Stability and Generalization for Randomized Coordinate Descent

Puyu Wang[1] , **Liang Wu**[2*] and **Yunwen Lei**[3]
[1]School of Mathematics, Northwest University, Xi'an 710127, China
[2]Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China
[3]School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK
wangpuyu@stumail.nwu.edu.cn, wuliang@swufe.edu.cn, y.lei@bham.ac.uk

Abstract

Randomized coordinate descent (RCD) is a popular optimization algorithm with wide applications in solving various machine learning problems, which motivates a lot of theoretical analysis on its convergence behavior. As a comparison, there is no work studying how the models trained by RCD would generalize to test examples. In this paper, we initialize the generalization analysis of RCD by leveraging the powerful tool of algorithmic stability. We establish argument stability bounds of RCD for both convex and strongly convex objectives, from which we develop optimal generalization bounds by showing how to early-stop the algorithm to tradeoff the estimation and optimization. Our analysis shows that RCD enjoys better stability as compared to stochastic gradient descent.

1 Introduction

Randomized coordinate descent (RCD) is a popular method for solving large-scale optimization problems which are ubiquitous in the big-data era [Nesterov, 2012; Richtárik and Takáč, 2014; Richtárik and Takáč, 2016; Zeng *et al.*, 2019]. As an iterative algorithm, it iteratively updates a single randomly chosen coordinate along the negative direction of the derivative while keeping the other coordinates fixed. Due to its ease of implementation and high efficiency, RCD has found wide applications in various areas such as compressed sensing, network problems and optimization [Richtárik and Takáč, 2014]. In particular, the conceptual and algorithmic simplicity makes it especially useful for large-scale problems for which even the simplest full-dimensional vector operations are very expensive [Nesterov, 2012].

The popularity of RCD motivates a lot of theoretical analysis to understand its empirical behavior. Specifically, iteration complexities of RCD are well studied in the literature under different settings (e.g., convex/strongly convex [Nesterov, 2012], smooth/nonsmooth cases [Richtárik and Takáč, 2014; Richtárik and Takáč, 2016; Lu and Xiao, 2015]) for different variants (e.g., distributed RCD [Richtárik and Takáč, 2016], accelerated RCD [Nesterov, 2012; Ren and Zhu, 2017; Gu *et al.*, 2018; Li and Lin, 2020; Chen and Gu, 2016] and RCD for primal-dual problems [Qu *et al.*, 2016]). These discussions concern how the empirical risks of the models trained by RCD would decay along the optimization process. As a comparison, there is little analysis on how these models would behave on testing examples, which is what really matters in machine learning. Actually, if the models are very complicated, it is very likely that the models would admit a small empirical risk or even interpolate the training examples but meanwhile suffer from a large test error. This discrepancy between training and testing, as referred to as overfitting, is a fundamental problem in machine learning [Bousquet and Elisseeff, 2002]. The existing convergence analysis of RCD is not enough to fully understand why models trained by RCD have a good prediction performance in real applications. In particular, it is not clear how the optimization and statistical behavior of RCD would change along the optimization process, which is useful for designing efficient models in practice. For example, generalization analysis provides a principled guideline on how to stop the algorithm appropriately for a best generalization.

In this paper, we aim to bridge the generalization and optimization of RCD by leveraging the celebrated concept of algorithmic stability. We establish stability bounds of RCD as measured by several concepts, including ℓ_1-argument stability, ℓ_2-argument stability and uniform stability. Under standard assumptions on smoothness, Lipschitz continuity and convexity of objective functions, we show clearly how the stability and the optimization error would behave along the learning process. This suggests a principled way to early-stop the algorithm to get a best generalization behavior. We consider convex, strongly convex and nonconvex objective functions. In the convex and strongly convex cases, we develop minimax optimal generalization bounds of the order $O(1/\sqrt{n})$ and $O(1/n)$ respectively, where n is the sample size. Our analysis not only suggests that RCD has a better stability than stochastic gradient descent (SGD), but also is able to exploit a low noise condition to get an optimistic bound $O(1/n)$ in the convex case. Finally, we develop generalization bounds with high probability which are useful to understand the robustness and variation of the training algorithm [Feldman and Vondrak, 2019].

We survey related work in Section 2 and formulate the problem in Section 3. We give stability and generalization

*Corresponding Author

bounds in Section 4 and Section 5. We report experimental results in Section 6 and present some proofs in Section 7. The full proofs can be found in the arXiv version of this paper.

2 Related Work

2.1 Randomized Coordinate Descent

RCD was widely used to solve large-scale optimization problems in machine learning, including linear SVMs [Chang *et al.*, 2008], ℓ_1-regularized models for sparse learning [Shalev-Shwartz and Tewari, 2009] and low-rank matrix learning [Hu and Kwok, 2019]. The convergence rate of RCD and its accelerated variant were studied in the seminal work [Nesterov, 2012], where the advantage of RCD over deterministic algorithms is clearly illustrated. These results were extended to structure optimization problems where the objective function consists of a smooth data-fitting term and a nonsmooth regularizer [Richtárik and Takáč, 2014; Lu and Xiao, 2015]. RCD was also adapted to distributed data analysis [Richtárik and Takáč, 2016; Sun *et al.*, 2017; Xiao *et al.*, 2019], primal-dual optimization [Qu *et al.*, 2016] and privacy-preserving problems [Damaskinos *et al.*, 2020]. All these discussions consider the convergence rate of optimization errors for RCD. As a comparison, we are interested in the generalization behavior of models trained by RCD, which is the ultimate goal in machine learning.

2.2 Stability and Generalization

We now review the related work on algorithmic stability and its application on generalization analysis. The framework of algorithmic stability was established in a seminal paper [Bousquet and Elisseeff, 2002], where the important uniform stability was introduced. This algorithmic stability was extended to study randomized algorithms in Elisseeff *et al.* [2005]. Other than uniform stability, several other stability measures including hypothesis stability [Bousquet and Elisseeff, 2002], on-average stability [Shalev-Shwartz *et al.*, 2010] and argument stability [Liu *et al.*, 2017] have been introduced in statistical learning theory, whose connection to learnability has been established [Mukherjee and Zhou, 2006; Shalev-Shwartz *et al.*, 2010]. The uniform stability of stochastic gradient descent (SGD) was established for learning with (strongly) convex, smooth and Lipschitz loss functions [Hardt *et al.*, 2016]. This motivates the recent work of studying generalization of stochastic optimization algorithms via several stability [Meng *et al.*, 2017; Charles and Papailiopoulos, 2018; Kuzborskij and Lampert, 2018; Yin *et al.*, 2018; Yuan *et al.*, 2019; Lei and Ying, 2020; Bassily *et al.*, 2020; Lei *et al.*, 2020; Lei and Ying, 2021; Wang *et al.*, 2021; Yang *et al.*, 2021]. For example, an on-average model stability [Lei and Ying, 2020] has been proposed to remove the smoothness assumption or Lipschitz continuity assumption in Hardt *et al.* [2016]. Recently, elegant concentration inequalities have been developed to get high-probability bounds via uniform stability [Feldman and Vondrak, 2019; Bousquet *et al.*, 2020]. To our knowledge, the algorithmic stability of RCD has not been studied yet, which is the topic of this paper.

3 Problem Formulation

Let ρ be a probability measure defined over a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is an input space and $\mathcal{Y} \subset \mathbb{R}$ is an output space. We aim to build a parametric model $h_{\mathbf{w}} : \mathcal{X} \mapsto \mathbb{R}$, where $\mathbf{w}$ is the model parameter which belongs to the parameter space $\mathcal{W} \subseteq \mathbb{R}^d$. The performance of the model $h_{\mathbf{w}}$ on a single example z can be measured by a nonnegative loss function $f(\mathbf{w}; z)$. The quality of a model can be quantified by a population risk $F(\mathbf{w}) = \mathbb{E}_z[f(\mathbf{w}; z)]$, where $\mathbb{E}_z$ denotes the expectation w.r.t. z. We wish to approximate the best model $\mathbf{w}^* \in \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$. However, the probability measure is often unknown and we only have access to a training sample $S = \{z_1, z_2, \ldots, z_n\}$ drawn independently from ρ. The empirical behavior of $h_{\mathbf{w}}$ on S can be measured by a empirical risk $F_S(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{w}; z_i)$.

We give necessary *notations*. For any $\mathbf{x} \in \mathbb{R}^d$, we denote the norm $\|\mathbf{x}\|_p = \big(\sum_{i=1}^d |\mathbf{x}_i|^p\big)^{1/p}$ for $p \geq 1$. For any $m \in \mathbb{N}$, we denote $[m] := \{1, \ldots, m\}$. We use the notation $B = O(\tilde{B})$ if there exists a constant $c_0 > 0$ such that $B \leq c_0\tilde{B}$, and use $B \asymp \widetilde{B}$ if there exist constants $c_1, c_2 > 0$ such that $c_1\widetilde{B} < B \leq c_2\widetilde{B}$. We say $g : \mathcal{W} \mapsto \mathbb{R}$ is L-smooth if $\|\nabla g(\mathbf{w}) - \nabla g(\mathbf{w}')\|_2 \leq L\|\mathbf{w} - \mathbf{w}'\|_2$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$.

We apply a randomized algorithm A to the sample S and get an output model $A(S) \in \mathcal{W}$. We are interested in studying the *excess generalization error* $F(A(S)) - F(\mathbf{w}^*)$. Since $\mathbb{E}[F_S(\mathbf{w}^*)] = F(\mathbf{w}^*)$, we have the decomposition

$$\begin{aligned}\mathbb{E}_{S,A}\big[F(A(S)) - F(\mathbf{w}^*)\big] &= \mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] \\ &+ \mathbb{E}_{S,A}\big[F_S(A(S)) - F_S(\mathbf{w}^*)\big]. \qquad (1)\end{aligned}$$

We refer to the first term $\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big]$ as the estimation error, and the second term $\mathbb{E}_{S,A}\big[F_S(A(S)) - F_S(\mathbf{w}^*)\big]$ as the optimization error. A standard approach to control estimation error is to study the algorithmic stability of the algorithm A, i.e., how the model would change if we change the training sample by a single example. There are several variants of stability measures including the uniform stability, hypothesis stability, on-average stability and argument stability [Bousquet and Elisseeff, 2002; Hardt *et al.*, 2016; Elisseeff *et al.*, 2005], among which the uniform stability is the most popular one.

Definition 1 (Uniform Stability)**.** A randomized algorithm A is ϵ-uniformly stable if for all datasets $S, \widetilde{S} \in \mathcal{Z}^n$ differing by one example, we have $\sup_z \big[f(A(S); z) - f(A(\widetilde{S}); z)\big] \leq \epsilon$.

We consider the on-average argument stability, an advantage of which is that it can imply better generalization bounds without a Lipschitz continuity assumption on loss functions.

Definition 2 (On-average Argument Stability)**.** Let $S = \{z_1, \ldots, z_n\}$ and $S' = \{z'_1, \ldots, z'_n\}$ be drawn independently from ρ. For any $i = 1, \ldots, n$, define $S^{(i)} = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n\}$ as the set formed from S by replacing z_i with z'_i. We say a randomized algorithm A is ℓ_1 on-average argument ϵ-stable if $\mathbb{E}_{S,S',A}\big[\frac{1}{n}\sum_{i=1}^n \|A(S) - A(S^{(i)})\|_2\big] \leq \epsilon$, and ℓ_2 on-average argument ϵ-stable if $\mathbb{E}_{S,S',A}\big[\frac{1}{n}\sum_{i=1}^n \|A(S) - A(S^{(i)})\|_2^2\big] \leq \epsilon^2$.

Lemma 1 gives a connection between on-average stability and generalization [Lei and Ying, 2020]. Assumption 1 holds for popular loss including logistic loss and Huber loss.

Assumption 1. Let $G_1, G_2 > 0$. Assume for all $\mathbf{w} \in \mathcal{W}$ and $z \in \mathcal{Z}$, $\|\nabla f(\mathbf{w}; z)\|_1 \leq G_1$ and $\|\nabla f(\mathbf{w}; z)\|_2 \leq G_2$.

Lemma 1. *Let S, S' and $S^{(i)}$ be constructed as Definition 2.*

(a) If Assumption 1 holds, then

$$\big|\mathbb{E}_{S,A}\big[F_S(A(S)) - F(A(S))\big]\big| \leq \frac{G_2}{n}\mathbb{E}_{S,S',A}\Big[\sum_{i=1}^{n}\|A(S) - A(S^{(i)})\|_2\Big].$$

(b) If for any z, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative and L-smooth, then for any $\gamma > 0$ we have

$$\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] \leq \frac{1}{\gamma}\mathbb{E}_{S,A}\big[F_S(A(S))\big] + \frac{L(1+\gamma)}{2n}\sum_{i=1}^{n}\mathbb{E}_{S,S',A}\big[\|A(S^{(i)}) - A(S)\|_2^2\big].$$

In this paper, we consider the specific RCD method widely used in large-scale learning problems. Let $\mathbf{w}_1 \in \mathcal{W}$ be the initial point. At the t-th iteration it first randomly selects a single coordinate $i_t \in [d]$, and then performs the update along the i_t-th coordinate as [Nesterov, 2012]

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla_{i_t} F_S(\mathbf{w}_t)\mathbf{e}_{i_t}, \tag{2}$$

where $\nabla_i g$ denotes the derivative of g w.r.t. the i-th coordinate and $\mathbf{e}_i$ is a vector in $\mathbb{R}^d$ with the i-th coordinate being 1 and other coordinates being 0. Here $\{\eta_t\}$ is a nonnegative stepsize sequence. It is clear that RCD sequentially updates a randomly selected coordinate while keeping others fixed. In this paper, we consider the update of only a single coordinate per iteration. Our discussions can be readily extended to randomized block coordinate descent where the coordinates are partitioned into blocks, and each block of coordinates is updated per iteration [Nesterov, 2012].

4 Stability of RCD

In this section, we present our stability bounds of RCD. To this aim, we first introduce several standard assumptions [Nesterov, 2012]. The first assumption is the convexity of the empirical risk. Note we do not require the convexity of each loss function, which is used in the stability analysis of SGD [Hardt *et al.*, 2016; Kuzborskij and Lampert, 2018].

Assumption 2. For any training dataset set S, F_S is convex.

Our second assumption is the coordinate-wise smoothness.

Definition 3. We say a differentiable function $g : \mathcal{W} \mapsto \mathbb{R}$ has coordinate-wise Lipschitz continuous gradients with parameter $\widetilde{L} > 0$ if the following inequality holds for all $\alpha \in \mathbb{R}, \mathbf{w} \in \mathcal{W}, i \in [d]$ $g(\mathbf{w}+\alpha\mathbf{e}_i) \leq g(\mathbf{w}) + \alpha\nabla_i g(\mathbf{w}) + \widetilde{L}\alpha^2/2$.

Assumption 3. For any training dataset S, F_S is L-smooth and has coordinate-wise Lipschitz continuous gradients with parameter $\widetilde{L} > 0$.

We first consider *convex* problems. Part (a) of Theorem 2 considers the ℓ_1 argument stability, while Part (b) considers ℓ_2 argument stability. The proof is given in Section 7.

Theorem 2. *Let Assumptions 2, 3 hold. Let $\{\mathbf{w}_t\}, \{\mathbf{w}_t^{(i)}\}$ be given by (2) with $\eta_t \leq 2/\widetilde{L}$ based on S and $S^{(i)}$, respectively.*

(a) If Assumption 1 holds, then

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,S',A}\Big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2\Big] \leq \frac{2G_1}{nd}\sum_{k=1}^{t}\eta_k. \tag{3}$$

(b) For any $p > 0$ the ℓ_2 on-average argument stability can be bounded by

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_A\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \leq \frac{4L(1+1/p)}{n^2 d}\times \sum_{j=1}^{t}\big(1+p\big)^{t-j}\eta_j^2\mathbb{E}_A\big[F_S(\mathbf{w}_j) + F_{S'}(\mathbf{w}_j)\big]. \tag{4}$$

Remark 1. We compare Theorem 2 with related work. Under Assumptions 1, 2 and 3, it was shown SGD with t iterations enjoys the ℓ_1 argument stability bound $O(\frac{G_2}{n}\sum_{k=1}^{t}\eta_k)$. Eq. (3) shows RCD admits a better stability since there is a d in the denominator. Note in the worst case we can choose $G_1 \leq \sqrt{d}G_2$ for which our stability bounds of RCD are of the order $O(\frac{G_2}{n\sqrt{d}}\sum_{k=1}^{t}\eta_k)$. A notable property of Part (b) is that the stability bound (4) does not require the Lipschitz condition as $\|\nabla f(\mathbf{w}; z)\|_2 \leq G_2$, which is widely used in the existing stability analysis [Hardt *et al.*, 2016; Charles and Papailiopoulos, 2018; Kuzborskij and Lampert, 2018]. Indeed, a key point here is that we replace the Lipschitz constant G_2 by empirical/population risks F and F_S. Since we are minimizing the empirical risk by RCD, it is reasonable that F and F_S would be small and in this case the algorithm would be more stable. This gives an intuitive connection between stability and optimization: a small optimization error is also beneficial to improve stability.

We now consider *strongly convex* case. Theorem 2 shows the stability becomes worse as we run more iterations. In the following theorem, we show the stability can be further improved if we impose a strong convexity assumption.

Assumption 4. Assume for all $S, i \in [d], \mathbf{w} \in \mathcal{W}$, the function $v \mapsto F_S(\mathbf{w} + v\mathbf{e}_i)$ is σ-strongly convex, i.e.,

$$F_S(\mathbf{w}+v\mathbf{e}_i) \geq F_S(\mathbf{w}+v'\mathbf{e}_i) + (v-v')\nabla_i F_S(\mathbf{w}+v'\mathbf{e}_i) + \sigma(v-v')^2/2, \quad \forall v, v' \in \mathbb{R}.$$

Theorem 3. *Let Assumptions 1, 3, 4 hold. Let $\{\mathbf{w}_t\}, \{\mathbf{w}_t^{(i)}\}$ be produced by (2) with $\eta_t \leq 1/\widetilde{L}$ based on S and $S^{(i)}$, respectively. Then we have $\mathbb{E}_A[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2] \leq \frac{4G_1}{n\sigma}$.*

Remark 2. Stability bounds of the order $O(1/(n\sigma))$ were established for SGD under a strong convexity assumption [Hardt *et al.*, 2016], which are extended to RCD here. Another difference is that the stability bounds in Hardt *et al.* [2016] are established for either the constant stepsize sequence $\eta_t \equiv \eta$ or the specific stepsize sequence $\eta_t = 1/(t\sigma)$. As a comparison, our results apply to general stepsizes.

We further consider *nonconvex* case. We now present stability bounds for nonconvex problems, which are ubiquitous in the modern machine learning. We denote $\prod_{k=t+1}^{t}\big(1+\widetilde{L}\eta_k d^{-\frac{1}{2}}\big)=1$ [Ying and Zhou, 2017].

Theorem 4. *Let Assumptions 1 and 3 hold. Let* $\{\mathbf{w}_t\},\{\mathbf{w}_t^{(i)}\}$ *be produced by* (2) *based on* S *and* $S^{(i)}$*, respectively. Then*

$$\mathbb{E}_A\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\big]\leq\frac{2G_1}{nd}\sum_{j=1}^{t}\eta_j\prod_{k=j+1}^{t}\big(1+\widetilde{L}\eta_k d^{-\frac{1}{2}}\big).$$

We finally consider *almost sure* bounds. The following theorem gives almost sure stability bounds, which is useful to develop high-probability generalization bounds. We need a coordinate-wise Lipschitz continuity assumption.

Assumption 5. For all S and $[i]\in[d]$, assume $|\nabla_i F_S(\mathbf{w})|\leq\widetilde{G}$ for all $\mathbf{w}\in\mathcal{W}$.

Theorem 5. *Let Assumptions 1, 2, 3, 5 hold. Then RCD with* T *iterations is* $\frac{2G_2\widetilde{G}}{n}\sum_{t=1}^{T}\eta_t$*-uniformly stable.*

5 Generalization of RCD

In this section, we use our stability bounds to develop generalization bounds for RCD. According to (1), our analysis requires to handle optimization errors by tools in convex optimization (details are given in the arXiv version).

We first consider *convex case*. We use the technique of ℓ_1 on-average argument stability to develop generalization bounds under a Lipschitz continuity assumption.

Theorem 6. *Let Assumptions 1, 2, 3 hold. Let* $\{\mathbf{w}_t\}$ *be produced by* (2) *with* $\eta_t\equiv\eta\leq 2/\widetilde{L}$*. Then*

$$\mathbb{E}_{S,A}\big[F(\mathbf{w}_T)-F(\mathbf{w}^*)\big]\leq\frac{2G_1G_2T\eta}{nd}+\frac{d\|\mathbf{w}_1-\mathbf{w}^*\|_2^2}{2T\eta}+\frac{dF(\mathbf{w}_1)}{T}.\quad(5)$$

If $T\asymp d\sqrt{n}$*, then* $\mathbb{E}_{S,A}\big[F(\mathbf{w}_T)-F(\mathbf{w}^*)\big]=O(1/\sqrt{n})$.

The first term on the right-hand side of Eq. (5) is related to estimation error, while the remaining two terms are related to optimization error. According to (5), we know that estimation error bounds increase as we run more and more iterations, while optimization errors decrease. This suggests that we should balance these two errors by stoping the algorithm at an appropriate iteration to enjoy a favorable generalization.

Remark 3. Under the same condition, it was shown that SGD with $T\asymp n$ can achieve the excess generalization bounds $O(1/\sqrt{n})$ [Hardt *et al.*, 2016]. Here we show that the same generalization bounds can be achieved by RCD.

In Theorem 6, we require the boundedness assumption of stochastic gradients (note the bounded gradient assumption does not hold for the least square loss). We now show that this boundedness assumption can be removed by using the ℓ_2-on-average argument stability. A nice property is that it incorporates the information of $F(\mathbf{w}^*)$ in the generalization bounds. This suggests that better generalization bounds can be achieved if $F(\mathbf{w}^*)$ is small, which are called optimistic bounds in the literature [Srebro *et al.*, 2010; Zhang and Zhou, 2019]. Here we introduce a parameter γ to balance different components of the generalization bounds.

Theorem 7. *Let Assumptions 2, 3 hold. Let* $\{\mathbf{w}_t\}$ *be produced by* (2) *with nonincreasing* $\eta_t\leq 2/\widetilde{L}$*. For any* $\gamma>0$ *such that* $(1+T)(1+\gamma)L^2e\sum_{t=1}^{T}\eta_t^2\leq n^2d/4$*, we have*

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T)-F_S(\mathbf{w}^*)]=O\Big(\frac{1}{\gamma}+\frac{L^2(\gamma+\gamma^{-1})T}{n^2d}\sum_{t=1}^{T}\eta_t^2\Big)\times F(\mathbf{w}^*)+O\Big(\frac{d+d\gamma^{-1}}{\sum_{t=1}^{T}\eta_t}+\frac{L^2(\gamma+\gamma^{-1})T}{n^2}\Big).\quad(6)$$

The following corollary gives a quantitative suggestion on how to stop the algorithm for a good generalization.

Corollary 8. *Let Assumptions 2, 3 hold and* $d=O(n^2)$*. Let* $\{\mathbf{w}_t\}$ *be produced by* (2) *with* $\eta_t\equiv\eta\leq 2/\widetilde{L}$*.*

(a) If $(1+T)(L+n\sqrt{d}/T)LeT\eta^2\leq n^2d/4$*, then we can choose* $T\asymp\sqrt{n}d^{\frac{3}{4}}$ *to get*

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T)-F_S(\mathbf{w}^*)]=O(d^{\frac{1}{4}}n^{-\frac{1}{2}}).$$

(b) If $F(\mathbf{w}^*)=O(d^{\frac{1}{2}}Ln^{-1})$ *and* $(1+T)L^2eT\eta^2\leq n^2d/8$*, we can choose* $T\asymp n\sqrt{d}$ *and get*

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T)-F_S(\mathbf{w}^*)]=O(d^{\frac{1}{2}}n^{-1}).$$

Remark 4. If $T\asymp\sqrt{n}d^{\frac{3}{4}}$, then $(1+T)(L+n\sqrt{d}/T)LeT\eta^2\asymp nd^{\frac{3}{2}}+n^{\frac{3}{2}}d^{\frac{5}{4}}$. Then the assumption in Part (a) holds if $d\leq cn^2$ for some appropriate $c>0$. If $T\asymp n\sqrt{d}$, then $(1+T)L^2eT\asymp n^2d$. In this case, the assumption $(1+T)L^2eT\eta^2\leq n^2d/8$ in Part (b) is also easy to satisfy.

Remark 5. As compared to Theorem 6, Part (a) admits a worse dependency on the dimensionality, which is the cost we pay for removing the Lipschitz continuity assumption. Furthermore, Part (b) shows that RCD is able to achieve a generalization bound as fast as $O(\sqrt{d}/n)$ if the best model has a small population risk, while Theorem 6 fails to exploit this low-noise assumption and can only imply at most the generalization bound $O(1/\sqrt{n})$.

Now, we present generalization bounds of RCD for *strongly convex* objective functions.

Theorem 9. *Let Assumptions 1, 2, 3, 4 hold. Let* $\{\mathbf{w}_t\}$ *be produced by* (2) *with* $\eta_t\equiv\eta\leq 1/\widetilde{L}$*. Then*

$$\mathbb{E}_{S,A}\big[F(\mathbf{w}_{T+1})-F(\mathbf{w}^*)\big]\leq\frac{4G_1G_2}{n\sigma}+\big(1-\eta\sigma/d\big)^TF(\mathbf{w}_1).$$

In particular, we can set $T\asymp d\sigma^{-1}\log 1/(n\sigma)$ *to get* $\mathbb{E}_{S,A}\big[F(\mathbf{w}_{T+1})-F(\mathbf{w}^*)\big]=O(1/(n\sigma))$.

Remark 6. Stability bounds of the order $O(1/(n\sigma))$ were established for SGD under a strongly convex setting [Hardt *et al.*, 2016], which together with optimization error bounds of the order $O(1/(T\sigma))$ [Rakhlin *et al.*, 2012], shows that SGD with n iterations can achieve excess risk bounds $O(1/(n\sigma))$. Here we show that this optimal generalization bound can also be achieved for RCD with $d\sigma^{-1}\log 1/(n\sigma)$ iterations.

Finally, we present *high-probability* bounds, which are more challenging than bounds in expectation and are important to understand the variation of algorithms in repeated runs.

Theorem 10. *Let Assumptions 1, 2, 3, 5 hold. Let* $\{\mathbf{w}_t\}$ *be produced by* (2) *with* $\eta_t \equiv \eta \le 2/\widetilde{L}$ *and* $\delta \in (0,1)$. *Assume* $\|\mathbf{w}_t\|_\infty \le R$ *and* $|f(\mathbf{w}_t; z)| \le R$ *for all* t. *If we choose* $T \asymp n^{\frac{2}{3}} d^{\frac{1}{3}} \log^{-\frac{2}{3}} n \log^{-\frac{1}{3}}(1/\delta)$, *then with probability* $1-\delta$

$$F(\bar{\mathbf{w}}_T) - F(\mathbf{w}^*) = O\Big(\big(d/n\big)^{\frac{1}{3}} \log^{\frac{1}{3}} n \log^{\frac{2}{3}}(1/\delta)\Big),$$

where $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$ *is an average of iterates.*

6 Experiments

In this section, we present some experimental results to illustrate our stability bounds. We follow the set up in Hardt *et al.* [2016], i.e., we consider two neighboring datasets and run RCD/SGD with $\eta_t \equiv 0.01$ on these neighboring datasets to produce two iterate sequences $\{\mathbf{w}_t\}, \{\mathbf{w}'_t\}$. We then plot the Euclidean distance between two iterate sequences as a function of the iteration number. We consider the least square regression for two datasets: ionosphere, svmguide3 and MNIST. We repeat the experiments 100 times and report the average of results. In Figure 1 we plot the Euclidean distance as a function of the number of iterations. Experimental results show that the Euclidean distance for RCD is much smaller than that with SGD, which is consistent with our theoretical results that RCD is more stable than SGD.

7 Proof of Theorem 2

Lemma 11. *Let* $g : \mathbb{R}^d \mapsto \mathbb{R}$ *be convex and have coordinate-wise Lipschitz continuous gradients with parameter* $\widetilde{L} > 0$. *Then for any* $\eta \le 2/\widetilde{L}$ *and any* $i \in [d]$ *we have the following inequality for any* $\mathbf{w}$ *and* $\tilde{\mathbf{w}}$

$$\|\mathbf{w} - \eta\nabla_i g(\mathbf{w})\mathbf{e}_i - \tilde{\mathbf{w}} + \eta\nabla_i g(\tilde{\mathbf{w}})\mathbf{e}_i\|_2 \le \|\mathbf{w} - \tilde{\mathbf{w}}\|_2. \quad (7)$$

Furthermore, if g *is* σ*-coordinate-wise strongly convex and* $\eta \le 1/\widetilde{L}$, *then (*w_i *denotes the* i*-th coordinate of* $\mathbf{w} \in \mathbb{R}^d$*)*

$$\|\mathbf{w} - \eta\nabla_i g(\mathbf{w})\mathbf{e}_i - \tilde{\mathbf{w}} + \eta\nabla_i g(\tilde{\mathbf{w}})\mathbf{e}_i\|_2^2 \le \|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2 - \eta\sigma|w_i - \tilde{w}_i|^2. \quad (8)$$

Proof of Theorem 2. By the update rule (2), we know

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2 \\
&= \|\mathbf{w}_t - \eta_t\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t)\mathbf{e}_{i_t} - \mathbf{w}_t^{(i)} + \eta_t\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t^{(i)})\mathbf{e}_{i_t} \\
&+ \eta_t\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t)\mathbf{e}_{i_t} - \eta_t\nabla_{i_t} F_S(\mathbf{w}_t)\mathbf{e}_{i_t}\|_2 \\
&\le \|\mathbf{w}_t - \eta_t\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t)\mathbf{e}_{i_t} - \mathbf{w}_t^{(i)} + \eta_t\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t^{(i)})\mathbf{e}_{i_t}\|_2 \\
&+ \eta_t\|\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t)\mathbf{e}_{i_t} - \nabla_{i_t} F_S(\mathbf{w}_t)\mathbf{e}_{i_t}\|_2 \qquad (9) \\
&\le \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2 + \eta_t\|\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t)\mathbf{e}_{i_t} - \nabla_{i_t} F_S(\mathbf{w}_t)\mathbf{e}_{i_t}\|_2, \qquad (10)
\end{aligned}$$

where we have used Lemma 11 in the last step. Since S and $S^{(i)}$ differ by the i-th example, we know

$$|\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t) - \nabla_{i_t} F_S(\mathbf{w}_t)| = \frac{1}{n}\big|\nabla_{i_t} f(\mathbf{w}_t; z_i) - \nabla_{i_t} f(\mathbf{w}_t; z'_i)\big| \le \frac{1}{n}\Big(\big|\nabla_{i_t} f(\mathbf{w}_t; z_i)\big| + \big|\nabla_{i_t} f(\mathbf{w}_t; z'_i)\big|\Big). \quad (11)$$

Note that i_t is uniformly drawn from $[d]$, we further know

$$\begin{aligned}
&\mathbb{E}_{i_t}\big[|\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t) - \nabla_{i_t} F_S(\mathbf{w}_t)|\big] \\
&\le \frac{1}{nd}\sum_{j=1}^d \Big(\big|\nabla_j f(\mathbf{w}_t; z_i)\big| + \big|\nabla_j f(\mathbf{w}_t; z'_i)\big|\Big) \\
&= \frac{1}{nd}\big(\|\nabla f(\mathbf{w}_t; z_i)\|_1 + \|\nabla f(\mathbf{w}_t; z'_i)\|_1\big) \le \frac{2G_1}{nd}, \qquad (12)
\end{aligned}$$

where we have used Assumption 1 in the last step. Plugging the above inequality back into (10), we get

$$\mathbb{E}_A\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2\big] \le \mathbb{E}_A\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2\big] + \frac{2G_1\eta_t}{nd}.$$

Applying the above inequality recursively gives the stated inequality. This completes the proof of Part (a).

We now prove Part (b). According to (2), we know

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2 \\
&= \|\mathbf{w}_t - \eta_t\nabla_{i_t} F_S(\mathbf{w}_t)\mathbf{e}_{i_t} - \mathbf{w}_t^{(i)} + \eta_t\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t^{(i)})\mathbf{e}_{i_t}\|_2^2 \\
&= \|\mathbf{w}_t - \eta_t\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t)\mathbf{e}_{i_t} - \mathbf{w}_t^{(i)} + \eta_t\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t^{(i)})\mathbf{e}_{i_t} \\
&+ \eta_t\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t)\mathbf{e}_{i_t} - \eta_t\nabla_{i_t} F_S(\mathbf{w}_t)\mathbf{e}_{i_t}\|_2^2.
\end{aligned}$$

By $(a+b)^2 \le (1+p)a^2 + (1+1/p)b^2$ we know

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2 \le (1+p)\times \\
&\|\mathbf{w}_t - \eta_t\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t)\mathbf{e}_{i_t} - \mathbf{w}_t^{(i)} + \eta_t\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t^{(i)})\mathbf{e}_{i_t}\|_2^2 \\
&+ (1+1/p)\eta_t^2\|\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t)\mathbf{e}_{i_t} - \nabla_{i_t} F_S(\mathbf{w}_t)\mathbf{e}_{i_t}\|_2^2.
\end{aligned}$$

It then follows from Lemma 11 that

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2 \le (1+p)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \\
&(1+1/p)\eta_t^2\|\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t)\mathbf{e}_{i_t} - \nabla_{i_t} F_S(\mathbf{w}_t)\mathbf{e}_{i_t}\|_2^2, \qquad (13)
\end{aligned}$$

Note that S and $S^{(i)}$ differ by the i-th example, we can analyze analogously to (11) and get

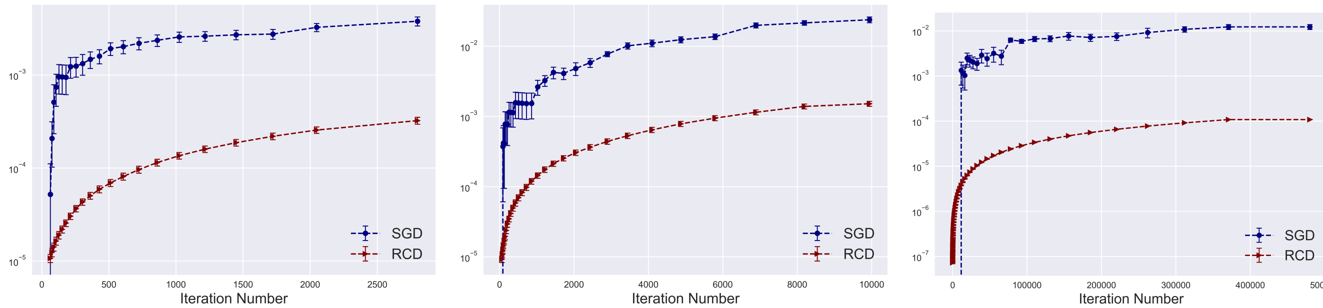
$$|\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t) - \nabla_{i_t} F_S(\mathbf{w}_t)|^2 \le \frac{2}{n^2}\Big(\big|\nabla_{i_t} f(\mathbf{w}_t; z_i)\big|^2 + \big|\nabla_{i_t} f(\mathbf{w}_t; z'_i)\big|^2\Big).$$

Since i_t is uniformly drawn from $[d]$, we further know

$$\begin{aligned}
&\mathbb{E}_{i_t}\big[|\nabla_{i_t} F_{S^{(i)}}(\mathbf{w}_t) - \nabla_{i_t} F_S(\mathbf{w}_t)|^2\big] \\
&\le \frac{2}{n^2 d}\sum_{j=1}^d \Big(\big|\nabla_j f(\mathbf{w}_t; z_i)\big|^2 + \big|\nabla_j f(\mathbf{w}_t; z'_i)\big|^2\Big) \\
&= \frac{2}{n^2 d}\big(\|\nabla f(\mathbf{w}_t; z_i)\|_2^2 + \|\nabla f(\mathbf{w}_t; z'_i)\|_2^2\big) \\
&\le \frac{4L}{n^2 d}\big(f(\mathbf{w}_t; z_i) + f(\mathbf{w}_t; z'_i)\big),
\end{aligned}$$

where we have used the self-bounding property according to the L-smoothness of f in the last step. Putting the above inequality back into (13) implies

$$\mathbb{E}_A\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \le (1+p)\mathbb{E}_A\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2\big] + \frac{4(1+1/p)L\eta_t^2}{n^2 d}\mathbb{E}_A\big[f(\mathbf{w}_t; z_i) + f(\mathbf{w}_t; z'_i)\big]. \quad (14)$$



(a) Ionosphere (b) Svmguide3 (c) MNIST

Figure 1: Euclidean distance between two iterate sequences of RCD/SGD on neighboring datasets.

It then follows that

$$\mathbb{E}_A\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big] \le$$
$$\frac{4L(1+1/p)}{n^2d}\sum_{j=1}^{t}\big(1+p\big)^{t-j}\eta_j^2\mathbb{E}_A\big[f(\mathbf{w}_j;z_i)+f(\mathbf{w}_j;z_i')\big].$$

Taking an average over i, we derive

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_A\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big] \le$$
$$\frac{4L(1+1/p)}{n^3d}\sum_{i=1}^{n}\sum_{j=1}^{t}\big(1+p\big)^{t-j}\eta_j^2\mathbb{E}_A\big[f(\mathbf{w}_j;z_i)+f(\mathbf{w}_j;z_i')\big]$$
$$=\frac{4L(1+1/p)}{n^2d}\sum_{j=1}^{t}\big(1+p\big)^{t-j}\eta_j^2\mathbb{E}_A\big[F_S(\mathbf{w}_j)+F_{S'}(\mathbf{w}_j)\big].$$

The proof is complete. □

8 Conclusions

In this paper, we initialize the generalization analysis of RCD based on the algorithmic stability. We establish upper bounds of argument stability and uniform stability for RCD, which further imply the optimal generalization bounds of the order $O(1/\sqrt{n})$ and $O(1/n)$ in the convex and strongly convex case, respectively. We also consider nonconvex case and develop high-probability bounds. Remarkably, our analysis can leverage the low-noise assumption to yield optimistic generalization bounds $O(1/n)$ in the convex case without a bounded gradient assumption.

There are several interesting future directions. First, it would be interesting to extend our analysis to other variants, such as distributed RCD and RCD for structure optimization. Second, here we assume the objectives are convex/strongly convex and each coordinate is sampled with the same probability during RCD updates. It is interesting to extend our discussion to nonconvex setting and importance sampling [Nesterov, 2012], which are popular in modern machine learning.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 61903309, 61806091, 11771012, U1811461) and the Fundamental Research Funds for the Central Universities (JBK1806002).

References

[Bassily *et al.*, 2020] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *NeurIPS*, volume 33, pages 4381–4391, 2020.

[Bousquet and Elisseeff, 2002] Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2(Mar):499–526, 2002.

[Bousquet *et al.*, 2020] Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *COLT*, pages 610–626, 2020.

[Chang *et al.*, 2008] Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. Coordinate descent method for large-scale l2-loss linear support vector machines. *JMLR*, 9(Jul):1369–1398, 2008.

[Charles and Papailiopoulos, 2018] Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *ICML*, pages 744–753, 2018.

[Chen and Gu, 2016] Jinghui Chen and Quanquan Gu. Accelerated stochastic block coordinate gradient descent for sparsity constrained nonconvex optimization. In *UAI*, 2016.

[Damaskinos *et al.*, 2020] Georgios Damaskinos, Celestine Mendler-Dünner, Rachid Guerraoui, Nikolaos Papandreou, and Thomas Parnell. Differentially private stochastic coordinate descent. *arXiv preprint arXiv:2006.07272*, 2020.

[Elisseeff *et al.*, 2005] Andre Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. *JMLR*, 6(Jan):55–79, 2005.

[Feldman and Vondrak, 2019] Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *COLT*, pages 1270–1279, 2019.

[Gu *et al.*, 2018] Bin Gu, Yingying Shan, Xiang Geng, and Guansheng Zheng. Accelerated asynchronous greedy coordinate descent algorithm for svms. In *IJCAI*, pages 2170–2176, 2018.

[Hardt *et al.*, 2016] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *ICML*, pages 1225–1234, 2016.

[Hu and Kwok, 2019] En-Liang Hu and James T Kwok. Low-rank matrix learning using biconvex surrogate minimization. *TNNLS*, 30(11):3517–3527, 2019.

[Kuzborskij and Lampert, 2018] Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *ICML*, pages 2820–2829, 2018.

[Lei and Ying, 2020] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *ICML*, pages 5809–5819, 2020.

[Lei and Ying, 2021] Yunwen Lei and Yiming Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *ICLR*, 2021.

[Lei *et al.*, 2020] Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper generalization bounds for pairwise learning. In *NeurIPS*, volume 33, 2020.

[Li and Lin, 2020] Huan Li and Zhouchen Lin. On the complexity analysis of the primal solutions for the accelerated randomized dual coordinate ascent. *JMLR*, 21(33):1–45, 2020.

[Liu *et al.*, 2017] Tongliang Liu, Gábor Lugosi, Gergely Neu, and Dacheng Tao. Algorithmic stability and hypothesis complexity. In *ICML*, pages 2159–2167, 2017.

[Lu and Xiao, 2015] Zhaosong Lu and Lin Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1-2):615–642, 2015.

[Meng *et al.*, 2017] Qi Meng, Yue Wang, Wei Chen, Taifeng Wang, Zhi-Ming Ma, and Tie-Yan Liu. Generalization error bounds for optimization algorithms via stability. In *AAAI*, volume 31, 2017.

[Mukherjee and Zhou, 2006] Sayan Mukherjee and Ding-Xuan Zhou. Learning coordinate covariances via gradients. *JMLR*, 7:519–549, 2006.

[Nesterov, 2012] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[Qu *et al.*, 2016] Zheng Qu, Peter Richtárik, Martin Takác, and Olivier Fercoq. Sdna: Stochastic dual newton ascent for empirical risk minimization. In *ICML*, pages 1823–1832, 2016.

[Rakhlin *et al.*, 2012] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, pages 449–456, 2012.

[Ren and Zhu, 2017] Yong Ren and Jun Zhu. Distributed accelerated proximal coordinate gradient methods. In *IJCAI*, pages 2655–2661, 2017.

[Richtárik and Takáč, 2014] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

[Richtárik and Takáč, 2016] Peter Richtárik and Martin Takáč. Distributed coordinate descent method for learning with big data. *JMLR*, 17(1):2657–2681, 2016.

[Shalev-Shwartz and Tewari, 2009] Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for ℓ_1 regularized loss minimization. In *ICML*, pages 929–936, 2009.

[Shalev-Shwartz *et al.*, 2010] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *JMLR*, 11(Oct):2635–2670, 2010.

[Srebro *et al.*, 2010] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *NeurIPS*, pages 2199–2207, 2010.

[Sun *et al.*, 2017] Tao Sun, Robert Hannah, and Wotao Yin. Asynchronous coordinate descent under more realistic assumptions. In *NeurIPS*, pages 6182–6190, 2017.

[Wang *et al.*, 2021] Puyu Wang, Yunwen Lei, Yiming Ying, and Hai Zhang. Differentially private SGD with non-smooth loss. *arXiv preprint arXiv:2101.08925*, 2021.

[Xiao *et al.*, 2019] Lin Xiao, Adams Wei Yu, Qihang Lin, and Weizhu Chen. Dscovr: Randomized primal-dual block coordinate algorithms for asynchronous distributed optimization. *JMLR*, 20:43–1, 2019.

[Yang *et al.*, 2021] Zhenhuan Yang, Yunwen Lei, Siwei Lyu, and Yiming Ying. Stability and differential privacy of stochastic gradient descent for pairwise learning with non-smooth loss. In *AISTATS*, pages 2026–2034. PMLR, 2021.

[Yin *et al.*, 2018] Dong Yin, Ashwin Pananjady, Max Lam, Dimitris Papailiopoulos, Kannan Ramchandran, and Peter Bartlett. Gradient diversity: a key ingredient for scalable distributed learning. In *AISTATS*, pages 1998–2007. PMLR, 2018.

[Ying and Zhou, 2017] Yiming Ying and Ding-Xuan Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224–244, 2017.

[Yuan *et al.*, 2019] Zhuoning Yuan, Yan Yan, Rong Jin, and Tianbao Yang. Stagewise training accelerates convergence of testing error over sgd. In *NeurIPS*, pages 2604–2614, 2019.

[Zeng *et al.*, 2019] Jinshan Zeng, Tim Tsz-Kit Lau, Shaobo Lin, and Yuan Yao. Global convergence of block coordinate descent in deep learning. In *International Conference on Machine Learning*, pages 7313–7323. PMLR, 2019.

[Zhang and Zhou, 2019] Lijun Zhang and Zhi-Hua Zhou. Stochastic approximation of smooth and strongly convex functions: Beyond the O(1/t) convergence rate. In *COLT*, pages 3160–3179, 2019.