2} - \frac{T(T-1)(T-2)}{6n^3}.\right)\right) \\ &= \frac{T(T-1)}{2n} - \frac{T(T-1)(T-2)}{6n^2} \\ &= \frac{T(T-1)}{2n}\left(1 - \frac{T-2}{3n}\right) \end{aligned}$$

Since $T < n$, we have $1 - \frac{T-2}{3n} \geq 1 - \frac{T}{3n} \geq \frac{2}{3}$. Furthermore, if $T \geq 2$, $T(T-1) \geq T^2/2$. Then, by combining this result with Eq. (5), we obtain that if $T < n$,

$$E(n,T) \geq \frac{1}{2}\left(\frac{T^2}{6n}\right) = \Omega\left(\frac{T^2}{n}\right).$$

Conclusively, we have

$$E(n,T) = \begin{cases} \Omega(T) & T \geq n \\ \Omega(T^2/n) & T < n \end{cases} = \Omega\left(\min\left\{1, \frac{T}{n}\right\} T\right),$$

which completes the proof.

A.7 Proof of Theorem 8

In this section, we prove Theorem 8. Let $S, S' \in \mathcal{Z}^n$ be the neighboring datasets, and $(w_t)_{t=1}^T$ and $(w'_t)_{t=1}^T$ be the coupled DR-SGD iterates from $A_{f,\theta}^{\text{DR-SGD}}(S)$ and $A_{f,\theta}^{\text{DR-SGD}}(S')$, respectively. Let $\delta_t = w_t - w'_t$ for each $t \in [T]$.

For each $t \in [T]$, let B_t be the event that there exists $\tau \leq t$ such that $z_{i_\tau} \neq z'_{i_\tau}$ and A_t be the event that $z_{i_t} \neq z'_{i_t}$. Since $w_T = w'_T$ conditioned on B_T^c, we have $\mathbb{E}[\|\delta_T\|_2] = P(B_T)\mathbb{E}[\|\delta_T\|_2|B_T]$.

Here, we have

$$P(B_T) = 1 - \left(1 - \frac{1}{n}\right)^T \leq \min\left\{1, \frac{T}{n}\right\}. \tag{6}$$

In the rest of the proof, we show the upper bound of $\mathbb{E}[\|\delta_T\|_2|B_T]$ by constructing a recurrence of $\mathbb{E}[\|\delta_t\|_2|B_t]$. For $t = 1$, we have

$$\begin{aligned}\mathbb{E}[\|\delta_1\|_2|B_1] = \mathbb{E}[\|\delta_1\|_2|A_1] &= \mathbb{E}[\|\mathsf{DR}(\eta\nabla f(w_0; z_{i_1})) - \mathsf{DR}(\eta\nabla f(w_0; z'_{i_1}))\|_2 | z_{i_1} \neq z'_{i_1}] \\ &\leq \|\mathsf{DR}(\eta\nabla f(w_0; z_{i_1}))\|_2 + \|\mathsf{DR}(\eta\nabla f(w_0; z'_{i_1}))\|_2 \leq 4\eta L\end{aligned}$$

where the second inequality follows from the triangle inequality, and the last inequality follows from Lemma 14.

We now consider general $t \geq 2$. Here, B_t can be regarded as the union of two mutually exclusive events, $B_{t-1}^c \cap A_t$ and B_{t-1}. Thus, we can decompose $\mathbb{E}[\|\delta_t\|_2|B_t]$ as follows:

$$\begin{aligned}\mathbb{E}[\|\delta_t\|_2|B_t] &= \frac{P(B_{t-1}^c \cap A_t)}{P(B_t)}\mathbb{E}[\|\delta_t\|_2|B_{t-1}^c \cap A_t] + \frac{P(B_{t-1})}{P(B_t)}\mathbb{E}[\|\delta_t\|_2|B_{t-1}] \\ &\leq \mathbb{E}[\|\delta_t\|_2|B_{t-1}^c \cap A_t] + \mathbb{E}[\|\delta_t\|_2|B_{t-1}].\end{aligned}$$

We now show the upper bounds for each term. Conditioned on $w_1, \cdots, w_{t-1}$ and $w'_1, \cdots, w'_{t-1}$, we have

$$\begin{aligned}\|\delta_t\|_2 &= \|\delta_{t-1} - (\mathsf{DR}(\eta\nabla f(w_{t-1}; z_{i_t})) - \mathsf{DR}(\eta\nabla f(w'_{t-1}; z'_{i_t})))\|_2 \\ &\leq \|\delta_{t-1}\|_2 + \|\mathsf{DR}(\eta\nabla f(w_{t-1}; z_{i_t}))\|_2 + \|\mathsf{DR}(\eta\nabla f(w'_{t-1}; z'_{i_t}))\|_2 \\ &\leq \|\delta_{t-1}\|_2 + 4\eta L\end{aligned}$$

where the last inequality follows from Lemma 14 and L-Lipschitzness of f. Then, we have

$$\begin{aligned}&\mathbb{E}[\|\delta_t\|_2|B_{t-1}^c \cap A_t] \leq 4\eta L, \\ &\mathbb{E}[\|\delta_t\|_2|B_{t-1}] \leq \mathbb{E}[\|\delta_{t-1}\|_2|B_{t-1}] + 4\eta L.\end{aligned}$$

Then, we have

$$\mathbb{E}[\|\delta_T\|_2|B_T] \leq \mathbb{E}[\|\delta_{T-1}\|_2|B_{T-1}] + 8\eta L \leq \cdots \leq 8\eta LT.$$

Conclusively, we have

$$\varepsilon_{us}(A_{f,\theta}^{\text{DR-SGD}}) \leq L\mathbb{E}[\|\delta_T\|_2] \leq L\mathbb{E}[\|\delta_T\|_2|B_T]P(B_T) \leq \min\left\{1, \frac{T}{n}\right\} 8\eta L^2 T$$

which completes the proof.

B Proofs for SR-GD and SR-SGD

In this section, we prove generalization and stability bounds of SR-GD and SR-SGD. We first prove Theorem 9 in Appendix B.1. We then show the upper bounds of uniform argument stability for SR-SGD and SR-GD (Theorems 10 and 11) in Appendices B.2 and B.3, respectively. We then present the proofs of uniform argument stability bounds for separable loss functions (Theorems 12 and 13) in Appendices B.4 and B.5, respectively.

B.1 Proof of Theorem 9

In this section, we prove Theorem 9. We use the following function, whose construction is in Appendix B.1.1.

Lemma 19. *Let $d \in \mathbb{N}$, $L, \Delta > 0$. Then, there exists $\alpha > 0$ and $\gamma > 0$ such that $\alpha > (e^2 - 1)\gamma$, and a function $g : \mathbb{R}^d \to \mathbb{R}$ such that g is convex, L-Lipschitz, $4L/\Delta$-smooth, and satisfies the following:*

$$\nabla g(\mathbf{0}_d) = -\frac{L}{\sqrt{d}}\mathbf{1}_d, \quad g(c) = L\Delta\left(\sqrt{\alpha k_c + \gamma} - \frac{\sqrt{\gamma}}{2}\right), \quad \forall c \in \{0, \Delta\}^d$$

where k_c denotes the number of non-zero entries in c, i.e., $k_c = \|c\|_1/\Delta$.

To prove the lower bounds for SR-GD and SR-SGD, we use the following lemma.

Lemma 20. *Let g be the function constructed in Appendix B.1.1, $Z \sim \mathcal{D}$, and $A = A(Z) \in [0, 1]$ satisfy*

$$\mathbb{P}(A \le 1/4) \ge \rho, \qquad \mathbb{P}(A \ge 1/2) \ge \rho \tag{7}$$

for a fixed universal constant $\rho > 0$. For $f(w; z) = A(z)g(w)$, $n \ge 2$, $\eta L < \Delta\sqrt{d}$, and $\theta = (\eta, T = 1, w_0 = \mathbf{0}_d)$,

$$\varepsilon_g(A^{\text{SR-GD}}_{f,\theta}; \mathcal{D}),\ \varepsilon_g(A^{\text{SR-SGD}}_{f,\theta}; \mathcal{D}) \gtrsim \frac{\sqrt{\min\{\Delta, \eta L d^{1/2}\}\eta L^3 d^{1/2}}}{n}.$$

To finish the proof of Theorem 9, take $A(i) = (i-1)/(m-1)$ and Z uniform on $[m]$. At least one quarter of the sample points satisfy $A(i) \le 1/4$, and at least one half satisfy $A(i) \ge 1/2$. Hence Eq. (7) holds with $\rho = 1/4$, uniformly in m, and Lemma 20 gives both bounds.

B.1.1 Proof of Lemma 19

We construct a Huber function (Huber [1964]) $g : \mathbb{R}^d \to \mathbb{R}$ defined by

$$g(x) = \begin{cases} \frac{2L}{\Delta}\left\|x - \frac{\Delta}{4\sqrt{d}}\mathbf{1}_d\right\|_2^2 & x \in \mathcal{B}_{\Delta/4}\left(\frac{\Delta}{4\sqrt{d}}\mathbf{1}_d\right), \\ L\left\|x - \frac{\Delta}{4\sqrt{d}}\mathbf{1}_d\right\|_2 - \frac{L\Delta}{8} & o.w. \end{cases}$$

Intuitively, g consists of a quadratic part within a ball of radius $\Delta/4$ centered at $(\Delta/(4\sqrt{d}))\mathbf{1}_d$, and a linear part outside this ball. It is easy to observe that g is convex, L-Lipschitz, and $4L/\Delta$-smooth. Furthermore, we have

$$\nabla g(\mathbf{0}_d) = -\frac{L}{\sqrt{d}}\mathbf{1}_d, \quad g(c) = L\Delta\left(\sqrt{\left(1 - \frac{1}{2\sqrt{d}}\right)k_c + \frac{1}{16}} - \frac{1}{8}\right), \quad \forall c \in \{0, \Delta\}^d$$

where k_c denotes the number of non-zero entries in c. We note that if we choose $\alpha = 1 - 1/(2\sqrt{d})$ and $\gamma = 1/16$, then our g satisfies the conditions for the gradient at the origin and the values at $c \in \{0, \Delta\}^d$ given in Lemma 19. Furthermore, $\alpha > (e^2 - 1)\gamma$ holds since

$$1 - \frac{1}{2\sqrt{d}} \ge \frac{1}{2} > \frac{e^2 - 1}{16}$$

for all $d \in \mathbb{N}$. This completes the proof.

B.1.2 Proof of Lemma 20

Since $A(z) \in [0,1]$, the loss f is convex, L-Lipschitz, and $(4L/\Delta)$-smooth. For notational simplicity, let $r = \eta L/(\Delta\sqrt{d}) \in (0,1)$ and

$$H(p) = \mathbb{E}_{K\sim Bin(d,p)}[\sqrt{\alpha K + \gamma}], \qquad \alpha = 1 - \frac{1}{2\sqrt{d}}, \quad \gamma = \frac{1}{16}.$$

The function H is nondecreasing. To lower-bound the generalization error, we introduce the following lemma, whose proof is presented in Appendix B.1.3

Lemma 21. *Let $a, b > 0$ be universal constants such that $a > b(e^2 - 1)$, $n \in \mathbb{N}$, $p, p' \in (0,1)$ such that $p > p'$, $X \sim Bin(n,p)$ and $Y \sim Bin(n,p')$. Then*

$$\mathbb{E}[\sqrt{aX+b} - \sqrt{aY+b}] = \Omega\left(\left(1 - \frac{p'}{p}\right)\min\left\{\sqrt{\mathbb{E}[X]}, \mathbb{E}[X]\right\}\right).$$

By Lemma 21, for $0 \le p' < p < 1$,

$$H(p) - H(p') \gtrsim \left(1 - \frac{p'}{p}\right)\min\{\sqrt{dp}, dp\}. \tag{8}$$

The case $p' = 0$ follows by continuity.

We first consider SR-GD. Let $A_i = A(z_i)$ and $\overline{A} = n^{-1}\sum_{i=1}^n A_i$. Conditioned on S, each coordinate of w_1 equals Δ with probability $r\overline{A}$, independently. Thus, we have

$$\mathbb{E}[g(w_1) \mid S] = L\Delta\left(H(r\overline{A}) - \frac{\sqrt{\gamma}}{2}\right).$$

Let $A_1, A_1', A_2, \cdots, A_n$ be independent copies of A, set $B = \sum_{i=2}^n A_i$, and write $\mu = \mathbb{E}[A] \ge \rho/2$. We first compute the expected generalization gap. Since $f(w;z) = A(z)g(w)$, we have

$$\begin{aligned}
\varepsilon_g(A_{f,\theta}^{\text{SR-GD}}; \mathcal{D}) = \left|\mathbb{E}\left[(\overline{A} - \mu)g(w_1)\right]\right| &= L\Delta\left|\mathbb{E}\left[(\overline{A} - \mu)H(r\overline{A})\right]\right| \\
&= \frac{L\Delta}{n}\left|\sum_{i=1}^n \mathbb{E}\left[(A_i - \mu)H(r\overline{A})\right]\right| \\
&= L\Delta\left|\mathbb{E}\left[(A_1 - \mu)H\left(\frac{r(B+A_1)}{n}\right)\right]\right| \\
&= L\Delta\left|\mathbb{E}\left[(A_1 - A_1')H\left(\frac{r(B+A_1)}{n}\right)\right]\right|.
\end{aligned}$$

Here, the second equality follows from the conditional expectation of $g(w_1)$ and $\mathbb{E}[\overline{A} - \mu] = 0$. The fourth equality follows from symmetry among the n samples. The last equality uses the independence of A_1' from (B, A_1) and $\mathbb{E}[A_1'] = \mu$.

Since A_1 and A_1' are independent copies and are independent of B, exchanging them gives

$$\begin{aligned}
\mathbb{E}\left[(A_1 - A_1')H\left(\frac{r(B+A_1)}{n}\right)\right] &= \mathbb{E}\left[(A_1' - A_1)H\left(\frac{r(B+A_1')}{n}\right)\right] \\
&= -\mathbb{E}\left[(A_1 - A_1')H\left(\frac{r(B+A_1')}{n}\right)\right].
\end{aligned}$$

Averaging the first and last expressions yields

$$\mathbb{E}\left[(A_1 - A_1')H\left(\frac{r(B+A_1)}{n}\right)\right] = \frac{1}{2}\mathbb{E}\left[(A_1 - A_1')\left\{H\left(\frac{r(B+A_1)}{n}\right) - H\left(\frac{r(B+A_1')}{n}\right)\right\}\right].$$

The integrand on the right-hand side is nonnegative since H is nondecreasing. Therefore, we can remove the absolute value and obtain

$$\varepsilon_g(A_{f,\theta}^{\text{SR-GD}}; \mathcal{D}) = \frac{L\Delta}{2}\mathbb{E}\left[(A_1 - A_1')\left\{H\left(\frac{r(B+A_1)}{n}\right) - H\left(\frac{r(B+A_1')}{n}\right)\right\}\right]. \tag{9}$$

We next show that the event $E_B = \{B \geq (n-1)\mu/2\}$ has probability bounded below by a constant. Let $N = n - 1$. Since $A^2 \leq A$, the Paley–Zygmund inequality gives

$$\mathbb{P}(E_B) \geq \frac{(\mathbb{E}[B])^2}{4\mathbb{E}[B^2]} \geq \frac{N\mu}{4(1+N\mu)} \geq \frac{\mu}{4(1+\mu)}.$$

Now consider the event $E_B \cap \{A_1 \geq 1/2, A_1' \leq 1/4\}$, whose probability is at least $\rho^2\mathbb{P}(E_B)$ by independence. On this event, $A_1 - A_1' \geq 1/4$, and, for $p = r(B + A_1)/n$ and $p' = r(B + A_1')/n$,

$$p \geq \frac{r\mu}{4}, \qquad 1 - \frac{p'}{p} = \frac{A_1 - A_1'}{B + A_1} \geq \frac{1}{4n}.$$

Applying Eq. (8) in Eq. (9) yields

$$\varepsilon_g(A_{f,\theta}^{\text{SR-GD}}; \mathcal{D}) \gtrsim \frac{L\Delta}{n} \min\{\sqrt{rd}, rd\}.$$

We now consider SR-SGD. Since $T = 1$, conditioning on the selected training sample gives

$$\varepsilon_g(A_{f,\theta}^{\text{SR-SGD}}; \mathcal{D}) = \frac{L\Delta}{2n}\mathbb{E}\big[(A - A')\{H(rA) - H(rA')\}\big],$$

where A' is an independent copy. On $\{A \geq 1/2, A' \leq 1/4\}$, we have $A - A' \geq 1/4$, $rA \geq r/2$, and $1 - A'/A \geq 1/2$. By applying Eq. (8) and using the nonnegativity of the integrand, we obtain the same lower bound. Then, we have

$$L\Delta \min\{\sqrt{rd}, rd\} = \sqrt{\min\{\Delta, \eta L d^{1/2}\}\eta L^3 d^{1/2}},$$

which completes the proof.

B.1.3 Proof of Lemma 21

In this section, we prove Lemma 21. Since $p > p'$, we can apply a monotone coupling such that $X \geq Y$ almost surely.

We first compute $\mathbb{E}[Y|X]$. Let $X_1, \cdots, X_n \sim Ber(p)$ and $Y_1, \cdots, Y_n \sim Ber(p')$ be two sets of i.i.d. random variables such that $X = \sum_{i=1}^n X_i$ and $Y = \sum_{i=1}^n Y_i$. Since X and Y are monotonically coupled, we have

$$\mathbb{P}(Y_i = 1|X_i = 0) = 0, \quad \mathbb{P}(Y_i = 1|X_i = 1) = \frac{p'}{p}.$$

Thus, $Y|X = \sum_{i=1}^n Y_i|X_i \sim Bin(X, p'/p)$ and $\mathbb{E}[Y|X] = (p'/p)X$.

We now show the lower bound of $\mathbb{E}[\sqrt{aX+b} - \sqrt{aY+b}]$.

$$\begin{aligned}
\mathbb{E}[\sqrt{aX+b} - \sqrt{aY+b}] &= \mathbb{E}\left[\frac{a(X-Y)}{\sqrt{aX+b} + \sqrt{aY+b}}\right] \\
&\geq \mathbb{E}\left[\frac{a(X-Y)}{2\sqrt{aX+b}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{a(X-Y)}{2\sqrt{aX+b}}\middle| X\right]\right] \\
&= \mathbb{E}\left[\frac{a(X - \mathbb{E}[Y|X])}{2\sqrt{aX+b}}\right] \\
&= \frac{a}{2}\left(1 - \frac{p'}{p}\right)\mathbb{E}\left[\frac{X}{\sqrt{aX+b}}\right].
\end{aligned}$$

It suffices to show that $\mathbb{E}\left[X/\sqrt{aX+b}\right] = \Omega(\min\{\sqrt{\mathbb{E}[X]}, \mathbb{E}[X]\})$. To this end, we use the following lemma. The proof of Lemma 22 is in Appendix B.1.4.

Lemma 22. *Let $a > 0$ and $b \geq 0$ be universal constants such that $a > b(e^2 - 1)$, $n \in \mathbb{N}$, $p \in (0, 1)$, and $X \sim Bin(n, p)$. Then it holds that*

$$\mathbb{E}[\sqrt{aX+b}] - \sqrt{b} = \Omega\left(\min\left\{\sqrt{\mathbb{E}[X]}, \mathbb{E}[X]\right\}\right).$$

Then, we have

$$\begin{aligned}\mathbb{E}\left[\frac{X}{\sqrt{aX+b}}\right] &= \mathbb{P}(X \geq 1)\mathbb{E}\left[\frac{X}{\sqrt{aX+b}}\middle| X \geq 1\right] \\ &\geq \mathbb{P}(X \geq 1)\mathbb{E}\left[\frac{X}{\sqrt{(a+b)X}}\middle| X \geq 1\right] \\ &\geq \frac{1}{\sqrt{a+b}}\mathbb{P}(X \geq 1)\mathbb{E}[\sqrt{X}|X \geq 1] \\ &= \frac{1}{\sqrt{a+b}}\mathbb{E}[\sqrt{X}] = \Omega\left(\min\left\{\sqrt{\mathbb{E}[X]}, \mathbb{E}[X]\right\}\right)\end{aligned}$$

where the last equality follows from Lemma 22. Conclusively, we have

$$\begin{aligned}\mathbb{E}[\sqrt{aX+b} - \sqrt{aY+b}] &= \Omega\left(\left(1 - \frac{p'}{p}\right)\mathbb{E}\left[\frac{X}{\sqrt{aX+b}}\right]\right) \\ &= \Omega\left(\left(1 - \frac{p'}{p}\right)\min\left\{\sqrt{\mathbb{E}[X]}, \mathbb{E}[X]\right\}\right)\end{aligned}$$

which completes the proof.

B.1.4 Proof of Lemma 22

In this section, we prove Lemma 22. Here, we assume $n \geq 2$; the case where $n = 1$ follows directly since $X \sim Ber(p)$ and $\mathbb{E}[\sqrt{aX+b}] - \sqrt{b} = p(\sqrt{a+b} - \sqrt{b}) = \Omega(p)$. To this end, we prove the following two statements: (1) if $\mathbb{E}[X] \geq 1$, then $\mathbb{E}[\sqrt{aX+b} - \sqrt{b}] = \Omega(\sqrt{\mathbb{E}[X]})$, and (2) if $\mathbb{E}[X] < 1$, then $\mathbb{E}[\sqrt{aX+b} - \sqrt{b}] = \Omega(\mathbb{E}[X])$. Then, we have $\mathbb{E}[\sqrt{aX+b} - \sqrt{b}] = \Omega(\min\{\sqrt{\mathbb{E}[X]}, \mathbb{E}[X]\})$.

We first show that the statement (1): if $\mathbb{E}[X] \geq 1$, then $\mathbb{E}[\sqrt{aX+b} - \sqrt{b}] = \Omega(\sqrt{\mathbb{E}[X]})$. Since $\sqrt{aX+b} - \sqrt{b} = 0$ if $X = 0$, we have

$$\mathbb{E}[\sqrt{aX+b}] - \sqrt{b} = \mathbb{P}(X \geq 1)\mathbb{E}[\sqrt{aX+b} - \sqrt{b}|X \geq 1].$$

Since $\mathbb{E}[X] \geq 1$, we have $\mathbb{P}(X \geq 1) = \Omega(1)$; $\mathbb{P}(X \geq 1) = 1 - (1-p)^n \geq 1 - e^{-np} \geq 1 - e^{-1}$. We now lower bound the expectation term. Given $X \geq 1$, we have

$$\begin{aligned}\sqrt{aX+b} - \sqrt{b} &= \frac{aX}{\sqrt{aX+b} + \sqrt{b}} \\ &\geq \frac{aX}{\sqrt{X}(\sqrt{a+b} + \sqrt{b})} \\ &= \frac{a}{\sqrt{a+b} + \sqrt{b}}\sqrt{X}.\end{aligned}$$

Thus, we have

$$\mathbb{P}(X \geq 1)\mathbb{E}[\sqrt{aX+b} - \sqrt{b}|X \geq 1] \geq \mathbb{P}(X \geq 1)\frac{a}{\sqrt{a+b} + \sqrt{b}}\mathbb{E}\left[\sqrt{X}|X \geq 1\right] = \Omega(\mathbb{E}[\sqrt{X}]).$$

We now show that $\mathbb{E}[\sqrt{X}] = \Omega(\sqrt{np}) = \Omega(\sqrt{\mathbb{E}[X]})$. We note that for any $\delta > 0$, it holds that

$$\mathbb{E}\left[\sqrt{X}\right] \geq \delta\mathbb{P}\left(\sqrt{X} \geq \delta\right) = \delta\mathbb{P}\left(X \geq \delta^2\right). \tag{$*$}$$

By choosing $\delta = \sqrt{np/2}$, we have

$$(*) = \sqrt{\frac{np}{2}}\mathbb{P}\left(X \geq \frac{np}{2}\right) \geq \frac{1}{4}\sqrt{\frac{np}{2}}\frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} = \frac{1}{4}\sqrt{\frac{np}{2}}\frac{np}{np+1-p} \geq \frac{1}{8\sqrt{2}}\sqrt{np}$$

where the second inequality follows from the Paley-Zygmund inequality, and the last inequality follows from the fact that $np \geq 1$ and $p < 1$.

We now prove the second statement: if $\mathbb{E}[X] < 1$, then $\mathbb{E}[\sqrt{aX+b} - \sqrt{b}] = \Omega(\mathbb{E}[X])$. To this end, we consider the lower bound of $\mathbb{E}[\sqrt{aX+b}]$:

$$\begin{aligned}
\mathbb{E}[\sqrt{aX+b}] &\geq \sqrt{b}\mathbb{P}(X=0) + \sqrt{a+b}\mathbb{P}(X=1) \\
&= \sqrt{b}\,(1-p)^n + \sqrt{a+b}np(1-p)^{n-1} \\
&\geq \sqrt{b}(1-np) + \sqrt{a+b}np(1-p)^{n-1} \\
&= \sqrt{b} + np(\sqrt{a+b}(1-p)^{n-1} - \sqrt{b}) \\
&\geq \sqrt{b} + np\left(\frac{\sqrt{a+b}}{e} - \sqrt{b}\right)
\end{aligned}$$

where the inequality in the third line follows from Bernoulli's inequality, and the last inequality follows from the fact that $\mathbb{E}[X] = np < 1$, which implies that

$$(1-p)^{n-1} > \left(1 - \frac{1}{n}\right)^{n-1} \geq \frac{1}{e}$$

for $n \geq 2$. Here, if $b = 0$, then we have

$$\mathbb{E}[\sqrt{aX}] \geq \frac{np\sqrt{a}}{e} = \Omega(\mathbb{E}[X]).$$

Suppose that $b > 0$. Since $a > b(e^2 - 1)$, we have

$$\frac{\sqrt{a+b}}{e} - \sqrt{b} = \frac{\sqrt{b}}{e}\left(\sqrt{1 + \frac{a}{b}} - e\right) > \frac{\sqrt{b}}{e}\left(\sqrt{1+e^2-1} - e\right) = 0$$

which implies

$$\mathbb{E}[\sqrt{aX+b}] - \sqrt{b} \geq np\left(\frac{\sqrt{a+b}}{e} - \sqrt{b}\right) = \Omega(\mathbb{E}[X]).$$

It completes the proof.

B.2 Proof of Theorem 10

In this section, we prove Theorem 10. Let $S = (z_1, \cdots, z_n)$ and $S' = (z'_1, \cdots, z'_n)$ be neighboring datasets. Let $(w_t)_{t=1}^T$ and $(w'_t)_{t=1}^T$ be the trajectories from $A_{f,\theta}^{\text{SR-SGD}}(S)$ and $A_{f,\theta}^{\text{SR-SGD}}(S')$, respectively. We note that the stochastic rounding operations in two SR-SGD parameters are monotonically coupled: for each iteration $t \in [T]$ and coordinate $j \in [d]$, they use the same uniform random variable $u_{t,j}$ in Eq. (3). For each $t \in [T]$, let $\delta_t = w_t - w'_t$ and B_t be the event that there exists $\tau \leq t$ such that $z_{i_\tau} \neq z'_{i_\tau}$ and A_t be the event that $z_{i_t} \neq z'_{i_t}$. Since $w_T = w'_T$ conditioned on B_T^c, we have $\mathbb{E}[\|\delta_T\|] = P(B_T)\mathbb{E}[\|\delta_T\||B_T]$.

Here, we have

$$P(B_T) = 1 - \left(1 - \frac{1}{n}\right)^T \leq \min\left\{1, \frac{T}{n}\right\}. \tag{10}$$

In the rest of the proof, we show the upper bound on $\mathbb{E}[\|\delta_T\|_2|B_T]$. To this end, we construct a recurrence with respect to $\mathbb{E}[\|\delta_t\|_2^2|B_t]$ and then reduce it to the upper bound of $\mathbb{E}[\|\delta_T\|_2|B_T]$.

We first consider the upper bound of $\mathbb{E}[\|\delta_1\|_2^2|B_1]$:

$$\mathbb{E}[\|\delta_1\|_2^2|B_1] = \mathbb{E}[\|\delta_1\|_2^2|A_1] = \mathbb{E}[\|\mathsf{SR}(\eta\nabla f(w_0; z_{i_1})) - \mathsf{SR}(\eta\nabla f(w'_0; z'_{i_1}))\|_2^2|z_{i_1} \neq z'_{i_1}].$$

To upper bound this expectation, we introduce the following lemma. The proof of Lemma 23 is in Appendix B.2.1.

Lemma 23. *Let $\Delta > 0$ and $x, y, p, q \in \mathbb{R}^d$. Then, under the coordinate-wise monotone coupling, it holds that*

$$\mathbb{E}[\|x - \mathsf{SR}(p) - (y - \mathsf{SR}(q))\|_2^2] \leq \|x - p - (y - q)\|_2^2 + \Delta\|p - q\|_1.$$

Since f is L-Lipschitz, we have $\|\eta\nabla f(w_t; z_{i_t}) - \eta\nabla f(w'_t; z'_{i_t})\|_2 \leq 2\eta L$ and $\|\eta\nabla f(w_t; z_{i_t}) - \eta\nabla f(w'_t; z'_{i_t})\|_1 \leq 2\eta L\sqrt{d}$. Therefore, by Lemma 23, we have

$$\mathbb{E}[\|\delta_1\|_2^2|B_1] \leq (2\eta L)^2 + 2\Delta\eta L\sqrt{d}. \tag{11}$$

We now consider the general $t = 2, \cdots, T$. Here, B_t can be considered as the union of two mutually exclusive events, $B_{t-1}^c \cap A_t$ and B_{t-1}. Then, we can decompose $\mathbb{E}[\|\delta_t\|_2^2|B_t]$ as follows:

$$\begin{aligned}\mathbb{E}[\|\delta_t\|_2^2|B_t] &= \frac{P(B_{t-1}^c \cap A_t)}{P(B_t)}\mathbb{E}[\|\delta_t\|_2^2|B_{t-1}^c \cap A_t] + \frac{P(B_{t-1})}{P(B_t)}\mathbb{E}[\|\delta_t\|_2^2|B_{t-1}] \\ &\leq \max\{\mathbb{E}[\|\delta_t\|_2^2|B_{t-1}^c \cap A_t], \mathbb{E}[\|\delta_t\|_2^2|B_{t-1}]\}.\end{aligned}$$

We now show the upper bounds of $\mathbb{E}[\|\delta_t\|_2^2|B_{t-1}^c \cap A_t]$ and $\mathbb{E}[\|\delta_t\|_2^2|B_{t-1}]$, respectively.

Conditioned on $w_1, \cdots, w_{t-1}$ and $w'_1, \cdots, w'_{t-1}$, by Lemma 23, we have

$$\mathbb{E}[\|\delta_t\|_2^2] \leq \|\delta_{t-1} - \eta(\nabla f(w_{t-1}; z_{i_t}) - \nabla f(w'_{t-1}; z'_{i_t}))\|_2^2 + 2\Delta\eta L\sqrt{d}.$$

We note that the first term of the right-hand side is equal to the squared norm of two SGD updates. Thus, one can apply the non-expansiveness of SGD. See the following lemma.

Lemma 24 (Lemma 3.7 in [Hardt et al., 2016])**.** *Let $d \in \mathbb{N}$, $\eta, M > 0$ with $\eta < 2/M$, $f : \mathbb{R}^d \to \mathbb{R}$ be a convex, M-smooth function. Then, for all $u, v \in \mathbb{R}^d$, it holds that*

$$\|u - \eta\nabla f(u) - (v - \eta\nabla f(v))\|_2 \leq \|u - v\|_2.$$

If $z_{i_t} = z'_{i_t}$, since $\eta < 2/M$, we have $\|\delta_{t-1} - \eta(\nabla f(w_{t-1}; z_{i_t}) - \nabla f(w'_{t-1}; z'_{i_t}))\|_2^2 \leq \|\delta_{t-1}\|_2^2$ by Lemma 24. If $z_{i_t} \neq z'_{i_t}$, by the triangle inequality, we have

$$\begin{aligned}\|\delta_{t-1} - \eta(\nabla f(w_{t-1}; z_{i_t}) - \nabla f(w'_{t-1}; z'_{i_t}))\|_2^2 &\leq (\|\delta_{t-1}\|_2 + \eta\|\nabla f(w_{t-1}; z_{i_t}) - \nabla f(w'_{t-1}; z'_{i_t})\|_2)^2 \\ &\leq (\|\delta_{t-1}\|_2 + 2\eta L)^2 \\ &\leq \|\delta_{t-1}\|_2^2 + 4\eta L\|\delta_{t-1}\|_2 + 4\eta^2 L^2.\end{aligned}$$

Conclusively, we have

$$\mathbb{E}[\|\delta_t\|_2^2] \leq \begin{cases} \|\delta_{t-1}\|_2^2 + 2\Delta\eta L\sqrt{d} & z_{i_t} = z'_{i_t}, \\ \|\delta_{t-1}\|_2^2 + 4\eta L\|\delta_{t-1}\|_2 + 4\eta^2 L^2 + 2\Delta\eta L\sqrt{d} & z_{i_t} \neq z'_{i_t}. \end{cases}$$

By conditioning on the events $B_{t-1}^c \cap A_t$ and B_{t-1}, we have

$$\mathbb{E}[\|\delta_t\|_2^2 | B_{t-1}^c \cap A_t] \leq 4\eta^2 L^2 + 2\Delta\eta L\sqrt{d},$$
$$\mathbb{E}[\|\delta_t\|_2^2 | B_{t-1}] \leq \mathbb{E}[\|\delta_{t-1}\|_2^2 | B_{t-1}] + \frac{4\eta L}{n}\mathbb{E}[\|\delta_{t-1}\|_2 | B_{t-1}] + \frac{4\eta^2 L^2}{n} + 2\Delta\eta L\sqrt{d}.$$

Thus, we have

$$\mathbb{E}[\|\delta_t\|_2^2 | B_t] \leq \max\left\{\mathbb{E}[\|\delta_{t-1}\|_2^2 | B_{t-1}] + \frac{4\eta L}{n}\mathbb{E}[\|\delta_{t-1}\|_2 | B_{t-1}] + \gamma, \omega\right\}$$

where $\gamma = 4\eta^2 L^2/n + 2\Delta\eta L\sqrt{d}$ and $\omega = 4\eta^2 L^2 + 2\Delta\eta L\sqrt{d}$.

We now solve this recurrence using the following lemma. The proof of Lemma 25 is in Appendix B.2.2.

Lemma 25. *Let $T \in \mathbb{N}$, $\omega, \xi \geq 0$ such that $\xi \geq \omega$, and $\alpha, \gamma > 0$. Suppose that the non-negative random variables $X_1, \cdots, X_T$ satisfy that $\mathbb{E}[X_1^2] \leq \xi$, and $\mathbb{E}[X_t^2] \leq \max\{\mathbb{E}[X_{t-1}^2] + \alpha\mathbb{E}[X_{t-1}] + \gamma, \omega\}$ for all $t \in \{2, \cdots, T\}$. Then it holds that*

$$\mathbb{E}[X_t] \leq \alpha(t-1) + \sqrt{\gamma(t-1) + \xi}$$

for all $t \in [T]$.

Recall Eq. (11) for the base case: $\mathbb{E}[\|\delta_1\|_2^2 | B_1] \leq 4\eta^2 L^2 + 2\Delta\eta L\sqrt{d}$. Then, by applying Lemma 25 with the choices of $\alpha = 4\eta L/n$, $\gamma = 4\eta^2 L^2/n + 2\Delta\eta L\sqrt{d}$, and $\xi = 4\eta^2 L^2 + 2\Delta\eta L\sqrt{d}$, we have

$$\begin{aligned}\mathbb{E}[\|\delta_T\|_2 | B_T] &\leq \frac{4\eta L(T-1)}{n} + \sqrt{\gamma(T-1) + \omega} \\ &= \frac{4\eta L(T-1)}{n} + \sqrt{4\eta^2 L^2 + \frac{4\eta^2 L^2 (T-1)}{n} + 2\Delta\eta L d^{1/2} T}.\end{aligned}$$

By combining this result with Eq. (10), we have

$$\begin{aligned}\mathbb{E}[\|\delta_T\|_2] &\leq \min\left\{1, \frac{T}{n}\right\}\left(\frac{4\eta L(T-1)}{n} + \sqrt{4\eta^2 L^2 + \frac{4\eta^2 L^2 (T-1)}{n} + 2\Delta\eta L d^{1/2} T}\right) \\ &\leq \frac{4\eta L(T-1)}{n} + \min\left\{1, \frac{T}{n}\right\}\left(2\eta L\left(1 + \sqrt{\frac{T-1}{n}}\right) + \sqrt{2\Delta\eta L d^{1/2} T}\right).\end{aligned}$$

We note that

$$\min\left\{1, \frac{T}{n}\right\}\left(1 + \sqrt{\frac{T-1}{n}}\right) \leq \frac{2T}{n}.$$

Indeed, if $T \leq n$, then $\sqrt{(T-1)/n} < 1$, yielding the above inequality, and if $T > n$, then $1 + \sqrt{(T-1)/n} < T/n + T/n = 2T/n$. Then, we obtain

$$\varepsilon_{uas}(A_{f,\theta}^{\mathsf{SR\text{-}SGD}}) \leq \frac{8\eta LT}{n} + \min\left\{1, \frac{T}{n}\right\}\sqrt{2\Delta\eta L d^{1/2} T}.$$

which completes the proof.

B.2.1 Proof of Lemma 23

In this section, we prove Lemma 23. We now expand the left side of the given expression:

$$\|x-y\|_2^2 - 2\mathbb{E}[\langle x-y, \mathsf{SR}(p)-\mathsf{SR}(q)\rangle] + \mathbb{E}[\|\mathsf{SR}(p)-\mathsf{SR}(q)\|_2^2].$$

Here, since $\mathbb{E}[\mathsf{SR}(p)] = p$ and $\mathbb{E}[\mathsf{SR}(q)] = q$ and $\mathbb{E}[\cdot]$ is a linear operator, we have

$$\mathbb{E}[\langle x-y, \mathsf{SR}(p)-\mathsf{SR}(q)\rangle] = \langle x-y, p-q\rangle .$$

Furthermore, we note that

$$\mathbb{E}[\|\mathsf{SR}(p)-\mathsf{SR}(q)\|_2^2] = \sum_{i=1}^{d} \mathbb{E}[(\mathsf{SR}(p_i)-\mathsf{SR}(q_i))^2]$$

where p_i and q_i are the i-th coordinate of p and q, respectively. For each $i \in [d]$, let $|p_i - q_i|/\Delta = m_i + r_i$ where $m_i \in \mathbb{Z}_{\geq 0}$ and $r_i \in [0,1)$. Then, under the monotone coupling, we have

$$|\mathsf{SR}(p_i)-\mathsf{SR}(q_i)| = \begin{cases} (m_i+1)\Delta & \text{with probability } r_i, \\ m_i\Delta & \text{with probability } 1-r_i. \end{cases}$$

Then, we have

$$\begin{aligned}
\mathbb{E}[|\mathsf{SR}(p_i)-\mathsf{SR}(q_i)|^2] &= \Delta^2(r_i(m_i+1)^2 + (1-r_i)m_i^2) \\
&= \Delta^2(m_i^2 + 2m_ir_i + r_i) \\
&= \Delta^2(m_i^2 + 2m_ir_i + r_i^2 - r_i^2 + r_i) \\
&= |p_i-q_i|^2 + \Delta^2 r_i(1-r_i).
\end{aligned}$$

Here, since $r_i(1-r_i) \leq r_i \leq |p_i-q_i|/\Delta$, we have $\mathbb{E}[|\mathsf{SR}(p_i)-\mathsf{SR}(q_i)|^2] \leq |p_i-q_i|^2 + \Delta|p_i-q_i|$. Hence, we obtain

$$\mathbb{E}[\|\mathsf{SR}(p)-\mathsf{SR}(q)\|_2^2] \leq \|p-q\|_2^2 + \Delta\|p-q\|_1.$$

Conclusively, we have

$$\begin{aligned}
\mathbb{E}[\|x-\mathsf{SR}(p)-y+\mathsf{SR}(q)\|_2^2] &\leq \|x-y\|_2^2 - 2\langle x-y, p-q\rangle + \|p-q\|_2^2 + \Delta\|p-q\|_1 \\
&= \|x-y-(p-q)\|_2^2 + \Delta\|p-q\|_1
\end{aligned}$$

which completes the proof.

B.2.2 Proof of Lemma 25

In this section, we prove Lemma 25. Here, we prove a stronger statement on the second moment $\mathbb{E}[X_{t+1}^2]$, and then use Jensen's inequality: if we prove

$$\mathbb{E}[X_{t+1}^2] \leq (\alpha t + \sqrt{\gamma t + \xi})^2, \tag{12}$$

then it holds that $\mathbb{E}[X_{t+1}] \leq \sqrt{\mathbb{E}[X_{t+1}^2]} \leq \alpha t + \sqrt{\gamma t + \xi}$, which completes the proof.

The base case ($t = 0$) can be easily obtained from the assumption: $\mathbb{E}[X_1^2] \leq \xi$.

We now fix $t \geq 1$, and assume the claim holds for $s = t-1$, i.e.,

$$\mathbb{E}[X_t^2] \leq \left(\alpha(t-1) + \sqrt{\gamma(t-1)+\xi}\right)^2.$$

We note that

$$\mathbb{E}[X_{t+1}^2] \leq \max\{\mathbb{E}[X_t^2] + \alpha\mathbb{E}[X_t] + \gamma, \omega\}.$$

Here, if $\mathbb{E}[X_{t+1}^2] \leq \omega$, then, since $\xi \geq \omega$, it holds that

$$\mathbb{E}[X_{t+1}^2] \leq \omega \leq \xi \leq (\alpha t + \sqrt{\gamma t + \xi})^2$$

since $\alpha, \gamma > 0$.

For the rest of the proof, we consider the case where $\mathbb{E}[X_{t+1}^2] \leq \mathbb{E}[X_t^2] + \alpha\mathbb{E}[X_t] + \gamma$. Let $A = \sqrt{\gamma(t-1)+\xi}$, $B = \alpha(t-1) + A$, and $C = \sqrt{\gamma t + \xi} - A$. Then, by the induction hypothesis and Jensen's inequality, it holds that

$$\mathbb{E}[X_{t+1}^2] \leq \mathbb{E}[X_t^2] + \alpha\mathbb{E}[X_t] + \gamma \leq \mathbb{E}[X_t^2] + \alpha\sqrt{\mathbb{E}[X_t^2]} + \gamma \leq B^2 + \alpha B + \gamma.$$

We now show that

$$B^2 + \alpha B + \gamma \leq (B + C + \alpha)^2 = (\alpha t + \sqrt{\gamma t + \xi})^2. \tag{13}$$

We note that since $A + C = \sqrt{\gamma t + \xi}$, we have

$$(A + C)^2 = \gamma t + \xi = \gamma + \gamma(t-1) + \xi = \gamma + A^2$$

which implies $\gamma = 2CA + C^2$. By applying this relation to the left side of Eq. (13), and by expanding the right side of Eq. (13), it can be written as

$$B^2 + \alpha B + 2CA + C^2 \leq B^2 + C^2 + \alpha^2 + 2BC + 2C\alpha + 2\alpha B$$

which implies

$$2CA \leq \alpha^2 + 2BC + 2C\alpha + \alpha B.$$

We now evaluate this inequality. Since $A \leq B$, we have $2CA \leq 2BC \leq 2BC + \alpha^2 + 2C\alpha + \alpha B$. Therefore, the inductive step holds, and this completes the proof.

B.3 Proof of Theorem 11

In this section, we prove Theorem 11. For each $t \in [T]$, let $\delta_t = w_t - w'_t$ where $(w_t)_{t=1}^T$ and $(w'_t)_{t=1}^T$ are the iterates of $A_{f,\theta}^{\text{SR-GD}}(S)$ and $A_{f,\theta}^{\text{SR-GD}}(S')$ where $S \simeq S'$, respectively. Note that the stochastic rounding operations of two SR-GD parameters are coupled coordinate-wise and monotonically. That is, for each iteration $t \in [T]$ and coordinate $j \in [d]$, two rounding operations in $w_{t,j}$ and $w'_{t,j}$ use the same uniform random variable, independently across t and j. Let B_t be an event that there exists $\tau \le t$ such that $w_\tau \neq w'_\tau$, and A_t be an event that $w_t \neq w'_t$; we define $B_0 = \varnothing$. Since $\mathbb{E}[\|\delta_T\|_2 | B_T^c] = 0$, we have $\mathbb{E}[\|\delta_T\|_2] = \mathbb{P}(B_T)\mathbb{E}[\|\delta_T\|_2 | B_T]$.

Here, the upper bound of $\mathbb{P}(B_T)$ can be obtained by the following lemma. We present the proof of Lemma 26 in Appendix B.3.1.

Lemma 26. *Let $R, \Delta > 0$ and $d, T \in \mathbb{N}$. Consider two sequences $(x_t)_{t=0}^T$ and $(y_t)_{t=0}^T$ such that for each $t \in [T]$, $x_t = x_{t-1} - \mathsf{SR}(p_t)$ and $y_t = y_{t-1} - \mathsf{SR}(q_t)$ where $x_0 = y_0$ and p_t, q_t are d-dimensional random vectors depending on x_{t-1} and y_{t-1}, respectively. For each $t \in [T]$, let B_t be the event that there exists $\tau \in [t]$ such that $x_\tau \neq y_\tau$. Suppose that $\|p_t - q_t\|_2 \le R$ on the event B_{t-1}^c. Then, under the coordinate-wise monotone coupling, it holds that*

$$\mathbb{P}(B_T) \le \min\left\{1, \frac{RT\sqrt{d}}{\Delta}\right\}.$$

On the event B_{t-1}^c, we have $w_{t-1} = w'_{t-1}$, so $\|\eta\nabla f(w_{t-1}; S) - \eta\nabla f(w'_{t-1}; S')\|_2 \le 2\eta L/n$ by the L-Lipschitzness of f. By applying Lemma 26 with the choices of $x_t = w_t$, $p_t = \eta\nabla f(w_{t-1}; S)$, $y_t = w'_t$, and $q_t = \eta\nabla f(w'_{t-1}; S')$ for all $t \in [T]$, we have

$$\mathbb{P}(B_T) \le \min\left\{1, \frac{2\eta LT\sqrt{d}}{n\Delta}\right\}. \tag{14}$$

In the rest of the proof, we show the upper bound on $\mathbb{E}[\|\delta_T\|_2 | B_T]$. To this end, we firstly construct a recurrence with respect to $\mathbb{E}[\|\delta_t\|_2^2 | B_t]$; we then reduce it to the upper bound on $\mathbb{E}[\|\delta_T\|_2 | B_T]$.

We note that for each $t \in [T]$, B_t is the union of two mutually exclusive events, $B_{t-1}^c \cap A_t$ and B_{t-1}. Then, we can decompose $\mathbb{E}[\|\delta_t\|_2^2 | B_t]$ as follows:

$$\begin{aligned}\mathbb{E}[\|\delta_t\|_2^2 | B_t] &= \frac{P(B_{t-1}^c \cap A_t)}{P(B_t)}\mathbb{E}[\|\delta_t\|_2^2 | B_{t-1}^c \cap A_t] + \frac{P(B_{t-1})}{P(B_t)}\mathbb{E}[\|\delta_t\|_2^2 | B_{t-1}] \\ &\le \max\{\mathbb{E}[\|\delta_t\|_2^2 | B_{t-1}^c \cap A_t], \mathbb{E}[\|\delta_t\|_2^2 | B_{t-1}]\}.\end{aligned}$$

To show the upper bound of $\mathbb{E}[\|\delta_t\|_2^2 | B_{t-1}^c \cap A_t]$, we introduce the following lemma. The proof of Lemma 27 is in Appendix B.3.3.

Lemma 27. *Let $\eta, L, \Delta > 0$, $T, d \in \mathbb{N}$, $w_0 \in \mathbb{Z}_\Delta^d$, f be a convex, L-Lipschitz loss function. For $\theta = (\eta, T, w_0)$ and $S \simeq S' \in \mathcal{Z}^n$, let $(w_t)_{t=1}^T$ and $(w'_t)_{t=1}^T$ be the trajectories from $A_{f,\theta}^{\text{SR-GD}}(S)$ and $A_{f,\theta}^{\text{SR-GD}}(S')$, respectively. For each $t \in [T]$, let $\delta_t = w_t - w'_t$ and B_t be an event that there exists $\tau \le t$ such that $w_\tau \neq w'_\tau$, and A_t be an event that $w_t \neq w'_t$. Then, under the coordinate-wise monotone coupling, it holds that*

$$\mathbb{E}[\|\delta_t\|_2^2 | B_{t-1}^c \cap A_t] \le \max\left\{\Delta, \frac{2\eta L}{n}\right\}^2 + \frac{2\Delta\eta L\sqrt{d}}{n}.$$

By Lemma 27, we have $\mathbb{E}[\|\delta_t\|_2^2 | B_{t-1}^c \cap A_t] \le \max\{\Delta, 2\eta L/n\}^2 + 2\Delta\eta L\sqrt{d}/n$.

We now show the upper bound of $\mathbb{E}[\|\delta_t\|_2^2 | B_{t-1}]$. To this end, we introduce the following lemma. The proof of Lemma 28 is in Appendix B.3.5.

Lemma 28. *Let $\eta > 0$, $T, d \in \mathbb{N}$, $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ be a convex, L-Lipschitz, and M-smooth function. Let $S, S' \in \mathcal{Z}^n$ be the neighboring datasets. Then, if $\eta < 2/M$, it holds that*

$$\|x - \eta\nabla f(x; S) - (y - \eta\nabla f(y; S'))\|_2^2 \le \|x - y\|_2^2 + \frac{4\eta L}{n}\|x - y\|_2 + \frac{4\eta^2 L^2}{n^2}.$$

Then, conditioned on B_{t-1}, we have

$$\begin{aligned}\mathbb{E}[\|\delta_t\|_2^2] &\le \mathbb{E}[\|\delta_{t-1} - \eta(\nabla f(w_{t-1}; S) - \nabla f(w'_{t-1}; S'))\|_2^2 + \eta\Delta\|\nabla f(w_{t-1}; S) - \nabla f(w'_{t-1}; S')\|_1] \\ &\le \mathbb{E}[\|\delta_{t-1}\|_2^2] + \frac{4\eta L}{n}\mathbb{E}[\|\delta_{t-1}\|_2] + \frac{4\eta^2 L^2}{n^2} + 2\Delta\eta L\sqrt{d}\end{aligned}$$

where the first inequality follows from Lemma 23 and the last inequality follows from Lemma 28.

Then, we have

$$\mathbb{E}[\|\delta_t\|_2^2|B_t] \le \max\left\{\mathbb{E}[\|\delta_{t-1}\|_2^2|B_{t-1}] + \frac{4\eta L}{n}\mathbb{E}[\|\delta_{t-1}\|_2|B_{t-1}] + \gamma, \omega\right\}$$

where $\gamma = 4\eta^2L^2/n^2 + 2\Delta\eta L\sqrt{d}$ and $\omega = \max\{\Delta, 2\eta L/n\}^2 + 2\Delta\eta L\sqrt{d}/n$.

We now solve the recurrence of $\mathbb{E}[\|\delta_t\|_2^2|B_t]$ using Lemma 25. By Lemma 27, we have $\mathbb{E}[\|\delta_1\|_2^2|B_1] = \mathbb{E}[\|\delta_1\|_2^2|B_0^c \cap A_1] \le \omega = \max\{\Delta, 2\eta L/n\}^2 + 2\Delta\eta L\sqrt{d}/n$. Then, by applying Lemma 25 with the choices of $\alpha = 4\eta L/n$, $\gamma = 4\eta^2L^2/n^2 + 2\Delta\eta L\sqrt{d}$, and $\xi = \max\{\Delta, 2\eta L/n\}^2 + 2\Delta\eta L\sqrt{d}/n$, we have

$$\begin{aligned}\mathbb{E}[\|\delta_T\|_2|B_T] &\le \frac{4\eta L(T-1)}{n} + \sqrt{\gamma(T-1)+\omega} \\ &= \frac{4\eta L(T-1)}{n} + \sqrt{\left(\frac{4\eta^2L^2}{n^2} + 2\Delta\eta L d^{1/2}\right)(T-1) + \max\left\{\Delta, \frac{2\eta L}{n}\right\}^2 + \frac{2\Delta\eta L d^{1/2}}{n}} \\ &\le \frac{6\eta LT}{n} + \sqrt{\Delta^2 + \frac{2\Delta\eta L d^{1/2}}{n} + 2\Delta\eta L d^{1/2}(T-1)}.\end{aligned}$$

where the last inequality follows from the sub-additivity of square root, $\sqrt{4\eta^2L^2(T-1)/n^2} \le 2\eta L(T-1)/n$ and $\max\{\Delta, 2\eta L/n\}^2 \le \Delta^2 + (2\eta L/n)^2$. By multiplying this result by $\mathbb{P}(B_T)$ in Eq. (14), we obtained the desired upper bound, and this completes the proof.

B.3.1 Proof of Lemma 26

In this section, we prove Lemma 26. For each $t \in [T]$, let A_t be the event that $x_t \neq y_t$. Then, since $B_T = \bigcup_{t=1}^T (A_t \cap B_{t-1}^c)$, we have $\mathbb{P}(B_T) \le \sum_{t=1}^T \mathbb{P}(A_t \cap B_{t-1}^c) \le \sum_{t=1}^T \mathbb{P}(A_t|B_{t-1}^c)$. Thus, it suffices to show that

$$\mathbb{P}(A_t|B_{t-1}^c) \le \frac{R\sqrt{d}}{\Delta} \tag{15}$$

for all $t \in [T]$. Then, we have $\mathbb{P}(B_T) \le \sum_{t=1}^T R\sqrt{d}/\Delta = RT\sqrt{d}/\Delta$, which completes the proof.

We now show that Eq. (15) follows. Specifically, we show that for fixed $t \in [T]$,

$$\mathbb{P}(A_t|x_1, \cdots, x_{t-1}, y_1, \cdots, y_{t-1}, B_{t-1}^c) \le \frac{R\sqrt{d}}{\Delta}$$

which directly implies that Eq. (15) follows. Conditioned on $x_1, \cdots, x_{t-1}, y_1, \cdots, y_{t-1}$ and the event B_{t-1}^c, two vectors p_t and q_t are fixed and $x_{t-1} = y_{t-1}$. Thus, the left-hand side of the above inequality is equal to $\mathbb{P}(\mathsf{SR}(p_t) \neq \mathsf{SR}(q_t))$. Here, if $\|p_t - q_t\|_\infty \ge \Delta$, then the desired inequality $\mathbb{P}(\mathsf{SR}(p_t) \neq \mathsf{SR}(q_t)) \le R\sqrt{d}/\Delta$ holds since $R \ge \Delta$ and $d \in \mathbb{N}$. Thus, we assume that $\|p_t - q_t\|_\infty < \Delta$.

To this end, we introduce the following lemma. The proof of Lemma 29 is in Appendix B.3.2.

Lemma 29. *Let $\Delta > 0$ and $x, y \in \mathbb{R}$ such that $|x - y| < \Delta$. Under the monotone coupling, it holds that*

$$|\mathsf{SR}(x) - \mathsf{SR}(y)|/\Delta \sim Ber(|x - y|/\Delta).$$

Let $p_t = (p_{t,1}, \cdots, p_{t,d})$ and $q_t = (q_{t,1}, \cdots, q_{t,d})$. Since $\|p_t - q_t\|_\infty < \Delta$, we have $|p_{t,i} - q_{t,i}| < \Delta$ for all $i \in [d]$. Then, by Lemma 29, we have

$$|\mathsf{SR}(p_{t,i}) - \mathsf{SR}(q_{t,i})| = \begin{cases}\Delta & w.p.\ |p_{t,i} - q_{t,i}|/\Delta, \\ 0 & w.p.\ 1 - |p_{t,i} - q_{t,i}|/\Delta\end{cases} \quad \text{for all } i \in [d].$$

Then, we have

$$\begin{aligned}\mathbb{P}(\mathsf{SR}(p_t) \neq \mathsf{SR}(q_t)) &= \mathbb{P}\left(\mathsf{SR}(p_{t,i}) \neq \mathsf{SR}(q_{t,i}) \text{ for some } i \in [d]\right) \\ &\leq \sum_{i=1}^{d} \mathbb{P}(\mathsf{SR}(p_{t,i}) \neq \mathsf{SR}(q_{t,i})) \\ &= \sum_{i=1}^{d} \frac{|p_{t,i} - q_{t,i}|}{\Delta} \\ &= \frac{\|p_t - q_t\|_1}{\Delta} \leq \frac{\sqrt{d}\|p_t - q_t\|_2}{\Delta} \leq \frac{R\sqrt{d}}{\Delta}.\end{aligned}$$

This completes the proof.

B.3.2 Proof of Lemma 29

In this section, we prove Lemma 29. Without loss of generality, we assume that $x > y$. From the definition of SR (See Section 2.3), we have

$$\mathsf{SR}(x) = \lfloor x \rfloor_\Delta + \Delta \mathbb{1}\left[U \leq \frac{r_x}{\Delta}\right], \quad \mathsf{SR}(y) = \lfloor y \rfloor_\Delta + \Delta \mathbb{1}\left[U \leq \frac{r_y}{\Delta}\right]$$

where $r_x = x - \lfloor x \rfloor_\Delta$, $r_y = y - \lfloor y \rfloor_\Delta$, $U \sim Unif([0,1])$. We note that under the monotone coupling, $\mathsf{SR}(x)$ and $\mathsf{SR}(y)$ use the same U. Then, we have

$$\mathbb{1}\left[U \leq \frac{r_x}{\Delta}\right] - \mathbb{1}\left[U \leq \frac{r_y}{\Delta}\right] = \begin{cases} 1 & \frac{r_y}{\Delta} < U \leq \frac{r_x}{\Delta} \\ -1 & \frac{r_x}{\Delta} < U \leq \frac{r_y}{\Delta} \\ 0 & o.w. \end{cases}$$

Then, we have

$$|\mathsf{SR}(x) - \mathsf{SR}(y)| = \left| \lfloor x \rfloor_\Delta - \lfloor y \rfloor_\Delta + \Delta \left(\mathbb{1}\left[U \leq \frac{r_x}{\Delta}\right] - \mathbb{1}\left[U \leq \frac{r_y}{\Delta}\right] \right) \right|.$$

Since $|x - y| < \Delta$, $\lfloor x \rfloor_\Delta - \lfloor y \rfloor_\Delta \in \{0, \Delta\}$. If $\lfloor x \rfloor_\Delta = \lfloor y \rfloor_\Delta$, then $r_x > r_y$. In this case, $\mathsf{SR}(x) - \mathsf{SR}(y) = \Delta$ if and only if $r_y/\Delta < U < r_x/\Delta$ which occurs with probability $(r_x - r_y)/\Delta = (x - y)/\Delta$.

On the other hand, if $\lfloor x \rfloor_\Delta - \lfloor y \rfloor_\Delta = \Delta$, then $r_x < r_y$. In this case, $|\mathsf{SR}(x) - \mathsf{SR}(y)| = \Delta$ if and only if $U \leq r_x/\Delta$ or $U > r_y/\Delta$, which occurs with probability $r_x/\Delta + (1 - r_y/\Delta) = 1 - (r_y - r_x)/\Delta = (x - y)/\Delta$.

Conclusively, in both cases, we have

$$|\mathsf{SR}(x) - \mathsf{SR}(y)| = \begin{cases} \Delta & w.p.\ (x - y)/\Delta, \\ 0 & w.p.\ 1 - (x - y)/\Delta. \end{cases}$$

which completes the proof.

B.3.3 Proof of Lemma 27

In this section, we prove Lemma 27. Let $t \in [T]$. Conditioned on $B_{t-1}^c \cap A_t$, we have $w_1 = w'_1, \cdots, w_{t-1} = w'_{t-1}$ and $w_t \neq w'_t$. Thus, it holds that

$$\delta_t = \mathsf{SR}(\eta \nabla f(w'_{t-1}; S')) - \mathsf{SR}(\eta \nabla f(w_{t-1}; S)).$$

We show the upper bound on $\mathbb{E}[\|\delta_t\|_2^2 | B_{t-1}^c \cap A_t]$ by applying the following lemma, whose proof is presented in Appendix B.3.4.

Lemma 30. *Let $R > 0$ and $x, y \in \mathbb{R}^d$ such that $0 < \|x - y\|_2 \leq R$. Under the coordinate-wise monotone coupling, we have*

$$\mathbb{E}[\|\mathsf{SR}(x) - \mathsf{SR}(y)\|_2^2 | \mathsf{SR}(x) \neq \mathsf{SR}(y)] \leq \max\{\Delta, R\}^2 + \Delta R\sqrt{d}.$$

Since $w_{t-1} = w'_{t-1}$ and f is L-Lipschitz, it holds that $\|\eta \nabla f(w'_{t-1}; S') - \eta \nabla f(w_{t-1}; S)\|_2 \leq 2\eta L/n$. Then, by applying Lemma 30 with the choice of $R = 2\eta L/n$, we have

$$\mathbb{E}[\|\delta_t\|_2^2 | B_{t-1}^c \cap A_t] \leq \max\left\{\Delta, \frac{2\eta L}{n}\right\}^2 + \frac{2\Delta \eta L \sqrt{d}}{n}.$$

This completes the proof.

B.3.4 Proof of Lemma 30

In this section, we prove Lemma 30. Since $x \neq y$ and $\|\mathsf{SR}(x) - \mathsf{SR}(y)\|_2^2 = 0$ if $\mathsf{SR}(x) = \mathsf{SR}(y)$, we have

$$\begin{aligned}\mathbb{E}[\|\mathsf{SR}(x) - \mathsf{SR}(y)\|_2^2 | \mathsf{SR}(x) \neq \mathsf{SR}(y)] &= \frac{\mathbb{E}[\|\mathsf{SR}(x) - \mathsf{SR}(y)\|_2^2 \mathbb{1}\{\mathsf{SR}(x) \neq \mathsf{SR}(y)\}]}{\mathbb{P}(\mathsf{SR}(x) \neq \mathsf{SR}(y))} \\ &= \frac{\mathbb{E}[\|\mathsf{SR}(x) - \mathsf{SR}(y)\|_2^2]}{\mathbb{P}(\mathsf{SR}(x) \neq \mathsf{SR}(y))}.\end{aligned}$$

We prove Lemma 30 by showing the lower bound of $\mathbb{P}(\mathsf{SR}(x) \neq \mathsf{SR}(y))$ and the upper bound of $\mathbb{E}[\|\mathsf{SR}(x) - \mathsf{SR}(y)\|_2^2]$. We first prove the lower bound of $\mathbb{P}(\mathsf{SR}(x) \neq \mathsf{SR}(y))$. Let $x = (x_1, \cdots, x_d)$ and $y = (y_1, \cdots, y_d)$. For each $i \in [d]$, let $|x_i - y_i| = m_i \Delta + r_i$ for $m_i \in \mathbb{Z}_{\geq 0}$ and $r_i \in [0, \Delta)$. If $m_i \neq 0$, then $\mathsf{SR}(x_i) = \mathsf{SR}(y_i)$ never occurs under monotone coupling. Then, we have

$$\mathbb{P}(\mathsf{SR}(x_i) = \mathsf{SR}(y_i)) = \begin{cases} 1 - r_i/\Delta & m_i = 0, \\ 0 & o.w. \end{cases}$$

We note that if $m_i = 0$, then $r_i = |x_i - y_i|$. If $m_i = 0$ for all $i \in [d]$, i.e., $\|x - y\|_\infty < \Delta$, we have

$$\begin{aligned}\mathbb{P}(\mathsf{SR}(x) \neq \mathsf{SR}(y)) &= 1 - \prod_{i=1}^d \mathbb{P}(\mathsf{SR}(x_i) = \mathsf{SR}(y_i)) \\ &= 1 - \prod_{i=1}^d \left(1 - \frac{|x_i - y_i|}{\Delta}\right) \\ &\geq 1 - \exp\left(-\sum_{i=1}^d \frac{|x_i - y_i|}{\Delta}\right) = 1 - \exp\left(-\frac{\|x - y\|_1}{\Delta}\right)\end{aligned}$$

where the inequality follows from the standard inequality $1 - x \leq \exp(-x)$. If there exists $i \in [d]$ such that $m_i > 0$, i.e., $|x_i - y_i| \geq \Delta$, then $\mathbb{P}(\mathsf{SR}(x) \neq \mathsf{SR}(y)) = 1$.

We now show the upper bound of $\mathbb{E}[\|\mathsf{SR}(x) - \mathsf{SR}(y)\|_2^2]$. If $\|x - y\|_\infty < \Delta$, then by Lemma 29, $|\mathsf{SR}(x_i) - \mathsf{SR}(y_i)|/\Delta \sim Ber(|x_i - y_i|/\Delta)$ for all $i \in [d]$. Then, we have

$$\begin{aligned}\mathbb{E}[\|\mathsf{SR}(x) - \mathsf{SR}(y)\|_2^2] &= \Delta^2 \sum_{i=1}^d \mathbb{E}[(|\mathsf{SR}(x_i) - \mathsf{SR}(y_i)|/\Delta)^2] \\ &= \Delta^2 \sum_{i=1}^d \left(\left(\frac{|x_i - y_i|}{\Delta}\right)^2 + \frac{|x_i - y_i|}{\Delta}\left(1 - \frac{|x_i - y_i|}{\Delta}\right)\right) \\ &= \Delta^2 \times \frac{\sum_{i=1}^d |x_i - y_i|}{\Delta} = \Delta \|x - y\|_1.\end{aligned}$$

Using these results, we now show the upper bound of $\mathbb{E}[\|\mathsf{SR}(x) - \mathsf{SR}(y)\|_2^2 | \mathsf{SR}(x) \neq \mathsf{SR}(y)]$. If $\|x - y\|_\infty \geq \Delta$, the upper bound can be easily obtained by

$$\begin{aligned}\mathbb{E}[\|\mathsf{SR}(x) - \mathsf{SR}(y)\|_2^2 | \mathsf{SR}(x) \neq \mathsf{SR}(y)] &= \frac{\mathbb{E}[\|\mathsf{SR}(x) - \mathsf{SR}(y)\|_2^2]}{\mathbb{P}(\mathsf{SR}(x) \neq \mathsf{SR}(y))} \\ &\leq \|x - y\|_2^2 + \Delta \|x - y\|_1 \\ &\leq R^2 + \Delta R \sqrt{d}.\end{aligned}$$

Thus, we assume that $\|x-y\|_\infty < \Delta$. Let $\gamma = \|x-y\|_1/\Delta$. Then, by applying algebraic fact that $\gamma/(1-\exp(-\gamma)) \le 1+\gamma$, we have

$$\begin{aligned}
\mathbb{E}[\|\mathsf{SR}(x)-\mathsf{SR}(y)\|_2^2|\mathsf{SR}(x)\neq\mathsf{SR}(y)] &= \frac{\mathbb{E}[\|\mathsf{SR}(x)-\mathsf{SR}(y)\|_2^2]}{\mathbb{P}(\mathsf{SR}(x)\neq\mathsf{SR}(y))} \\
&\le \frac{\Delta\|x-y\|_1}{\mathbb{P}(\mathsf{SR}(x)\neq\mathsf{SR}(y))} \\
&\le \Delta^2\frac{\gamma}{1-\exp(-\gamma)} \\
&\le \Delta^2(1+\gamma) \\
&= \Delta^2+\Delta\|x-y\|_1 \le \Delta^2+\Delta\sqrt{d}R.
\end{aligned}$$

Conclusively, we have

$$\mathbb{E}[\|\mathsf{SR}(x)-\mathsf{SR}(y)\|_2^2|\mathsf{SR}(x)\neq\mathsf{SR}(y)] \le \max\{\Delta^2,R^2\}+\Delta R\sqrt{d}$$

which completes the proof.

B.3.5 Proof of Lemma 28

In this section, we prove Lemma 28. To this end, we use the fact that GD on the convex and M smooth function is non-expansive when the learning rate $\eta < 2/M$ (Lemma 24). Without loss of generality, suppose that $S=(z_1,\cdots,z_n)$ and $S'=(z_1',z_2,\cdots,z_n)$ where $z_1\neq z_1'$. Then, since $\nabla f(y;S') = \nabla f(y;S)+(\nabla f(y;z_1')-\nabla f(y;z_1))/n$ and $\|\nabla f(y;z_1)-\nabla f(y;z_1')\|_2 \le 2L$, by the triangle inequality, we have

$$\begin{aligned}
\|x-\eta\nabla f(x;S)-(y-\eta\nabla f(y;S'))\|_2 &\le \|x-\eta\nabla f(x;S)-(y-\eta\nabla f(y;S))\|_2+\frac{2\eta L}{n} \\
&\le \|x-y\|_2+\frac{2\eta L}{n}
\end{aligned}$$

By squaring both sides, we have

$$\|x-\eta\nabla f(x;S)-(y-\eta\nabla f(y;S'))\|_2^2 \le \|x-y\|_2^2+\frac{4\eta L}{n}\|x-y\|_2+\frac{4\eta^2L^2}{n^2}$$

which completes the proof.

B.4 Proof of Theorem 12

In this section, we prove Theorem 12. We first show the upper bound for SR-GD. Let $S = (z_1, \cdots, z_n)$ and $S' = (z'_1, \cdots, z'_n)$ be the neighboring datasets. Let $(w_t)_{t=1}^T$ and $(w'_t)_{t=1}^T$ be the trajectories from $A_{f,\theta}^{\text{SR-GD}}(S)$ and $A_{f,\theta}^{\text{SR-GD}}(S')$, respectively, and $\delta_t = w_t - w'_t$. In this proof, we couple the stochastic rounding operations coordinate-wise and monotonically; the two rounding operations use the same random variable $u_{t,j}$ in Eq. (3) for each $t \in [T]$ and $j \in [d]$.

Conditioned on w_t and w'_t, we have

$$\begin{aligned}\mathbb{E}[\|\delta_{t+1}\|_2^2] &= \mathbb{E}[\|\delta_t - \mathsf{SR}(\eta\nabla f(w_t; S)) + \mathsf{SR}(\eta\nabla f(w'_t; S'))\|_2^2] \\ &\le \|\delta_t - \eta(\nabla f(w_t; S) - \nabla f(w'_t; S'))\|_2^2 + \eta\Delta\|\nabla f(w_t; S) - \nabla f(w'_t; S')\|_1\end{aligned}$$

where the inequality in the second line follows from Lemma 23. Let $a_t = \nabla f(w_t; S) - \nabla f(w'_t; S)$ and $b_t = \nabla f(w'_t; S) - \nabla f(w'_t; S')$. Then, we have $\nabla f(w_t; S) - \nabla f(w'_t; S') = a_t + b_t$. Thus, we have

$$\begin{aligned}\mathbb{E}[\|\delta_{t+1}\|_2^2] &\le \|\delta_t - \eta(a_t + b_t)\|_2^2 + \eta\Delta\|a_t + b_t\|_1 \\ &\le (\|\delta_t - \eta a_t\|_2 + \eta\|b_t\|_2)^2 + \eta\Delta(\|a_t\|_1 + \|b_t\|_1) \\ &\le \|\delta_t - \eta a_t\|_2^2 + \eta\Delta\|a_t\|_1 + 2\eta\|\delta_t - \eta a_t\|_2\|b_t\|_2 + \eta^2\|b_t\|_2^2 + \eta\Delta\|b_t\|_1.\end{aligned}$$

We now introduce the following lemma, where the proof is deferred to Appendix B.4.1.

Lemma 31. *Let $\Delta, L, M, \eta > 0$ such that $\eta < 1/M$. Suppose that a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex, M-smooth, and separable. Then, for any $x, y \in \mathbb{Z}_\Delta^d$, it holds that*

$$\|x - y - \eta(\nabla f(x) - \nabla f(y))\|_2^2 + \eta\Delta\|\nabla f(x) - \nabla f(y)\|_1 \le \|x - y\|_2^2.$$

By Lemma 31, we have $\|\delta_t - \eta a_t\|_2^2 + \eta\Delta\|a_t\|_1 \le \|\delta_t\|_2^2$. Furthermore, since f is convex and M smooth with $\eta < 1/M$, $\|\delta_t - \eta a_t\|_2 \le \|\delta_t\|_2$ by Lemma 24. Moreover, since f is L-Lipschitz and $S \simeq S'$, it holds that $\|b_t\|_2 \le 2L/n$ and hence $\|b_t\|_1 \le 2L\sqrt{d}/n$. Then, we obtain

$$\begin{aligned}\mathbb{E}[\|\delta_{t+1}\|_2^2] &\le \mathbb{E}[\|\delta_t\|_2^2] + \frac{4\eta L}{n}\mathbb{E}[\|\delta_t\|_2] + \frac{4\eta^2L^2}{n^2} + \frac{2\Delta\eta L\sqrt{d}}{n} \\ &= \max\left\{\mathbb{E}[\|\delta_t\|_2^2] + \frac{4\eta L}{n}\mathbb{E}[\|\delta_t\|_2] + \frac{4\eta^2L^2}{n^2} + \frac{2\Delta\eta L\sqrt{d}}{n}, 0\right\}.\end{aligned}$$

We now solve this recursion using Lemma 25. To this end, we evaluate the upper bound of $\mathbb{E}[\|\delta_1\|_2^2]$: by Lemma 23, we have

$$\mathbb{E}[\|\delta_1\|_2^2] = \mathbb{E}[\|\mathsf{SR}(\eta\nabla f(w_0; S)) - \mathsf{SR}(\eta\nabla f(w_0; S'))\|_2^2] \le \frac{4\eta^2L^2}{n^2} + \frac{2\Delta\eta L\sqrt{d}}{n}.$$

Then, by applying Lemma 25 with $\alpha = 4\eta L/n$ and $\gamma = \xi = 4\eta^2L^2/n^2 + 2\Delta\eta L\sqrt{d}/n$, we have

$$\begin{aligned}\mathbb{E}[\|\delta_T\|_2] &\le \frac{4\eta L(T-1)}{n} + \sqrt{\frac{4\eta^2L^2T}{n^2} + \frac{2\Delta\eta Ld^{1/2}T}{n}} \\ &\le \frac{6\eta LT}{n} + \sqrt{\frac{2\Delta\eta Ld^{1/2}T}{n}}\end{aligned}$$

which completes the proof for SR-GD.

We now show the upper bound for SR-SGD. Let $\delta'_t = v_t - v'_t$ where $(v_t)_{t=1}^T$ and $(v'_t)_{t=1}^T$ are the coupled SR-SGD iterates from $A_{f,\theta}^{\text{SR-SGD}}(S)$ and $A_{f,\theta}^{\text{SR-SGD}}(S')$, respectively. Then, we have

$$\mathbb{E}[\|\delta'_t\|_2] = \left(1 - \frac{1}{n}\right)\mathbb{E}[\|\delta'_t\|_2 | z_{i_t} = z'_{i_t}] + \frac{1}{n}\mathbb{E}[\|\delta'_t\|_2 | z_{i_t} \ne z'_{i_t}]. \tag{16}$$

We now bound each term in Eq. (16) using the following lemmas. The proofs of Lemmas 32 and 33 are in Appendices B.4.2 and B.4.3, respectively.

Lemma 32. *Let $d \in \mathbb{N}$, $\eta, \Delta > 0$ and $x, y \in \mathbb{Z}_\Delta^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ be a convex, M-smooth, and separable function and $\eta < 1/M$. Then, under the coordinate-wise monotone coupling, it holds that*

$$\mathbb{E}[\|x - \mathsf{SR}(\eta \nabla f(x)) - (y - \mathsf{SR}(\eta \nabla f(y)))\|_2] \leq \|x - y\|_2$$

Lemma 33. *Let $d \in \mathbb{N}$, $\Delta, \eta, M > 0$ such that $\eta < 1/M$, $f, g : \mathbb{R}^d \to \mathbb{R}$ be convex, L-Lipschitz, M-smooth, and separable functions. Then, under the coordinate-wise monotone coupling, for any $x, y \in \mathbb{Z}_\Delta^d$, it holds that*

$$\mathbb{E}[\|x - \mathsf{SR}(\eta \nabla f(x)) - (y - \mathsf{SR}(\eta \nabla g(y)))\|_2] \leq \|x - y\|_2 + 2\eta L + \sqrt{2\Delta \eta L d^{1/2}}.$$

Lemma 32 implies that if the loss function is convex and separable, and the learning rate η is sufficiently small, then SR-SGD is non-expansive. Furthermore, Lemma 33 implies that if the loss function is convex and separable, and the learning rate η is sufficiently small, then the expected distance between SR-SGD updates is upper bounded by their original distance plus $2\eta L + \sqrt{2\Delta \eta L d^{1/2}}$. By applying Lemmas 32 and 33, we have

$$\begin{aligned}
\mathbb{E}[\|\delta_t'\|_2] &= \left(1 - \frac{1}{n}\right) \mathbb{E}[\|\delta_t'\|_2 | z_{i_t} = z_{i_t}'] + \frac{1}{n} \mathbb{E}[\|\delta_t'\|_2 | z_{i_t} \neq z_{i_t}'] \\
&\leq \mathbb{E}[\|\delta_{t-1}'\|_2] + \frac{2\eta L}{n} + \frac{\sqrt{2\Delta \eta L d^{1/2}}}{n} \leq \frac{2\eta L T}{n} + \frac{T\sqrt{2\Delta \eta L d^{1/2}}}{n}
\end{aligned}$$

which completes the proof.

B.4.1 Proof of Lemma 31

In this section, we prove Lemma 31. For notational convenience, let $F(x) := \nabla f(x)$ for all $x \in \mathbb{R}^d$. By expanding the given expression in Lemma 31, we have

$$\begin{aligned}
&\|x - y\|_2^2 - \eta \langle x - y, F(x) - F(y) \rangle + \eta^2 \|F(x) - F(y)\|_2^2 \\
&\qquad - \eta \langle x - y, F(x) - F(y) \rangle + \eta \Delta \|F(x) - F(y)\|_1.
\end{aligned}$$

Here, it is easy to verify that $-\eta \langle x - y, F(x) - F(y) \rangle + \eta^2 \|F(x) - F(y)\|_2^2 \leq 0$ since gradient for convex and smooth function is co-coercive: $\langle x - y, F(x) - F(y) \rangle \geq 1/M \|F(x) - F(y)\|_2^2$. Since $\eta < 1/M$, we have

$$-\eta \langle x - y, F(x) - F(y) \rangle + \eta^2 \|F(x) - F(y)\|_2^2 \leq \eta \left(\eta - \frac{1}{M}\right) \|F(x) - F(y)\|_2^2 \leq 0.$$

In the rest of the proof, we show that

$$-\eta \langle x - y, F(x) - F(y) \rangle + \eta \Delta \|F(x) - F(y)\|_1 \tag{17}$$

is less than or equal to 0. Since f is separable, there exists $f_1, \cdots, f_d : \mathbb{R} \to \mathbb{R}$ such that $f(x) = \sum_{i=1}^d f_i(x_i)$ for all $(x_1, \cdots, x_d) \in \mathbb{R}^d$. Then, Eq. (17) can be expressed by

$$\eta \sum_{i=1}^d (\Delta |f_i'(x_i) - f_i'(y_i)| - (x_i - y_i)(f_i'(x_i) - f_i'(y_i))).$$

Here, we note that since f is convex, each univariate function f_i is convex. Thus, f_i' is monotone and, therefore $(x_i - y_i)(f_i'(x_i) - f_i'(y_i)) \geq 0$. Furthermore, since $x_i, y_i \in \mathbb{Z}_\Delta$, $|x_i - y_i| \geq \Delta$ if $x_i \neq y_i$. Hence, we have

$$(x_i - y_i)(f_i'(x_i) - f_i'(y_i)) \geq \Delta |f_i'(x_i) - f_i'(y_i)|.$$

Conclusively, we have

$$\|x - y - \eta(\nabla f(x) - \nabla f(y))\|_2^2 + \eta \Delta \|\nabla f(x) - \nabla f(y)\|_1 \leq \|x - y\|_2^2$$

which completes the proof.

B.4.2 Proof of Lemma 32

In this section, we prove Lemma 32. For notational convenience, let $F(x) := \nabla f(x)$ for all $x \in \mathbb{R}^d$. By Lemma 23, we have

$$\mathbb{E}[\|x - y - \mathsf{SR}(\eta F(x)) + \mathsf{SR}(\eta F(y))\|_2^2] \leq \|x - y - \eta(F(x) - F(y))\|_2^2 + \eta\Delta\|F(x) - F(y)\|_1.$$

Then, by Lemma 31, we have

$$\mathbb{E}[\|x - y - \mathsf{SR}(\eta F(x)) + \mathsf{SR}(\eta F(y))\|_2^2] \leq \|x - y\|_2^2.$$

Then, by Jensen's inequality, we have the desired inequality:

$$\begin{aligned}\mathbb{E}[\|x - \mathsf{SR}(\eta\nabla f(x)) - (y - \mathsf{SR}(\eta\nabla f(y)))\|_2] &\leq \sqrt{\mathbb{E}[\|x - \mathsf{SR}(\eta\nabla f(x)) - (y - \mathsf{SR}(\eta\nabla f(y)))\|_2^2]} \\ &\leq \|x - y\|_2\end{aligned}$$

which completes the proof.

B.4.3 Proof of Lemma 33

In this section, we prove Lemma 33. By the triangle inequality, $\mathbb{E}[\|x - y - (\mathsf{SR}(\eta\nabla f(x)) - \mathsf{SR}(\eta\nabla g(y)))\|_2]$ is upper bounded by

$$\mathbb{E}[\|x - y - (\mathsf{SR}(\eta\nabla f(x)) - \mathsf{SR}(\eta\nabla f(y)))\|_2] + \mathbb{E}[\|\mathsf{SR}(\eta\nabla f(y)) - \mathsf{SR}(\eta\nabla g(y))\|_2].$$

Here, by Lemma 32, we have

$$\mathbb{E}[\|x - y - (\mathsf{SR}(\eta\nabla f(x)) - \mathsf{SR}(\eta\nabla f(y)))\|_2] \leq \|x - y\|_2.$$

Furthermore, it holds that

$$\begin{aligned}\mathbb{E}[\|\mathsf{SR}(\eta\nabla f(y)) - \mathsf{SR}(\eta\nabla g(y))\|_2] &\leq \sqrt{\mathbb{E}[\|\mathsf{SR}(\eta\nabla f(y)) - \mathsf{SR}(\eta\nabla g(y))\|_2^2]} \\ &\leq \sqrt{\|\eta\nabla f(y) - \eta\nabla g(y)\|_2^2 + \Delta\|\eta\nabla f(y) - \eta\nabla g(y)\|_1} \\ &\leq \|\eta\nabla f(y) - \eta\nabla g(y)\|_2 + \sqrt{\Delta d^{1/2}\|\eta\nabla f(y) - \eta\nabla g(y)\|_2} \\ &\leq 2\eta L + \sqrt{2\Delta\eta L d^{1/2}}\end{aligned}$$

where the first inequality follows from Jensen's inequality, the inequality in the second line follows from Lemma 23, the inequality in the third line follows from sub-additivity of the square root and the fact that $\|x\|_1 \leq \sqrt{d}\|x\|_2$ for any $x \in \mathbb{R}^d$, and the last inequality follows from the L-Lipschitzness of f and g.

Combining these results, we have

$$\mathbb{E}[\|x - \mathsf{SR}(\eta\nabla f(x)) - (y - \mathsf{SR}(\eta\nabla g(y)))\|_2] \leq \|x - y\|_2 + 2\eta L + \sqrt{2\Delta\eta L d^{1/2}}$$

which completes the proof.

B.5 Proof of Theorem 13

In this section, we prove Theorem 13. To this end, we prove a stronger theorem as stated in the following theorem.

Theorem 34. *Let $\mathcal{Z}$ be a sample space, $\eta, \Delta, L > 0$ and $n, T, d \in \mathbb{N}$. There exists an L-Lipschitz affine loss function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ such that for $\theta = (\eta, T, w_0 = \mathbf{0}_d)$, the following holds: if $\eta L/(n\sqrt{d}) = m_0\Delta + r_0$ where $m_0 \in \mathbb{Z}_{\geq 0}$ and $r_0 \in [0, \Delta)$, then,*

$$\varepsilon_{uas}(A^{\mathsf{SR\text{-}GD}}_{f,\theta}) \gtrsim \frac{\eta LT}{n} + \sqrt{\min\{r_0 Td, \Delta\} r_0 Td}.$$

Furthermore, if $\eta L/\sqrt{d} = m_1\Delta + r_1$ where $m_1 \in \mathbb{Z}_{\geq 0}$ and $r_1 \in [0, \Delta)$, then

$$\varepsilon_{uas}(A^{\mathsf{SR\text{-}SGD}}_{f,\theta}) \gtrsim \min\left\{1, \frac{T}{n}\right\}\left(\eta L + \sqrt{\min\{r_1 d, \Delta\} r_1 d}\right).$$

Theorem 34 provides lower bounds of uniform argument stability of SR-GD and SR-SGD for general $\eta > 0$. We note that when $\eta < n\Delta\sqrt{d}/L$, then $m_0 = 0$ and $r_0 = \eta L/(n\sqrt{d})$, yielding the bound for SR-GD in Theorem 13. Likewise, if $\eta < \Delta\sqrt{d}/L$, then $m_1 = 0$ and $r_1 = \eta L/\sqrt{d}$, resulting in the bound for SR-SGD in Theorem 13.

We now prove Theorem 34. To this end, we fix two distinct samples of $\mathcal{Z}$ and denote them by 0 and 1 for notational simplicity. We consider neighboring datasets $S = (0, \cdots, 0)$, $S' = (1, 0, \cdots, 0)$ and a loss function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ defined by

$$f(w; 0) = 0, \qquad f(w; 1) = -\left\langle \frac{L}{\sqrt{d}}\mathbf{1}_d, w \right\rangle.$$

We note that $\nabla f(w; 0) = \mathbf{0}_d$ and $\nabla f(w; 1) = -(L/\sqrt{d})\mathbf{1}_d$ for all $w \in \mathbb{R}^d$. We now show the lower bound of uniform argument stability of SR-GD and SR-SGD, respectively.

We first consider SR-GD. For $\theta = (\eta, T, w_0 = \mathbf{0}_d)$, let $(w_t)_{t=1}^T$ and $(w'_t)_{t=1}^T$ be the SR-GD iterates from $A^{\mathsf{SR\text{-}GD}}_{f,\theta}(S)$ and $A^{\mathsf{SR\text{-}GD}}_{f,\theta}(S')$, respectively. Obviously, $w_t = \mathbf{0}_d$ almost surely for all $t \in [T]$. Thus, we have $\mathbb{E}[\|w'_T - w_T\|_2] = \mathbb{E}[\|w'_T\|_2]$. To lower bound this expectation, we introduce the following lemma. The proof of Lemma 35 is in Appendix B.5.1.

Lemma 35. *Let $d, T \in \mathbb{N}$, $c, \Delta > 0$ such that $c = m\Delta + r$ for some $m \in \mathbb{Z}_{\geq 0}$ and $r \in [0, \Delta)$. Suppose that the random vectors $x_0, \cdots, x_T \in \mathbb{Z}^d_\Delta$ are recursively defined as follows:*

$$x_0 = \mathbf{0}_d, \quad x_t = x_{t-1} + \mathsf{SR}(c\mathbf{1}_d)$$

for each $t \in [T]$. Then, it holds that for any $t \in [T]$,

$$\mathbb{E}[\|x_t\|_2] = \Omega\left(ct\sqrt{d} + \min\{\sqrt{rtd\Delta}, rtd\}\right).$$

Lemma 35 provides the lower bound of expected 2-norm of SR-GD iterates initialized at the origin, where the loss function is linear and has a gradient in a diagonal direction. Since $\nabla f(w; S') = -L/(n\sqrt{d})\mathbf{1}_d$ for all $w \in \mathbb{R}^d$, we have the update rule for w'_t: for all $t \in [T]$,

$$w'_t = w'_{t-1} - \mathsf{SR}\left(-\frac{\eta L}{n\sqrt{d}}\mathbf{1}_d\right).$$

Here, we note that for any $x \in \mathbb{R}^d$, $\mathsf{SR}(x)$ and $-\mathsf{SR}(-x)$ have the same distribution. Therefore, if we define $(\tilde{w}_t)_{t=1}^T$ such that $\tilde{w}_0 = w_0$ and

$$\tilde{w}_t = \tilde{w}_{t-1} + \mathsf{SR}\left(\frac{\eta L}{n\sqrt{d}}\mathbf{1}_d\right),$$

then $\mathbb{E}[\|w'_T\|_2] = \mathbb{E}[\|\tilde{w}_T\|_2]$. Then, by applying Lemma 35, we have

$$\mathbb{E}[\|w'_T\|_2] = \mathbb{E}[\|\tilde{w}_T\|_2] = \Omega\left(\frac{\eta LT}{n} + \sqrt{\min\{\Delta, r_0 Td\} r_0 Td}\right)$$

which completes the proof for SR-GD.

We now consider SR-SGD. For $\theta = (\eta, T, w_0 = \mathbf{0}_d)$, let $(v_t)_{t=1}^T$ and $(v'_t)_{t=1}^T$ be coupled SR-SGD iterates from $A_{f,\theta}^{\text{SR-SGD}}(S)$ and $A_{f,\theta}^{\text{SR-SGD}}(S')$. Obviously, $v_t = \mathbf{0}_d$ almost surely for all $t \in [T]$. Then, we have $\mathbb{E}[\|v'_T - v_T\|_2] = \mathbb{E}[\|v'_T\|_2]$.

Let $K = |\{t \in [T] : z'_{i_t} = 1\}|$. Then, we have $K \sim Bin(T, 1/n)$. We note that if $z'_{i_t} = 0$ is observed, then $v'_t = v'_{t-1}$ almost surely. Furthermore, if $z'_{i_t} = 1$ is observed, then we have $v'_t = v'_{t-1} - \mathsf{SR}(-(\eta L/\sqrt{d})\mathbf{1}_d)$. Then, it holds that $\mathbb{E}[\|v'_T\|_2] = \mathbb{P}(K \geq 1)\mathbb{E}[\|v'_T\|_2 | K \geq 1]$; if $k = 0$, i.e., $z'_{i_1} = \cdots = z'_{i_T} = 0$, then we obtain $v'_T = 0$. We now show the lower bound on $\mathbb{P}(K \geq 1)$ and $\mathbb{E}[\|v'_T\|_2 | K \geq 1]$.

The lower bound of $\mathbb{P}(K \geq 1)$ is obtained by

$$\mathbb{P}(K \geq 1) = 1 - \left(1 - \frac{1}{n}\right)^T = \Omega\left(\min\left\{1, \frac{T}{n}\right\}\right). \tag{18}$$

We now consider $\mathbb{E}[\|v'_T\|_2 | K \geq 1]$. Given $K = k \geq 1$, let $1 \leq \tau_1 < \cdots < \tau_k \leq T$ be the iteration times where $i_{\tau_j} = 1$. We consider a sequence $(\tilde{v}_t)_{t=1}^k$ such that $\tilde{v}_0 = v_0$ and $\tilde{v}_j = \tilde{v}_{j-1} + \mathsf{SR}((\eta L/\sqrt{d})\mathbf{1}_d)$ for all $j \in [k]$. Then, $\tilde{v}_j$ and v'_{τ_j} have the same distribution for all $j \in [k]$. Then, by Lemma 35, we have

$$\mathbb{E}[\|v'_T\|_2 | K = k] = \mathbb{E}[\|\tilde{v}_k\|_2 | K = k] = \Omega\left(\eta L k + \min\left\{\sqrt{r_1 k d \Delta}, r_1 k d\right\}\right)$$

Since the right-hand side does not decrease in k, on the event $K \geq 1$, it has the minimum value at $k = 1$. By combining this result with Eq. (18), we have

$$\mathbb{E}[\|v'_T\|_2] = \mathbb{P}(K \geq 1)\mathbb{E}[\|v'_T\|_2 | K \geq 1] \gtrsim \min\left\{1, \frac{T}{n}\right\}\left(\eta L + \sqrt{\min\{\Delta, r_1 d\} r_1 d}\right).$$

which completes the proof.

B.5.1 Proof of Lemma 35

In this section, we prove Lemma 35. Since stochastic rounding is applied to any vector coordinate-wise independently, x_t is updated symmetrically with respect to coordinates. Let y_t be a first coordinate of x_t for all $t \in [T]$.

We now show that for all $t \in [T]$, $y_t = m\Delta t + \Delta X_t$ where $X_t \sim Bin(t, r/\Delta)$ via mathematical induction on $t \in [T]$.

We first consider the base case. Since $y_0 = 0$, we have $y_1 = y_0 + \mathsf{SR}(c) = \mathsf{SR}(m\Delta + r) = m\Delta + \mathsf{SR}(r)$. Since $r \in [0, \Delta)$, we have

$$\mathsf{SR}(r) = \begin{cases} 0 & w.p\ 1 - r/\Delta, \\ \Delta & w.p\ r/\Delta. \end{cases}$$

Thus, $y_1 = m\Delta + \Delta X_1$ where $X_1 \sim Ber(r/\Delta) = Bin(1, r/\Delta)$.

We now consider the general case. To this end, assume that $y_t = m\Delta t + \Delta X_t$. Since $\mathsf{SR}(c) = m\Delta + \Delta B_{t+1}$ where $B_{t+1} \sim Ber(r/\Delta)$, we have the following from the update rule of y_t:

$$\begin{aligned} y_{t+1} = y_t + \mathsf{SR}(c) &= m\Delta t + \Delta X_t + m\Delta + \Delta B_{t+1} \\ &= m\Delta(t+1) + \Delta(X_t + B_{t+1}). \end{aligned}$$

Here, since $X_t \sim Bin(t, r/\Delta)$ and $B_{t+1} \sim Ber(r/\Delta)$ are independent, we have $X_t + B_{t+1} \triangleq X_{t+1} \sim Bin(t+1, r/\Delta)$. This completes the mathematical induction.

We now show the lower bound in Lemma 35. Here, if $r = 0$, then the stochastic term is zero, yielding $\mathbb{E}[\|x_t\|_2] = m\Delta t\sqrt{d} = ct\sqrt{d}$. We now assume that $r \in (0, \Delta)$. To this end, we introduce the following technical lemma. The proof of Lemma 36 is in Appendix B.5.2

Lemma 36. *Let $n, d \in \mathbb{N}$, $p \in (0,1)$ and $X_1, \cdots, X_d \sim Bin(n, p)$ be independent. Then, we have*

$$\mathbb{E}\left[\sqrt{\sum_{i=1}^d X_i^2}\right] = \Omega(np\sqrt{d} + \min\{\sqrt{dnp}, dnp\}).$$

By directly computing the expectation of the 2-norm of x_t, we have

$$\begin{aligned}
\mathbb{E}[\|x_t\|_2] = \mathbb{E}\left[\sqrt{\sum_{i=1}^{d} x_{t,i}^2}\right] &= \mathbb{E}\left[\sqrt{\sum_{i=1}^{d} (m\Delta t + \Delta X_{t,i})^2}\right] \\
&\geq \mathbb{E}\left[\sqrt{\sum_{i=1}^{d} (m\Delta t)^2 + (\Delta X_{t,i})^2}\right] \\
&\gtrsim m\sqrt{d}\Delta t + \Delta\mathbb{E}\left[\sqrt{\sum_{i=1}^{d} X_{t,i}^2}\right] \\
&\gtrsim m\sqrt{d}\Delta t + \Delta\left(\frac{rt\sqrt{d}}{\Delta} + \min\left\{\sqrt{\frac{rtd}{\Delta}}, \frac{rtd}{\Delta}\right\}\right) \\
&= (m\Delta + r)t\sqrt{d} + \min\{\sqrt{rtd\Delta}, rtd\} \\
&= ct\sqrt{d} + \min\{\sqrt{rtd\Delta}, rtd\}
\end{aligned}$$

where the inequality in the second line follows from $(a+b)^2 \geq a^2 + b^2$ if $a, b \geq 0$, the asymptotic relation in the third line follows from $\sqrt{a+b} \geq (\sqrt{a}+\sqrt{b})/\sqrt{2}$, the asymptotic relation in the fourth line follows from Lemma 36. This completes the proof.

B.5.2 Proof of Lemma 36

In this section, we prove Lemma 36. Let $Y_1 = \sum_{i=1}^{d} X_i$ and $Y_2 = \sum_{i=1}^{d} X_i^2$. Since $X_1, \cdots, X_d \sim Bin(n,p)$ are independent, it holds that $Y_1 \sim Bin(dn, p)$. We now derive the lower bound of

$$\mathbb{E}\left[\sqrt{\sum_{i=1}^{d} X_i^2}\right] = \mathbb{E}[\sqrt{Y_2}]$$

by showing that $\mathbb{E}[\sqrt{Y_2}] = \Omega(np\sqrt{d})$ and $\mathbb{E}[\sqrt{Y_2}] = \Omega\left(\min\{\sqrt{dnp}, dnp\}\right)$, respectively.

Firstly, by the Cauchy-Schwarz inequality, we have

$$\mathbb{E}[\sqrt{Y_2}] \geq \frac{1}{\sqrt{d}}\mathbb{E}[Y_1] = \frac{dnp}{\sqrt{d}} = np\sqrt{d}.$$

We now show that $\mathbb{E}[\sqrt{Y_2}] = \Omega\left(\min\{\sqrt{dnp}, dnp\}\right)$. Since each X_i takes nonnegative integer values, $\mathbb{E}[\sqrt{Y_2}] \geq \mathbb{E}[\sqrt{Y_1}]$ follows.

Since $Y_1 \sim Bin(dn, p)$, by applying Lemma 22, we have

$$\mathbb{E}[\sqrt{Y_1}] = \Omega\left(\min\left\{\mathbb{E}[Y_1], \sqrt{\mathbb{E}[Y_1]}\right\}\right) = \Omega(\min\{dnp, \sqrt{dnp}\}),$$

which completes the proof.

C Extensions to continuous and supervised sample spaces

In this section, we extend the generalization bounds to continuous sample spaces. We then give supervised-learning constructions on Cartesian product spaces. The stability bounds already hold on arbitrary sample spaces containing at least two distinct elements.

C.1 Generalization lower bounds on $[0, 1]$

Lemma 37. *The generalization lower bounds of Theorems 3 and 9 hold for the uniform distribution on $[0, 1]$, with losses independent of n. The original assumptions and regularity constants are preserved, and the losses can be chosen continuous in the sample argument.*

Proof. We first consider Theorem 3. Let μ, b be the parameters in Eq. (4) and set $A(z) = \mu+b(2z-1)$ and $f(w; z) = -A(z)w$. For uniform $Z \in [0, 1]$, the variable $Y = 2Z - 1$ is centered, bounded by one in absolute value, and has variance $1/3$. Thus, Lemma 15 applies directly, with the same $b \gtrsim L$.

We now consider Theorem 9. Let $f(w; z) = zg(w)$, where g is constructed in Appendix B.1.1. It is easy to observe that the coefficient $A(Z) = Z$ satisfies Eq. (7) with $\rho = 1/4$. The conclusions for both SR-GD and SR-SGD therefore follow from Lemma 20. This completes the proof. □

We note that the same constructions apply to any non-atomic probability space. Indeed, successive equal-probability partitions give a measurable uniform random variable on $[0, 1]$. Composing the losses with this variable preserves the coefficient distributions used in Lemmas 15 and 20.

C.2 Cartesian product spaces and supervised losses

In this section, we show that the lower bounds also hold for supervised losses on Cartesian product sample spaces. To this end, we prove the following lemma.

Lemma 38. *Under the same assumptions on the losses and parameters, the lower bounds of Theorems 3, 4, 6, 9 and 13 can also be attained on a Cartesian product sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with a loss of the form*

$$f(w; (x, y)) = \ell(h(x, w), y).$$

The generalization lower bounds in Lemma 37 can also be attained with non-atomic joint distributions.

Proof. We first consider the affine losses used in Theorems 3, 4, 6 and 13. Let $\mathcal{A} \subseteq [-L, L]$ be a set of coefficients and v be a unit vector, which we specify below for each bound. Define

$$\mathcal{X} = \{\sigma a v : a \in \mathcal{A},\ \sigma \in \{-1, 1\}\}, \qquad \mathcal{Y} = \{-1, 1\},$$

and let $h(x, w) = \langle x, w\rangle$ and $\ell(u, y) = -yu$. Then, for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have

$$f(w; (x, y)) = -y \langle x, w\rangle , \qquad \nabla_w f(w; (x, y)) = -yx.$$

Since $|y| = 1$ and $\|x\|_2 \le L$, the loss is affine and L-Lipschitz on the entire sample space. In particular, it is convex, smooth, and separable.

We now show the generalization lower bounds in Theorems 3 and 6. Set $v = 1$. For Theorem 3, let A have the coefficient distribution in Appendix A.2; for Theorem 6, use the distribution in Appendix A.5. Take $\mathcal{A}$ to be the support of A. Let σ be an independent uniform sign and define $\mathcal{D}$ as the joint distribution of

$$X = \sigma A, \qquad Y = \sigma.$$

Then, for every realization of (A, σ), we have

$$f(w; (X, Y)) = -Aw, \qquad \nabla_w f(w; (X, Y)) = -A.$$

Thus, the iterates and the empirical and population losses coincide with those in the original constructions. Hence, the generalization errors are equal, which proves the lower bounds in Theorems 3 and 6. Since the coefficient sets are finite, the resulting Cartesian product spaces are also finite. For the DR-GD bound in Lemma 37, the same argument applies with $A = \mu + b(2U - 1)$, where U

is uniform on $[0,1]$ and μ, b are defined in Eq. (4). In this case, the resulting joint distribution is non-atomic.

We next show the stability lower bounds in Theorems 4 and 13. For Theorem 4, choose $(a_0, a_1) = (-a, L)$ and $v = 1$, where a is defined in Appendix A.3. For Theorem 13, choose $(a_0, a_1) = (0, L)$ and $v = \mathbf{1}_d/\sqrt{d}$. In both cases, take $\mathcal{A} = \{a_0, a_1\}$ and define

$$z_0 = (a_0 v, 1), \qquad z_1 = (a_1 v, 1), \qquad S = (z_0, \ldots, z_0), \qquad S' = (z_1, z_0, \ldots, z_0).$$

Then $S \simeq S'$ and $f(w; z_i) = -a_i \langle v, w \rangle$ for $i \in \{0, 1\}$. These are exactly the losses used in Appendices A.3 and B.5. Therefore, the coupled iterates and the loss values on these samples are unchanged, and both stability lower bounds follow.

We now consider Theorem 9. Let g be the function constructed in Appendix B.1.1, and define

$$h(x, w) = xg(w), \qquad \ell(u, y) = yu, \qquad f(w; (x, y)) = xyg(w).$$

For a finite sample space, let $\mathcal{X} = \mathcal{Y} = \{0, 1/(m-1), \ldots, 1\}$ for some integer $m \geq 2$ and let $\mathcal{D}$ be uniform on $\mathcal{X} \times \mathcal{Y}$. For the continuous bounds in Lemma 37, take $\mathcal{X} = \mathcal{Y} = [0, 1]$ and let $\mathcal{D}$ be uniform on $[0, 1]^2$. In both cases, X and Y are independent and uniform on their respective spaces. Since $xy \in [0, 1]$, the loss is convex, L-Lipschitz, and $(4L/\Delta)$-smooth for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Furthermore, the coefficient $A = XY$ satisfies

$$\mathbb{P}(A \leq 1/4) \geq \mathbb{P}(X \leq 1/4) \geq 1/4,$$
$$\mathbb{P}(A \geq 1/2) \geq \mathbb{P}(X \geq 3/4, Y \geq 3/4) \geq 1/16.$$

Thus, Eq. (7) holds with $\rho = 1/16$. By applying Lemma 20, we obtain the desired generalization lower bounds for both SR-GD and SR-SGD, which completes the proof. □

D Varying learning rates

In this section, we consider a deterministic learning-rate schedule $(\eta_t)_{t=1}^T$, fixed independently of the data and the algorithmic randomness. We replace η in the t-th update by η_t and write $\theta = ((\eta_t)_{t=1}^T, T, w_0)$. Set

$$H_1 = \sum_{t=1}^{T} \eta_t, \qquad H_2 = \sum_{t=1}^{T} \eta_t^2, \qquad \eta_{\max} = \max_{t\in[T]} \eta_t.$$

The schedule is shared by the two trajectories used in stability analysis.

D.1 Deterministic rounding

Lemma 39 (Generalization upper bound for DR-GD)**.** *Under the assumptions of Theorem 2, with a positive learning rate η_t at each iteration,*

$$\varepsilon_g(A_{f,\theta}^{\text{DR-GD}}; \mathcal{D}) \le 2L^2 H_1 \sqrt{\frac{m-1}{n}}.$$

Proof. By Lemma 14, every update satisfies

$$\|w_t - w_{t-1}\|_2 \le 2\eta_t L, \qquad \|w_T - w_0\|_2 \le 2LH_1.$$

Applying the argument in Appendix A.1 with $2LH_1$ in place of $2\eta LT$ completes the proof. □

Lemma 40 (Stability upper bound for DR-SGD when $d = 1$)**.** *Under the assumptions of Theorem 5, suppose that $0 < \eta_t < 2/M$ for every $t \in [T]$. Then*

$$\varepsilon_{us}(A_{f,\theta}^{\text{DR-SGD}}) \le \frac{L}{n} \sum_{t=1}^{T} (2\eta_t L + \Delta).$$

Proof. Let S, S' be neighboring datasets and use the same sampled indices for the two iterates. For each $t \in [T]$, let $\delta_t = |w_t - w_t'|$. On a shared example, Lemma 16, applied with learning rate η_t, gives $\delta_t \le \delta_{t-1}$. On the differing example, the rounding error bound and Lipschitzness give

$$\delta_t \le \delta_{t-1} + 2\eta_t L + \Delta.$$

The differing index is selected with probability $1/n$, independently of the history. Therefore,

$$\mathbb{E}[\delta_t] \le \mathbb{E}[\delta_{t-1}] + \frac{2\eta_t L + \Delta}{n}.$$

Since $\delta_0 = 0$, summing over $t \in [T]$ and using L-Lipschitzness completes the proof. □

Lemma 41 (General-dimensional stability upper bound for DR-SGD)**.** *Under the assumptions of Theorem 8, for any positive deterministic schedule,*

$$\varepsilon_{us}(A_{f,\theta}^{\text{DR-SGD}}) \le 8L^2 H_1 \min\left\{1, \frac{T}{n}\right\}.$$

Proof. Let B_T denote the event that the differing index is selected by time T. The coupled trajectories are identical on B_T^c, and

$$\mathbb{P}(B_T) = 1 - \left(1 - \frac{1}{n}\right)^T \le \min\left\{1, \frac{T}{n}\right\}.$$

By Lemma 14, each trajectory is at distance at most $2LH_1$ from w_0. Thus, for every sampled index, $\|w_T - w_T'\|_2 \le 4LH_1$. Thus, we have

$$\varepsilon_{us}(A_{f,\theta}^{\text{DR-SGD}}) \le L\mathbb{E}[\|w_T - w_T'\|_2] \le 4L^2 H_1 \mathbb{P}(B_T) \le 8L^2 H_1 \min\left\{1, \frac{T}{n}\right\}.$$

This completes the proof. □

D.2 Stochastic rounding

For stochastic rounding, both the first separation and the distance accumulated after separation can depend on the schedule. We use the following lemma to prove upper bounds in terms of H_1, H_2, and $\eta_{\max}$.

Lemma 42. *Let $a_0 = 0$ and $a_t, \alpha_t, \gamma_t, \omega_t \geq 0$ satisfy*

$$a_t \leq \max\{a_{t-1} + \alpha_t\sqrt{a_{t-1}} + \gamma_t, \omega_t\}, \qquad t \in [T].$$

With $\Omega = \max_{t\in[T]} \omega_t$, we have

$$\sqrt{a_T} \leq \sum_{t=1}^{T} \alpha_t + \sqrt{\Omega + \sum_{t=1}^{T} \gamma_t}.$$

Proof. For notational simplicity, let $A_t = \sum_{s=1}^{t} \alpha_s$ and $C_t = \Omega + \sum_{s=1}^{t} \gamma_s$. We show that $a_t \leq (A_t + \sqrt{C_t})^2$ by mathematical induction on t. The base case is immediate. If $a_{t-1} \leq (A_{t-1} + \sqrt{C_{t-1}})^2$, monotonicity of $x + \alpha_t\sqrt{x}$ gives

$$\begin{aligned} a_{t-1} + \alpha_t\sqrt{a_{t-1}} + \gamma_t &\leq (A_{t-1} + \sqrt{C_{t-1}})^2 + \alpha_t(A_{t-1} + \sqrt{C_{t-1}}) + \gamma_t \\ &\leq (A_{t-1} + \alpha_t + \sqrt{C_{t-1} + \gamma_t})^2. \end{aligned}$$

The second inequality follows by expansion and $\sqrt{C_{t-1} + \gamma_t} \geq \sqrt{C_{t-1}}$. Furthermore, $\omega_t \leq \Omega \leq C_t$. This proves the induction step and completes the proof. □

Lemma 43 (Uniform argument stability upper bounds for stochastic rounding)**.** *Let $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ be convex, L-Lipschitz, and M-smooth, with $w_0 \in \mathbb{Z}_\Delta^d$ and $0 < \eta_t < 2/M$ for all $t \in [T]$. Define*

$$P_{\text{GD}} = \min\left\{1, \frac{2L\sqrt{d}\,H_1}{n\Delta}\right\}, \qquad P_{\text{SGD}} = 1 - \left(1 - \frac{1}{n}\right)^T,$$

$$\Omega_{\text{GD}} = \max\left\{\Delta, \frac{2L\eta_{\max}}{n}\right\}^2 + \frac{2\Delta L\sqrt{d}\,\eta_{\max}}{n}, \qquad \Omega_{\text{SGD}} = 4L^2\eta_{\max}^2 + 2\Delta L\sqrt{d}\,\eta_{\max}.$$

Then, it holds that

$$\varepsilon_{uas}(A_{f,\theta}^{\text{SR-GD}}) \leq P_{\text{GD}}\left(\frac{4LH_1}{n} + \sqrt{\Omega_{\text{GD}} + \frac{4L^2H_2}{n^2} + 2\Delta L\sqrt{d}\,H_1}\right),$$

$$\varepsilon_{uas}(A_{f,\theta}^{\text{SR-SGD}}) \leq P_{\text{SGD}}\left(\frac{4LH_1}{n} + \sqrt{\Omega_{\text{SGD}} + \frac{4L^2H_2}{n} + 2\Delta L\sqrt{d}\,H_1}\right).$$

Proof. Let S, S' be neighboring datasets with differing index i_*, and let $\delta_t = w_t - w_t'$ denote the difference between their coupled iterates. For SR-GD, let B_t denote the event that the trajectories have separated at least once by time t. For SR-SGD, let B_t denote the event that i_* has been selected by time t. In both cases, $B_{t-1} \subseteq B_t$ and $\delta_t = 0$ on B_t^c.

Set $B_0 = \emptyset$, $a_0 = 0$, and

$$a_t = \mathbb{E}[\|\delta_t\|_2^2 \mid B_t], \qquad t \in [T],$$

with $a_t = 0$ when $\mathbb{P}(B_t) = 0$. We first bound $\mathbb{P}(B_T)$ and the conditional second moments on $B_{t-1}^c \cap B_t$ and B_{t-1} for each algorithm.

We first consider SR-GD. Define

$$p_t = \eta_t\nabla f(w_{t-1}; S), \qquad q_t = \eta_t\nabla f(w_{t-1}'; S').$$

On B_{t-1}^c, the previous iterates are equal. Since S, S' differ at one index and f is L-Lipschitz, we have

$$\|p_t - q_t\|_2 \leq \frac{2\eta_t L}{n}, \qquad \|p_t - q_t\|_1 \leq \frac{2\eta_t L\sqrt{d}}{n}.$$

Under monotone stochastic rounding, the probability that the rounded inputs differ in a coordinate is at most their absolute difference divided by Δ. A union bound over coordinates therefore gives

$$\mathbb{P}(B_t \mid B_{t-1}^c) \le \frac{2\eta_t L\sqrt{d}}{n\Delta}.$$

Summing over the possible first separation times yields

$$\mathbb{P}(B_T) = \sum_{t=1}^{T} \mathbb{P}(B_{t-1}^c \cap B_t) \le \min\left\{1, \sum_{t=1}^{T} \frac{2\eta_t L\sqrt{d}}{n\Delta}\right\} = P_{\text{GD}}.$$

On $B_{t-1}^c \cap B_t$, we have $\delta_{t-1} = 0$ and $\delta_t \neq 0$. Applying Lemma 27 with η_t in place of η gives

$$\mathbb{E}[\|\delta_t\|_2^2 \mid B_{t-1}^c \cap B_t] \le \max\left\{\Delta, \frac{2\eta_t L}{n}\right\}^2 + \frac{2\Delta\eta_t L\sqrt{d}}{n} \triangleq \omega_t^{\text{GD}}.$$

We next consider B_{t-1}. Since $\eta_t < 2/M$, by Lemma 28, we have

$$\|\delta_{t-1} - p_t + q_t\|_2^2 \le \|\delta_{t-1}\|_2^2 + \frac{4\eta_t L}{n}\|\delta_{t-1}\|_2 + \frac{4\eta_t^2 L^2}{n^2}.$$

Furthermore, L-Lipschitzness gives $\|p_t - q_t\|_1 \le 2\eta_t L\sqrt{d}$. Thus, by Lemma 23, conditioning first on the previous iterates and then on B_{t-1} yields

$$\begin{aligned}\mathbb{E}[\|\delta_t\|_2^2 \mid B_{t-1}] &\le \mathbb{E}[\|\delta_{t-1} - p_t + q_t\|_2^2 + \Delta\|p_t - q_t\|_1 \mid B_{t-1}] \\ &\le a_{t-1} + \frac{4\eta_t L}{n}\mathbb{E}[\|\delta_{t-1}\|_2 \mid B_{t-1}] + \gamma_t^{\text{GD}},\end{aligned}$$

where

$$\gamma_t^{\text{GD}} = \frac{4\eta_t^2 L^2}{n^2} + 2\Delta\eta_t L\sqrt{d}.$$

We now consider SR-SGD. Since the indices are sampled independently and uniformly from $[n]$,

$$\mathbb{P}(B_T) = 1 - \left(1 - \frac{1}{n}\right)^T = P_{\text{SGD}}.$$

In this case, define

$$p_t = \eta_t \nabla f(w_{t-1}; z_{i_t}), \qquad q_t = \eta_t \nabla f(w'_{t-1}; z'_{i_t}).$$

By L-Lipschitzness,

$$\|p_t - q_t\|_2 \le 2\eta_t L, \qquad \|p_t - q_t\|_1 \le 2\eta_t L\sqrt{d}.$$

On $B_{t-1}^c \cap B_t$, the previous iterates are equal and $i_t = i_*$. Since this event does not depend on the rounding randomness at iteration t, Lemma 23 gives

$$\mathbb{E}[\|\delta_t\|_2^2 \mid B_{t-1}^c \cap B_t] \le 4\eta_t^2 L^2 + 2\Delta\eta_t L\sqrt{d} \triangleq \omega_t^{\text{SGD}}.$$

We next consider B_{t-1}. If $i_t \neq i_*$, the two updates use the same loss, so Lemma 24 applies. If $i_t = i_*$, the triangle inequality and L-Lipschitzness apply instead. Consequently,

$$\|\delta_{t-1} - p_t + q_t\|_2^2 \le \begin{cases} \|\delta_{t-1}\|_2^2, & i_t \neq i_*, \\ (\|\delta_{t-1}\|_2 + 2\eta_t L)^2, & i_t = i_*. \end{cases}$$

The current index is independent of the previous indices and rounding randomness. Hence, even conditioned on B_{t-1}, the differing index is selected with probability $1/n$. Applying Lemma 23 and averaging over the current index gives

$$\begin{aligned}\mathbb{E}[\|\delta_t\|_2^2 \mid B_{t-1}] &\le \left(1 - \frac{1}{n}\right) a_{t-1} + \frac{1}{n}\mathbb{E}[(\|\delta_{t-1}\|_2 + 2\eta_t L)^2 \mid B_{t-1}] + 2\Delta\eta_t L\sqrt{d} \\ &= a_{t-1} + \frac{4\eta_t L}{n}\mathbb{E}[\|\delta_{t-1}\|_2 \mid B_{t-1}] + \gamma_t^{\text{SGD}},\end{aligned}$$

where

$$\gamma_t^{\mathrm{SGD}} = \frac{4\eta_t^2 L^2}{n} + 2\Delta\eta_t L\sqrt{d}.$$

We now complete the proof for both algorithms. Write (γ_t, ω_t) for the corresponding pair defined above. By Jensen's inequality,

$$\mathbb{E}[\|\delta_{t-1}\|_2 \mid B_{t-1}] \le \sqrt{a_{t-1}}.$$

Since B_t is the union of the disjoint events B_{t-1} and $B_{t-1}^c \cap B_t$, the law of total expectation gives

$$\begin{aligned} a_t &= \frac{\mathbb{P}(B_{t-1})}{\mathbb{P}(B_t)}\mathbb{E}[\|\delta_t\|_2^2 \mid B_{t-1}] + \frac{\mathbb{P}(B_{t-1}^c \cap B_t)}{\mathbb{P}(B_t)}\mathbb{E}[\|\delta_t\|_2^2 \mid B_{t-1}^c \cap B_t] \\ &\le \max\left\{ a_{t-1} + \frac{4\eta_t L}{n}\sqrt{a_{t-1}} + \gamma_t, \omega_t \right\}. \end{aligned}$$

Applying Lemma 42 with $\alpha_t = 4\eta_t L/n$ and $\Omega = \max_{t\in[T]} \omega_t$, we obtain

$$\sqrt{a_T} \le \sum_{t=1}^{T} \frac{4\eta_t L}{n} + \sqrt{\Omega + \sum_{t=1}^{T} \gamma_t} = \frac{4LH_1}{n} + \sqrt{\Omega + \sum_{t=1}^{T} \gamma_t}.$$

Finally, since $\delta_T = 0$ on B_T^c, we have

$$\mathbb{E}[\|\delta_T\|_2] = \mathbb{P}(B_T)\mathbb{E}[\|\delta_T\|_2 \mid B_T] \le \mathbb{P}(B_T)\sqrt{a_T}.$$

For SR-GD, we have

$$\Omega = \Omega_{\mathrm{GD}}, \qquad \sum_{t=1}^{T} \gamma_t = \frac{4L^2 H_2}{n^2} + 2\Delta L\sqrt{d}\, H_1.$$

For SR-SGD, we have

$$\Omega = \Omega_{\mathrm{SGD}}, \qquad \sum_{t=1}^{T} \gamma_t = \frac{4L^2 H_2}{n} + 2\Delta L\sqrt{d}\, H_1.$$

Substituting these expressions and the respective bounds on $\mathbb{P}(B_T)$ gives the two stated bounds. Taking the supremum over neighboring datasets S, S' completes the proof.

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contribution of the paper, and these claims are formally supported by the theoretical results in Sections 3 and 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The paper states the assumption on the loss function, rounding scheme, and optimization in Section 2 and each theorem statement in Sections 3 and 4. The main proof ideas are described in Sections 3 and 4, and the full proofs are provided in the appendices.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: This paper does not include experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: This paper does not include experimental results requiring data or code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [N/A]

Justification: Our paper does not include experiments. All algorithmic settings related to the theoretical results are specified in each theorem statement and Section 2.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification: This paper does not include experiments, so statistical significance tests or error bars are not applicable.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This paper does not include computational experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This paper is a theoretical study of generalization and stability under fixed-point parameter spaces and does not involve any harmful consequences.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is a foundational and theoretical study of generalization and stability properties of algorithms over a discrete space. We could not find notable potential societal consequences of our work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not release data, model, or systems that pose a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This work does not use existing code, datasets, pretrained models, or other external assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not introduce or release new assets such as datasets, code, model.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing, user studies, or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology and results of this paper do not involve LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.