Learning Theory of Stochastic Regularized Dual Averaging

Yunwen Lei, Zimeng Wang, Xiaoming Yuan

Abstract

Machine learning often involves structured optimization problems with an ℓ_1-regularizer to encourage either sparsity or interpretability. While the stochastic regularized dual averaging (RDA) method addresses the limitations of stochastic gradient descent in structured optimization, its theoretical analysis focuses primarily on optimization errors, leaving the generalization behavior to testing examples unexplored. This work bridges this gap by initializing the stability and generalization analysis of stochastic RDA through the lens of algorithmic stability. We establish novel stability and generalization error bounds for three distinct scenarios: convex and smooth losses, strongly convex and smooth losses, and nonsmooth, Lipschitz and convex losses. Additionally, we derive new optimization error bounds for stochastic RDA by removing the bounded gradient assumption for smooth problems. We also develop optimal excess risk bounds for ℓ_1-regularized problems, which can be fast under a low-noise condition.

Index Terms

Stochastic regularized dual averaging, learning theory, algorithmic stability, generalization analysis.

I. Introduction

MODERN machine learning problems often involve models and datasets of extremely large scale, posing significant computational challenges for learning algorithms. The classic stochastic gradient descent (SGD) method [43], while simple to implement, is known to have a key limitation: it struggles to effectively exploit problem structure when the loss function includes a nonsmooth regularizer, such as the ℓ_1-norm [17]. To address this limitation, the stochastic regularized dual averaging (RDA) method was developed in [53], and has proven particularly effective in online learning and reinforcement learning [18, 35]. Stochastic RDA operates by solving a minimization problem that incorporates the running average of all past gradients and a potentially nonsmooth regularizer.

A substantial body of work has been dedicated to investigating the theoretical properties of stochastic RDA and its variants [9, 10, 18, 22, 53]. In the original work [53], optimization error bounds for stochastic RDA were established for both convex and strongly convex loss functions under the assumption of bounded gradients. These results were later extended in [9] to encompass both convex and nonconvex differentiable loss functions. An adaptive variant of stochastic RDA was introduced in [18], which leverages historical data to enable more informed gradient-based learning. For the specific case of regularized least-squares regression, [22] demonstrated that stochastic RDA achieves a refined convergence rate without requiring strong convexity assumptions. Furthermore, accelerated versions of stochastic RDA, which incorporate Nesterov's momentum [38], have been proposed in [10, 53]. These accelerated variants exhibit improved convergence rates compared to the standard stochastic RDA method.

The aforementioned works primarily focus on the optimization errors of stochastic RDA, i.e., the performance of models on the training dataset. However, in machine learning, the generalization behavior of models on unseen testing samples is of greater interest. To our best knowledge, there is no generalization analysis of stochastic RDA. Motivated by this gap, we initiate the generalization analysis of stochastic RDA by leveraging the concept of algorithmic stability [5], which is a fundamental concept measuring the robustness of a learning algorithm under the perturbation of training examples. Our main contributions are summarized as follows.

1. We consider a generalized stochastic RDA method with time-varying regularizers and derive stability bounds for stochastic RDA under three distinct settings: (i) convex and smooth loss functions, (ii) strongly convex and smooth loss functions, and (iii) convex, nonsmooth but Lipschitz continuous loss functions. Building on these stability bounds, we establish, for the first time in the literature, corresponding generalization error bounds for stochastic RDA.
2. We establish optimization error bounds for stochastic RDA in both convex and strongly convex settings, removing the bounded gradient assumption required in the original work [53]. By combining these optimization error bounds with our generalization error results, we further derive excess population risk bounds for ℓ_1-regularized problems.
3. The derived excess population risk bounds of stochastic RDA are optimal for all three settings. Our bounds are of order $O(1/\sqrt{n})$ for convex loss functions and of order $O(1/(n\mu))$ for μ-strongly convex and smooth loss functions, where n

Y. Lei and X. Yuan are with the Department of Mathematics, The University of Hong Kong, Hong Kong, China (e-mail: leiyw@hku.hk, xmyuan@hku.hk). Z. Wang is with the Department of Mathematics and Mathematical Statistics, Umeå University, Sweden (e-mail: zimeng.wang@umu.se).

denotes the sample size. These results match the existing minimax lower bounds for statistical guarantees [1]. For convex and smooth losses, our results are optimistic and can be improved to the order $O(1/n)$ under a low-noise condition, i.e., there exists a model with a vanishing risk.

Technical contributions. It is important to note that existing stability analyses of SGD and its variants [23, 30] do not apply to stochastic RDA due to the presence of nonsmooth regularizers and fundamentally different gradient aggregation schemes. To address these challenges, we develop several new technical tools.

Our first technical contribution is a refined nonexpansiveness property for the proximal operator $\mathrm{prox}_{\alpha r}(\cdot)$ associated with the ℓ_1-regularizer $r(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$ for any $\alpha, \lambda > 0$, where $\mathrm{prox}_{\alpha r}(\mathbf{w}) := \arg\min_{\tilde{\mathbf{w}}} \big\{\alpha r(\tilde{\mathbf{w}}) + \frac{1}{2}\|\tilde{\mathbf{w}} - \mathbf{w}\|^2\big\}$. We prove that (see Lemma 2)

$$\|\mathrm{prox}_{\alpha r}(\alpha\mathbf{w}) - \mathrm{prox}_{\alpha r}(\alpha\mathbf{w}')\|^2 \leq \alpha^2\big(\|\mathbf{w} - \mathbf{w}'\|^2 - \|r'(\mathrm{prox}_{\alpha r}(\alpha\mathbf{w})) - r'(\mathrm{prox}_{\alpha r}(\alpha\mathbf{w}'))\|^2\big),\ \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d, \tag{I.1}$$

where r' denotes a subgradient of r. Compared with the classical nonexpansiveness property [23]

$$\|\mathrm{prox}_{\alpha r}(\alpha\mathbf{w}) - \mathrm{prox}_{\alpha r}(\alpha\mathbf{w}')\| \leq \alpha\|\mathbf{w} - \mathbf{w}'\|,\ \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d, \tag{I.2}$$

the improved bound Eq. (I.1) introduces an additional non-positive term that captures the difference between subgradients at the proximal outputs.

This refinement is crucial for obtaining sharp stability bounds for stochastic RDA. To see this, note that stochastic RDA can be written as $\mathbf{w}_{t+1} = \mathrm{prox}_{t\bar{r}_t/\beta_t}(-ts_t/\beta_t)$, where $\bar{r}_t$ is a scaled ℓ_1-regularizer, s_t is the running average of stochastic gradients, and $\beta_t > 0$ is a parameter in the algorithm. In the stability analysis, we compare the iterates generated from two neighboring datasets (i.e., datasets that differ by at most a single example). To this end, we may first estimate the change of $\mathbb{E}[\|s_t - s_t'\|^2]$ and then use some nonexpansiveness property of proximal operators to control $\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t'\|^2]$, where $\mathbb{E}[\cdot]$ denotes the expectation, $\{s_t'\}$ and $\{\mathbf{w}_t'\}$ are produced by the same algorithm on a neighboring dataset. We show in our analysis that $\mathbb{E}[\|s_{t+1} - s_{t+1}'\|^2] \lesssim \sum_{k=1}^t \frac{k^2}{\beta_k(t+1)^2}\mathbb{E}[\|\bar{r}_k'(\mathbf{w}_{k+1}) - \bar{r}_k'(\mathbf{w}_{k+1}')\|^2] + \frac{1}{nt}\hat{L}_t$ (see Eq. (V.9)), where we ignore some constant factors in $\lesssim$, and $\hat{L}_t$ is a running average of training errors during the optimization process. Then if we directly apply the classic nonexpansiveness property Eq. (I.2), we get $\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}'\|^2] \lesssim \frac{1}{\beta_t^2}\sum_{k=1}^{t-1}\frac{k^2}{\beta_k}\mathbb{E}[\|\bar{r}_k'(\mathbf{w}_{k+1}) - \bar{r}_k'(\mathbf{w}_{k+1}')\|^2] + \frac{t\hat{L}_t}{n\beta_t^2}$. A problem of this approach is that the term $\frac{1}{\beta_t^2}\sum_{k=1}^{t-1}\frac{k^2}{\beta_k}\mathbb{E}[\|\bar{r}_k'(\mathbf{w}_{k+1}) - \bar{r}_k'(\mathbf{w}_{k+1}')\|^2]$ is difficult to control and may dominate the bound unless we choose extremely small regularization parameters. To address this difficulty, we apply the improved nonexpansiveness property Eq. (I.1) on $\mathbf{w}_{t+1} = \mathrm{prox}_{t\bar{r}_t/\beta_t}(-ts_t/\beta_t)$ to obtain (see Theorem 3)

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}'\|^2] \lesssim \frac{1}{\beta_t^2}\sum_{k=1}^{t-1}\frac{k^2}{\beta_k}\mathbb{E}[\|\bar{r}_k'(\mathbf{w}_{k+1}) - \bar{r}_k'(\mathbf{w}_{k+1}')\|^2] + \frac{t\hat{L}_t}{n\beta_t^2} - \frac{t^2}{\beta_t^2}\mathbb{E}[\|\bar{r}_t'(\mathbf{w}_{t+1}) - \bar{r}_t'(\mathbf{w}_{t+1}')\|^2].$$

With the above improved bound, our second technical contribution is to show that $\frac{1}{\beta_t^2}\sum_{k=1}^t\frac{k^2}{\beta_k}\mathbb{E}[\|\bar{r}_k'(\mathbf{w}_{k+1}) - \bar{r}_k'(\mathbf{w}_{k+1}')\|^2]$ can be partially offset by the negative term $-\frac{t^2}{\beta_t^2}\mathbb{E}[\|\bar{r}_t'(\mathbf{w}_{t+1}) - \bar{r}_t'(\mathbf{w}_{t+1}')\|^2]$ if we focus on stability at averaged iterates. This cancellation mechanism is the key to deriving effective stability bounds, which in turn yield optimal excess risk bounds for averaged iterates under more flexible choices of the regularization parameters.

The rest of the paper is organized as follows: we discuss the related work in Section II and formulate the problem in Section III. Then the stability and generalization analysis of stochastic RDA is established in Section IV, with the proofs for convex and smooth losses provided in Section V. Section VI is devoted to numerical experiments, and Section VII concludes the paper.

II. Related Work

In this section, we review related work on algorithmic stability and dual averaging methods.

Algorithmic stability. Algorithmic stability is a fundamental concept in statistical learning theory to study the generalization behavior of learning algorithms. The seminal work [5] introduced the uniform stability concept which is widely used to derive high-probability generalization bounds [6, 20, 21]. This stability concept has been relaxed from different aspects, leading to various stability measures such as the hypothesis stability, on-average stability and the on-average model stability [27, 28, 30, 46]. The work [23] pioneered the generalization analysis of SGD via algorithmic stability, which motivates increasing interests in stability and generalization analysis of stochastic optimization algorithms [14, 26, 39, 40, 45, 54]. Recent progress justifies algorithmic stability as an important tool to understand the generalization behavior of neural networks [42, 49, 51] and transformers [15, 32]. Other than generalization, stability also has close connections to other performance metrics such as privacy, fairness, and replicability [2, 7].

Dual averaging. The dual averaging (DA) method was initially introduced in [37] to optimize nonsmooth functions with potentially convex constraints. It was later generalized by Xiao [53] to solve stochastic composite problems, leading to the development of the stochastic RDA method. Stochastic RDA achieves convergence rates of order $O(1/\sqrt{t})$ with shrinking step sizes $O(1/\sqrt{k})$ for convex problems and $O(\ln t/t)$ with a constant step size $1/\mu$ for μ-strongly convex problems, where t

denotes the iteration number. By incorporating Nesterov's momentum technique [38], these rates can be further improved to $O(1/t^2 + 1/\sqrt{t})$ for convex problems and $O(1/t^2 + 1/t)$ for strongly convex problems [10, 53], which are optimal according to the complexity theory [36]. For the specific case of least-squares regression, the work [22] demonstrates that stochastic RDA can achieve the convergence rate of $O(1/t)$ without requiring strong convexity assumptions. Additionally, stochastic RDA can be combined with the alternating direction method of multipliers (ADMM) for solving convex online learning problems [48]. More recently, DA has been extensively studied in distributed and online settings [11–13, 19, 25, 34] for both convex and nonconvex problems and has proven highly effective in manifold identification [16, 29].

III. Problem Setup

Let $\mathbb{P}$ be a probability measure defined on a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is an input space and $\mathcal{Y}$ is an output space. Let $S = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ be a sequence of training examples drawn independently from $\mathbb{P}$, from which we aim to learn a model $h : \mathcal{X} \mapsto \mathcal{Y}$ for prediction. We consider parametric models where each model is parameterized by a parameter $\mathbf{w}$ in a parameter space $\mathcal{W} := \mathbb{R}^d$. Let $\ell(\mathbf{w}; \mathbf{z})$ be the loss incurred by applying $\mathbf{w}$ to do the prediction on $\mathbf{z}$. We are interested in minimizing the population risk

$$L(\mathbf{w}) := \mathbb{E}_{\mathbf{z}\sim\mathbb{P}}[\ell(\mathbf{w}; \mathbf{z})],$$

where $\mathbb{E}_{\mathbf{z}}[\cdot]$ denotes the expectation w.r.t. $\mathbf{z}$. Since the probability measure $\mathbb{P}$ is unknown, we consider its empirical approximation

$$L_S(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}; \mathbf{z}_i),$$

which is referred to as the empirical risk (training error). We always mean the Euclidean norm when mentioning $\|\cdot\|$.

A. Stability and Generalization

We often apply a learning algorithm to approximately minimize the training error. We denote by $A(S)$ the output model by applying an algorithm A to the dataset S. In statistical learning theory (SLT), we are especially interested in the relative behavior of the output model $A(S)$ as compared to the best model $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} L(\mathbf{w})$, which is measured by the excess population risk $L(A(S)) - L(\mathbf{w}^*)$. A standard approach to tackle the excess population risk is to consider the following error decomposition [4]

$$\mathbb{E}_{S,A}[L(A(S)) - L(\mathbf{w}^*)] = \mathbb{E}_{S,A}[L(A(S)) - L_S(A(S))] + \mathbb{E}_{S,A}[L_S(A(S)) - L_S(\mathbf{w}^*)],$$

where we use the identity $\mathbb{E}[L_S(\mathbf{w}^*)] = L(\mathbf{w}^*)$ since $\mathbf{w}^*$ is independent of S. We refer to $L(A(S)) - L_S(A(S))$ as the generalization gap, which is the difference between testing and training on the output model. We refer to $L_S(A(S)) - L_S(\mathbf{w}^*)$ as the optimization error, which measures the suboptimality of the output model as compared to $\mathbf{w}^*$ in training. Generalization gap is a term of central interest in SLT, while optimization error has been extensively studied in optimization theory [3].

We will leverage the algorithmic stability to study the generalization gap. Intuitively, we say an algorithm is stable if a change of a single example does not lead to a significant change in the behavior of output models. Various stability concepts have been introduced, including uniform stability [5], on-average stability [46] and on-average model stability [30]. In this paper, we focus on the on-average model stability since it can imply fast rates without a Lipschitzness assumption.

Definition 1 (On-average Model Stability [30])**.** Let $S = \{z_1, \ldots, z_n\}$ and $S' = \{z'_1, \ldots, z'_n\}$ be drawn independently from $\mathbb{P}$. For any $i \in [n] := \{1, \ldots, n\}$, 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 its i-th element with z'_i. Let $\epsilon > 0$. We say a randomized algorithm A is on-average model ϵ-stable if $\mathbb{E}_{S,S',A}[\frac{1}{n}\sum_{i=1}^{n} \|A(S) - A(S^{(i)})\|^2] \le \epsilon^2$.

Next, we recall the definitions of convexity, Lipschitz continuity, and smoothness as follows.

Definition 2 (Convexity, Lipschitzness and smoothness)**.** Let $g : \mathcal{W} \mapsto \mathbb{R}$ be differentiable. Let $\sigma \ge 0, \alpha > 0$, and $G > 0$.

1) We say g is σ-strongly convex if

$$g(\mathbf{w}) \ge g(\mathbf{w}') + \langle \mathbf{w} - \mathbf{w}', \nabla g(\mathbf{w}')\rangle + \frac{\sigma}{2}\|\mathbf{w} - \mathbf{w}'\|^2, \quad \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}.$$

   We say g is convex if the above inequality holds with $\sigma = 0$.
2) We say g is G-Lipschitz continuous if $|g(\mathbf{w}) - g(\mathbf{w}')| \le G\|\mathbf{w} - \mathbf{w}'\|$, $\forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}$.
3) We say g is α-smooth if $\|\nabla g(\mathbf{w}) - \nabla g(\mathbf{w}')\| \le \alpha\|\mathbf{w} - \mathbf{w}'\|$, $\forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}$.

Remark 1. We always assume that the loss function $\ell(\cdot; \mathbf{z})$ is convex for all $\mathbf{z} \in \mathcal{Z}$ in the upcoming sections. It should be mentioned that our analysis applies directly to kernel learning problems and potentially to neural networks. In particular,

consider a feature map $\phi : \mathcal{X} \to \mathcal{H}$ and a model $h_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle$ for the kernel setting, where $\mathcal{H}$ is a reproducing kernel Hilbert space and $\mathbf{w} \in \mathcal{H}$. Let the loss be $\ell(\mathbf{w}; \mathbf{z}) = \psi(h_{\mathbf{w}}(\mathbf{x}), y)$, where $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ and $\mathbf{z} := (\mathbf{x}, y)$. If ψ is convex and smooth with respect to its first argument, then $\ell(\cdot, \mathbf{z})$ is also convex and smooth. This condition is satisfied by many commonly used losses, such as the logistic loss $\psi(a, y) = \log(1 + \exp(-ay))$ and the least square loss $\psi(a, y) = \frac{1}{2}(a - y)^2$. Therefore, our generalization results apply to kernel learning accordingly. Furthermore, although neural networks are typically nonconvex, the analysis can be extended to stochastic RDA for neural networks by exploiting their weak convexity. A key observation is that the weak convexity parameter decreases as the network width increases [33]. This weakly convex property has already been used to extend stability-based generalization analyses of GD and SGD from convex problems (e.g., [23]) to overparameterized shallow neural networks (SNNs) (e.g., [31, 42, 49]). Similarly, our stability analysis for stochastic RDA in the convex setting may also be extended to overparameterized SNNs. We leave a complete treatment of this extension as future work.

Lemma 1 gives a quantitative connection between the generalization gap and the on-average stability for smooth problems and Lipschitz continuous problems. Note that Lemma 1 includes a free parameter γ, which can be selected to optimize the bounds.

Lemma 1 ([30]). *Let S, S', and $S^{(i)}$ be constructed as in Definition 1, and $\gamma > 0$.*

(a) *If for any $\mathbf{z} \in \mathcal{Z}$, the function $\mathbf{w} \mapsto \ell(\mathbf{w}; \mathbf{z})$ is nonnegative and α-smooth, then*

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

(b) *If for any $\mathbf{z} \in \mathcal{Z}$, the function $\mathbf{w} \mapsto \ell(\mathbf{w}; \mathbf{z})$ is G-Lipschitz continuous, then*

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

B. Stochastic Regularized Dual Averaging

In this paper, we consider the stochastic regularized dual averaging (RDA) method [53], which is especially useful for solving an optimization problem of a composite structure

$$F_S(\mathbf{w}) := L_S(\mathbf{w}) + r(\mathbf{w}), \tag{III.1}$$

where $r : \mathcal{W} \mapsto \mathbb{R}_+$ is a regularizer introduced to incorporate some prior information into the training process. A representative regularizer is $r(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$, where $\lambda > 0$ is a regularization parameter and $\|\cdot\|_1$ denotes the ℓ_1-norm, i.e., $\|\mathbf{w}\|_1 = \sum_{i=1}^{d}|w_i|$ for $\mathbf{w} = (w_1, \ldots, w_d)^\top$. The ℓ_1 regularizer is effective in promoting a sparse solution. If $L_S(\mathbf{w})$ has a sparse minimizer, then we can build an objective in Eq. (III.1) with $r(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$ and apply stochastic RDA to find such a minimizer. A sparse solution is preferable due to its good interpretability as most of the components are zero [24].

We consider a generalized stochastic regularized dual averaging method, which allows time-varying regularizers $r_k : \mathcal{W} \mapsto \mathbb{R}_+, k \in [n]$. For simplicity, we denote $R_t(\mathbf{w}) := \sum_{k=1}^{t} r_k(\mathbf{w})$ and $\bar{r}_t(\mathbf{w}) := R_t(\mathbf{w})/t$. Let $\Psi : \mathcal{W} \mapsto \mathbb{R}$ be σ-strongly convex w.r.t. a norm $\|\cdot\|$.

Definition 3 (Stochastic regularized dual averaging)**.** Let $\mathbf{w}_1 \in \mathcal{W}$ be an initialization point and $\{\beta_t\}$ be a sequence of positive numbers. At the t-th iteration, we randomly select j_t according to the uniform distribution over $[n] := \{1, \ldots, n\}$ and build $\mathbf{w}_{t+1}$ by solving the following minimization problem

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}}\Big\{\Big\langle \frac{1}{t}\sum_{\tau=1}^{t}\nabla\ell(\mathbf{w}_\tau; \mathbf{z}_{j_\tau}), \mathbf{w}\Big\rangle + \bar{r}_t(\mathbf{w}) + \frac{\beta_t\Psi(\mathbf{w})}{t}\Big\}. \tag{III.2}$$

If $r_k(\cdot) = r(\cdot)$ for all $k \in \mathbb{N}$, then Eq. (III.2) becomes

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}}\Big\{\Big\langle \frac{1}{t}\sum_{\tau=1}^{t}\nabla\ell(\mathbf{w}_\tau; \mathbf{z}_{j_\tau}), \mathbf{w}\Big\rangle + r(\mathbf{w}) + \frac{\beta_t\Psi(\mathbf{w})}{t}\Big\},$$

which recovers the stochastic RDA considered in [53]. In this paper, we only consider $\Psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$ with $\|\cdot\|$ being the Euclidean norm and r_k being a scaled ℓ_1-norm, i.e., $r_k(\mathbf{w}) = \lambda_k\|\mathbf{w}\|_1$ for some $\lambda_k \geq 0$. Then, $\sigma = 1$.

IV. Main Results

In this section, we present our main results on the stability and generalization of stochastic RDA. We consider three classes of problems: (i) convex and smooth losses, (ii) strongly convex and smooth losses, and (iii) convex, nonsmooth, and Lipschitz losses. Table I provides a comprehensive summary of our main results across these different settings.

TABLE I: Summary of main results for stochastic RDA under different settings

Setting			Main Results			EPR Bound		
Convex	**Smooth**	**Lipschitz**	**Stability**	**Optimization**	**EPR**	**General**	**Low-Noise**	**Lower Bound [1]**
✓	✓	×	Thm 3	Thm 4	Thm 5, Cor. 6	$O\big(\sqrt{L(\mathbf{w}^*)/n}\big)$	$O\big(1/n\big)$	$\Omega\big(\sqrt{1/n}\big)$
μ-SC	✓	×	Thm 7	Thm 8	Thm 9, Cor. 10	$\widetilde{O}\big(1/(\mu n)\big)$	–	$\Omega\big(1/(\mu n)\big)$
✓	×	✓	Thm 11	[53]	Thm 12, Cor. 13	$O\big(\sqrt{1/n}\big)$	–	$\Omega\big(\sqrt{1/n}\big)$

Note: μ-SC = μ-strongly convex, EPR = excess population risk, n = number of training samples. $\widetilde{O}(\cdot)$ hides logarithmic terms. All results assume $r_k(\mathbf{w}) = \lambda_k\|\mathbf{w}\|_1$ for EPR bounds. The lower bound is for a general case, i.e., without low-noise assumptions.

A. Convex and Smooth Loss

We first study the behavior of stochastic RDA when the loss function $\ell(\cdot;\mathbf{z})$ is convex and α-smooth for all $\mathbf{z}\in\mathcal{Z}$. Since $\Psi(\mathbf{w}) = \|\mathbf{w}\|^2/2$, we know

$$\begin{aligned}\mathbf{w}_{t+1} &= \arg\min_{\mathbf{w}\in\mathcal{W}}\Big\{\langle s_t,\mathbf{w}\rangle + \bar{r}_t(\mathbf{w}) + \frac{\beta_t}{2t}\|\mathbf{w}\|^2\Big\}\\ &= \arg\min_{\mathbf{w}\in\mathcal{W}}\Big\{\bar{r}_t(\mathbf{w}) + \frac{\beta_t}{2t}\big\|\mathbf{w} + ts_t/\beta_t\big\|^2\Big\} = \mathrm{prox}_{t\bar{r}_t/\beta_t}\big(-\frac{t}{\beta_t}s_t\big),\end{aligned} \tag{IV.1}$$

where $s_t := \frac{1}{t}\sum_{\tau=1}^{t}\nabla\ell(\mathbf{w}_\tau; z_{j_\tau})$ is the average of the first t stochastic gradients and the proximal operator associated with a regularizer $r:\mathcal{W}\mapsto\mathbb{R}_+$ is defined as

$$\mathrm{prox}_r(\mathbf{w}) := \arg\min_{\mathbf{w}'\in\mathcal{W}}\Big\{r(\mathbf{w}') + \frac{1}{2}\|\mathbf{w}'-\mathbf{w}\|^2\Big\},\ \forall\mathbf{w}\in\mathcal{W}.$$

Therefore, stochastic RDA can be efficiently implemented via the proximal operator, which can have a closed-form solution, e.g., the soft-thresholding operator if $r(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$. For the stability analysis, we need to understand how the proximal operator preserves the stability, which is achieved in the following lemma.

Lemma 2. *Let $r(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$ and $s, s'\in\mathcal{W}$. Let*

$$\mathbf{w} = \mathrm{prox}_{tr/\beta}\big(-\frac{t}{\beta}s\big),\quad \mathbf{w}' = \mathrm{prox}_{tr/\beta}\big(-\frac{t}{\beta}s'\big).$$

Then, we have

$$\|\mathbf{w}-\mathbf{w}'\|^2 \le \frac{t^2}{\beta^2}\Big(\|s-s'\|^2 - \|r'(\mathbf{w}) - r'(\mathbf{w}')\|^2\Big),$$

where r' is a subgradient of r.

Remark 2. If r is convex, the proximal operator $\mathrm{prox}_r(\cdot)$ exhibits the nonexpansive property [23]

$$\|\mathrm{prox}_r(\mathbf{w}_1) - \mathrm{prox}_r(\mathbf{w}_2)\| \le \|\mathbf{w}_1-\mathbf{w}_2\|,\ \forall\mathbf{w}_1,\mathbf{w}_2\in\mathcal{W},$$

which implies that the distance between two points does not increase after taking a proximal operator. This shows that $\|\mathrm{prox}_{tr/\beta}(t\mathbf{w}_1/\beta) - \mathrm{prox}_{tr/\beta}(t\mathbf{w}_2/\beta)\| \le \frac{t}{\beta}\|\mathbf{w}_1-\mathbf{w}_2\|$ for general convex r. Lemma 2 improves this result for $r(\cdot) = \lambda\|\cdot\|_1$ by showing that the square distance actually decreases by at least $\|r'(\mathbf{w}) - r'(\mathbf{w}')\|^2$ after a proximal operator. As we will see in the stability analysis, this improvement plays a crucial role in developing excess risk bounds with a larger class of regularizers.

We now present stability bounds for stochastic RDA. The upper bounds involve a summation of the empirical errors at the iterates, and improve if we find models with small training errors, which is consistent with existing analysis on SGD [30]. Furthermore, the upper bounds also involve a weighted summation of $\|R_k'(\mathbf{w}_{k+1}) - R_k'(\mathbf{w}_{k+1}^{(i)})\|^2$ (recall R_t is defined above Definition 3). This shows that the regularizers affect the stability of stochastic RDA. To our knowledge, this gives the first stability analysis for stochastic RDA. For brevity, we denote $\sum_{k=1}^{t-2}a_k = 0$ if $t\le 2$.

Theorem 3 (Stability bounds)**.** *Let $S, S^{(i)}$ be defined as in Definition 1. Let $\{\mathbf{w}_t\}_{t\geq 1}$ and $\{\mathbf{w}_t^{(i)}\}_{t\geq 1}$ be the sequence produced by stochastic RDA applied to S and $S^{(i)}$, respectively. Assume each $r_k(\mathbf{w})$ is a scaled ℓ_1-norm, and for any $\mathbf{z}\in\mathcal{Z}$, the map $\mathbf{w}\mapsto\ell(\mathbf{w};\mathbf{z})$ is nonnegative, convex and α-smooth. If $\beta_k\geq\alpha$ for all k, then for any $t\geq 1$, we have*

$$\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|^2]\leq\frac{2\alpha}{\beta_{t-1}^2}\sum_{k=1}^{t-2}\frac{1}{\beta_k}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2]-\frac{1}{\beta_{t-1}^2}\mathbb{E}[\|R_{t-1}'(\mathbf{w}_t)-R_{t-1}'(\mathbf{w}_t^{(i)})\|^2]$$
$$+\frac{16\alpha}{n\beta_{t-1}^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t-1}\mathbb{E}[L_S(\mathbf{w}_k)]. \tag{IV.2}$$

Remark 3. Our basic idea is to first control $\mathbb{E}[\|s_t-s_t^{(i)}\|^2]$ and then apply Lemma 2 to relate it to $\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|^2]$, where $s_t^{(i)}=\frac{1}{t}\sum_{\tau=1}^t\nabla\ell(\mathbf{w}_\tau^{(i)};S_{j_\tau}^{(i)})$ ($S_j^{(i)}$ denotes the j-th element in $S^{(i)}$). This explains why we have the term $-\frac{1}{\beta_{t-1}^2}\mathbb{E}[\|R_{t-1}'(\mathbf{w}_t)-R_{t-1}'(\mathbf{w}_t^{(i)})\|^2]$ in the stability bound (by Lemma 2). To control $\mathbb{E}[\|s_t-s_t^{(i)}\|^2]$, we use the following recurrent relationship

$$\big\|s_{t+1}-s_{t+1}^{(i)}\big\|^2=\Big\|\frac{t}{t+1}(s_t-s_t^{(i)})+\frac{1}{t+1}\big(\nabla\ell(\mathbf{w}_{t+1};S_{j_{t+1}})-\nabla\ell(\mathbf{w}_{t+1}^{(i)};S_{j_{t+1}}^{(i)})\big)\Big\|^2.$$

If we expand the square function, we will encounter a term $\big\langle s_t-s_t^{(i)},\nabla\ell(\mathbf{w}_{t+1};S_{j_{t+1}})-\nabla\ell(\mathbf{w}_{t+1}^{(i)};S_{j_{t+1}}^{(i)})\big\rangle$, which is further related to $-\big\langle\bar{r}_t'(\mathbf{w}_{t+1})-\bar{r}_t'(\mathbf{w}_{t+1}^{(i)}),\nabla\ell(\mathbf{w}_{t+1};S_{j_{t+1}})-\nabla\ell(\mathbf{w}_{t+1}^{(i)};S_{j_{t+1}}^{(i)})\big\rangle$ by using an optimality condition. This term challenges the stability analysis, and we have to use the Cauchy-Schartz inequality

$$-\big\langle\bar{r}_t'(\mathbf{w}_{t+1})-\bar{r}_t'(\mathbf{w}_{t+1}^{(i)}),\nabla\ell(\mathbf{w}_{t+1};S_{j_{t+1}})-\nabla\ell(\mathbf{w}_{t+1}^{(i)};S_{j_{t+1}}^{(i)})\big\rangle$$
$$\leq\big\|\bar{r}_t'(\mathbf{w}_{t+1})-\bar{r}_t'(\mathbf{w}_{t+1}^{(i)})\big\|\big\|\nabla\ell(\mathbf{w}_{t+1};S_{j_{t+1}})-\nabla\ell(\mathbf{w}_{t+1}^{(i)};S_{j_{t+1}}^{(i)})\big\|.$$

This explains why we involve a weighted summation of $\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2]$ in the stability bound. Our stability analysis is substantially different from existing stability analysis of SGD or its variants [23, 30]. We introduce several techniques such as the improved nonexpansiveness of the proximal operator, the building of a recursion inequality by optimality condition related to coercivity of gradient maps, as well as the resolution of a quadratic inequality.

Remark 4. The recent work [41] established notable deterministic stability bounds for minibatch SGD that apply to general batch schedules, going beyond uniform sampling strategies. A natural question is whether similar deterministic bounds can be derived for stochastic RDA with general sampling schemes. While one can obtain a deterministic bound using the identity $\sum_{i=1}^n\mathbb{I}_{[j_t=i]}=1$ for any $t\geq 1$ as in [41], where $\mathbb{I}_{[\cdot]}\in\{0,1\}$ is the indicator function, the resulting bound is unfortunately not tight. The key difficulty is that the analysis in Nikolakakis et al. [41] crucially relies on the nonexpansiveness of the gradient descent operator $\mathbf{w}\mapsto\mathbf{w}-\eta\nabla\ell(\mathbf{w};\mathbf{z})$, which does not directly extend to stochastic RDA due to the presence of time-varying regularizers and the proximal mapping structure in Eq. (IV.1). One may instead build a recursive relation for $\|s_t-s_t^{(i)}\|$ via the classical nonexpansiveness property Eq. (I.2), which results in the following deterministic bound for convex, α-smooth, and G-Lipschitz losses with ℓ_1-regularizers $r_k(\mathbf{w})=\lambda_k\|\mathbf{w}\|_1$ for all $k\geq 1$:

$$\frac{1}{n}\sum_{i=1}^n\|\mathbf{w}_{t+2}-\mathbf{w}_{t+2}^{(i)}\|\leq\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^t\frac{2(d\alpha)^{1/2}\sum_{j=1}^k\lambda_j}{\beta_k^{1/2}\beta_{t+1}}+\frac{2G(t+1)}{n\beta_{t+1}}.$$

However, this bound is not tight, as the first term on the right-hand side can dominate unless the regularization parameters $\{\lambda_j\}$ are chosen to be sufficiently small. In contrast, working with squared differences $\|s_t-s_t^{(i)}\|^2$ allows us to exploit the improved nonexpansiveness property in Lemma 2, which potentially leads to tighter stability bounds. Unfortunately, this approach does not yield a deterministic bound for $\frac{1}{n}\sum_{i=1}^n\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|^2$, because the resulting bound involves the quadratic term $\big(\sum_{k=1}^t\mathbb{I}_{[j_{k+1}=i]}\big)^2$. Unlike the linear case, one cannot simply use the identity $\sum_{i=1}^n\mathbb{I}_{[j_{t+1}=i]}=1$ to derive an effective deterministic bound for the sum $\sum_{i=1}^n\big(\sum_{k=1}^t\mathbb{I}_{[j_{k+1}=i]}\big)^2$.

Next, we study the decay of optimization errors for stochastic RDA, which applies to any convex regularizers.

Theorem 4 (Optimization error)**.** *Assume for any k and $\mathbf{z}$, the map $\mathbf{w}\mapsto r_k(\mathbf{w})$ is convex, and $\mathbf{w}\mapsto\ell(\mathbf{w};\mathbf{z})$ is nonnegative, convex, and α-smooth. Let $\{\mathbf{w}_t\}_{t\geq 1}$ be the sequence produced by stochastic RDA applied to S. If $\{\beta_k\}$ is nondecreasing, then we have*

$$\sum_{k=1}^t\Big(1-\frac{\alpha}{\beta_{k-1}}\Big)\mathbb{E}_A[L_S(\mathbf{w}_k)-L_S(\mathbf{w})]\leq\sum_{k=1}^t\mathbb{E}_A[r_k(\mathbf{w})-r_k(\mathbf{w}_{k+1})]+\frac{\beta_t\|\mathbf{w}\|^2}{2}+\sum_{k=1}^t\frac{\alpha L_S(\mathbf{w})}{\beta_{k-1}}.$$

Remark 5. Theorem 4 implies that

$$\sum_{k=1}^{t}\mathbb{E}_A[L_S(\mathbf{w}_k)-L_S(\mathbf{w})]\le\sum_{k=1}^{t}\frac{\alpha\mathbb{E}_A[L_S(\mathbf{w}_k)]}{\beta_{k-1}}+\sum_{k=1}^{t}\mathbb{E}_A[r_k(\mathbf{w})-r_k(\mathbf{w}_{k+1})]+\frac{\beta_t\|\mathbf{w}\|^2}{2}.$$

If we set $r_k(\cdot)=r(\cdot)$ for all $k\ge 1$, it follows that

$$\frac{1}{t}\sum_{k=1}^{t}\mathbb{E}_A[L_S(\mathbf{w}_k)+r(\mathbf{w}_k)-L_S(\mathbf{w})-r(\mathbf{w})]\le\sum_{k=1}^{t}\frac{\alpha\mathbb{E}_A[L_S(\mathbf{w}_k)]}{t\beta_{k-1}}+\frac{r(\mathbf{w}_1)}{t}+\frac{\beta_t\|\mathbf{w}\|^2}{2t}. \tag{IV.3}$$

The convergence rates of $\frac{1}{t}\sum_{k=1}^{t}\mathbb{E}_A[L_S(\mathbf{w}_k)+r(\mathbf{w}_k)-L_S(\mathbf{w})-r(\mathbf{w})]$ were studied in [53]. If ℓ is G-Lipschitz continuous, the following bound was established in [53]

$$\frac{1}{t}\sum_{k=1}^{t}\mathbb{E}_A[L_S(\mathbf{w}_k)+r(\mathbf{w}_k)-L_S(\mathbf{w})-r(\mathbf{w})]\le\frac{\beta_t\|\mathbf{w}\|^2}{t}+\frac{G^2}{2t}\sum_{k=1}^{t}\frac{1}{\beta_k}. \tag{IV.4}$$

Eq. (IV.3) and Eq. (IV.4) achieve similar rates, with the main difference that Eq. (IV.3) replaces G^2 in Eq. (IV.4) by the quantity $\alpha\mathbb{E}[L_S(\mathbf{w}_k)]$. In this way, instead of imposing a Lipschitzness assumption, we assume smoothness of the loss function, which enables us to relate the gradient norm to the function value via the self-bounding property (Lemma 15). Furthermore, Eq. (IV.4) implies a convergence rate of at most $O(1/\sqrt{t})$. In contrast, our analysis involves a weighted summation of training errors in the upper bound and implies fast convergence rates under a low-noise condition. Specifically, consider $\beta_k\equiv\beta>0$ for simplicity. Then, applying $r_k(\cdot)=r(\cdot)$ in Theorem 4 gives

$$\sum_{k=1}^{t}\Big(1-\frac{\alpha}{\beta}\Big)\mathbb{E}_A[L_S(\mathbf{w}_k)]\le\sum_{k=1}^{t}L_S(\mathbf{w})+\sum_{k=1}^{t}\mathbb{E}_A[r(\mathbf{w})-r(\mathbf{w}_{k+1})]+\frac{\beta\|\mathbf{w}\|^2}{2}. \tag{IV.5}$$

If we assume $\beta\ge 2\alpha$, $r(\cdot)\ge 0$, and that there exists a $\mathbf{w}_S\in\mathcal{W}$ satisfying $F_S(\mathbf{w}_S):=L_S(\mathbf{w}_S)+r(\mathbf{w}_S)\lesssim 1/t$ and $\|\mathbf{w}_S\|\lesssim 1$, then setting $\mathbf{w}=\mathbf{w}_S$ in Eq. (IV.5) implies

$$\begin{aligned}\sum_{k=1}^{t}\mathbb{E}_A[L_S(\mathbf{w}_k)]&\le 2\sum_{k=1}^{t}L_S(\mathbf{w}_S)+2\sum_{k=1}^{t}\mathbb{E}_A[r(\mathbf{w}_S)-r(\mathbf{w}_{k+1})]+\beta\|\mathbf{w}_S\|^2\\&\le 2t\big[L_S(\mathbf{w}_S)+r(\mathbf{w}_S)\big]+\beta\|\mathbf{w}_S\|^2\lesssim 1.\end{aligned}$$

We plug the above inequality back into Eq. (IV.3) to obtain

$$\frac{1}{t}\sum_{k=1}^{t}\mathbb{E}_A[L_S(\mathbf{w}_k)+r(\mathbf{w}_k)-L_S(\mathbf{w})-r(\mathbf{w})]\le\frac{\alpha}{t\beta}\sum_{k=1}^{t}\mathbb{E}_A[L_S(\mathbf{w}_k)]+\frac{r(\mathbf{w}_1)}{t}+\frac{\beta\|\mathbf{w}\|^2}{2t}\lesssim\frac{1}{t\beta}+\frac{1+\beta\|\mathbf{w}\|^2}{t}.$$

Then we can choose $\beta\asymp 1$ and $\mathbf{w}=\mathbf{w}_S$ to show that

$$\frac{1}{t}\sum_{k=1}^{t}\mathbb{E}_A[L_S(\mathbf{w}_k)+r(\mathbf{w}_k)]\lesssim 1/t.$$

This gives a convergence rate faster than $O(1/\sqrt{t})$ in Eq. (IV.4). Generally, we are minimizing F_S by stochastic RDA, and therefore it is expected that $\mathbb{E}_A[L_S(\mathbf{w}_k)]$ becomes smaller as the algorithm proceeds. Then our convergence rate involving $\mathbb{E}_A[L_S(\mathbf{w}_k)]$ benefits from small training errors.

We can plug the stability bounds into Lemma 1 to derive generalization bounds, which, together with the optimization error bounds in Theorem 4, imply the following excess risk bounds. We consider an average of iterates, which is a standard choice for stochastic RDA [53]. The risk bound is optimistic in the sense that it involves $L(\mathbf{w}^*)$, which can be fast if $L(\mathbf{w}^*)$ is small.

Theorem 5 (Excess population risk)**.** *Assume that each r_k is a scaled ℓ_1-norm, and for any $\mathbf{z}\in\mathcal{Z}$, the map $\mathbf{w}\mapsto\ell(\mathbf{w};\mathbf{z})$ is nonnegative, convex, and α-smooth. Let $\{\mathbf{w}_t\}_{t\ge1}$ be the sequence produced by stochastic RDA applied to S with $\{\beta_k\}$ being nondecreasing and $\beta_k\ge 2\alpha$ for all k. For simplicity, let $T_k^{(i)}:=\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2]$, and define*

$$A_t=\sum_{k=1}^{t}\frac{\alpha}{\beta_{k-1}},\quad\theta_k=\frac{1-\alpha/\beta_{k-1}}{t-A_t},\quad\bar{\mathbf{w}}_t=\sum_{k=1}^{t}\theta_k\mathbf{w}_k,\quad\xi_k:=\sum_{j=k}^{t-2}\frac{\alpha\theta_{j+2}}{\beta_{j+1}^2\beta_k}-\frac{\theta_{k+1}}{2\beta_k^2}.$$

Then for any $t\ge1$ and $\gamma>0$, we have

$$\begin{aligned}\mathbb{E}[L(\bar{\mathbf{w}}_t)-L(\mathbf{w}^*)]&\lesssim\frac{R_t(\mathbf{w}^*)+\beta_t\|\mathbf{w}^*\|^2+A_tL(\mathbf{w}^*)}{t-A_t}+\frac{R_t(\mathbf{w}^*)+\beta_t\|\mathbf{w}^*\|^2+tL(\mathbf{w}^*)}{(t-A_t)\gamma}\\&+(1+\gamma)\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\xi_kT_k^{(i)}+\frac{1}{n\beta_t^2}\Big(1+\frac{t}{n}\Big)\big(R_t(\mathbf{w}^*)+\beta_t\|\mathbf{w}^*\|^2+tL(\mathbf{w}^*)\big)\Big].\end{aligned}$$

Theorem 5 considers a general setting with general choices of β_k. To simplify the results, we assume $\beta_k \equiv \beta > 0$ and derive Corollary 6.

Corollary 6. *Let the assumptions in Theorem 5 hold and $r_k(\mathbf{w}) = \lambda_k \|\mathbf{w}\|_1$ for all $k \geq 1$.*

1. *If $L(\mathbf{w}^*) \geq \frac{1}{n}$, we can take $t \asymp n$, $\beta = 2\alpha\sqrt{L(\mathbf{w}^*)n} \geq 2\alpha$, $\gamma = \sqrt{nL(\mathbf{w}^*)}$, and*
$$\lambda_k \leq \begin{cases} n^{-3/2}\sqrt{L(\mathbf{w}^*)/d}, & \text{if } k < t - \sqrt{nL(\mathbf{w}^*)}, \\ 1, & \text{if } t - \sqrt{nL(\mathbf{w}^*)} \leq k \leq t \end{cases}$$
to derive $\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] \lesssim \sqrt{L(\mathbf{w}^*)/n}$.
2. *If $L(\mathbf{w}^*) < \frac{1}{n}$, we can take $t \asymp n$, $\beta = 2\alpha$, $\gamma \asymp 1$, and*
$$\lambda_k \leq \begin{cases} n^{-2}d^{-1/2}, & \text{if } k < t - 2, \\ 1, & \text{if } t - 2 \leq k \leq t \end{cases}$$
to derive $\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] \lesssim 1/n$.

Remark 6. Corollary 6 specifies how to choose the parameters t, β, γ, and $\{\lambda_k\}$ to obtain excess population risk (EPR) bounds of order $O(\sqrt{L(\mathbf{w}^*)/n})$ when $L(\mathbf{w}^*) \geq 1/n$, and a faster rate of order $O(1/n)$ when $L(\mathbf{w}^*) < 1/n$. Thus, the bound improves as $L(\mathbf{w}^*)$ decreases in the regime $L(\mathbf{w}^*) \geq 1/n$, and reaches the limit rate of order $O(1/n)$ once $L(\mathbf{w}^*) < 1/n$. Bounds that depend explicitly on $L(\mathbf{w}^*)$ are often referred to as optimistic bounds [47]. Such bounds were first established for empirical risk minimization with smooth losses [47] and later derived for SGD in convex and smooth settings via stability analysis [30]. Our result shows that similar optimistic rates also hold for stochastic RDA. The intuition behind these rates is as follows. If $L(\mathbf{w})$ is small, then we know from the self-bounding property of α-smooth and nonnegative functions $\mathbb{E}_{\mathbf{z}}[\|\nabla \ell(\mathbf{w}; \mathbf{z})\|^2] \leq 2\alpha L(\mathbf{w})$ that the variance of stochastic gradients is also small (low noise). This small variance allows stochastic learning algorithms to achieve faster convergence rates in the low-noise case; e.g., SGD achieves an $O(1/t)$ convergence rate in interpolation settings [50]. Our stability bound in Theorem 3 depends on the cumulative quantity $\sum_{k=1}^{t-1} \mathbb{E}[L_S(\mathbf{w}_k)]$, which benefits from such faster decay of the training loss. Consequently, we can achieve faster EPR bounds for problems with a smaller $L(\mathbf{w}^*)$. The risk bound of order $O(1/\sqrt{n})$ is minimax optimal in the general case. Therefore, our stability and optimization analysis imply optimal EPR bounds.

Remark 7. The two-phase selection of λ_k in Corollary 6 arises naturally from the structure of the EPR bound in Theorem 5, which involves a term $(1+\gamma)\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\xi_k T_k^{(i)}$. The coefficient ξ_k transits from positive to negative as k increases. Specifically, when $\beta_k \equiv \beta$ is constant, we have
$$\xi_k = \frac{1}{t\beta^2}\Big(\frac{\alpha(t-k-1)}{\beta} - \frac{1}{2}\Big),$$
which decreases in k and becomes negative for sufficiently large k. In particular, in the general case $L(\mathbf{w}^*) \geq 1/n$, choosing $\beta = 2\alpha\sqrt{L(\mathbf{w}^*)n}$ implies that the transition ($\xi_k = 0$) occurs at $k_0 = t - \sqrt{nL(\mathbf{w}^*)}$. In the low-noise case $L(\mathbf{w}^*) < 1/n$, setting $\beta = 2\alpha$ yields a transition at $k_1 = t - 2$. In the first phase (when $\xi_k > 0$), the term $T_k^{(i)}$ is active in the EPR bound. Since $T_k^{(i)} \leq 4(\sum_{j=1}^{k}\lambda_j)^2 d$ for ℓ_1 regularizers (see Eq. (V.22)), we need to choose small λ_k's to control the contribution of $\sum_{k=1}^{t-2}\xi_k T_k^{(i)}$ to the overall bound. In the second phase (when $\xi_k \leq 0$), the term $T_k^{(i)}$ becomes inactive and can be ignored, allowing us to use large λ_k's without affecting the desired rate.

B. Strongly Convex and Smooth Loss

In this section, we focus on the behavior of stochastic RDA when the loss function $\ell(\cdot; \mathbf{z})$ is μ-strongly convex and α-smooth for all $\mathbf{z} \in \mathcal{Z}$. For simplicity, we assume
$$\ell(\mathbf{w}; \mathbf{z}) \geq \frac{\mu}{2}\|\mathbf{w}\|^2, \ \forall \mathbf{w} \in \mathcal{W}, \mathbf{z} \in \mathcal{Z}. \tag{IV.6}$$
Define $\tilde{\ell}(\mathbf{w}; \mathbf{z}) := \ell(\mathbf{w}; \mathbf{z}) - \frac{\mu}{2}\|\mathbf{w}\|^2$ and $\widetilde{L}_S(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(\mathbf{w}; \mathbf{z}_i)$. Then in view of Eq. (III.1), we can rewrite it as
$$F_S(\mathbf{w}) = \underbrace{L_S(\mathbf{w}) - \frac{\mu}{2}\|\mathbf{w}\|^2}_{\widetilde{L}_S(\mathbf{w})} + \underbrace{r(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{w}\|^2}_{\tilde{r}(\mathbf{w})}. \tag{IV.7}$$
Denote $\tilde{s}_t := \frac{1}{t}\sum_{\tau=1}^{t}\nabla\tilde{\ell}(\mathbf{w}_\tau; S_{j_\tau})$, $\tilde{r}_k(\mathbf{w}) := r_k(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{w}\|^2$, and $\widetilde{R}_t(\mathbf{w}) := \sum_{k=1}^{t}\tilde{r}_k(\mathbf{w})$. Inspired by the reformulation (IV.7), we employ the stochastic RDA in the following way
$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}}\Big\{\langle \tilde{s}_t, \mathbf{w}\rangle + \frac{\widetilde{R}_t(\mathbf{w})}{t} + \frac{\beta_t}{2t}\|\mathbf{w}\|^2\Big\}. \tag{IV.8}$$

Theorem 7 (Stability bounds)**.** *Let $S, S^{(i)}$ be defined as in Definition 1. Let $\{\mathbf{w}_t\}_{t\geq 1}$ and $\{\mathbf{w}_t^{(i)}\}_{t\geq 1}$ be the sequence produced by stochastic RDA applied to S and $S^{(i)}$, respectively. Assume each $r_k(\mathbf{w})$ is a scaled ℓ_1-norm, and for any $\mathbf{z}\in\mathcal{Z}$, the map $\mathbf{w}\mapsto\ell(\mathbf{w};\mathbf{z})$ is nonnegative, μ-strongly convex, α-smooth, and satisfies Eq.* (IV.6)*. If $\beta_k\geq\alpha-\mu$ for all k, then for any $t\geq 1$, we have*

$$\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|^2]\leq\frac{2(\alpha-\mu)}{\big(\beta_{t-1}+\mu(t-1)\big)^2}\sum_{k=1}^{t-2}\frac{1}{\beta_k+\mu k}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2]$$
$$-\frac{1}{\big(\beta_{t-1}+\mu(t-1)\big)^2}\mathbb{E}[\|R_{t-1}'(\mathbf{w}_t)-R_{t-1}'(\mathbf{w}_t^{(i)})\|^2]+\frac{16(\alpha-\mu)}{n\big(\beta_{t-1}+\mu(t-1)\big)^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t-1}\mathbb{E}[L_S(\mathbf{w}_k)].$$

Remark 8. With appropriate choices of regularizers $\{r_k(\cdot)\}$, we can show that $\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|^2]\lesssim 1/(\mu n)^2$ for $n\lesssim t$, which implies that stochastic RDA is on-average model stable regardless of t. To see this, let $\beta_k\equiv\beta=\max\{\mu,2(\alpha-\mu)\}$, we can show the following result based on Theorem 8

$$\sum_{k=1}^t\mathbb{E}[L_S(\mathbf{w}_k)]\lesssim R_t(\mathbf{w}^*)+\beta\|\mathbf{w}^*\|^2+\mu\|\mathbf{w}_1\|^2+tL(\mathbf{w}^*).$$

Assume $r_k(\mathbf{w}):=\lambda_k\|\mathbf{w}\|_1$ and $n\lesssim t$. If the sequence $\{\lambda_k\}$ satisfies that $\sum_{j=1}^k\lambda_j\lesssim\sqrt{\mu}k/n$ for all k, then $\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2]\lesssim(\sum_{j=1}^k\lambda_j)^2\lesssim\mu k^2/n^2$ and $R_t(\mathbf{w}^*)\lesssim\sum_{k=1}^t\lambda_k\lesssim t$. Then we know

$$\frac{1}{\big(\beta_{t-1}+\mu(t-1)\big)^2}\sum_{k=1}^{t-2}\frac{1}{\beta_k+\mu k}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2]\lesssim\frac{1}{\mu^2t^2}\sum_{k=1}^t\frac{\mu k^2}{\mu kn^2}\lesssim\frac{1}{(\mu n)^2},$$
$$\frac{1}{n\big(\beta_{t-1}+\mu(t-1)\big)^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t-2}\mathbb{E}[L_S(\mathbf{w}_k)]\lesssim\frac{1}{n\mu^2t^2}\frac{t}{n}(t+1+t)\lesssim\frac{1}{(\mu n)^2}.$$

Therefore, we derive from Theorem 7 that $\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|^2]\lesssim\frac{1}{(\mu n)^2}$. This is consistent with the stability bounds of SGD in the strongly convex case [23], and we remove the Lipschitzness assumption there.

Theorem 8 (Optimization error)**.** *Assume for any k and $\mathbf{z}$, the map $\mathbf{w}\mapsto r_k(\mathbf{w})$ is convex, and $\mathbf{w}\mapsto\ell(\mathbf{w};\mathbf{z})$ is nonnegative, μ-strongly convex, α-smooth, and satisfies Eq.* (IV.6)*. Let $\{\mathbf{w}_t\}_{t\geq 1}$ be the sequence produced by stochastic RDA applied to S. If $\{\beta_k\}$ is nondecreasing, then*

$$\sum_{k=1}^t\Big(1-\frac{\alpha-\mu}{\beta_{k-1}+\mu(k-1)}\Big)\mathbb{E}_A[L_S(\mathbf{w}_k)-L_S(\mathbf{w})]$$
$$\leq\sum_{k=1}^t\mathbb{E}_A[r_k(\mathbf{w})-r_k(\mathbf{w}_{k+1})]+\frac{\beta_t\|\mathbf{w}\|^2+\mu\|\mathbf{w}_1\|^2}{2}+\sum_{k=1}^t\frac{(\alpha-\mu)L_S(\mathbf{w})}{\beta_{k-1}+\mu(k-1)}.$$

Remark 9. The optimization error bound established here differs from that obtained in the convex and smooth setting in several aspects: (i) the coefficient α in Theorem 4 is replaced by $\alpha-\mu$ due to the use of the self-bounding property of $\tilde{\ell}$ rather than ℓ; (ii) the coefficient β_{k-1} in Theorem 4 is substituted with $\beta_{k-1}+\mu(k-1)$ since the operator $\max_{\mathbf{w}}\{\langle x,\mathbf{w}\rangle-\frac{\beta_t}{2}\|\mathbf{w}\|^2-\widetilde{R}_t(\mathbf{w})\}$ becomes $\frac{1}{\beta_t+\mu t}$-smooth due to the strong convexity of $\widetilde{R}_t(\mathbf{w})$; and (iii) an additional term $\mu\|\mathbf{w}_1\|^2/2$ is introduced on the RHS of the optimization error. This happens because using the formulation Eq. (IV.8), we have a summation $\frac{\mu}{2}\sum_{k=1}^t(\|\mathbf{w}_k\|^2-\|\mathbf{w}_{k+1}\|^2)$, which simplifies to

$$\frac{\mu}{2}(\|\mathbf{w}_1\|^2-\|\mathbf{w}_{t+1}\|^2)\leq\frac{\mu\|\mathbf{w}_1\|^2}{2}.$$

Furthermore, if we set $r_k(\cdot)=r(\cdot)$ for all k, then Theorem 8 implies that

$$\frac{1}{t}\sum_{k=1}^t\mathbb{E}_A[L_S(\mathbf{w}_k)+r(\mathbf{w}_k)-L_S(\mathbf{w})-r(\mathbf{w})]\leq\sum_{k=1}^t\frac{(\alpha-\mu)\mathbb{E}_A[L_S(\mathbf{w}_k)]}{t(\beta_{k-1}+\mu(k-1))}+\frac{r(\mathbf{w}_1)}{t}+\frac{\beta_t\|\mathbf{w}\|^2+\mu\|\mathbf{w}_1\|^2}{2t}.$$

If we treat $\mathbb{E}_A[L_S(\mathbf{w}_k)]$ as constants, then by taking $\beta_k\equiv\mu$, we recover the convergence rate of $O(\ln(t)/t)$ established in [53] without the Lipschitzness assumption.

Theorem 9 (Excess population risk)**.** *Assume that each r_k is a scaled ℓ_1-norm, and for any $\mathbf{z}\in\mathcal{Z}$, the map $\mathbf{w}\mapsto\ell(\mathbf{w};\mathbf{z})$ is nonnegative, μ-strongly convex, α-smooth, and satisfies Eq.* (IV.6)*. Let $\{\mathbf{w}_t\}_{t\geq 1}$ be the sequence produced by stochastic*

RDA applied to S with $\{\beta_k\}$ being nondecreasing and $\beta_k \geq 2(\alpha-\mu)$ for all k. For simplicity, let $T_k^{(i)} := \mathbb{E}[\|R_k'(\mathbf{w}_{k+1}) - R_k'(\mathbf{w}_{k+1}^{(i)})\|^2]$, $\tilde{\beta}_k := \beta_k + \mu k$, and define

$$\tilde{A}_t = \sum_{k=1}^{t} \frac{\alpha-\mu}{\tilde{\beta}_{k-1}}, \quad \tilde{\theta}_k = \frac{1-(\alpha-\mu)/\tilde{\beta}_{k-1}}{t-\tilde{A}_t}, \quad \bar{\mathbf{w}}_t = \sum_{k=1}^{t} \tilde{\theta}_k \mathbf{w}_k, \quad \tilde{\xi}_k := \sum_{j=k}^{t-2} \frac{(\alpha-\mu)\tilde{\theta}_{j+2}}{\tilde{\beta}_{j+1}^2 \tilde{\beta}_k} - \frac{\tilde{\theta}_{k+1}}{2\tilde{\beta}_k^2}.$$

Then for any $t \geq 1$ and $\gamma > 0$, we have

$$\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] \lesssim \frac{R_t(\mathbf{w}^*) + \beta_t\|\mathbf{w}^*\|^2 + \mu\|\mathbf{w}_1\|^2 + \tilde{A}_t L(\mathbf{w}^*)}{t-\tilde{A}_t} + \frac{R_t(\mathbf{w}^*) + \beta_t\|\mathbf{w}^*\|^2 + \mu\|\mathbf{w}_1\|^2 + tL(\mathbf{w}^*)}{(t-\tilde{A}_t)\gamma}$$
$$+ (1+\gamma)\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\tilde{\xi}_k T_k^{(i)} + \frac{1}{n\tilde{\beta}_t^2}\Big(1+\frac{t}{n}\Big)\big(R_t(\mathbf{w}^*) + \beta_t\|\mathbf{w}^*\|^2 + \mu\|\mathbf{w}_1\|^2 + tL(\mathbf{w}^*)\big)\Big].$$

Then we proceed to consider $\beta_k \equiv \beta$ and $r_k(\mathbf{w}) = \lambda_k\|\mathbf{w}\|_1$, and establish Corollary 10, where we assume μ is not too small and satisfies $n^{-1}\sqrt{\ln(n)} \lesssim \mu$.

Corollary 10. *Let the assumptions in Theorem 9 hold and $r_k(\mathbf{w}) = \lambda_k\|\mathbf{w}\|_1$ for all $k \geq 1$. If $n^{-1}\sqrt{\ln(n)} \lesssim \mu$, then we can take $t \asymp n$, $\beta = \max\{2(\alpha-\mu), \mu\}$, $\gamma \asymp \mu n$, and*

$$\lambda_k \leq \begin{cases} \dfrac{\sqrt{\mu\ln(n)}}{n\sqrt{d}}, \ \textit{if } k < \dfrac{8(\alpha-\mu)}{\mu+8(\alpha-\mu)}t, \\ \dfrac{\sqrt{\ln(n)}}{\mu^2 n}, \ \textit{if } \dfrac{8(\alpha-\mu)}{\mu+8(\alpha-\mu)}t \leq k \leq t \end{cases}$$

to derive $\mathbb{E}[L(\bar{\mathbf{w}}_t)] - L(\mathbf{w}^*) \lesssim \frac{\ln(n)}{\mu n}$.

Remark 10. Corollary 10 demonstrates the selection criteria for t, β, γ and $\{\lambda_k\}$ to achieve an excess risk bound of order $O(\ln(n)/(\mu n))$ for learning with strongly convex and smooth loss functions. Note that a practical choice of μ is $\mu \asymp 1/\sqrt{n}$. This choice implies that $\lambda_k \lesssim \sqrt{\ln(n)}$ holds for the latter half of λ_k's. As compared with Corollary 6 where the last $O(\sqrt{t})$ terms of λ_k's are relatively large, the transition here occurs at a time step linear in t, thereby providing greater flexibility in choosing $\{\lambda_k\}$. This enhanced flexibility stems from the observation that the weight $\frac{1}{\beta_k+\mu k}$ in the summation $\sum_{k=1}^{t-2}\frac{1}{\beta_k+\mu k}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1}) - R_k'(\mathbf{w}_{k+1}^{(i)})\|^2]$ decreases more rapidly due to the additional linear term μk. Moreover, if we omit the logarithmic term, the risk bound of order $O(1/\mu n)$ is minimax optimal in strongly convex settings. Consequently, our stability and optimization analysis suggests nearly optimal excess risk bounds.

C. Convex, Lipschitz and Nonsmooth Loss

In this section, we focus on the behavior of stochastic RDA when the loss function $\ell(\cdot;\mathbf{z})$ is convex but possibly nonsmooth, in which case $\nabla\ell$ represents a subgradient. To this end, we assume the subgradients are uniformly bounded, that is, there exists a constant $G > 0$, such that

$$\|\nabla\ell(\mathbf{w};\mathbf{z})\| \leq G, \ \forall \mathbf{w} \in \mathcal{W}, \mathbf{z} \in \mathcal{Z}. \tag{IV.9}$$

Theorem 11 (Stability bounds)**.** *Let $S, S^{(i)}$ be defined in Definition 1. Let $\{\mathbf{w}_t\}_{t\geq 1}$ and $\{\mathbf{w}_t^{(i)}\}_{t\geq 1}$ be the sequence produced by stochastic RDA applied to S and $S^{(i)}$, respectively. Assume each $r_k(\mathbf{w})$ is a scaled ℓ_1-norm, and for any $\mathbf{z} \in \mathcal{Z}$, the map $\mathbf{w} \mapsto \ell(\mathbf{w};\mathbf{z})$ is convex and satisfies the bounded gradient assumption in Eq.* (IV.9)*. Then for any $t \geq 1$, we have*

$$\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|^2] \leq \frac{8G}{\beta_{t-1}^2}\sum_{k=1}^{t-2}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1}) - R_k'(\mathbf{w}_{k+1}^{(i)})\|] - \frac{1}{\beta_{t-1}^2}\mathbb{E}[\|R_{t-1}'(\mathbf{w}_t) - R_{t-1}'(\mathbf{w}_t^{(i)})\|^2]$$
$$+ \frac{16G^2t^2}{n^2\beta_{t-1}^2} + \frac{8G^2t}{\beta_{t-1}^2}. \tag{IV.10}$$

Remark 11. In contrast to the smooth case, we do not have the self-bounding property and the coercivity of gradients in the nonsmooth setting. This adds difficulty to the stability analysis of stochastic RDA. To address this, we leverage the monotonicity of the (sub)gradient for convex ℓ:

$$\big\langle \mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}, \nabla\ell(\mathbf{w}_{t+1};\mathbf{z}) - \nabla\ell(\mathbf{w}_{t+1}^{(i)};\mathbf{z})\big\rangle \geq 0, \ \forall \mathbf{z} \in \mathcal{Z}.$$

This approach, however, leads to a weaker stability bound that includes an additional term $8G^2t/\beta_{t-1}^2$.

We combine the stability bounds in Theorem 11 and existing optimization error bounds [53] together to derive excess risk bounds in the nonsmooth and Lipschitz case.

Theorem 12 (Excess population risk)**.** *Assume that each r_k is a scaled ℓ_1-norm, and for any $\mathbf{z} \in \mathcal{Z}$, the map $\mathbf{w} \mapsto \ell(\mathbf{w}; \mathbf{z})$ is convex and satisfies the bounded gradient assumption in Eq.* (IV.9)*. Let $\{\mathbf{w}_t\}_{t\geq 1}$ be the sequence produced by stochastic RDA applied to S with $\{\beta_k\}$ being nondecreasing. For simplicity, denote $P_k^{(i)} := \mathbb{E}[\|R_k'(\mathbf{w}_{k+1}) - R_k'(\mathbf{w}_{k+1}^{(i)})\|]$ and $\bar{\mathbf{w}}_t = \frac{1}{t}\sum_{k=1}^t \mathbf{w}_k$. Then for any $t \geq 1$ and $\gamma > 0$, we have*

$$\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] \lesssim \frac{1}{\gamma} + \gamma\Big[\frac{1}{nt}\sum_{i=1}^n\sum_{k=1}^{t-2}\Big[\big(\sum_{j=k}^{t-2}\frac{4G}{\beta_{j+1}^2}\big)P_k^{(i)} - \frac{1}{2\beta_k^2}\big(P_k^{(i)}\big)^2\Big] + \frac{t^2}{n^2\beta_{t-1}^2} + \frac{t}{\beta_{t-1}^2}\Big]$$
$$+ \bar{r}_t(\mathbf{w}^*) + \frac{\beta_t\|\mathbf{w}^*\|^2}{t} + \frac{1}{t}\sum_{k=1}^t \frac{1}{\beta_{k-1}}.$$

Theorem 12 considers a general setting with general choices of β_k. In what follows, we assume $\beta_k \equiv \beta$ and derive Corollary 13.

Corollary 13. *Let the assumptions in Theorem 12 hold and $r_k(\mathbf{w}) = \lambda_k\|\mathbf{w}\|_1$ for all $k \geq 1$. Then we can take $t \asymp n^2$, $\beta \asymp n^{3/2}$, $\gamma = \sqrt{n}$, and*

$$\lambda_k \leq \begin{cases} n^{-2}d^{-1/2}, & \text{if } k \leq t-2, \\ n^{3/2}, & \text{if } t-2 < k \leq t, \end{cases}$$

to derive $\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] \lesssim 1/\sqrt{n}$.

Remark 12. Corollary 13 provides guidelines for selecting t, β, γ and $\{\lambda_k\}$ to attain an excess risk bound of order $O(1/\sqrt{n})$, which is minimax optimal. While the result established in Corollary 6 requires $t \asymp n$, achieving an excess risk bound of the same order necessities $t \asymp n^2$ in nonsmooth contexts. This discrepancy arises from the absence of the self-bounding property and the coercivity of gradients, causing the excess population risk established in Theorem 12 to include an additional term $\gamma t/\beta_{t-1}^2$, which deteriorates the overall bound if $t \asymp n$.

V. Proofs for Convex and Smooth Problems

A. Proofs on Stability Bounds

We first prove Lemma 2 to show the stability-promoting property for the proximal operator of ℓ_1-norm. Note that this proximal operator associated with $r(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$ can be expressed by the soft-thresholding operator.

Proof of Lemma 2. Note that both $\|\cdot\|^2$ and $\|\cdot\|_1$ are decomposable functions along the coordinates. Therefore, it suffices to consider the one-dimensional case $\mathcal{W} = \mathbb{R}^1$. According to the definition of $\mathbf{w}$ and $\mathbf{w}'$, we know that

$$\mathbf{w} = \begin{cases} -ts/\beta - t\lambda/\beta, & \text{if } -s \geq \lambda \\ 0, & \text{if } |s| \leq \lambda \\ -ts/\beta + t\lambda/\beta, & \text{if } -s \leq -\lambda. \end{cases} \qquad \mathbf{w}' = \begin{cases} -ts'/\beta - t\lambda/\beta, & \text{if } -s' \geq \lambda \\ 0, & \text{if } |s'| \leq \lambda \\ -ts'/\beta + t\lambda/\beta, & \text{if } -s' \leq -\lambda. \end{cases} \tag{V.1}$$

Note that

$$\mathbf{w} = \arg\min_{\tilde{\mathbf{w}}} \Big\{\frac{tr(\tilde{\mathbf{w}})}{\beta} + \frac{1}{2}\big\|\tilde{\mathbf{w}} + \frac{ts}{\beta}\big\|^2\Big\}.$$

By the first-order optimality condition, we know that $tr'(\mathbf{w})/\beta + \mathbf{w} + ts/\beta = 0$ and $tr'(\mathbf{w}')/\beta + \mathbf{w}' + ts'/\beta = 0$. This together with Eq. (V.1) implies

$$r'(\mathbf{w}) = \begin{cases} \lambda, & \text{if } -s \geq \lambda \\ -s, & \text{if } |s| \leq \lambda \\ -\lambda, & \text{if } -s \leq -\lambda. \end{cases} \qquad r'(\mathbf{w}') = \begin{cases} \lambda, & \text{if } -s' \geq \lambda \\ -s', & \text{if } |s'| \leq \lambda \\ -\lambda, & \text{if } -s' \leq -\lambda \end{cases} \tag{V.2}$$

and

$$\begin{aligned} \beta^2(\mathbf{w} - \mathbf{w}')^2/t^2 &= \big(s + r'(\mathbf{w}) - s' - r'(\mathbf{w}')\big)^2 \\ &= (s-s')^2 + (r'(\mathbf{w}) - r'(\mathbf{w}'))^2 + 2(s-s')(r'(\mathbf{w}) - r'(\mathbf{w}')). \end{aligned}$$

It suffices to show that

$$(s-s')(r'(\mathbf{w}) - r'(\mathbf{w}')) \leq -(r'(\mathbf{w}) - r'(\mathbf{w}'))^2. \tag{V.3}$$

We prove this inequality by considering several cases. Due to the symmetry, we can assume that $s \leq s'$.

1) If $|s| \leq \lambda$ and $|s'| \leq \lambda$, then Eq. (V.2) implies $r'(\mathbf{w}) = -s$ and $r'(\mathbf{w}') = -s'$. Therefore, we know

$$(s-s')(r'(\mathbf{w}) - r'(\mathbf{w}')) = -(s-s')^2 = -(r'(\mathbf{w}) - r'(\mathbf{w}'))^2.$$

2) If $s > \lambda$ and $s' > \lambda$, then Eq. (V.2) implies $r'(\mathbf{w}) - r'(\mathbf{w}') = 0$ and Eq. (V.3) holds.
3) If $s < -\lambda$ and $s' < -\lambda$, then Eq. (V.2) implies $r'(\mathbf{w}) - r'(\mathbf{w}') = 0$ and Eq. (V.3) holds.
4) If $-s > \lambda$ and $|s'| \le \lambda$, then Eq. (V.2) implies $r'(\mathbf{w}) = \lambda$ and $r'(\mathbf{w}') = -s'$. Therefore, we know
$$(s - s')(r'(\mathbf{w}) - r'(\mathbf{w}')) = (s - s')(\lambda + s') \le (-\lambda - s')(\lambda + s') = -(r'(\mathbf{w}) - r'(\mathbf{w}'))^2,$$
where we have used the inequality $s < -\lambda$ and $\lambda + s' \ge 0$.
5) If $-s > \lambda$ and $-s' < -\lambda$, then Eq. (V.2) implies $r'(\mathbf{w}) = \lambda$ and $r'(\mathbf{w}') = -\lambda$. Therefore, we know
$$(s - s')(r'(\mathbf{w}) - r'(\mathbf{w}')) = 2(s - s')\lambda \le -4\lambda^2 = -(r'(\mathbf{w}) - r'(\mathbf{w}'))^2.$$
6) If $|s| \le \lambda$ and $s' > \lambda$, then Eq. (V.2) implies $r'(\mathbf{w}) = -s$ and $r'(\mathbf{w}') = -\lambda$. Therefore, we know
$$(s - s')(r'(\mathbf{w}) - r'(\mathbf{w}')) = -(s' - s)(\lambda - s) \le -(\lambda - s)^2 = -(r'(\mathbf{w}) - r'(\mathbf{w}'))^2,$$
where we have used the inequality $\lambda - s > 0$.

Therefore, Eq. (V.3) holds for all the cases. The proof is completed. □

To prove Theorem 3, we introduce two lemmas. The following lemma controls the solution of a quadratic inequality. We omit the proof for brevity.

Lemma 14. *Let $a, b \ge 0$. If $x^2 \le ax + b$, then $x^2 \le a^2 + 2b$.*

The following lemma establishes the self-bounding property for smooth and nonnegative functions, meaning that the gradients can be bounded by the function values.

Lemma 15 (*Self-bounding* property [47])**.** *For $\mathbf{z} \in \mathcal{Z}$, if the function $\mathbf{w} \to \ell(\mathbf{w}; \mathbf{z})$ is nonnegative and α-smooth, then $\|\nabla \ell(\mathbf{w}; \mathbf{z})\|^2 \le 2\alpha \ell(\mathbf{w}; \mathbf{z})$.*

Proof of Theorem 3. For simplicity, we denote
$$s_t = \frac{1}{t}\sum_{\tau=1}^{t} \nabla \ell(\mathbf{w}_\tau; S_{j_\tau}), \quad s_t^{(i)} = \frac{1}{t}\sum_{\tau=1}^{t} \nabla \ell(\mathbf{w}_\tau^{(i)}; S_{j_\tau}^{(i)}),$$
where $S_j = \mathbf{z}_j$, $S_{j_\tau}^{(i)} = \mathbf{z}_{j_\tau}$ if $j_\tau \ne i$, and $S_{j_\tau}^{(i)} = \mathbf{z}_i'$ if $j_\tau = i$. Similar to Eq. (IV.1), we have
$$\mathbf{w}_{t+1}^{(i)} = \arg\min_{\mathbf{w}\in\mathcal{W}} \Big\{ \langle s_t^{(i)}, \mathbf{w}\rangle + \bar{r}_t(\mathbf{w}) + \frac{\beta_t}{2t}\|\mathbf{w}\|^2 \Big\} = \mathrm{prox}_{t\bar{r}_t/\beta_t}\big(-\frac{t}{\beta_t} s_t^{(i)}\big).$$
From Lemma 2, we know that (recall that R_t is defined above Definition 3)
$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|^2 \le \frac{t^2}{\beta_t^2}\|s_t - s_t^{(i)}\|^2 - \frac{1}{\beta_t^2}\|R_t'(\mathbf{w}_{t+1}) - R_t'(\mathbf{w}_{t+1}^{(i)})\|^2. \tag{V.4}$$
Note that Eq. (IV.2) holds for $t = 1$ as $\mathbf{w}_1 = \mathbf{w}_1^{(i)}$. According to the definition of s_t and $s_t^{(i)}$, we know for any $t \ge 1$
$$s_{t+1} = \frac{t}{t+1} s_t + \frac{1}{t+1}\nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}})$$
and therefore
$$\begin{aligned}
\big\|s_{t+1} - s_{t+1}^{(i)}\big\|^2 &= \Big\|\frac{t}{t+1}(s_t - s_t^{(i)}) + \frac{1}{t+1}\big(\nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big)\Big\|^2 \\
&= \frac{t^2}{(t+1)^2}\|s_t - s_t^{(i)}\|^2 + \frac{1}{(t+1)^2}\big\|\nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\|^2 \\
&\quad + \frac{2t}{(t+1)^2}\big\langle s_t - s_t^{(i)}, \nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\rangle.
\end{aligned} \tag{V.5}$$
From the optimality condition, it is clear that
$$s_t + \bar{r}_t'(\mathbf{w}_{t+1}) + \frac{\beta_t}{t}\mathbf{w}_{t+1} = 0, \quad s_t^{(i)} + \bar{r}_t'(\mathbf{w}_{t+1}^{(i)}) + \frac{\beta_t}{t}\mathbf{w}_{t+1}^{(i)} = 0,$$
where $\bar{r}_t'(\cdot)$ denotes a subgradient. Therefore,
$$\begin{aligned}
&\big\langle s_t - s_t^{(i)}, \nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\rangle \\
=& -\big\langle \bar{r}_t'(\mathbf{w}_{t+1}) - \bar{r}_t'(\mathbf{w}_{t+1}^{(i)}), \nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\rangle \\
&- \frac{\beta_t}{t}\big\langle \mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}, \nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\rangle \\
\le& \|\bar{r}_t'(\mathbf{w}_{t+1}) - \bar{r}_t'(\mathbf{w}_{t+1}^{(i)})\|\Delta_t^{(i)} - \frac{\beta_t}{t}\big\langle \mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}, \nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\rangle,
\end{aligned}$$

where we introduce

$$\Delta_t^{(i)} = \|\nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\|.$$

We plug this inequality back into Eq. (V.5), and derive

$$\begin{aligned}\big\|s_{t+1} - s_{t+1}^{(i)}\big\|^2 \leq & \frac{t^2}{(t+1)^2}\|s_t - s_t^{(i)}\|^2 + \frac{(\Delta_t^i)^2}{(t+1)^2} + \frac{2t\|\bar{r}_t'(\mathbf{w}_{t+1}) - \bar{r}_t'(\mathbf{w}_{t+1}^{(i)})\|\Delta_t^{(i)}}{(t+1)^2} \\ & - \frac{2\beta_t}{(t+1)^2}\big\langle \mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}, \nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\rangle. \end{aligned} \tag{V.6}$$

By Young's inequality, we know

$$\frac{2t\|\bar{r}_t'(\mathbf{w}_{t+1}) - \bar{r}_t'(\mathbf{w}_{t+1}^{(i)})\|\Delta_t^{(i)}}{(t+1)^2} \leq \frac{\alpha t^2}{\beta_t(t+1)^2}\|\bar{r}_t'(\mathbf{w}_{t+1}) - \bar{r}_t'(\mathbf{w}_{t+1}^{(i)})\|^2 + \frac{\beta_t(\Delta_t^i)^2}{\alpha(t+1)^2}.$$

It then follows that

$$\begin{aligned}\big\|s_{t+1} - s_{t+1}^{(i)}\big\|^2 \leq & \frac{t^2}{(t+1)^2}\|s_t - s_t^{(i)}\|^2 + \frac{(\Delta_t^i)^2}{(t+1)^2} + \frac{\alpha t^2}{\beta_t(t+1)^2}\|\bar{r}_t'(\mathbf{w}_{t+1}) - \bar{r}_t'(\mathbf{w}_{t+1}^{(i)})\|^2 \\ & + \frac{\beta_t(\Delta_t^i)^2}{\alpha(t+1)^2} - \frac{2\beta_t}{(t+1)^2}\big\langle \mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}, \nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\rangle. \end{aligned}$$

We first assume $j_{t+1} \neq i$ and then $S_{j_{t+1}} = S_{j_{t+1}}^{(i)}$. By the coercivity of ℓ [23], we know that

$$\big\langle \mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}, \nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\rangle \geq \frac{1}{\alpha}\big(\Delta_t^{(i)}\big)^2.$$

We combine the above two inequalities to derive that

$$\big\|s_{t+1} - s_{t+1}^{(i)}\big\|^2 \leq \frac{t^2}{(t+1)^2}\|s_t - s_t^{(i)}\|^2 + \frac{\alpha t^2}{\beta_t(t+1)^2}\|\bar{r}_t'(\mathbf{w}_{t+1}) - \bar{r}_t'(\mathbf{w}_{t+1}^{(i)})\|^2, \tag{V.7}$$

where we have used the following inequality due to $\beta_t \geq \alpha$

$$\begin{aligned}\frac{(\Delta_t^i)^2}{(t+1)^2} + \frac{\beta_t(\Delta_t^i)^2}{\alpha(t+1)^2} &\leq \Big(\frac{\alpha}{(t+1)^2} + \frac{\beta_t}{(t+1)^2}\Big)\big\langle \mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}, \nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\rangle \\ &\leq \frac{2\beta_t}{(t+1)^2}\big\langle \mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}, \nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\rangle. \end{aligned}$$

We then consider $j_{t+1} = i$. In this case, we have

$$\begin{aligned}\big\|s_{t+1} - s_{t+1}^{(i)}\big\| &= \Big\|\frac{t}{t+1}(s_t - s_t^{(i)}) + \frac{1}{t+1}\big(\nabla \ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big)\Big\| \\ &\leq \frac{t}{t+1}\|s_t - s_t^{(i)}\| + \frac{1}{t+1}\tilde{\triangle}_t^{(i)}, \end{aligned} \tag{V.8}$$

where we introduce

$$\tilde{\triangle}_t^{(i)} := \|\nabla \ell(\mathbf{w}_{t+1}; \mathbf{z}_i) - \nabla \ell(\mathbf{w}_{t+1}^{(i)}; \mathbf{z}_i')\|.$$

It then follows that

$$\big\|s_{t+1} - s_{t+1}^{(i)}\big\|^2 \leq \frac{t^2}{(t+1)^2}\|s_t - s_t^{(i)}\|^2 + \frac{(\tilde{\triangle}_t^{(i)})^2}{(t+1)^2} + \frac{2t}{(t+1)^2}\|s_t - s_t^{(i)}\|\tilde{\triangle}_t^{(i)}.$$

We combine the above two cases and take expectations over both sides to get (note the event $j_{t+1} = i$ happens with probability $1/n$)

$$\begin{aligned}\mathbb{E}[\|s_{t+1} - s_{t+1}^{(i)}\|^2] \leq & \frac{t^2}{(t+1)^2}\mathbb{E}[\|s_t - s_t^{(i)}\|^2] + \frac{\alpha}{\beta_t(t+1)^2}\mathbb{E}[\|R_t'(\mathbf{w}_{t+1}) - R_t'(\mathbf{w}_{t+1}^{(i)})\|^2] \\ & + \frac{\mathbb{E}[(\tilde{\triangle}_t^{(i)})^2]}{(t+1)^2 n} + \frac{2t}{n(t+1)^2}\mathbb{E}[\|s_t - s_t^{(i)}\|\tilde{\triangle}_t^{(i)}]. \end{aligned}$$

Since $\mathbb{E}[XY] \leq (\mathbb{E}[X^2])^{\frac{1}{2}}(\mathbb{E}[Y^2])^{\frac{1}{2}}$ for any random variables X and Y, we further get

$$\begin{aligned}\mathbb{E}[\|s_{t+1} - s_{t+1}^{(i)}\|^2] \leq & \frac{t^2}{(t+1)^2}\mathbb{E}[\|s_t - s_t^{(i)}\|^2] + \frac{\alpha}{\beta_t(t+1)^2}\mathbb{E}[\|R_t'(\mathbf{w}_{t+1}) - R_t'(\mathbf{w}_{t+1}^{(i)})\|^2] \\ & + \frac{\mathbb{E}[(\tilde{\triangle}_t^{(i)})^2]}{(t+1)^2 n} + \frac{2t}{n(t+1)^2}\big(\mathbb{E}[\|s_t - s_t^{(i)}\|^2]\big)^{\frac{1}{2}}\big(\mathbb{E}[(\tilde{\triangle}_t^{(i)})^2]\big)^{\frac{1}{2}}. \end{aligned}$$

We apply the above inequality recursively and derive

$$\mathbb{E}[\|s_{t+1}-s_{t+1}^{(i)}\|^2] \le \frac{1}{(t+1)^2}\mathbb{E}[\|s_1-s_1^{(i)}\|^2] + \sum_{k=1}^{t}\frac{\alpha}{\beta_k(k+1)^2}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2]\prod_{k'=k+1}^{t}\frac{(k')^2}{(k'+1)^2}$$
$$+\sum_{k=1}^{t}\frac{\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]}{(k+1)^2 n}\prod_{k'=k+1}^{t}\frac{(k')^2}{(k'+1)^2} + \sum_{k=1}^{t}\frac{2k}{n(k+1)^2}\big(\mathbb{E}[\|s_k-s_k^{(i)}\|^2]\big)^{\frac{1}{2}}\big(\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]\big)^{\frac{1}{2}}\prod_{k'=k+1}^{t}\frac{(k')^2}{(k'+1)^2}.$$

Since $\mathbf{w}_1 = \mathbf{w}_1^{(i)}$, we have

$$\mathbb{E}[\|s_1-s_1^{(i)}\|^2] = \mathbb{E}[\|\nabla\ell(\mathbf{w}_1;S_{j_1})-\nabla\ell(\mathbf{w}_1^{(i)};S_{j_1}^{(i)})\|^2] = \frac{1}{n}\mathbb{E}[(\tilde{\triangle}_0^{(i)})^2].$$

It then follows that

$$\mathbb{E}[\|s_{t+1}-s_{t+1}^{(i)}\|^2] \le \sum_{k=1}^{t}\frac{\alpha}{\beta_k(t+1)^2}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2]$$
$$+\sum_{k=0}^{t}\frac{\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]}{(t+1)^2 n} + \sum_{k=1}^{t}\frac{2k}{n(t+1)^2}\big(\mathbb{E}[\|s_k-s_k^{(i)}\|^2]\big)^{\frac{1}{2}}\big(\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]\big)^{\frac{1}{2}}.$$

We can rewrite the above inequality as

$$(t+1)^2\mathbb{E}[\|s_{t+1}-s_{t+1}^{(i)}\|^2] \le \sum_{k=1}^{t}\frac{\alpha}{\beta_k}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2]$$
$$+\sum_{k=0}^{t}\frac{\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]}{n} + \sum_{k=1}^{t}\frac{2k}{n}\big(\mathbb{E}[\|s_k-s_k^{(i)}\|^2]\big)^{\frac{1}{2}}\big(\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]\big)^{\frac{1}{2}}.$$

Denote $\mathfrak{C}_{t,i} := \max_{k\in[t+1]}\big(k^2\mathbb{E}[\|s_k-s_k^{(i)}\|^2]\big)^{\frac{1}{2}}$. It then follows that

$$(t+1)^2\mathbb{E}[\|s_{t+1}-s_{t+1}^{(i)}\|^2] \le \sum_{k=1}^{t}\frac{\alpha}{\beta_k}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2] + \sum_{k=0}^{t}\frac{\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]}{n} + \sum_{k=1}^{t}\frac{2\mathfrak{C}_{t,i}}{n}\big(\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]\big)^{\frac{1}{2}}.$$

Since the right-hand side is an increasing function of t, we further get

$$\mathfrak{C}_{t,i}^2 \le \sum_{k=1}^{t}\frac{\alpha}{\beta_k}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2] + \sum_{k=0}^{t}\frac{\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]}{n} + \sum_{k=1}^{t}\frac{2\mathfrak{C}_{t,i}}{n}\big(\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]\big)^{\frac{1}{2}}.$$

Solving the above quadratic inequality of $\mathfrak{C}_{t,i}$ gives (by Lemma 14)

$$\begin{aligned}\mathfrak{C}_{t,i}^2 &\le \sum_{k=1}^{t}\frac{2\alpha}{\beta_k}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2] + \sum_{k=0}^{t}\frac{2\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]}{n} + \Big(\sum_{k=1}^{t}\frac{2}{n}\big(\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]\big)^{\frac{1}{2}}\Big)^2\\ &\le \sum_{k=1}^{t}\frac{2\alpha}{\beta_k}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2] + \sum_{k=0}^{t}\frac{2\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]}{n} + \frac{4t}{n^2}\Big(\sum_{k=1}^{t}\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2]\Big)\\ &\le \sum_{k=1}^{t}\frac{2\alpha}{\beta_k}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2] + \Big(\frac{2}{n}+\frac{4t}{n^2}\Big)\sum_{k=0}^{t}\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2],\end{aligned}$$

where we use the Cauchy-Schwarz inequality to derive the second inequality. By the self-bounding property (Lemma 15), we know that

$$\begin{aligned}\mathbb{E}[(\tilde{\triangle}_k^{(i)})^2] &\le 2\mathbb{E}[\|\nabla\ell(\mathbf{w}_{k+1};z_i)\|^2 + \|\nabla\ell(\mathbf{w}_{k+1}^{(i)};z_i')\|^2]\\ &\le 4\alpha\mathbb{E}[\ell(\mathbf{w}_{k+1};z_i)+\ell(\mathbf{w}_{k+1}^{(i)};z_i')] = 8\alpha\mathbb{E}[\ell(\mathbf{w}_{k+1};z_i)] = 8\alpha\mathbb{E}[L_S(\mathbf{w}_{k+1})],\end{aligned}$$

where we have used the following identity due to the symmetry of the algorithm

$$\mathbb{E}[\ell(\mathbf{w}_{k+1};z_i)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\ell(\mathbf{w}_{k+1};z_i)] := \mathbb{E}[L_S(\mathbf{w}_{k+1})].$$

It then follows that

$$\mathbb{E}[\|s_{t+1}-s_{t+1}^{(i)}\|^2] \le \sum_{k=1}^{t} \frac{2\alpha}{\beta_k (t+1)^2} \mathbb{E}[\|R_k'(\mathbf{w}_{k+1}) - R_k'(\mathbf{w}_{k+1}^{(i)})\|^2] + \frac{16\alpha}{(t+1)^2 n}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t+1}\mathbb{E}[L_S(\mathbf{w}_k)], \tag{V.9}$$

which together with Eq. (V.4) proves Eq. (IV.2). We can slightly improve the result if $t = 2$. Indeed, by Eq. (V.4) and $\mathbf{w}_1 = \mathbf{w}_1^{(i)}$, we know

$$\begin{aligned}
\mathbb{E}[\|\mathbf{w}_2 - \mathbf{w}_2^{(i)}\|^2] &\le \frac{1}{\beta_1^2}\mathbb{E}[\|s_1 - s_1^{(i)}\|^2] - \frac{1}{\beta_1^2}\mathbb{E}[\|R_1'(\mathbf{w}_2) - R_1'(\mathbf{w}_2^{(i)})\|^2] \\
&= \frac{1}{\beta_1^2}\mathbb{E}[\|\nabla \ell(\mathbf{w}_1; S_{j_1}) - \nabla \ell(\mathbf{w}_1^{(i)}; S_{j_1}^{(i)})\|^2] - \frac{1}{\beta_1^2}\mathbb{E}[\|R_1'(\mathbf{w}_2) - R_1'(\mathbf{w}_2^{(i)})\|^2] \\
&\le \frac{8\alpha}{n\beta_1^2}\mathbb{E}[\ell(\mathbf{w}_1; z_i)] - \frac{1}{\beta_1^2}\mathbb{E}[\|R_1'(\mathbf{w}_2) - R_1'(\mathbf{w}_2^{(i)})\|^2] \\
&= \frac{8\alpha}{n\beta_1^2}\mathbb{E}[L_S(\mathbf{w}_1)] - \frac{1}{\beta_1^2}\mathbb{E}[\|R_1'(\mathbf{w}_2) - R_1'(\mathbf{w}_2^{(i)})\|^2],
\end{aligned} \tag{V.10}$$

where we use the self-bounding property to obtain the second inequality. The proof is now completed. □

B. Proofs on Optimization Error Bounds

For the convergence analysis, we follow the analysis in [53] to prove Theorem 4. We first introduce several useful concepts and tools in optimization theory. The Moreau envelope of $r : \mathcal{W} \mapsto \mathbb{R}$ is defined by

$$\mathcal{M}_r(\mathbf{w}) := \min_{\mathbf{w}' \in \mathcal{W}} \Big\{ r(\mathbf{w}') + \frac{1}{2}\|\mathbf{w}' - \mathbf{w}\|^2 \Big\},\ \forall \mathbf{w} \in \mathcal{W}.$$

It has been shown in [44] that $\mathcal{M}_r(\mathbf{w})$ is differentiable everywhere, with its gradient given by

$$\nabla \mathcal{M}_r(\mathbf{w}) = \mathbf{w} - \mathrm{prox}_r(\mathbf{w}),\ \forall \mathbf{w} \in \mathcal{W}. \tag{V.11}$$

Proof of Theorem 4. Let $g_0 = 0$ and $g_t = \sum_{k=1}^{t} \nabla \ell(\mathbf{w}_k; \mathbf{z}_{j_k})$ for $t \ge 1$. Define

$$M_t(x) := \max_{\mathbf{w} \in \mathcal{W}} \Big\{ \langle x, \mathbf{w}\rangle - \frac{\beta_t}{2}\|\mathbf{w}\|^2 - R_t(\mathbf{w}) \Big\}.$$

By the definition of $M_t(\cdot)$, we have that

$$\langle -g_t, \mathbf{w}\rangle - R_t(\mathbf{w}) = \langle -g_t, \mathbf{w}\rangle - R_t(\mathbf{w}) - \frac{\beta_t}{2}\|\mathbf{w}\|^2 + \frac{\beta_t}{2}\|\mathbf{w}\|^2 \le M_t(-g_t) + \frac{\beta_t}{2}\|\mathbf{w}\|^2. \tag{V.12}$$

Note that

$$\begin{aligned}
M_t(x) &= -\min_{\mathbf{w} \in \mathcal{W}} \Big\{ \frac{\beta_t}{2}\|\mathbf{w} - \frac{1}{\beta_t}x\|^2 + R_t(\mathbf{w}) \Big\} + \frac{1}{2\beta_t}\|x\|^2 \\
&= -\beta_t \mathcal{M}_{R_t/\beta_t}(x/\beta_t) + \frac{1}{2\beta_t}\|x\|^2,
\end{aligned}$$

where $\mathcal{M}_{R_t/\beta_t}(\cdot)$ denotes the Moreau envelope of the function $\frac{1}{\beta_t}R_t(\cdot)$. Therefore, it follows from Eq. (V.11) that

$$\nabla M_t(x) = \mathrm{prox}_{R_t/\beta_t}(x/\beta_t) - \frac{x}{\beta_t} + \frac{x}{\beta_t} = \mathrm{prox}_{R_t/\beta_t}(x/\beta_t),$$

and hence $M_t(\cdot)$ is $(1/\beta_t)$-smooth by the nonexpansive property of the proximal operator. From the definition of $\mathbf{w}_{t+1}$, we know that

$$\begin{aligned}
\mathbf{w}_{t+1} &= \arg\min_{\mathbf{w} \in \mathcal{W}} \Big\{ \langle g_t, \mathbf{w}\rangle + \frac{\beta_t}{2}\|\mathbf{w}\|^2 + R_t(\mathbf{w}) \Big\} \\
&= \arg\min_{\mathbf{w} \in \mathcal{W}} \Big\{ \frac{\beta_t}{2}\|\mathbf{w} + \frac{1}{\beta_t}g_t\|^2 + R_t(\mathbf{w}) \Big\} \\
&= \mathrm{prox}_{R_t/\beta_t}(-g_t/\beta_t) = \nabla M_t(-g_t).
\end{aligned}$$

On the other hand, it follows from the definition of $\mathbf{w}_{t+1}$ and $\beta_t \ge \beta_{t-1}$ that

$$\begin{aligned}
M_t(-g_t) &= \langle -g_t, \mathbf{w}_{t+1}\rangle - \frac{\beta_t}{2}\|\mathbf{w}_{t+1}\|^2 - R_t(\mathbf{w}_{t+1}) \\
&= \langle -g_t, \mathbf{w}_{t+1}\rangle - \frac{\beta_{t-1}}{2}\|\mathbf{w}_{t+1}\|^2 - R_{t-1}(\mathbf{w}_{t+1}) + \frac{1}{2}(\beta_{t-1} - \beta_t)\|\mathbf{w}_{t+1}\|^2 - r_t(\mathbf{w}_{t+1}) \\
&\le \max_{\mathbf{w} \in \mathcal{W}} \Big\{ \langle -g_t, \mathbf{w}\rangle - \frac{\beta_{t-1}}{2}\|\mathbf{w}\|^2 - R_{t-1}(\mathbf{w}) \Big\} + \frac{1}{2}(\beta_{t-1} - \beta_t)\|\mathbf{w}_{t+1}\|^2 - r_t(\mathbf{w}_{t+1}) \\
&\le M_{t-1}(-g_t) - r_t(\mathbf{w}_{t+1}).
\end{aligned} \tag{V.13}$$

Rearranging the terms in Eq. (V.13), we have from the $\frac{1}{\beta_{k-1}}$-smoothness of $M_{k-1}(\cdot)$ and self-bounding property of ℓ that

$$\begin{aligned} M_k(-g_k) + r_k(\mathbf{w}_{k+1}) &\leq M_{k-1}(-g_k) \\ &\leq M_{k-1}(-g_{k-1}) + \langle \nabla M_{k-1}(-g_{k-1}), g_{k-1} - g_k \rangle + \frac{1}{2\beta_{k-1}} \|g_{k-1} - g_k\|^2 \\ &= M_{k-1}(-g_{k-1}) - \langle \nabla \ell(\mathbf{w}_k; \mathbf{z}_{j_k}), \mathbf{w}_k \rangle + \frac{1}{2\beta_{k-1}} \|\nabla \ell(\mathbf{w}_k; \mathbf{z}_{j_k})\|^2 \\ &\leq M_{k-1}(-g_{k-1}) - \langle \nabla \ell(\mathbf{w}_k; \mathbf{z}_{j_k}), \mathbf{w}_k \rangle + \frac{\alpha}{\beta_{k-1}} \ell(\mathbf{w}_k; \mathbf{z}_{j_k}). \end{aligned} \tag{V.14}$$

Rearranging the terms in Eq. (V.14) gives

$$\langle \nabla \ell(\mathbf{w}_k; \mathbf{z}_{j_k}), \mathbf{w}_k \rangle + r_k(\mathbf{w}_{k+1}) \leq M_{k-1}(-g_{k-1}) - M_k(-g_k) + \frac{\alpha}{\beta_{k-1}} \ell(\mathbf{w}_k; \mathbf{z}_{j_k}). \tag{V.15}$$

Summing up Eq. (V.15) for $k = 1, \ldots, t$, and noting $M_0(-g_0) = M_0(0) = 0$, we get

$$\begin{aligned} \sum_{k=1}^{t} (\langle \nabla \ell(\mathbf{w}_k; \mathbf{z}_{j_k}), \mathbf{w}_k \rangle + r_k(\mathbf{w}_{k+1})) &\leq M_0(-g_0) - M_t(-g_t) + \sum_{k=1}^{t} \frac{\alpha}{\beta_{k-1}} \ell(\mathbf{w}_k; \mathbf{z}_{j_k}) \\ &= -M_t(-g_t) + \sum_{k=1}^{t} \frac{\alpha}{\beta_{k-1}} \ell(\mathbf{w}_k; \mathbf{z}_{j_k}) \\ &\leq \langle g_t, \mathbf{w} \rangle + R_t(\mathbf{w}) + \frac{\beta_t \|\mathbf{w}\|^2}{2} + \sum_{k=1}^{t} \frac{\alpha}{\beta_{k-1}} \ell(\mathbf{w}_k; \mathbf{z}_{j_k}), \end{aligned} \tag{V.16}$$

where we have used Eq. (V.12) in the last step. By the convexity of ℓ, we further get

$$\begin{aligned} \sum_{k=1}^{t} \big(\ell(\mathbf{w}_k; \mathbf{z}_{j_k}) - \ell(\mathbf{w}; \mathbf{z}_{j_k})\big) &\leq \sum_{k=1}^{t} \langle \nabla \ell(\mathbf{w}_k; \mathbf{z}_{j_k}), \mathbf{w}_k - \mathbf{w} \rangle = \sum_{k=1}^{t} \langle \nabla \ell(\mathbf{w}_k; \mathbf{z}_{j_k}), \mathbf{w}_k \rangle - \langle g_t, \mathbf{w} \rangle \\ &\leq R_t(\mathbf{w}) + \frac{\beta_t \|\mathbf{w}\|^2}{2} + \sum_{k=1}^{t} \frac{\alpha}{\beta_{k-1}} \ell(\mathbf{w}_k; \mathbf{z}_{j_k}) - \sum_{k=1}^{t} r_k(\mathbf{w}_{k+1}). \end{aligned}$$

We take expectations on both sides and get

$$\sum_{k=1}^{t} \mathbb{E}_A[L_S(\mathbf{w}_k) - L_S(\mathbf{w})] \leq R_t(\mathbf{w}) + \frac{\beta_t \|\mathbf{w}\|^2}{2} + \sum_{k=1}^{t} \frac{\alpha \mathbb{E}_A[L_S(\mathbf{w}_k)]}{\beta_{k-1}} - \sum_{k=1}^{t} \mathbb{E}_A[r_k(\mathbf{w}_{k+1})].$$

That is, we have

$$\sum_{k=1}^{t} \Big(1 - \frac{\alpha}{\beta_{k-1}}\Big) \mathbb{E}_A[L_S(\mathbf{w}_k) - L_S(\mathbf{w})] \leq R_t(\mathbf{w}) + \frac{\beta_t \|\mathbf{w}\|^2}{2} + \sum_{k=1}^{t} \frac{\alpha L_S(\mathbf{w})}{\beta_{k-1}} - \sum_{k=1}^{t} \mathbb{E}_A[r_k(\mathbf{w}_{k+1})].$$

The proof is completed. □

C. Proofs on Excess Risk Bounds

We combine the stability and convergence analysis to prove the excess risk bounds of stochastic RDA.

Proof of Theorem 5. We first bound $\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_t^{(i)}\|^2]$. It follows from Theorem 3 that for $t \geq 3$

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|^2] \leq \frac{2\alpha}{n\beta_{t-1}^2} \sum_{i=1}^{n} \sum_{k=1}^{t-2} \frac{T_k^{(i)}}{\beta_k} - \frac{1}{n\beta_{t-1}^2} \sum_{i=1}^{n} T_{t-1}^{(i)} + \frac{16\alpha}{n\beta_{t-1}^2} \Big(1 + \frac{2t}{n}\Big) \sum_{k=1}^{t} \mathbb{E}[L_S(\mathbf{w}_k)].$$

Moreover, we have $\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_1^{(i)}\|^2] = 0$ and

$$\mathbb{E}[\|\mathbf{w}_2 - \mathbf{w}_2^{(i)}\|^2] \leq \frac{8\alpha}{n\beta_1^2} \mathbb{E}[L_S(\mathbf{w}_1)] - \frac{T_1^{(i)}}{\beta_1^2}$$

from Eq. (V.10). Therefore, by the convexity of $\|\cdot\|^2$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\bar{\mathbf{w}}_t-\bar{\mathbf{w}}_t^{(i)}\|^2]\le\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{t}\theta_j\mathbb{E}[\|\mathbf{w}_j-\mathbf{w}_j^{(i)}\|^2]\\
&\le\frac{1}{n}\sum_{i=1}^{n}\sum_{j=3}^{t}\Big[\frac{2\alpha\theta_j}{\beta_{j-1}^2}\sum_{k=1}^{j-2}\frac{T_k^{(i)}}{\beta_k}-\frac{\theta_j}{\beta_{j-1}^2}T_{j-1}^{(i)}\Big]-\frac{\theta_2}{n\beta_1^2}\sum_{i=1}^{n}T_1^{(i)}+\frac{16\alpha}{n\beta_{t-1}^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t}\mathbb{E}[L_S(\mathbf{w}_k)]\\
&=\frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{j=3}^{t}\sum_{k=1}^{j-2}\frac{2\alpha\theta_jT_k^{(i)}}{\beta_{j-1}^2\beta_k}-\sum_{k=1}^{t-1}\frac{\theta_{k+1}}{\beta_k^2}T_k^{(i)}\Big]+\frac{16\alpha}{n\beta_{t-1}^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t}\mathbb{E}[L_S(\mathbf{w}_k)]\\
&=\frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{j=1}^{t-2}\sum_{k=1}^{j}\frac{2\alpha\theta_{j+2}T_k^{(i)}}{\beta_{j+1}^2\beta_k}-\sum_{k=1}^{t-1}\frac{\theta_{k+1}}{\beta_k^2}T_k^{(i)}\Big]+\frac{16\alpha}{n\beta_{t-1}^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t}\mathbb{E}[L_S(\mathbf{w}_k)]\\
&=\frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{k=1}^{t-2}\Big(\sum_{j=k}^{t-2}\frac{2\alpha\theta_{j+2}}{\beta_{j+1}^2\beta_k}\Big)T_k^{(i)}-\sum_{k=1}^{t-1}\frac{\theta_{k+1}}{\beta_k^2}T_k^{(i)}\Big]+\frac{16\alpha}{n\beta_{t-1}^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t}\mathbb{E}[L_S(\mathbf{w}_k)]\\
&\le\frac{2}{n}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\xi_kT_k^{(i)}+\frac{16\alpha}{n\beta_{t-1}^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t}\mathbb{E}[L_S(\mathbf{w}_k)].
\end{aligned}$$

Then combining Lemma 1 Part (a) and the above inequality together, we get (note $\beta_k\ge 2\alpha$)

$$\begin{aligned}
&\mathbb{E}[L(\bar{\mathbf{w}}_t)-L_S(\bar{\mathbf{w}}_t)]\\
&\le\frac{\alpha}{\gamma}\mathbb{E}[L_S(\bar{\mathbf{w}}_t)]+\frac{\alpha+\gamma}{2n}\sum_{i=1}^{n}\mathbb{E}[\|\bar{\mathbf{w}}_t-\bar{\mathbf{w}}_t^{(i)}\|^2]\\
&\le\frac{\alpha}{\gamma}\mathbb{E}[L_S(\bar{\mathbf{w}}_t)]+(\alpha+\gamma)\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\xi_kT_k^{(i)}+\frac{8\alpha}{n\beta_{t-1}^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t}\mathbb{E}[L_S(\mathbf{w}_k)]\Big]\\
&\le\frac{\alpha}{\gamma}\mathbb{E}[L_S(\bar{\mathbf{w}}_t)]+(\alpha+\gamma)\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\xi_kT_k^{(i)}+\frac{16\alpha}{n\beta_{t-1}^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t}\Big(1-\frac{\alpha}{\beta_{k-1}}\Big)\mathbb{E}[L_S(\mathbf{w}_k)]\Big].
\end{aligned}\tag{V.17}$$

By Theorem 4 and $\mathbb{E}[L_S(\mathbf{w})]=L(\mathbf{w})$, we know

$$\sum_{k=1}^{t}\Big(1-\frac{\alpha}{\beta_{k-1}}\Big)\mathbb{E}[L_S(\mathbf{w}_k)-L_S(\mathbf{w})]\le R_t(\mathbf{w})+\frac{\beta_t\|\mathbf{w}\|^2}{2}+\sum_{k=1}^{t}\frac{\alpha L(\mathbf{w})}{\beta_{k-1}}.\tag{V.18}$$

Setting $\mathbf{w}=\mathbf{w}^*$ leads to

$$\sum_{k=1}^{t}\Big(1-\frac{\alpha}{\beta_{k-1}}\Big)\mathbb{E}[L_S(\mathbf{w}_k)]\lesssim R_t(\mathbf{w}^*)+\beta_t\|\mathbf{w}^*\|^2+tL(\mathbf{w}^*).\tag{V.19}$$

By the convexity of L_S, we have from Eq. (V.19) that

$$\mathbb{E}[L_S(\bar{\mathbf{w}}_t)]\le\sum_{k=1}^{t}\theta_k\mathbb{E}[L_S(\mathbf{w}_k)]\lesssim\frac{R_t(\mathbf{w}^*)+\beta_t\|\mathbf{w}^*\|^2+tL(\mathbf{w}^*)}{t-A_t}.$$

Plugging Eq. (V.19) and the above inequality into Eq. (V.17) implies

$$\begin{aligned}
\mathbb{E}[L(\bar{\mathbf{w}}_t)-L_S(\bar{\mathbf{w}}_t)]&\lesssim\frac{R_t(\mathbf{w}^*)+\beta_t\|\mathbf{w}^*\|^2+tL(\mathbf{w}^*)}{(t-A_t)\gamma}\\
&+(1+\gamma)\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\xi_kT_k^{(i)}+\frac{1}{n\beta_t^2}\Big(1+\frac{t}{n}\Big)\big(R_t(\mathbf{w}^*)+\beta_t\|\mathbf{w}^*\|^2+tL(\mathbf{w}^*)\big)\Big].
\end{aligned}\tag{V.20}$$

On the other hand, we have from Eq. (V.18) that

$$\mathbb{E}[L_S(\bar{\mathbf{w}}_t)]-L(\mathbf{w}^*)\lesssim\frac{R_t(\mathbf{w}^*)+\beta_t\|\mathbf{w}^*\|^2+A_tL(\mathbf{w}^*)}{t-A_t}.\tag{V.21}$$

Combining Eq. (V.20) and Eq. (V.21), we have

$$\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] \lesssim \frac{R_t(\mathbf{w}^*) + \beta_t\|\mathbf{w}^*\|^2 + A_t L(\mathbf{w}^*)}{t - A_t} + \frac{R_t(\mathbf{w}^*) + \beta_t\|\mathbf{w}^*\|^2 + tL(\mathbf{w}^*)}{(t - A_t)\gamma}$$
$$+ (1+\gamma)\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\xi_k T_k^{(i)} + \frac{1}{n\beta_t^2}\Big(1+\frac{t}{n}\Big)\big(R_t(\mathbf{w}^*) + \beta_t\|\mathbf{w}^*\|^2 + tL(\mathbf{w}^*)\big)\Big].$$

The proof is completed. □

For $k \geq 1$, let $r_k(\mathbf{w}) = \lambda_k\|\mathbf{w}\|_1$ with $\lambda_k > 0$, then $R_t(\mathbf{w}) = \big(\sum_{k=1}^t \lambda_k\big)\|\mathbf{w}\|_1$. From the fact $\|R_t'(\mathbf{w})\|_\infty \leq \sum_{k=1}^t \lambda_k$, we know

$$\|R_t'(\mathbf{w})\|^2 \leq \big(\sum_{k=1}^{t}\lambda_k\big)^2 d,\ \forall \mathbf{w} \in \mathcal{W}$$

which further implies that

$$T_k^{(i)} \leq 4\big(\sum_{j=1}^{k}\lambda_j\big)^2 d. \tag{V.22}$$

Proof of Corollary 6. Since $\beta_k \equiv \beta$, we know $\theta_k \equiv \frac{1}{t}$, $A_t = \frac{\alpha t}{\beta}$ and

$$\xi_k = \frac{1}{t\beta^2}\big(\frac{\alpha(t-k-1)}{\beta} - \frac{1}{2}\big). \tag{V.23}$$

It follows from Theorem 5 that

$$\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] \lesssim \frac{\sum_{k=1}^t \lambda_k + \beta + \alpha t L(\mathbf{w}^*)/\beta}{t - \alpha t/\beta} + \frac{\sum_{k=1}^t \lambda_k + \beta + tL(\mathbf{w}^*)}{(t-\alpha t/\beta)\gamma}$$
$$+ (1+\gamma)\Big[\frac{d}{t\beta^2}\sum_{k=1}^{t-2}\max\Big\{\frac{\alpha(t-k-1)}{\beta} - \frac{1}{2}, 0\Big\}\big(\sum_{j=1}^{k}\lambda_j\big)^2 + \frac{1}{n\beta^2}\Big(1+\frac{t}{n}\Big)\big(\sum_{k=1}^{t}\lambda_k + \beta + tL(\mathbf{w}^*)\big)\Big], \tag{V.24}$$

where we have used Eq. (V.22) and Eq. (V.23). We first consider the case $L(\mathbf{w}^*) \geq \frac{1}{n}$. From the condition on $\{\lambda_k\}$ and t, we know that

$$\sum_{k=1}^{t}\lambda_k \leq (t - \sqrt{nL(\mathbf{w}^*)})n^{-\frac{3}{2}}\sqrt{L(\mathbf{w}^*)/d} + \sqrt{nL(\mathbf{w}^*)} \lesssim \sqrt{nL(\mathbf{w}^*)}.$$

Therefore, we have

$$\frac{\sum_{k=1}^t \lambda_k + \beta + \alpha t L(\mathbf{w}^*)/\beta}{t - \alpha t/\beta} \lesssim \frac{\sqrt{nL(\mathbf{w}^*)}}{n - \frac{\sqrt{n}}{2\sqrt{L(\mathbf{w}^*)}}} \leq \frac{\sqrt{nL(\mathbf{w}^*)}}{n - n/2} \asymp \sqrt{\frac{L(\mathbf{w}^*)}{n}}, \tag{V.25}$$

$$\frac{\sum_{k=1}^t \lambda_k + \beta + tL(\mathbf{w}^*)}{(t-\alpha t/\beta)\gamma} \lesssim \frac{\sqrt{nL(\mathbf{w}^*)} + nL(\mathbf{w}^*)}{n\sqrt{nL(\mathbf{w}^*)}} = \frac{1}{n} + \sqrt{\frac{L(\mathbf{w}^*)}{n}} \asymp \sqrt{\frac{L(\mathbf{w}^*)}{n}}, \tag{V.26}$$

$$\frac{1+\gamma}{n\beta^2}\Big(1+\frac{t}{n}\Big)\big(\sum_{k=1}^{t}\lambda_k + \beta + tL(\mathbf{w}^*)\big) \lesssim \frac{\sqrt{nL(\mathbf{w}^*)}}{n^2 L(\mathbf{w}^*)}\big(\sqrt{nL(\mathbf{w}^*)} + nL(\mathbf{w}^*)\big) \asymp \sqrt{\frac{L(\mathbf{w}^*)}{n}}. \tag{V.27}$$

Moreover, note that for $k \geq t - \sqrt{nL(\mathbf{w}^*)}$, we have $\frac{\alpha(t-k-1)}{\beta} \leq \frac{\alpha\sqrt{nL(\mathbf{w}^*)}}{2\alpha\sqrt{nL(\mathbf{w}^*)}} \leq \frac{1}{2}$. Therefore, we have

$$\frac{(1+\gamma)d}{t\beta^2}\sum_{k=1}^{t-2}\max\Big\{\frac{\alpha(t-k-1)}{\beta} - \frac{1}{2}, 0\Big\}\big(\sum_{j=1}^{k}\lambda_j\big)^2$$
$$= \frac{(1+\gamma)d}{t\beta^2}\sum_{k=1}^{\lfloor t - \sqrt{nL(\mathbf{w}^*)}\rfloor}\Big(\frac{\alpha(t-k-1)}{\beta} - \frac{1}{2}\Big)\big(\sum_{j=1}^{k}\lambda_j\big)^2$$
$$\lesssim \frac{d\sqrt{nL(\mathbf{w}^*)}}{n^2 L(\mathbf{w}^*)} \cdot \frac{\alpha n^2}{2\alpha\sqrt{nL(\mathbf{w}^*)}} \cdot \Big(nn^{-\frac{3}{2}}\sqrt{L(\mathbf{w}^*)/d}\Big)^2$$
$$\asymp \frac{1}{n} \leq \sqrt{\frac{L(\mathbf{w}^*)}{n}}, \tag{V.28}$$

Dataset	$a9a$	dna	$gisette$	$mnist$	$mushrooms$	$phishing$	$poker$	$rcv1$	$w6a$
n	32561	2000	6000	60000	8124	11055	25010	23149	17188
d	123	180	5000	780	112	68	10	47236	300

TABLE II: Summary of datasets used in the experiments. Here, 'n' stands for the sample size and 'd' denotes the number of features.

where we have used $\lambda_j = n^{-3/2}\sqrt{L(\mathbf{w}^*)/d},$ if $j < t - \sqrt{nL(\mathbf{w}^*)}$. Plugging Eq. (V.25) - (V.28) into Eq. (V.24) gives

$$\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] \lesssim \sqrt{L(\mathbf{w}^*)/n}.$$

Next, we consider the case $L(\mathbf{w}^*) < \frac{1}{n}$. From the condition on $\{\lambda_k\}$ and t, we know that

$$\sum_{k=1}^{t} \lambda_k \leq (t-3)n^{-2}d^{-1/2} + 3 \lesssim 1.$$

Therefore, we have

$$\frac{\sum_{k=1}^{t}\lambda_k + \beta + \alpha t L(\mathbf{w}^*)/\beta}{t - \alpha t/\beta} \asymp \frac{\sum_{k=1}^{t}\lambda_k + \beta + tL(\mathbf{w}^*)}{(t-\alpha t/\beta)\gamma} \lesssim \frac{1 + nL(\mathbf{w}^*)}{n - n/2} \asymp \frac{1}{n}, \tag{V.29}$$

$$\frac{(1+\gamma)d}{t\beta^2}\sum_{k=1}^{t-2}\max\Big\{\frac{\alpha(t-k-1)}{\beta} - \frac{1}{2}, 0\Big\}\big(\sum_{j=1}^{k}\lambda_j\big)^2 \lesssim \frac{d}{n}\sum_{k=1}^{t-3}\frac{t-k-2}{2}\big(\sum_{j=1}^{k}\lambda_j\big)^2 \lesssim \frac{d}{n}n^2(nn^{-2}d^{-1/2})^2 \asymp \frac{1}{n}, \tag{V.30}$$

$$\frac{1+\gamma}{n\beta^2}\Big(1 + \frac{t}{n}\Big)\big(\sum_{k=1}^{t}\lambda_k + \beta + tL(\mathbf{w}^*)\big) \lesssim \frac{1}{n}(1 + nL(\mathbf{w}^*)) \asymp \frac{1}{n}. \tag{V.31}$$

Plugging Eq. (V.29) - (V.31) in Eq. (V.24) gives

$$\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] \lesssim 1/n,$$

which completes the proof. □

VI. Simulation Results

In this section, we conduct some experiments on stochastic RDA to support our theory. Specifically, we consider the ℓ_1-regularized logistic regression problem, where the loss function is given by

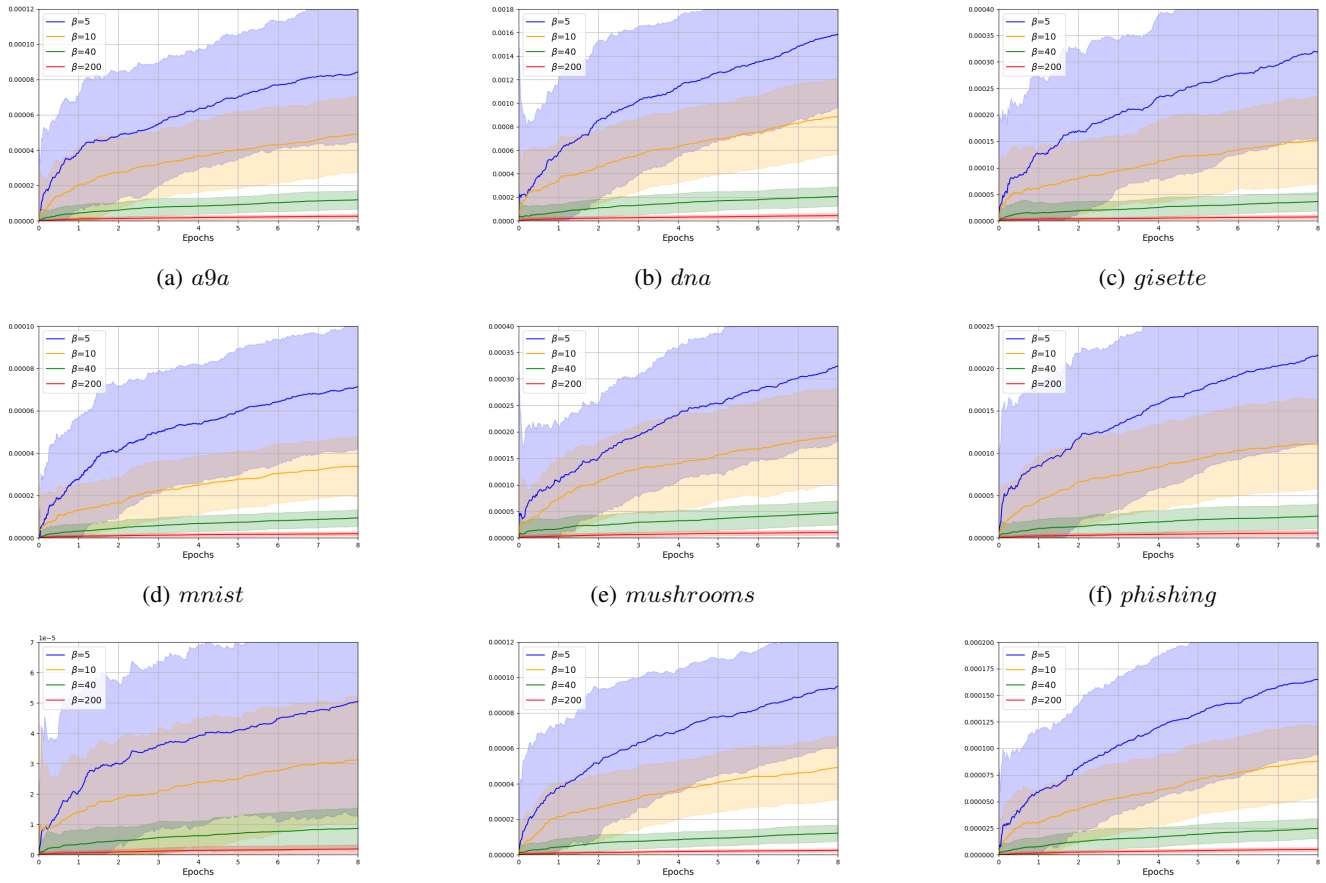
$$\ell(\mathbf{w}; \mathbf{z}) := \log(1 + \exp(-y\langle \mathbf{w}, \mathbf{x}\rangle)).$$

Here, $\mathbf{z} = (\mathbf{x}, y) \in \mathbb{R}^d \times \{-1, 1\}$ represents the data sample, where $\mathbf{x} \in \mathbb{R}^d$ corresponds to the feature vector and $y \in \{-1, 1\}$ represents the label. We use 9 real datasets from the LIBSVM library [8], whose information is summarized in Table II. We convert multi-class labeled datasets into binary class labeled datasets by assigning the first half of class labels to positive labels and the second half to negative labels. Subsequently, 80% of each dataset is randomly selected as the training set S, from which we perturb a single example to generate a neighboring dataset S'. We apply stochastic RDA (see Eq. (III.2)) with the same parameters on datasets S and S' and generate two sequences of iterates $\{\mathbf{w}_t\}_{t\geq 1}$ and $\{\mathbf{w}'_t\}_{t\geq 1}$. Subsequently, we compute the empirical distance $d_t := \|\mathbf{w}_t - \mathbf{w}'_t\|_2$ at each iteration to verify the stability of stochastic RDA. For parameter setting, we choose $\beta_k = \beta\sqrt{k}$ with $\beta \in \{5, 10, 40, 200\}$ and set $\beta_0 = \beta_1$. Moreover, we fix the regularization parameters as $\lambda_k \equiv \lambda = 10^{-4}$. To ensure robustness, we conduct 100 repetitions of the experiments for each dataset and report the average and standard deviation of $\{d_t\}$ against the number of passes (defined as the iteration number t divided by the sample size n).

In Figure 1, we present the evolution of $\{d_t\}$ for the logistic loss with four sequences of $\{\beta_k\}$. We observe that d_t is increasing w.r.t. t and decreasing w.r.t. β, which is consistent with our stability bounds established in Theorem 3.

To provide stronger validation of Theorem 3, we conduct additional experiments that directly compare the theoretical stability bound with the empirical distance d_t. In particular, we train the ℓ_1-regularized logistic regression model on datasets dna and $mushrooms$ using stochastic RDA for 100 epochs. We set $\beta_k = \beta\sqrt{k}$ with $\beta = 10$ and $\lambda_k \equiv 10^{-4}$, and approximate expectations using 100 independent runs. All other experimental settings are kept consistent with those in the previous experiment.

For each dataset, Figure 2 reports the comparison between the empirical distance d_t and the theoretical upper bound, with both mean values and standard deviations. From the results, we observe that the theoretical bound consistently upper bounds d_t throughout the training process. Moreover, the difference decreases as the iteration progresses, indicating that the bound becomes tighter over time. These findings provide strong empirical validation of our stability analysis in Theorem 3.

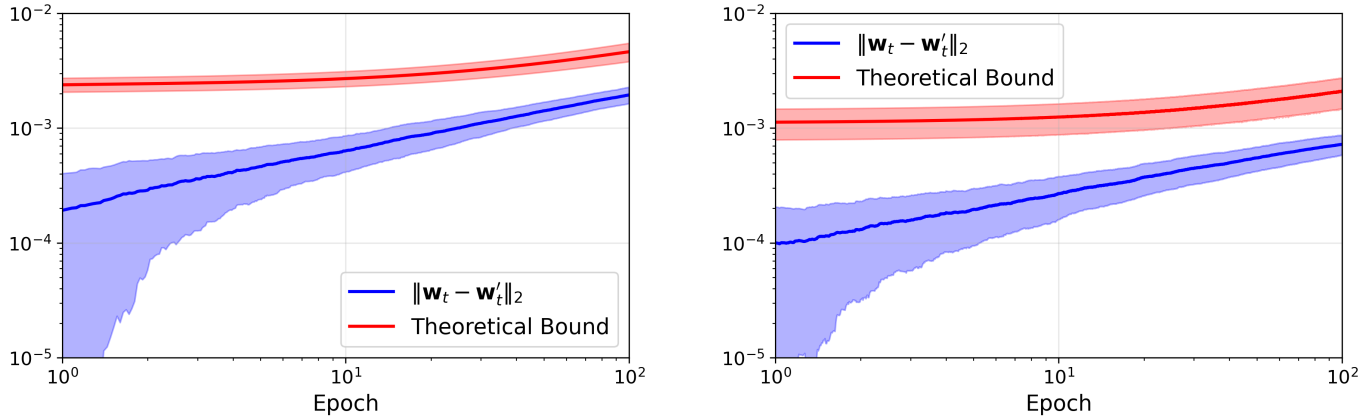


(a) *a9a* (b) *dna* (c) *gisette*

(d) *mnist* (e) *mushrooms* (f) *phishing*

(g) *poker* (h) *rcv1* (i) *w6a*

Fig. 1: Empirical distance $\|\mathbf{w}_t - \mathbf{w}'_t\|_2$ as a function of the number of passes for the logistic loss.



(a) *dna* (b) *mushrooms*

Fig. 2: Comparison between the empirical distance $\|\mathbf{w}_t - \mathbf{w}'_t\|_2$ and the theoretical stability bound established in Theorem 3.

VII. Conclusion

In this paper, we initiate the generalization analysis of the stochastic RDA method by leveraging the concept of algorithmic stability. We derive stability bounds of stochastic RDA for three distinct classes of loss functions: (i) convex and smooth, (ii) strongly convex and smooth, and (iii) convex, Lipschitz and nonsmooth losses. These stability bounds enable us to establish, for the first time, generalization error bounds for stochastic RDA. Furthermore, we provide new optimization error bounds for stochastic RDA in both convex and strongly convex settings, eliminating the bounded gradient assumption in the existing work. By unifying the generalization and optimization analyses, we derive excess population risk bounds for ℓ_1-regularized

problems that match existing minimax lower bounds. These contributions provide a deeper understanding of the generalization behavior of stochastic RDA and pave the way for further research into the stability and generalization of stochastic learning algorithms in machine learning.

ACKNOWLEDGEMENTS

We are grateful to the associate editor and reviewers for their thoughtful comments and constructive suggestions. The work of Y. Lei was supported by the Hong Kong Research Grants Council under the GRF projects 17302624 and 17305425. The work of Z. Wang was supported by the Wallenberg AI, Autonomous Systems, and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation. The work of X. Yuan was supported by the Hong Kong Research Grants Council under the GRF project 17309824.

APPENDIX

A. Proofs on Strongly Convex and Smooth Problems

In this section, we present proofs for strongly convex and smooth problems. Our basic idea is to reduce the problem into a convex and smooth problem, and then we can apply our previous results to derive the stated bounds. To this end, we introduce the following lemma.

Lemma A.16 ([52, Exercise 8, Chapter 2])**.** *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is a μ-strongly convex function with L-Lipschitz continuous gradient. Then the function $q(x) := f(x) - \frac{\mu}{2}\|x\|^2$ is convex with $(L-\mu)$-Lipschitz continuous gradients.*

1) Proofs on Stability Bounds: We first present the proof of Theorem 7 on stability analysis of stochastic RDA.

Proof of Theorem 7. It follows from Lemma A.16 and Eq. (IV.6) that $\tilde{\ell}$ is nonnegative, convex, and $(\alpha-\mu)$-smooth. Hence, the self-bounding property of $\tilde{\ell}$ (Lemma 15) implies that

$$\|\nabla\tilde{\ell}(\mathbf{w}_k;\mathbf{z}_{j_k})\|^2 \leq 2(\alpha-\mu)\tilde{\ell}(\mathbf{w}_k;\mathbf{z}_{j_k}). \tag{A.1}$$

Besides, notice that Eq. (IV.8) can be equivalently written as

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}}\Big\{\langle\tilde{s}_t,\mathbf{w}\rangle + \bar{r}_t(\mathbf{w}) + \frac{\beta_t+\mu t}{2t}\|\mathbf{w}\|^2\Big\}.$$

Then imitating the proof of Theorem 3, and replacing β_t by $\beta_t+\mu t$, we can derive

$$\begin{aligned}\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|^2] \leq& \frac{2(\alpha-\mu)}{\big(\beta_{t-1}+\mu(t-1)\big)^2}\sum_{k=1}^{t-2}\frac{1}{\beta_k+\mu k}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|^2]\\ -\frac{1}{\big(\beta_{t-1}+\mu(t-1)\big)^2}&\mathbb{E}[\|R_{t-1}'(\mathbf{w}_t)-R_{t-1}'(\mathbf{w}_t^{(i)})\|^2] + \frac{16(\alpha-\mu)}{n\big(\beta_{t-1}+\mu(t-1)\big)^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t-1}\mathbb{E}[\widetilde{L}_S(\mathbf{w}_k)].\end{aligned}$$

Hence the result follows by noting that $\widetilde{L}_S(\mathbf{w}) \leq L_S(\mathbf{w})$. The proof is completed. □

2) Proofs on Optimization Error Bounds: We now turn to the convergence analysis. As for the stability analysis, we also leverage the optimization analysis in Theorem 4 to conduct the convergence analysis for strongly convex problems.

Proof of Theorem 8. Let $\tilde{g}_0 = 0$ and $\tilde{g}_t = \sum_{k=1}^t \nabla\tilde{\ell}(\mathbf{w}_k;\mathbf{z}_{j_k})$ for $t \geq 1$. Define

$$\widetilde{M}_t(x) := \max_{\mathbf{w}\in\mathcal{W}}\Big\{\langle x,\mathbf{w}\rangle - \frac{\beta_t}{2}\|\mathbf{w}\|^2 - \widetilde{R}_t(\mathbf{w})\Big\} = \max_{\mathbf{w}\in\mathcal{W}}\Big\{\langle x,\mathbf{w}\rangle - \frac{\beta_t+\mu t}{2}\|\mathbf{w}\|^2 - R_t(\mathbf{w})\Big\}.$$

Following similar arguments as in the proof of Theorem 4, we know that $\widetilde{M}_t(\cdot)$ is $\frac{1}{\beta_t+\mu t}$-smooth. Then imitating Eq. (V.13)-(V.16) and noting the self bounding property in Eq. (A.1), we can correspondingly establish that

$$\sum_{k=1}^t(\langle\nabla\tilde{\ell}(\mathbf{w}_k;\mathbf{z}_{j_k}),\mathbf{w}_k\rangle + \tilde{r}_k(\mathbf{w}_{k+1})) \leq \langle\tilde{g}_t,\mathbf{w}\rangle + \widetilde{R}_t(\mathbf{w}) + \frac{\beta_t\|\mathbf{w}\|^2}{2} + \sum_{k=1}^t\frac{\alpha-\mu}{\beta_{k-1}+\mu(t-1)}\tilde{\ell}(\mathbf{w}_k;\mathbf{z}_{j_k}).$$

By the convexity of $\tilde{\ell}$, we further get

$$\begin{aligned}\sum_{k=1}^t\big(\tilde{\ell}(\mathbf{w}_k;\mathbf{z}_{j_k}) - \tilde{\ell}(\mathbf{w};\mathbf{z}_{j_k})\big) &\leq \sum_{k=1}^t\langle\nabla\tilde{\ell}(\mathbf{w}_k;\mathbf{z}_{j_k}),\mathbf{w}_k-\mathbf{w}\rangle = \sum_{k=1}^t\langle\nabla\tilde{\ell}(\mathbf{w}_k;\mathbf{z}_{j_k}),\mathbf{w}_k\rangle - \langle\tilde{g}_t,\mathbf{w}\rangle\\ &\leq \widetilde{R}_t(\mathbf{w}) + \frac{\beta_t\|\mathbf{w}\|^2}{2} + \sum_{k=1}^t\frac{\alpha-\mu}{\beta_{k-1}+\mu(t-1)}\tilde{\ell}(\mathbf{w}_k;\mathbf{z}_{j_k}) - \sum_{k=1}^t\tilde{r}_k(\mathbf{w}_{k+1}).\end{aligned}$$

Taking expectations over both sides leads to

$$\sum_{k=1}^{t}\mathbb{E}_A[\widetilde{L}_S(\mathbf{w}_k)-\widetilde{L}_S(\mathbf{w})]\leq\widetilde{R}_t(\mathbf{w})+\frac{\beta_t\|\mathbf{w}\|^2}{2}+\sum_{k=1}^{t}\frac{(\alpha-\mu)\mathbb{E}_A[\widetilde{L}_S(\mathbf{w}_k)]}{\beta_{k-1}+\mu(t-1)}-\sum_{k=1}^{t}\mathbb{E}[\tilde{r}_k(\mathbf{w}_{k+1})]. \tag{A.2}$$

Plugging $\widetilde{L}_S(\mathbf{w})=L_S(\mathbf{w})-\frac{\mu}{2}\|\mathbf{w}\|^2$, $\widetilde{R}_t(\mathbf{w})=R_t(\mathbf{w})+\frac{\mu t}{2}\|\mathbf{w}\|^2$, and $\tilde{r}_k(\mathbf{w})=r_k(\mathbf{w})+\frac{\mu}{2}\|\mathbf{w}\|^2$ into Eq. (A.2) further gives

$$\begin{aligned}
&\sum_{k=1}^{t}\mathbb{E}_A[L_S(\mathbf{w}_k)-L_S(\mathbf{w})]\\
\leq&R_t(\mathbf{w})+\frac{\beta_t\|\mathbf{w}\|^2}{2}+\sum_{k=1}^{t}\frac{(\alpha-\mu)\mathbb{E}_A[\widetilde{L}_S(\mathbf{w}_k)]}{\beta_{k-1}+\mu(t-1)}-\sum_{k=1}^{t}\mathbb{E}[r_k(\mathbf{w}_{k+1})]+\frac{\mu}{2}\sum_{k=1}^{t}(\|\mathbf{w}_k\|^2-\|\mathbf{w}_{k+1}\|^2)\\
\leq&\sum_{k=1}^{t}\mathbb{E}_A[r_k(\mathbf{w})-r_k(\mathbf{w}_{k+1})]+\frac{\beta_t\|\mathbf{w}\|^2+\mu\|\mathbf{w}_1\|^2}{2}+\sum_{k=1}^{t}\frac{(\alpha-\mu)\mathbb{E}_A[L_S(\mathbf{w}_k)]}{\beta_{k-1}+\mu(k-1)}.
\end{aligned}$$

Then the result follows after rearranging the terms. The proof is completed. □

3) Proofs on Excess Risk Bounds: In this subsection, we present the proof on excess risk bounds. We first prove Theorem 9, and then derive Corollary 6 by taking specific values for the parameters.

Proof of Theorem 9. We first bound $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\bar{\mathbf{w}}_t-\bar{\mathbf{w}}_t^{(i)}\|^2]$. It follows from Theorem 7 that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|^2]\leq\frac{2(\alpha-\mu)}{n\tilde{\beta}_{t-1}^2}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\frac{T_k^{(i)}}{\tilde{\beta}_k}-\frac{1}{n\tilde{\beta}_{t-1}^2}\sum_{i=1}^{n}T_{t-1}^{(i)}+\frac{16(\alpha-\mu)}{n\tilde{\beta}_{t-1}^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t}\mathbb{E}[L_S(\mathbf{w}_k)].$$

Then, following similar arguments as in the proof of Theorem 5, we can derive

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\bar{\mathbf{w}}_t-\bar{\mathbf{w}}_t^{(i)}\|^2]\leq\frac{2}{n}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\tilde{\xi}_kT_k^{(i)}+\frac{16(\alpha-\mu)}{n\tilde{\beta}_{t-1}^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t}\mathbb{E}[L_S(\mathbf{w}_k)].$$

Then combining Lemma 1, Part (a) and the above inequality together, we get (note $\tilde{\beta}_k\geq\beta_k\geq2(\alpha-\mu)$)

$$\begin{aligned}
&\mathbb{E}[L(\bar{\mathbf{w}}_t)-L_S(\bar{\mathbf{w}}_t)]\leq\frac{\alpha}{\gamma}\mathbb{E}[L_S(\bar{\mathbf{w}}_t)]+\frac{\alpha+\gamma}{2n}\sum_{i=1}^{n}\mathbb{E}[\|\bar{\mathbf{w}}_t-\bar{\mathbf{w}}_t^{(i)}\|^2]\\
\leq&\frac{\alpha}{\gamma}\mathbb{E}[L_S(\bar{\mathbf{w}}_t)]+(\alpha+\gamma)\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\tilde{\xi}_kT_k^{(i)}+\frac{8(\alpha-\mu)}{n\tilde{\beta}_{t-1}^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t}\mathbb{E}[L_S(\mathbf{w}_k)]\Big]\\
\leq&\frac{\alpha}{\gamma}\mathbb{E}[L_S(\bar{\mathbf{w}}_t)]+(\alpha+\gamma)\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\tilde{\xi}_kT_k^{(i)}+\frac{16(\alpha-\mu)}{n\tilde{\beta}_{t-1}^2}\Big(1+\frac{2t}{n}\Big)\sum_{k=1}^{t}\Big(1-\frac{\alpha-\mu}{\tilde{\beta}_{k-1}}\Big)\mathbb{E}[L_S(\mathbf{w}_k)]\Big].
\end{aligned}\tag{A.3}$$

By Theorem 8 and $\mathbb{E}[L_S(\mathbf{w})]=L(\mathbf{w})$, we know

$$\sum_{k=1}^{t}\Big(1-\frac{\alpha-\mu}{\tilde{\beta}_{k-1}}\Big)\mathbb{E}[L_S(\mathbf{w}_k)-L_S(\mathbf{w})]\leq R_t(\mathbf{w})+\frac{\beta_t\|\mathbf{w}\|^2+\mu\|\mathbf{w}_1\|^2}{2}+\sum_{k=1}^{t}\frac{(\alpha-\mu)L(\mathbf{w})}{\tilde{\beta}_{k-1}}.$$

Setting $\mathbf{w}=\mathbf{w}^*$ gives

$$\sum_{k=1}^{t}\Big(1-\frac{\alpha-\mu}{\tilde{\beta}_{k-1}}\Big)\mathbb{E}[L_S(\mathbf{w}_k)]\lesssim R_t(\mathbf{w}^*)+\beta_t\|\mathbf{w}^*\|^2+\mu\|\mathbf{w}_1\|^2+tL(\mathbf{w}^*). \tag{A.4}$$

By the convexity of L_S, we have from Eq. (A.4) that

$$\mathbb{E}[L_S(\bar{\mathbf{w}}_t)]\leq\sum_{k=1}^{t}\tilde{\theta}_k\mathbb{E}[L_S(\mathbf{w}_k)]\lesssim\frac{R_t(\mathbf{w}^*)+\beta_t\|\mathbf{w}^*\|^2+\mu\|\mathbf{w}_1\|^2+tL(\mathbf{w}^*)}{t-\tilde{A}_t}. \tag{A.5}$$

Plugging Eq. (A.4) and (A.5) into Eq. (A.3) further gives

$$\begin{aligned}
&\mathbb{E}[L(\bar{\mathbf{w}}_t)-L_S(\bar{\mathbf{w}}_t)]\lesssim\frac{R_t(\mathbf{w}^*)+\beta_t\|\mathbf{w}^*\|^2+\mu\|\mathbf{w}_1\|^2+tL(\mathbf{w}^*)}{(t-\tilde{A}_t)\gamma}\\
&+(1+\gamma)\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\tilde{\xi}_kT_k^{(i)}+\frac{1}{n\tilde{\beta}_t^2}\Big(1+\frac{t}{n}\Big)\big(R_t(\mathbf{w}^*)+\beta_t\|\mathbf{w}^*\|^2+\mu\|\mathbf{w}_1\|^2+tL(\mathbf{w}^*)\big)\Big].
\end{aligned}\tag{A.6}$$

On the other hand, we have from Eq. (A.5) that

$$\mathbb{E}[L_S(\bar{\mathbf{w}}_t)] - L(\mathbf{w}^*) \lesssim \frac{R_t(\mathbf{w}^*) + \beta_t\|\mathbf{w}^*\|^2 + \mu\|\mathbf{w}_1\|^2 + \tilde{A}_t L(\mathbf{w}^*)}{t - \tilde{A}_t}. \tag{A.7}$$

Combining Eq. (A.6) and Eq. (A.7), we have

$$\begin{aligned}\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] &\lesssim \frac{R_t(\mathbf{w}^*) + \beta_t\|\mathbf{w}^*\|^2 + \mu\|\mathbf{w}_1\|^2 + \tilde{A}_t L(\mathbf{w}^*)}{t - \tilde{A}_t} + \frac{R_t(\mathbf{w}^*) + \beta_t\|\mathbf{w}^*\|^2 + \mu\|\mathbf{w}_1\|^2 + tL(\mathbf{w}^*)}{(t - \tilde{A}_t)\gamma} \\ &\quad + (1+\gamma)\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{t-2}\tilde{\xi}_k T_k^{(i)} + \frac{1}{n\tilde{\beta}_t^2}\Big(1 + \frac{t}{n}\Big)\big(R_t(\mathbf{w}^*) + \beta_t\|\mathbf{w}^*\|^2 + \mu\|\mathbf{w}_1\|^2 + tL(\mathbf{w}^*)\big)\Big].\end{aligned}$$

The proof is completed. □

Proof of Corollary 10. We start with the following inequality

$$\begin{aligned}\tilde{A}_t &= \sum_{k=1}^{t}\frac{\alpha-\mu}{\beta + \mu(k-1)} = \frac{\alpha-\mu}{\mu}\sum_{k=1}^{t}\frac{1}{\beta/\mu - 1 + k} \leq \frac{\alpha-\mu}{\beta} + \frac{\alpha-\mu}{\mu}\int_1^t \frac{1}{\beta/\mu - 1 + x}dx \\ &\leq \frac{\alpha-\mu}{\beta} + \frac{\alpha-\mu}{\mu}\ln\Big(\frac{\mu(t-1)}{\beta} + 1\Big).\end{aligned} \tag{A.8}$$

Then it follows from the basic inequality $\ln(1+x) \leq x$ and $\beta \geq 2(\alpha-\mu)$ that

$$\frac{1}{t - \tilde{A}_t} \leq \frac{1}{t - \frac{\alpha-\mu}{\beta} - \frac{\alpha-\mu}{\mu}\ln(\frac{\mu(t-1)}{\beta} + 1)} \leq \frac{1}{t - \frac{\alpha-\mu}{\beta} - \frac{\alpha-\mu}{\mu}\frac{\mu(t-1)}{\beta}} = \frac{1}{(1 - \frac{\alpha-\mu}{\beta})t} \leq \frac{2}{t} \tag{A.9}$$

Plugging $\beta \geq \mu$ in Eq. (A.8), we know $\tilde{A}_t \lesssim \frac{\ln(t)}{\mu}$. Then applying Theorem 9, together with Eq. (V.22) and Eq. (A.9), we can derive

$$\begin{aligned}\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] &\lesssim \frac{\mu\sum_{k=1}^{t}\lambda_k + \mu + \ln(t)L(\mathbf{w}^*)}{\mu t} + \frac{\sum_{k=1}^{t}\lambda_k + 1 + tL(\mathbf{w}^*)}{t\gamma} \\ &\quad + (1+\gamma)\Big[d\sum_{k=1}^{t-2}\tilde{\xi}_k\big(\sum_{j=1}^{k}\lambda_j\big)^2 + \frac{1}{n\tilde{\beta}_t^2}\Big(1+\frac{t}{n}\Big)\big(\sum_{k=1}^{t}\lambda_k + 1 + tL(\mathbf{w}^*)\big)\Big].\end{aligned} \tag{A.10}$$

From the condition on $\{\lambda_k\}$ and t, we know that

$$\sum_{k=1}^{t}\lambda_k \leq \frac{\sqrt{\mu\ln(n)}}{n\sqrt{d}}t + \frac{\sqrt{\ln(n)}}{n\mu^2}\frac{\mu}{\mu + 8(\alpha-\mu)}t \lesssim \frac{\sqrt{\ln(n)}}{\mu}.$$

Therefore, we have (note $n^{-1}\sqrt{\ln(n)} \lesssim \mu$)

$$\frac{\mu\sum_{k=1}^{t}\lambda_k + \mu + \ln(t)L(\mathbf{w}^*)}{\mu t} \lesssim \frac{\sqrt{\ln(n)} + \mu + \ln(n)}{\mu n} \asymp \frac{\ln(n)}{\mu n}, \tag{A.11}$$

$$\frac{\sum_{k=1}^{t}\lambda_k + 1 + tL(\mathbf{w}^*)}{t\gamma} \lesssim \frac{\mu^{-1}\sqrt{\ln(n)} + 1 + n}{\mu n^2} \asymp \frac{1}{\mu n} + \frac{\sqrt{\ln n}}{\mu^2 n^2} \lesssim \frac{1}{\mu n}, \tag{A.12}$$

$$\frac{1+\gamma}{n\tilde{\beta}_t^2}\Big(1+\frac{t}{n}\Big)\big(\sum_{k=1}^{t}\lambda_k + 1 + tL(\mathbf{w}^*)\big) \lesssim \frac{\mu n}{\mu^2 n t^2}(\mu^{-1}\sqrt{\ln(n)} + 1 + n) \lesssim \frac{1}{\mu n}. \tag{A.13}$$

On the other hand, we know from Eq. (A.9) that $\tilde{\theta}_k \leq \frac{1}{t-\tilde{A}_t} \leq \frac{2}{t}$ for all $k \leq t$. Moreover, the condition $\beta \geq 2(\alpha-\mu)$ implies that

$$\tilde{\theta}_k \geq \frac{1 - (\alpha-\mu)/\beta}{t} \geq \frac{1 - 1/2}{t} = \frac{1}{2t}.$$

Hence, we further derive that (note $\tilde{\beta}_{j+1} \geq \tilde{\beta}_k$ for $j \geq k$)

$$\tilde{\xi}_k \leq \frac{2(\alpha-\mu)}{t}\sum_{j=k}^{t-2}\frac{1}{\tilde{\beta}_{j+1}^2\tilde{\beta}_k} - \frac{1}{4t\tilde{\beta}_k^2} \leq \frac{2(\alpha-\mu)}{t\tilde{\beta}_k^2}\Big(\sum_{j=k}^{t-2}\frac{1}{\tilde{\beta}_{j+1}} - \frac{1}{8(\alpha-\mu)}\Big). \tag{A.14}$$

Since $\tilde{\beta}_k \geq \mu + \mu k = \mu(k+1)$, we have from the inequality $\ln(1+x) \leq x$ that

$$\sum_{j=k}^{t-2}\frac{1}{\tilde{\beta}_{j+1}} \leq \sum_{j=k+2}^{t}\frac{1}{\mu j} \leq \int_{k+1}^{t}\frac{1}{\mu x}dx = \frac{1}{\mu}\ln\Big(\frac{t}{k+1}\Big) \leq \frac{t-k}{\mu k}.$$

Then if $k \geq \frac{8(\alpha-\mu)}{\mu+8(\alpha-\mu)}t$, we know from Eq. (A.14) that $\tilde{\xi}_k \leq 0$. For $k < \frac{8(\alpha-\mu)}{\mu+8(\alpha-\mu)}t$, it follows from $\tilde{\theta}_k \leq 2/t$ and $\tilde{\beta}_k \geq \mu(k+1)$ that

$$\tilde{\xi}_k \lesssim \frac{1}{t\tilde{\beta}_k}\sum_{j=k}^{t-2}\frac{1}{\tilde{\beta}_{j+1}^2} \leq \frac{1}{\mu^2 t\tilde{\beta}_k}\sum_{j=k}^{t-2}\frac{1}{(j+2)^2} \leq \frac{1}{\mu^3 kt}\int_{k+1}^{t} x^{-2}dx \lesssim \frac{1}{\mu^3 k^2 t}.$$

Therefore, we can further derive that

$$(1+\gamma)d\sum_{k=1}^{t-2}\tilde{\xi}_k\big(\sum_{j=1}^{k}\lambda_j\big)^2 \lesssim \mu n d\sum_{k=1}^{\lfloor\frac{8(\alpha-\mu)}{\mu+8(\alpha-\mu)}t\rfloor}\frac{1}{\mu^3 k^2 t}\Big(k\frac{\sqrt{\mu\ln(n)}}{n\sqrt{d}}\Big)^2 \lesssim \frac{\ln(n)}{\mu n}. \tag{A.15}$$

Plugging Eq. (A.11) - (A.13) and Eq. (A.15) into Eq. (A.10) gives

$$\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] \lesssim \frac{\ln(n)}{\mu n},$$

which completes the proof. □

B. Proofs on Convex, Lipschitz and Nonsmooth Problems

In this section, we present the proofs for convex, Lipschitz and nonsmooth problems.

1) Proofs on Stability Bounds: We first present the proof of stability bounds. Unlike the smooth case, we use the monotonicity of gradients for convex functions.

Proof of Theorem 11. Recall the notations s_t, $s_t^{(i)}$, and $\Delta_t^{(i)}$ introduced in the proof of Theorem 3. We start from the following inequality (i.e., Eq. (V.6))

$$\begin{aligned}\big\|s_{t+1} - s_{t+1}^{(i)}\big\|^2 \leq& \frac{t^2}{(t+1)^2}\|s_t - s_t^{(i)}\|^2 + \frac{(\Delta_t^{(i)})^2}{(t+1)^2} + \frac{2t\|\bar{r}_t'(\mathbf{w}_{t+1}) - \bar{r}_t'(\mathbf{w}_{t+1}^{(i)})\|\Delta_t^{(i)}}{(t+1)^2}\\ &- \frac{2\beta_t}{(t+1)^2}\big\langle \mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}, \nabla\ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla\ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\rangle.\end{aligned}$$

Note we have from Eq. (IV.9) that

$$\Delta_t^{(i)} = \big\|\nabla\ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla\ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\| \leq \big\|\nabla\ell(\mathbf{w}_{t+1}; S_{j_{t+1}})\big\| + \big\|\nabla\ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\| \leq 2G.$$

Then it follows that

$$\begin{aligned}\big\|s_{t+1} - s_{t+1}^{(i)}\big\|^2 \leq& \frac{t^2}{(t+1)^2}\|s_t - s_t^{(i)}\|^2 + \frac{4G^2}{(t+1)^2} + \frac{4Gt}{(t+1)^2}\|\bar{r}_t'(\mathbf{w}_{t+1}) - \bar{r}_t'(\mathbf{w}_{t+1}^{(i)})\|\\ &- \frac{2\beta_t}{(t+1)^2}\big\langle \mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}, \nabla\ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla\ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\rangle.\end{aligned}$$

We first assume $j_{t+1} \neq i$ and then $S_{j_{t+1}} = S_{j_{t+1}}^{(i)}$. The convexity of ℓ implies that

$$\big\langle \mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}, \nabla\ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla\ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big\rangle \geq 0,$$

and hence we derive

$$\big\|s_{t+1} - s_{t+1}^{(i)}\big\|^2 \leq \frac{t^2}{(t+1)^2}\|s_t - s_t^{(i)}\|^2 + \frac{4G^2}{(t+1)^2} + \frac{4Gt}{(t+1)^2}\|\bar{r}_t'(\mathbf{w}_{t+1}) - \bar{r}_t'(\mathbf{w}_{t+1}^{(i)})\|.$$

We then consider $j_{t+1} = i$. In this case, we have

$$\begin{aligned}\big\|s_{t+1} - s_{t+1}^{(i)}\big\| &= \Big\|\frac{t}{t+1}(s_t - s_t^{(i)}) + \frac{1}{t+1}\big(\nabla\ell(\mathbf{w}_{t+1}; S_{j_{t+1}}) - \nabla\ell(\mathbf{w}_{t+1}^{(i)}; S_{j_{t+1}}^{(i)})\big)\Big\|\\ &\leq \frac{t}{t+1}\|s_t - s_t^{(i)}\| + \frac{2G}{t+1},\end{aligned}$$

which gives the following inequality

$$\big\|s_{t+1} - s_{t+1}^{(i)}\big\|^2 \leq \frac{t^2}{(t+1)^2}\|s_t - s_t^{(i)}\|^2 + \frac{4G^2}{(t+1)^2} + \frac{4Gt}{(t+1)^2}\|s_t - s_t^{(i)}\|.$$

We combine the above two cases and take expectations over both sides to get

$$\mathbb{E}\big[\big\|s_{t+1}-s_{t+1}^{(i)}\big\|^2\big] \le \frac{t^2}{(t+1)^2}\mathbb{E}[\|s_t-s_t^{(i)}\|^2] + \frac{4G^2}{(t+1)^2} + \frac{4G}{(t+1)^2}\mathbb{E}[\|R_t'(\mathbf{w}_{t+1})-R_t'(\mathbf{w}_{t+1}^{(i)})\|] + \frac{4Gt}{n(t+1)^2}\big(\mathbb{E}[\|s_t-s_t^{(i)}\|^2]\big)^{\frac{1}{2}}.$$

We apply the above inequality recursively and derive

$$\mathbb{E}[\|s_{t+1}-s_{t+1}^{(i)}\|^2] \le \sum_{k=1}^t \frac{4G^2}{(k+1)^2}\prod_{k'=k+1}^t\frac{(k')^2}{(k'+1)^2} + \sum_{k=1}^t\frac{4G}{(k+1)^2}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|]\prod_{k'=k+1}^t\frac{(k')^2}{(k'+1)^2} + \sum_{k=1}^t\frac{4Gk}{n(k+1)^2}\big(\mathbb{E}[\|s_k-s_k^{(i)}\|^2]\big)^{\frac{1}{2}}\prod_{k'=k+1}^t\frac{(k')^2}{(k'+1)^2} + \frac{1}{(t+1)^2}\mathbb{E}[\|s_1-s_1^{(i)}\|^2].$$

It then follows from $\mathbb{E}[\|s_1-s_1^{(i)}\|^2]=\mathbb{E}[\|\nabla\ell(\mathbf{w}_1;S_{j_1})-\nabla\ell(\mathbf{w}_1^{(i)};S_{j_1}^{(i)})\|^2]\le 4G^2$ that

$$\mathbb{E}[\|s_{t+1}-s_{t+1}^{(i)}\|^2] \le \frac{4G^2(t+1)}{(t+1)^2} + \sum_{k=1}^t\frac{4G}{(t+1)^2}\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|] + \frac{4G}{n(t+1)^2}\sum_{k=1}^t k\big(\mathbb{E}[\|s_k-s_k^{(i)}\|^2]\big)^{\frac{1}{2}}.$$

We can rewrite the above inequality as

$$(t+1)^2\mathbb{E}[\|s_{t+1}-s_{t+1}^{(i)}\|^2] \le 4G^2(t+1)+4G\sum_{k=1}^t\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|] + \frac{4G}{n}\sum_{k=1}^t k\big(\mathbb{E}[\|s_k-s_k^{(i)}\|^2]\big)^{\frac{1}{2}}. \quad \text{(A.16)}$$

Denote $\mathfrak{C}_{t,i}:=\max_{k\in[t+1]}\big(k^2\mathbb{E}[\|s_k-s_k^{(i)}\|^2]\big)^{\frac{1}{2}}$. Since the right-hand side of Eq. (A.16) is an increasing function of t, we further get

$$\mathfrak{C}_{t,i}^2 \le 4G\sum_{k=1}^t\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|]+4G^2(t+1)+\frac{4Gt\mathfrak{C}_{t,i}}{n}.$$

Solving the above quadratic inequality of $\mathfrak{C}_{t,i}$ gives (by Lemma 14)

$$\mathfrak{C}_{t,i}^2 \le 8G\sum_{k=1}^t\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|]+8G^2(t+1)+\frac{16G^2t^2}{n^2}.$$

It then follows that

$$\mathbb{E}[\|s_{t+1}-s_{t+1}^{(i)}\|^2] \le \frac{8G}{(t+1)^2}\sum_{k=1}^t\mathbb{E}[\|R_k'(\mathbf{w}_{k+1})-R_k'(\mathbf{w}_{k+1}^{(i)})\|]+\frac{16G^2}{n^2}+\frac{8G^2}{t+1},$$

which together with Eq. (V.4) proves the result Eq. (IV.10). We can slightly improve the result if $t=2$. Indeed, by Eq. (V.4) and $\mathbf{w}_1=\mathbf{w}_1^{(i)}$, we know

$$\begin{aligned}\mathbb{E}[\|\mathbf{w}_2-\mathbf{w}_2^{(i)}\|^2] &\le \frac{1}{\beta_1^2}\mathbb{E}[\|s_1-s_1^{(i)}\|^2]-\frac{1}{\beta_1^2}\mathbb{E}[\|R_1'(\mathbf{w}_2)-R_1'(\mathbf{w}_2^{(i)})\|^2]\\ &= \frac{1}{\beta_1^2}\mathbb{E}[\|\nabla\ell(\mathbf{w}_1;S_{j_1})-\nabla\ell(\mathbf{w}_1^{(i)};S_{j_1}^{(i)})\|^2]-\frac{1}{\beta_1^2}\mathbb{E}[\|R_1'(\mathbf{w}_2)-R_1'(\mathbf{w}_2^{(i)})\|^2]\\ &\le \frac{1}{n\beta_1^2}\big(2\mathbb{E}[\|\nabla\ell(\mathbf{w}_1;\mathbf{z}_i)\|^2]+2\mathbb{E}[\|\nabla\ell(\mathbf{w}_1^{(i)};\mathbf{z}_i')\|^2]\big)-\frac{1}{\beta_1^2}\mathbb{E}[\|R_1'(\mathbf{w}_2)-R_1'(\mathbf{w}_2^{(i)})\|^2]\\ &= \frac{4G^2}{n\beta_1^2}-\frac{1}{\beta_1^2}\mathbb{E}[\|R_1'(\mathbf{w}_2)-R_1'(\mathbf{w}_2^{(i)})\|^2]. \end{aligned} \quad \text{(A.17)}$$

The proof is now completed. □

2) Proofs on Excess Risk Bounds: We recall the following optimization error bounds for stochastic RDA in the convex and nonsmooth case.

Lemma A.17 (Optimization error [53]). *Assume the map $\mathbf{w}\mapsto r_k(\mathbf{w})$ is convex for any k, and for any $\mathbf{z}\in\mathcal{Z}$, $\mathbf{w}\mapsto\ell(\mathbf{w};\mathbf{z})$ is convex and satisfies the bounded gradient assumption in Eq.* (IV.9). *Let $\mathbf{w}_t$ be the sequence produced by stochastic RDA applied to S. If $\{\beta_k\}$ is nondecreasing, then*

$$\sum_{k=1}^t\mathbb{E}_A[L_S(\mathbf{w}_k)-L_S(\mathbf{w})] \le \sum_{k=1}^t\mathbb{E}_A[r_k(\mathbf{w})-r_k(\mathbf{w}_{k+1})]+\beta_t\|\mathbf{w}\|^2+\sum_{k=1}^t\frac{G^2}{2\beta_{k-1}}.$$

We combine the stability bounds in Theorem 11 and Lemma A.17 to get the excess risk bounds. We first prove Theorem 12 for general cases, and then prove Corollary 13 for excess risk bounds with specific parameters.

Proof of Theorem 12. We first bound $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_t^{(i)}\|^2]$. Note that $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ for any random variable X. Then it follows from Theorem 11 that for $t \geq 3$

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|^2] \leq \frac{8G}{n\beta_{t-1}^2}\sum_{i=1}^n\sum_{k=1}^{t-2} P_k^{(i)} - \frac{1}{n\beta_{t-1}^2}\sum_{i=1}^n \big(P_{t-1}^{(i)}\big)^2 + \frac{16G^2t^2}{n^2\beta_{t-1}^2} + \frac{8G^2t}{\beta_{t-1}^2}.$$

Moreover, we have $\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_1^{(i)}\|^2] = 0$ and

$$\mathbb{E}[\|\mathbf{w}_2 - \mathbf{w}_2^{(i)}\|^2] \leq \frac{4G^2}{n\beta_1^2} - \frac{\big(P_1^{(i)}\big)^2}{\beta_1^2}$$

from Eq. (A.17). Therefore, by convexity of $\|\cdot\|^2$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_t^{(i)}\|^2] \leq \frac{1}{nt}\sum_{i=1}^n\sum_{j=1}^t \mathbb{E}[\|\mathbf{w}_j - \mathbf{w}_j^{(i)}\|^2] \\
&\leq \frac{1}{nt}\sum_{i=1}^n\sum_{j=3}^t \Big[\frac{8G}{\beta_{j-1}^2}\sum_{k=1}^{j-2} P_k^{(i)} - \frac{1}{\beta_{j-1}^2}\big(P_{j-1}^{(i)}\big)^2\Big] - \frac{1}{nt\beta_1^2}\sum_{i=1}^n \big(P_1^{(i)}\big)^2 + \frac{16G^2t^2}{n^2\beta_{t-1}^2} + \frac{8G^2t}{\beta_{t-1}^2} \\
&= \frac{1}{nt}\sum_{i=1}^n \Big[\sum_{j=3}^t\sum_{k=1}^{j-2} \frac{8GP_k^{(i)}}{\beta_{j-1}^2} - \sum_{k=1}^{t-1}\frac{1}{\beta_k^2}\big(P_k^{(i)}\big)^2\Big] + \frac{16G^2t^2}{n^2\beta_{t-1}^2} + \frac{8G^2t}{\beta_{t-1}^2} \\
&= \frac{1}{nt}\sum_{i=1}^n \Big[\sum_{j=1}^{t-2}\sum_{k=1}^{j} \frac{8GP_k^{(i)}}{\beta_{j+1}^2} - \sum_{k=1}^{t-1}\frac{1}{\beta_k^2}\big(P_k^{(i)}\big)^2\Big] + \frac{16G^2t^2}{n^2\beta_{t-1}^2} + \frac{8G^2t}{\beta_{t-1}^2}.
\end{aligned}$$

Exchanging the order of summation in the above inequality further gives

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_t^{(i)}\|^2] \leq& \frac{1}{nt}\sum_{i=1}^n \Big[\sum_{k=1}^{t-2}\big(\sum_{j=k}^{t-2}\frac{8G}{\beta_{j+1}^2}\big)P_k^{(i)} - \sum_{k=1}^{t-1}\frac{1}{\beta_k^2}\big(P_k^{(i)}\big)^2\Big] + \frac{16G^2t^2}{n^2\beta_{t-1}^2} + \frac{8G^2t}{\beta_{t-1}^2} \\
\leq& \frac{1}{nt}\sum_{i=1}^n\sum_{k=1}^{t-2} \Big[\big(\sum_{j=k}^{t-2}\frac{8G}{\beta_{j+1}^2}\big)P_k^{(i)} - \frac{1}{\beta_k^2}\big(P_k^{(i)}\big)^2\Big] + \frac{16G^2t^2}{n^2\beta_{t-1}^2} + \frac{8G^2t}{\beta_{t-1}^2}.
\end{aligned}$$

Then combining Lemma 1 Part (b) and the above inequality together, we get

$$\begin{aligned}
&\mathbb{E}[L(\bar{\mathbf{w}}_t) - L_S(\bar{\mathbf{w}}_t)] \leq \frac{G^2}{2\gamma} + \frac{\gamma}{2n}\sum_{i=1}^n \mathbb{E}[\|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_t^{(i)}\|^2] \\
\leq& \frac{G^2}{2\gamma} + \gamma\Big[\frac{1}{nt}\sum_{i=1}^n\sum_{k=1}^{t-2}\Big[\big(\sum_{j=k}^{t-2}\frac{4G}{\beta_{j+1}^2}\big)P_k^{(i)} - \frac{1}{2\beta_k^2}\big(P_k^{(i)}\big)^2\Big] + \frac{8G^2t^2}{n^2\beta_{t-1}^2} + \frac{4G^2t}{\beta_{t-1}^2}\Big].
\end{aligned} \tag{A.18}$$

On the other hand, we have from Lemma A.17 and convexity of L_S that

$$\mathbb{E}[L_S(\bar{\mathbf{w}}_t)] - L(\mathbf{w}^*) \leq \frac{1}{t}\sum_{k=1}^t \mathbb{E}[L_S(\mathbf{w}_k) - L_S(\mathbf{w}^*)] \leq \bar{r}_t(\mathbf{w}^*) + \frac{\beta_t\|\mathbf{w}^*\|^2}{t} + \frac{G^2}{2t}\sum_{k=1}^t \frac{1}{\beta_{k-1}}. \tag{A.19}$$

Combining Eq. (A.18) and Eq. (A.19), we have

$$\begin{aligned}
\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] \lesssim& \frac{1}{\gamma} + \gamma\Big[\frac{1}{nt}\sum_{i=1}^n\sum_{k=1}^{t-2}\Big[\big(\sum_{j=k}^{t-2}\frac{4G}{\beta_{j+1}^2}\big)P_k^{(i)} - \frac{1}{2\beta_k^2}\big(P_k^{(i)}\big)^2\Big] + \frac{t^2}{n^2\beta_{t-1}^2} + \frac{t}{\beta_{t-1}^2}\Big] \\
&+ \bar{r}_t(\mathbf{w}^*) + \frac{\beta_t\|\mathbf{w}^*\|^2}{t} + \frac{1}{t}\sum_{k=1}^t \frac{1}{\beta_{k-1}}.
\end{aligned}$$

The proof is completed. □

Proof of Corollary 13. By Eq. (V.22), we know $P_k^{(i)} \leq 2\sqrt{d}\sum_{j=1}^k \lambda_j$. Then it follows from Theorem 12 that

$$\mathbb{E}[L(\bar{\mathbf{w}}_t) - L(\mathbf{w}^*)] \lesssim \frac{1}{\gamma} + \gamma\Big[\frac{\sqrt{d}}{t}\sum_{k=1}^{t-2}\frac{t-k-1}{\beta^2}\sum_{j=1}^k \lambda_j + \frac{t^2}{n^2\beta^2} + \frac{t}{\beta^2}\Big] + \frac{1}{t}\sum_{k=1}^t \lambda_k + \frac{\beta\|\mathbf{w}^*\|^2}{t} + \frac{1}{\beta}. \tag{A.20}$$

From the condition on $\{\lambda_k\}$ and t, we know that

$$\frac{1}{t}\sum_{k=1}^{t}\lambda_k \leq \frac{1}{tn^2\sqrt{d}}(t-2)+\frac{2n^{\frac{3}{2}}}{t} \lesssim n^{-1/2}.$$

Hence, we derive

$$\frac{1}{t}\sum_{k=1}^{t}\lambda_k+\frac{\beta\|\mathbf{w}^*\|^2}{t}+\frac{1}{\beta} \lesssim n^{-1/2}+\frac{n^{3/2}}{n^2}+n^{-3/2} \asymp \frac{1}{\sqrt{n}}. \tag{A.21}$$

For $k \leq t-2$, we can derive

$$\frac{\gamma\sqrt{d}}{t}\sum_{k=1}^{t-2}\frac{t-k-1}{\beta^2}\sum_{j=1}^{k}\lambda_j \leq \frac{\gamma\sqrt{d}}{t\beta^2}\sum_{k=1}^{t}t(kn^{-2}d^{-1/2}) \asymp \frac{\sqrt{n}}{tn^3}t^3n^{-2} \asymp \frac{1}{\sqrt{n}}. \tag{A.22}$$

Moreover, from the condition on γ and t, we further have

$$\gamma\Big(\frac{t^2}{n^2\beta^2}+\frac{t}{\beta^2}\Big) \asymp \sqrt{n}(\frac{n^4}{n^2n^3}+\frac{n^2}{n^3}) \asymp \frac{1}{\sqrt{n}}. \tag{A.23}$$

Plugging Eq. (A.21)-(A.23) into Eq. (A.20) and noting that $1/\gamma = 1/\sqrt{n}$ gives

$$\mathbb{E}[L(\bar{\mathbf{w}}_t)-L(\mathbf{w}^*)] \lesssim 1/\sqrt{n},$$

which completes the proof. □

REFERENCES

[1] A. Agarwal, M. J. Wainwright, P. L. Bartlett, and P. K. Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009.

[2] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

[3] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[4] O. Bousquet and L. Bottou. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

[5] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

[6] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626, 2020.

[7] M. Bun, M. Gaboardi, M. Hopkins, R. Impagliazzo, R. Lei, T. Pitassi, S. Sivakumar, and J. Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 520–527, 2023.

[8] C.-C. Chang. Libsvm data: Classification, regression, and multi-label. *http://www. csie. ntu. edu. tw/˜ cjlin/libsvmtools/datasets/*, 2008.

[9] S.-K. Chao and G. Cheng. A generalization of regularized dual averaging and its dynamics. *arXiv preprint arXiv:1909.10072*, 2019.

[10] X. Chen, Q. Lin, and J. Pena. Optimal regularized dual averaging methods for stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 395–403, 2012.

[11] I. Colin, A. Bellet, J. Salmon, and S. Clémençon. Gossip dual averaging for decentralized optimization of pairwise functions. In *International conference on machine learning*, pages 1388–1396. PMLR, 2016.

[12] A. Defazio and S. Jelassi. A momentumized, adaptive, dual averaged gradient method. *Journal of Machine Learning Research*, 23 (144):1–34, 2022.

[13] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012.

[14] X. Deng, L. Shen, S. Li, T. Sun, D. Li, and D. Tao. Towards understanding the generalizability of delayed stochastic gradient descent. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[15] P. Deora, R. Ghaderi, H. Taheri, and C. Thrampoulidis. On the optimization and generalization of multi-head attention. *Transactions on Machine Learning Research*, 2024.

[16] J. Duchi and F. Ruan. Local asymptotics for some stochastic optimization problems: Optimality, constraint identification, and dual averaging. *arXiv preprint arXiv:1612.05612*, 2016.

[17] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. In *Advances in Neural Information Processing Systems*, pages 495–503, 2009.

[18] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[19] J. C. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

[20] J. Fan and Y. Lei. High-probability generalization bounds for pointwise uniformly stable algorithms. *Applied and Computational Harmonic Analysis*, 70:101632, 2024.

[21] V. Feldman and J. Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279, 2019.

[22] N. Flammarion and F. Bach. Stochastic composite least-squares regression with convergence rate $o(1/n)$. In *Conference on Learning Theory*, pages 831–875. PMLR, 2017.

[23] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

[24] T. Hastie, R. Tibshirani, and M. Wainwright. Statistical learning with sparsity. *Monographs on Statistics and Applied Probability*, 143 (143):8, 2015.

[25] A. Héliou, M. Martin, P. Mertikopoulos, and T. Rahier. Online non-convex optimization with imperfect feedback. *Advances in Neural Information Processing Systems*, 33:17224–17235, 2020.

[26] K. Ji, Y. Zhou, and Y. Liang. Understanding estimation and generalization error of generative adversarial networks. *IEEE Transactions on Information Theory*, 67(5):3114–3129, 2021.

[27] I. Kuzborskij and C. Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2820–2829, 2018.

[28] I. Kuzborskij and F. Orabona. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pages 942–950. PMLR, 2013.

[29] S. Lee, S. J. Wright, and L. Bottou. Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(6), 2012.

[30] Y. Lei and Y. Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819, 2020.

[31] Y. Lei, R. Jin, and Y. Ying. Stability and generalization analysis of gradient methods for shallow neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pages 38557–38570, 2022.

[32] Y. Li, M. E. Ildiz, D. Papailiopoulos, and S. Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pages 19565–19594, 2023.

[33] C. Liu, L. Zhu, and M. Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. *Advances in Neural Information Processing Systems*, 33:15954–15964, 2020.

[34] C. Liu, X. Wu, X. Yi, Y. Shi, and K. H. Johansson. Rate analysis of dual averaging for nonconvex distributed optimization. *IFAC-PapersOnLine*, 56(2):5209–5214, 2023.

[35] S. Mahadevan, B. Liu, P. Thomas, W. Dabney, S. Giguere, N. Jacek, I. Gemp, and J. Liu. Proximal reinforcement learning: A new theory of sequential decision making in primal-dual spaces. *arXiv preprint arXiv:1405.6757*, 2014.

[36] A. S. Nemirovskij and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* John Wiley & Sons, 1983.

[37] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[38] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

[39] G. Neu, G. K. Dziugaite, M. Haghifam, and D. M. Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pages 3526–3545. PMLR, 2021.

[40] K. Nikolakakis, F. Haddadpour, A. Karbasi, and D. Kalogerias. Beyond lipschitz: Sharp generalization and excess risk bounds for full-batch gd. In *The Eleventh International Conference on Learning Representations*, 2023.

[41] K. E. Nikolakakis, A. Karbasi, and D. Kalogerias. Select without fear: Almost all minibatch schedules generalize optimally. *SIAM Journal on Mathematics of Data Science*, 7(3):965–992, 2025.

[42] D. Richards and I. Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. *Advances in Neural Information Processing Systems*, 34, 2021.

[43] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, pages 400–407, 1951.

[44] R. T. Rockafellar. *Convex Analysis*, volume 11. Princeton University Press, 1997.

[45] M. Schliserman and T. Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In *Conference on Learning Theory*, pages 3380–3394, 2022.

[46] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

[47] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2010.

[48] T. Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *International Conference on Machine Learning*, pages 392–400, 2013.

[49] H. Taheri and C. Thrampoulidis. Generalization and stability of interpolating neural networks with minimal width. *Journal of Machine Learning Research*, 25(156):1–41, 2024.

[50] S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *Advances in Neural Information Processing Systems*, 32, 2019.

[51] P. Wang, Y. Lei, D. Wang, Y. Ying, and D.-X. Zhou. Generalization guarantees of gradient descent for shallow neural networks. *Neural Computation*, 37(2):344–402, 2025.

[52] S. Wright and B. Recht. *Optimization for Data Analysis*. Cambridge University Press, 2022. ISBN 9781316518984. URL https://books.google.com.hk/books?id=Kzk9zgEACAAJ.

[53] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

[54] X.-T. Yuan and P. Li. Stability and risk bounds of iterative hard thresholding. *IEEE Transactions on Information Theory*, 68(10): 6663–6681, 2022.