Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 922–951
https://doi.org/10.22067/ijnao.2026.95961.1753
https://ijnao.um.ac.ir/

Research Article

Robust stochastic gradient descent for linearly constrained problems via adaptive barrier amplification

A. Fillali*, S. Addoune and R. Zeghdane

Abstract

Methods based on stochastic gradient descent using relaxed barrier functions provide a powerful framework for linearly constrained optimization but often suffer from sensitivity to hyperparameter settings and slow recovery from infeasible regions. Our investigation reveals that the optimal performance of standard relaxed barrier methods is confined to a narrow and unstable, "special-case" optimization path, limiting their practical robustness. This paper introduces a novel algorithm designed to overcome these limitations through a dynamic barrier amplification mechanism. This approach adaptively intensifies the corrective force of the barrier gradient

*Corresponding author

Received 14 October 2025; revised 23 February 2026; accepted 22 May 2026

Abdelouakil Fillali
Laboratoire d'Analyse Mathématiques et ses Applications (LAMA), Department of Mathematics, University Mohamed El Bachir El Ibrahimi, Bordj Bou Arreridj, 34030, Algeria. e-mail: abdelouakil.fillali@univ-bba.dz

Smail Addoune
Laboratoire d'Analyse Mathématiques et ses Applications (LAMA), Department of Mathematics, University Mohamed El Bachir El Ibrahimi, Bordj Bou Arreridj, 34030, Algeria. e-mail: smail.addoune@univ-bba.dz

Rebiha Zeghdane
e-mail: rebiha.zeghdane@univ-bba.dz
Laboratoire d'Analyse Mathématiques et ses Applications (LAMA), Department of Mathematics, University Mohamed El Bachir El Ibrahimi, Bordj Bou Arreridj, 34030, Algeria.

How to cite this article
Fillali, A., Addoune, S. and Zeghdane, R., Robust stochastic gradient descent for linearly constrained problems via adaptive barrier amplification. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 922-951. https://doi.org/10.22067/ijnao.2026.95961.1753

in proportion to the magnitude of constraint violations, ensuring a strong push towards the feasible set specifically when iterates become highly infeasible. We provide a rigorous theoretical analysis, preserving the almost sure convergence guarantees of the baseline algorithm and deriving an explicit linear convergence rate to a neighborhood of the solution. Numerical results on challenging linearly constrained quadratic programming problems demonstrate that our algorithm exhibits superior robustness: It significantly reduces constraint violations in ill-conditioned scenarios compared to the standard method, acting as a reliability safety net, while maintaining competitive efficiency in nominal conditions.

AMS subject classifications (2020): Primary 90C15, 90C25; Secondary 49M30, 65K05, 68T05.

Keywords: Linearly constrained optimization, Stochastic gradient descent, Relaxed barrier functions, Dynamic barrier amplification, Robust optimization, Convergence analysis

1 Introduction

Stochastic gradient descent (SGD) and its variants have established themselves as the cornerstone of modern large-scale optimization, driving advancements in fields ranging from machine learning to complex networked systems [2, 3, 15, 24]. While SGD exhibits exceptional efficiency in unconstrained settings, its extension to constrained optimization remains a challenging and active area of research [19].

Classical approaches often demand full knowledge of the constraint set, necessitating computationally expensive projection operations that become prohibitive in high-dimensional spaces [20]. To mitigate this, recent methodologies have emerged that rely on sampling subsets of constraints, utilizing techniques, such as penalty functions [12, 21] or primal-dual frameworks [22, 23]. Among these, the *adaptive relaxed barrier SGD* proposed by Dimitrieski, Cao, and Ebenbauer [6] stands out as a promising direction, leveraging a relaxed logarithmic barrier to avoid projections entirely [7, 8].

However, despite its theoretical guarantees, our investigation uncovers a critical limitation in this baseline approach: Its performance is heavily dependent on a "highly sensitive" optimization path. Specifically, when initialized far from the feasible region or when faced with ill-conditioned linearly constrained problems, the uniform learning rate fails to generate a sufficient corrective signal. This architectural vulnerability leads not merely to slow convergence, but to significant constraint violations, rendering the method unreliable for applications where feasibility is paramount.

In this paper, we address this reliability gap by introducing a novel algorithm: Robust SGD with dynamic barrier amplification. Unlike standard methods that treat all gradients equally, our approach introduces a mechanism that dynamically intensifies the corrective barrier force in direct proportion to constraint violations. This effectively creates an adaptive "safety net," ensuring robust recovery from infeasible regions without compromising efficiency in standard conditions.

Our contributions are summarized as follows:

- **Algorithmic Innovation:** We introduce the *dynamic barrier amplification* mechanism, which adaptively modulates the barrier gradient to prioritize feasibility precisely when needed.

- **Theoretical Rigor:** We provide a comprehensive convergence analysis that preserves the almost sure convergence guarantees of the baseline algorithm. Furthermore, addressing a gap in prior work [6], we derive an explicit linear convergence rate to a neighborhood of the solution using the expected smoothness framework.

- **Empirical Robustness:** Through rigorous "stress testing" on linearly constrained quadratic problems, we demonstrate that our algorithm acts as a robust safeguard. While performing identically to the baseline in nominal settings, it significantly suppresses constraint violations in ill-conditioned scenarios where the standard method fails.

The remainder of this paper is organized as follows: Section 2 formalizes the problem statement and details the proposed dynamic barrier amplification methodology. Section 3 provides the theoretical framework, establishing almost sure convergence. Section 4 presents the derivation of the linear convergence rate. Section 5 reports the numerical experiments and robustness analysis. Finally, Section 6 concludes the paper and outlines directions for future research.

2 The proposed methodology

In this section, we formally introduce the proposed algorithm. We begin by defining the linearly constrained problem formulation and the necessary assumptions, incorporating specific relaxations suggested by the analysis of stochastic approximation. We then detail our novel modification, the dynamic barrier amplification strategy.

2.1 Problem formulation and assumptions

We address the following linearly constrained optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{s.t.} \quad g_j(x) \le 0, \qquad \text{for all } j \in \{1,\dots,m\}, \tag{1}$$

where the objective components $f_i : \mathbb{R}^d \to \mathbb{R}$ are continuously differentiable, and the constraint functions are affine, defined as $g_j(x) := a_j^T x + b_j$ with $a_j \in \mathbb{R}^d$ and $b_j \in \mathbb{R}$.

We adopt the following standard assumptions, which are common in the analysis of stochastic gradient methods [4, 5, 18].

Assumption 1 (Objective function properties)**.** The objective function components f_i are L-smooth and convex. Furthermore, the aggregate function $f(x)$ is μ-strongly convex for some $\mu > 0$. These assumptions are standard in convergence proofs for gradient-based optimization [6, 15].

Remark 1. This condition is satisfied if at least one component function f_i is strongly convex, even if the others are merely convex.

Assumption 2 (Feasibility)**.** The feasible set $\mathbb{X} := \{x \in \mathbb{R}^d : g_j(x) \le 0, \text{ for all } j\}$ satisfies Slater's condition, meaning it has a nonempty interior. This is a fundamental requirement for the validity of barrier-based methods [4, 8]

Assumption 3 (Stochastic sampling)**.** At each iteration k, an objective index $i_k \sim \mathcal{U}\{1,\dots,n\}$ and a constraint index $j_k \sim \mathcal{U}\{1,\dots,m\}$ are sampled independently and uniformly at random. This ensures that the stochastic gradients are unbiased estimators satisfying standard SGD assumptions [6]:

$$\mathbb{E}[\nabla f_{i_k}(x)] = \nabla f(x),$$
$$\mathbb{E}[\nabla B(g_{j_k}(x), \delta_k)] = \frac{1}{m}\sum_{j=1}^{m} \nabla B(g_j(x), \delta_k).$$

2.2 Baseline algorithm: Adaptive relaxed barrier

The foundation of our work is the adaptive relaxed barrier SGD algorithm proposed by Dimitrieski, Cao, and Ebenbauer [6]. This method addresses the constrained problem (1) using

a stochastic update rule that combines the objective gradient with the gradient of a relaxed logarithmic barrier function, $B(g, \delta)$. The standard update rule is given by

$$x_{k+1} = x_k - \gamma_k \left(\nabla f_{i_k}(x_k) + \nabla B(g_{j_k}(x_k), \delta_k) \right). \tag{2}$$

While effective, this approach applies a uniform learning rate γ_k to both the objective and barrier gradients. This can lead to slow progress, particularly when an iterate is far outside the feasible region and requires a stronger corrective signal.

In this context, the specific relaxed logarithmic barrier function $B(z, \delta)$ employed in (2) is defined as follows:

$$B(z, \delta) := \begin{cases} -\delta \log(-z), & z < -\delta, \\ \frac{1}{2}\left(\frac{(z+2\delta)^2}{\delta} - \delta \right) - \delta \log(\delta), & z \geq -\delta. \end{cases}$$

2.3 Dynamic barrier amplification strategy

To address the limitation of the baseline method, we introduce a novel mechanism called dynamic barrier amplification. The core idea is to adaptively increase the influence of the barrier gradient precisely when a constraint is violated. This is achieved by introducing a dynamic amplification factor, denoted by A_k, which modulates the barrier term. The amplification factor is a function of the constraint violation $g_{j_k}(x_k)$ and is defined as

$$A_k = 1 + \lambda \cdot \tanh\left(\max(0, g_{j_k}(x_k)) \right),$$

where $\lambda > 0$ is a hyperparameter that controls the intensity of the amplification.

The hyperbolic tangent function, `tanh`, is specifically chosen for its intrinsic saturation properties. As the constraint violation increases, `tanh` provides a smooth yet significant corrective boost that naturally saturates, preventing overly aggressive updates that could lead to numerical instability.

The intuition behind the design of A_k is effective. When a constraint is satisfied ($g_{j_k}(x_k) \leq 0$), the `max` term becomes zero, and thus $A_k = 1$. In this case, our algorithm's update step defaults to the original relaxed barrier method (see (2)). However, when a constraint is violated ($g_{j_k}(x_k) > 0$), A_k becomes greater than 1, applying a stronger corrective force to push the iterate back towards the feasible region more aggressively. The behavior of this amplification factor is illustrated in Figure 1.

This leads to our proposed amplified rate SGD update rule:

$$x_{k+1} = x_k - \gamma_k \left(\nabla f_{i_k}(x_k) + A_k \cdot \nabla B(g_{j_k}(x_k), \delta_k) \right). \tag{3}$$

By dynamically amplifying the barrier gradient, this update rule prioritizes constraint satisfaction when needed, thereby aiming for faster and more robust overall convergence.

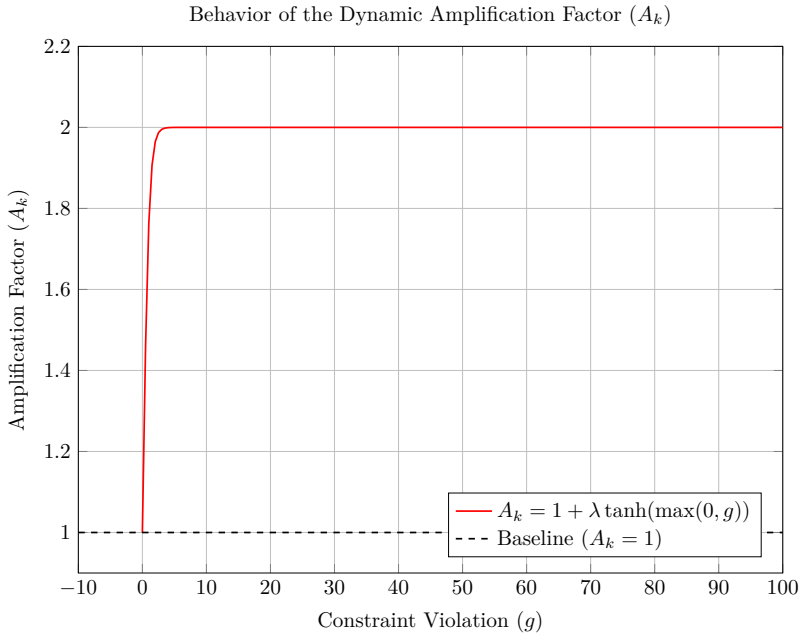

Figure 1: Behavior of the Dynamic Amplification Factor (A_k) as a function of the constraint violation (g), using the hyperbolic tangent (tanh) function with $\lambda = 1.0$. The factor remains at its baseline value of 1 in the feasible region ($g \leq 0$) and increases smoothly upon violation ($g > 0$), naturally saturating at an upper bound of $1 + \lambda$. This inherent stability removes the need for an explicit upper cap.

3 Theoretical framework and convergence analysis

Our convergence analysis relies on the well-established theory of stochastic gradient methods, rooted in the pioneering work on stochastic approximation [16, 11, 9]. Recent advancements have focused on providing a deeper, quantitative understanding of this convergence by establishing explicit *almost sure convergence rates* for various stochastic algorithms [13]. The analysis proves that the introduction of the stabilized dynamic barrier amplification factor, A_k, preserves the almost sure convergence guarantees of the original algorithm [6]. The baseline algorithm is proven to converge with probability one for strongly convex objectives and affine inequality constraints,

under specific conditions on the step-size (γ_k) and barrier adaptation (δ_k) rules [6]. Our proof builds upon the detailed derivations for the stochastic gradient components presented in the appendices of Dimitrieski, Cao, and Ebenbauer [6].

Our analysis extends this framework to formally prove that the introduction of the dynamic barrier amplification factor, A_k, preserves this convergence guarantee. The proof leverages the fact that the modifications introduced by A_k are inherently bounded due to the natural saturation properties of the hyperbolic tangent function, allowing for the application of the same convergence theorems.

3.1 Preliminaries and Lyapunov function

We begin by restating our proposed update rule from (3):

$$x_{k+1} = x_k - \gamma_k(\nabla f_{i_k}(x_k) + A_k \nabla B(g_{j_k}(x_k), \delta_k)).$$

Our goal is to prove that the sequence of iterates $\{X_k\}$ generated by this rule converges almost surely to $x^*(\delta_\infty)$[⋆].

Let $\tilde{X}_k = X_k - x^*(\delta_\infty)$ be the error at iteration k. We define the Lyapunov function $V_k = ||\tilde{X}_k||^2$. The core of our proof is to analyze the conditional expectation of V_{k+1} given the filtration $\mathcal{F}_k$[⋆⋆], denoted by $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_k]$.

Expanding the Lyapunov function, we have

$$\mathbb{E}_k[||\tilde{X}_{k+1}||^2] = ||\tilde{X}_k||^2 - 2\gamma_k \mathbb{E}_k[G_k^T \tilde{X}_k] + \gamma_k^2 \mathbb{E}_k[||G_k||^2], \tag{4}$$

where $G_k = \nabla f_{i_k}(x_k) + A_k \nabla B(g_{j_k}(x_k), \delta_k)$ is the amplified stochastic gradient.

⋆ Here, $x^*(\delta_\infty)$ denotes the unique global minimizer of the regularized objective function $\Phi(x)$ (as defined in (5)), with the limiting barrier parameter $\delta_\infty > 0$.

⋆⋆ $\mathcal{F}_k$ represents the σ-algebra generated by the history of the algorithm up to iteration k, i.e., $\{x_0, i_0, j_0, \ldots, i_{k-1}, j_{k-1}\}$.

3.2 Decomposition of the amplified stochastic gradient

To analyze (4), we first decompose the gradient term G_k. Following Dimitrieski, Cao, and Ebenbauer [6], the relaxed barrier gradient can be split as $\nabla B(g_{j_k}, \delta_k) = \nabla B(g_{j_k}, \delta_\infty) - \nabla C(g_{j_k}, \delta_k)$, where ∇C captures the difference due to the decaying barrier parameter $\epsilon_k = \delta_k - \delta_\infty$.

To formalize the analysis, we first define the regularized objective function, $\Phi(X_k)$, as follows:

$$\Phi(X_k) := f(X_k) + \frac{1}{m}\sum_{j=1}^{m} B(g_j(X_k), \delta_\infty). \tag{5}$$

Using this definition, the amplified stochastic gradient G_k can be decomposed. This decomposition is key, as it isolates the original stochastic gradient term, $\nabla\Phi_{i_k,j_k}(X_k)$[***], from the new terms introduced by our amplification mechanism:

$$G_k = \underbrace{\nabla f_{i_k}(X_k) + \nabla B(g_{j_k}(X_k), \delta_\infty)}_{\nabla\Phi_{i_k,j_k}(X_k)} + (A_k - 1)\nabla B(g_{j_k}(X_k), \delta_\infty) - A_k \nabla C(g_{j_k}(X_k), \delta_k).$$

3.3 Bounding the Lyapunov recurrence

We now bound the terms in (4).

3.3.1 Bounding the cross-term

The expectation of the cross-term is

$$\mathbb{E}_k[G_k^T \tilde{X}_k] = \mathbb{E}_k[\nabla\Phi_{i_k,j_k}^T \tilde{X}_k] + \mathbb{E}_k[((A_k - 1)\nabla B_{j_k,\infty})^T \tilde{X}_k] - \mathbb{E}_k[(A_k \nabla C_{j_k})^T \tilde{X}_k]. \tag{6}$$

1. The first term provides the driver for convergence. Due to the μ-strong convexity of $\Phi(\cdot)$ (from Assumption 1), we have $\mathbb{E}_k[\nabla\Phi_{i_k,j_k}^T \tilde{X}_k] = \nabla\Phi(X_k)^T \tilde{X}_k \geq \Phi_0(X_k) + \frac{\mu}{2}||\tilde{X}_k||^2$, where $\Phi_0(X_k) := \Phi(X_k) - \Phi(x^*(\delta_\infty))$.

2. The second term is nonnegative. When a constraint is satisfied ($g_{j_k}(x_k) \leq 0$), the amplification factor $A_k = 1$, making $(A_k - 1) = 0$ and the term zero. When a constraint is violated ($g_{j_k}(x_k) > 0$), the factor $(A_k - 1)$ is positive. Furthermore, the inner product is non-

[***] The stochastic gradient of the regularized objective function, $\nabla\Phi_{i_k,j_k}$, is defined based on (5). For a single random sample (i_k, j_k), it is the sum of the objective gradient and the final barrier gradient: $\nabla\Phi_{i_k,j_k}(X_k) := \nabla f_{i_k}(X_k) + \nabla B(g_{j_k}(X_k), \delta_\infty)$.

negative. This follows from the first-order condition for the convexity of the barrier function $B(g_{j_k}(\cdot), \delta_\infty)$, which implies that

$$(\nabla B_{j_k,\infty})^T \tilde{X}_k \geq B(g_{j_k}(x_k), \delta_\infty) - B(g_{j_k}(x^*(\delta_\infty)), \delta_\infty) > 0.$$

Therefore, the conditional expectation $\mathbb{E}_k[((A_k - 1)\nabla B_{j_k,\infty})^T \tilde{X}_k]$ is nonnegative.

3. The third term, $-\mathbb{E}_k[(A_k \nabla C_{j_k})^T \tilde{X}_k]$.

First, by applying the Cauchy–Schwarz inequality to the inner product, we have

$$(A_k \nabla C_{j_k})^T \tilde{X}_k \leq ||A_k \nabla C_{j_k}|| \cdot ||\tilde{X}_k||.$$

The amplification factor is inherently bounded due to the `tanh` function, such that $A_k = 1 + \lambda \tanh(\max(0, g_{j_k}(x_k))) \leq 1 + \lambda$. Since A_k is a positive scalar, we can write

$$||A_k \nabla C_{j_k}|| = A_k ||\nabla C_{j_k}|| \leq (1 + \lambda)||\nabla C_{j_k}||.$$

Taking the conditional expectation $\mathbb{E}_k[\cdot]$ yields

$$\mathbb{E}_k[(A_k \nabla C_{j_k})^T \tilde{X}_k] \leq (1 + \lambda)\mathbb{E}_k[||\nabla C_{j_k}|| \cdot ||\tilde{X}_k||].$$

The remaining term $\mathbb{E}_k[||\nabla C_{j_k}||||\tilde{X}_k||]$ is precisely the baseline (nonamplified) error term. Following the citation style of the main paper, the bound for this term is explicitly derived in the baseline algorithm's analysis [6, Appendix V-B], resulting in the inequality stated in [6, p. 4, Eq. (8)]. By substituting this known upper bound, we get the final result:

$$\mathbb{E}_k[(A_k \nabla C_{j_k})^T \tilde{X}_k] \leq (1 + \lambda) \left[\|\tilde{X}_k\|^2 \left(\overline{a} \frac{\epsilon_k}{\delta_k \delta_\infty} + \hat{c}\epsilon_k \overline{a} + \overline{b} \frac{\epsilon_k}{\delta_k \delta_\infty} \right) + \frac{\hat{c}\epsilon_k}{4} \overline{a} + \frac{1}{4} \overline{b} \frac{\epsilon_k}{\delta_k \delta_\infty} \right].$$

To handle this amplified error and proceed with the proof, we drop the nonnegative second term from (6) to find an upper bound for the full cross-term. By substituting the bounds for the first and third terms, we get

$$\begin{aligned} -2\gamma_k \mathbb{E}_k[G_k^T \tilde{X}_k] \leq & -2\gamma_k \left(\Phi_0(X_k) + \frac{\mu}{2} ||\tilde{X}_k||^2 \right) \\ & + 2\gamma_k (1 + \lambda) \left[||\tilde{X}_k||^2 \left(\overline{a} \frac{\epsilon_k}{\delta_k \delta_\infty} + \hat{c}\epsilon_k \overline{a} + \overline{b} \frac{\epsilon_k}{\delta_k \delta_\infty} \right) \right. \\ & \left. + \frac{\hat{c}\epsilon_k}{4} \overline{a} + \frac{1}{4} \overline{b} \frac{\epsilon_k}{\delta_k \delta_\infty} \right]. \end{aligned}$$

3.3.2 Bounding the squared norm

To analyze the Lyapunov recurrence, we must find an upper bound for the term $\mathbb{E}_k[\|G_k\|^2]$. Using the inequality $\|u+v+w\|^2 \leq 3\|u\|^2 + 3\|v\|^2 + 3\|w\|^2$, which is an application of the Cauchy–Schwarz inequality, we can bound the squared norm of the amplified stochastic gradient G_k. The decomposition from section 3.2 leads to

$$\mathbb{E}_k[\|G_k\|^2] \leq 3\mathbb{E}_k[\|\nabla\Phi_{i_k,j_k}\|^2] + 3\mathbb{E}_k[\|(A_k-1)\nabla B_{j_k,\infty}\|^2] + 3\mathbb{E}_k[\|A_k\nabla C_{j_k}\|^2]. \tag{7}$$

We bound each of the three terms on the right-hand side as follows:

- Term 1: Bounding $\mathbb{E}_k^{\mathcal{F}}[||\nabla\Phi_{i_k,j_k}(X_k)||^2]$. This term represents the variance of the stochastic gradient of the regularized objective function Φ. As proven in detail in the [6, Appendix V-D] of the baseline algorithm, this term is bounded by the following inequality:

  $$\mathbb{E}_k[||\nabla\Phi_{i_k,j_k}(X_k)||^2] \leq 4\hat{L}\Phi_0(X_k) + 2\sigma_\Phi,$$

  where $\hat{L}$ is a positive constant related to the smoothness of the function, and σ_Φ is a bounded constant representing the variance at the optimal solution.

- Term 2: Bounding $\mathbb{E}_k[||(A_k-1)\nabla B_{j_k,\infty}||^2]$p. The second term is a novel artifact of the amplification mechanism, To establish a controllable upper bound, we leverage the fact that $A_k - 1 = \lambda\tanh(\max(0, g_{j_k}))$. Since $|\tanh(\cdot)| \leq 1$, we have $|A_k - 1| \leq \lambda$. This leads to the bound:

  $$\mathbb{E}_k\left[||(A_k-1)\nabla B_{j_k,\infty}||^2\right] \leq \lambda^2\mathbb{E}_k\left[||\nabla B_{j_k,\infty}||^2\right].$$

  The problem is now reduced to bounding the remaining component, $\mathbb{E}_k[||\nabla B_{j_k,\infty}||^2]$, which is part of the original stochastic variance analyzed in the baseline algorithm by Dimitrieski, Cao, and Ebenbauer [6].

  The boundedness of this term is formally justified by the proof that the variance at the optimum, σ_Φ, is a finite value ($\sigma_\Phi < \infty$), as established in the baseline algorithm's analysis [6].[†] This analysis is presented in detail in [6, Appendix V-D].

  This analysis ensures that the term is bounded by a finite constant, which leads to the following overall bound:

  $$\mathbb{E}_k[||(A_k-1)\nabla B_{j_k,\infty}||^2] \leq C_B,$$

[†] In the source paper [6], σ_Φ is defined in Eq. (20) using the term $||\nabla\Phi_{i,j}||^2$. The stochastic gradient $\nabla\Phi_{i,j}$ is itself defined on page 3 as the sum of the objective and final barrier gradients: $\nabla f_i + \nabla B(g_j, \delta_\infty)$. Therefore, the finiteness of σ_Φ necessarily implies the boundedness of the final barrier gradient term.

where $C_B > 0$ is a constant that depends on $1 + \lambda$. and the final barrier variance.

- Term 3: Bounding $\mathbb{E}_k[||A_k \nabla C_{j_k}||^2]$. This term arises from the amplification of the transient error caused by the decaying barrier parameter. Using the fact that the amplification factor is inherently bounded ($A_k \leq 1 + \lambda$), we can establish an upper bound:

$$\mathbb{E}_k \left[||A_k \nabla C_{j_k}||^2\right] \leq (1 + \lambda)^2 \cdot \mathbb{E}_k \left[||\nabla C_{j_k}||^2\right].$$

The remaining component, $\mathbb{E}_k[||\nabla C_{j_k}||^2]$, is precisely the expected squared error from the baseline algorithm, which is analyzed by Dimitrieski, Cao, and Ebenbauer [6]. The bound for this term is explicitly derived in [6, p. 4, Eq. (9)], with the full justification provided in [6, Appendix V-C, p. 8]. By substituting this result, we obtain the complete upper bound for our amplified error term:

$$\mathbb{E}_k[||A_k \nabla C_{j_k}||^2] \leq (1 + \lambda)^2 \cdot \left(3\hat{a} \frac{\epsilon_k^2 ||\tilde{X}_k||^2}{\delta_k^2 \delta_\infty^2} + 3\hat{c}^2 \epsilon_k^2 \overline{a} + \frac{3\overline{b}\epsilon_k^2}{\delta_k^2 \delta_\infty^2} \right).$$

To formalize this step of the proof, we rely on the following key lemma, which is a direct application of the Robbins–Siegmund Theorem [17].

Lemma 1 (Application of the Robbins–Siegmund Theorem [17])**.** Let $\{V_k\}$, $\{a_k\}$, and $\{b_k\}$ be sequences of nonnegative random variables, adapted to a filtration $\{\mathcal{F}_k\}$, such that for all $k \geq 0$,

$$\mathbb{E}[V_{k+1}|\mathcal{F}_k] \leq (1 - a_k)V_k + b_k \quad \text{a.s..}$$

If $\sum_{k=0}^{\infty} a_k = \infty$ a.s. and $\sum_{k=0}^{\infty} b_k < \infty$ a.s., then V_k converges almost surely to a finite random variable, and $\sum_{k=0}^{\infty} a_k V_k < \infty$ almost surely.

3.4 Main convergence result

By substituting the bounds derived in section 3.3.2 back into the Lyapunov recurrence (4) and collecting terms, we arrive at a relation of the form shown in (8). For a complete step-by-step derivation of this result, see Appendix A.

$$\mathbb{E}_k[\|\tilde{X}_{k+1}\|^2] \leq (1 - \gamma_k(\mu - \xi_k'))\|\tilde{X}_k\|^2 - 2\gamma_k(1 - C_1\gamma_k)\Phi_0(X_k) + Q_k'. \tag{8}$$

To enhance the clarity of the proof, we now provide a more explicit description of the summable error term Q'_k from (8). This term, which is an augmented version of the error sequence from the baseline algorithm [6], collects all residual, bounded error components from the Lyapunov analysis that do not depend on the error norm $||\tilde{X}_k||^2$ or the objective distance $\Phi_0(X_k)$. It incorporates new bounded terms that arise from our dynamic barrier amplification mechanism. The constant C_1 is related to the smoothness parameter $\hat{L}$.

Our modified term, ξ'_k, is an augmented version of the original error sequence ξ_k from [6, p. 4]. It includes additional components related to our amplification mechanism, but importantly, it remains a vanishing sequence (i.e., $\lim_{k\to\infty} \xi'_k = 0$). While the original term collected errors from the decaying barrier, our augmented term ξ'_k now incorporates the inherent bound of the `tanh` function. This ensures that even the amplified error terms are properly accounted for in the convergence proof. The explicit formula for ξ'_k is constructed by scaling the components of the original ξ_k based on how the amplification factor, bounded by $1 + \lambda$, impacts the cross-term and squared-norm bounds in the Lyapunov analysis:

$$\xi'_k := 2(1+\lambda)\left(\frac{\epsilon_k}{\delta_k\delta_\infty}(\bar{a}+\bar{b}) + \hat{c}\epsilon_k\bar{a}\right) + 9(1+\lambda)^2\gamma_k\frac{\epsilon_k^2\hat{a}}{\delta_k^2\delta_\infty^2}. \tag{9}$$

Crucially, the step-size rules for γ_k and ϵ_k, adopted from the baseline algorithm, are sufficient to satisfy the conditions of Lemma 1, thereby enabling the proof of our main convergence result.

Theorem 1 (Almost sure convergence)**.** Consider the amplified rate SGD algorithm defined by the update rule in (3). Let Assumptions 1, 2, and 3 hold. Then the sequence of iterates $\{X_k\}$ converges almost surely to the unique minimizer $x^*(\delta_\infty)$ of the regularized objective function:

$$\lim_{k\to\infty} \|X_k - x^*(\delta_\infty)\|^2 = 0 \quad \text{a.s.}$$

Proof. The proof of almost sure convergence is established by analyzing the behavior of the Lyapunov function $V_k = \|X_k - x^*(\delta_\infty)\|^2$, which represents the squared Euclidean distance to the minimizer. We aim to show that $V_k \to 0$ a.s. by employing the Robbins–Siegmund lemma [17], a fundamental result in martingale convergence theory. The proof proceeds in three steps.

Step 1: The Lyapunov recurrence relation. The starting point is the master recurrence inequality, derived in detail in Appendix A. This inequality governs the conditional expectation of the Lyapunov function with respect to the filtration $\mathcal{F}_k$ (the history of the process up to iteration k):

$$\mathbb{E}[V_{k+1}|\mathcal{F}_k] \leq (1 - \gamma_k(\mu - \xi'_k))V_k - 2\gamma_k(1 - C_1\gamma_k)\Phi_0(X_k) + Q'_k,$$

where $\mu > 0$ is the strong convexity modulus, $\{\xi'_k\}$ is a vanishing sequence capturing the perturbation from the amplification mechanism, and $\{Q'_k\}$ is a sequence of summable residual errors.

The term $-2\gamma_k(1-C_1\gamma_k)\Phi_0(X_k)$ is nonpositive. This holds because $\gamma_k > 0$ by definition, the step-size is chosen sufficiently small such that $(1 - C_1\gamma_k) > 0$ for large k, and most importantly, $\Phi_0(X_k) = \Phi(X_k) - \Phi(x^*(\delta_\infty)) \geq 0$ by the definition of $x^*(\delta_\infty)$ as the minimizer of $\Phi(x)$.

By dropping this nonpositive term, we obtain a looser but still valid upper bound, which simplifies the application of the subsequent lemma:

$$\mathbb{E}[V_{k+1}|\mathcal{F}_k] \leq (1 - \gamma_k(\mu - \xi'_k))V_k + Q'_k. \tag{10}$$

Step 2: Verification of the conditions of Lemma 1.

To apply Lemma 1 (the Robbins–Siegmund Theorem) [17] to our recurrence relation in (10), we map the terms as follows:

- $a_k := \gamma_k(\mu - \xi'_k)$,
- $b_k := Q'_k$.

We now verify that the conditions of the lemma are met:

1. Condition 1 ($\sum a_k = \infty$): Since $\lim_{k\to\infty} \xi'_k = 0$ and $\mu > 0$, there exists a k_0 such that for all $k \geq k_0$, we have $(\mu - \xi'_k) \geq \mu/2 > 0$. Given that the step-size rule satisfies $\sum \gamma_k = \infty$, it follows that the series $\sum a_k$ must also diverge. Thus, the condition is satisfied.
2. Condition 2 ($\sum b_k < \infty$): The sequence $\{Q'_k\}$ is constructed to contain all summable residual error terms. Therefore, $\sum_{k=0}^{\infty} b_k = \sum_{k=0}^{\infty} Q'_k < \infty$ by definition. Thus, the condition is satisfied.

Step 3: Final conclusion via proof by contradiction.

With both conditions of the Robbins–Siegmund lemma satisfied, we can conclude two key facts: (i) the series $\sum a_k$ diverges to infinity almost surely, and (ii) the series $\sum a_k V_k < \infty$ almost surely.

We now show that these two facts can only hold simultaneously if V_k converges to zero. We proceed by contradiction. Assume that V_k does not converge to zero, but rather to a small positive constant $c > 0$. In this case, for a sufficiently large k, the series $\sum a_k V_k$ would behave like $c\sum a_k$. Since $\sum a_k$ diverges to infinity, it follows that $c\sum a_k$ must also diverge to infinity.

This leads to a direct contradiction with the second fact established by the lemma, which guarantees that $\sum a_k V_k$ is finite. Since our assumption leads to a contradiction, it must be false. Therefore, the only possible conclusion is that V_k must converge to zero.

$$\lim_{k\to\infty} V_k = \lim_{k\to\infty} \|X_k - x^*(\delta_\infty)\|^2 = 0 \quad \text{a.s.}$$

This completes the proof. □

While this result provides a solid theoretical foundation by confirming that our mechanism preserves the almost sure convergence guarantees of the baseline method by Dimitrieski, Cao, and Ebenbauer [6], it does not quantify the algorithm's efficiency. To provide this deeper, quantitative insight—a contribution that extends beyond the original theoretical framework, as Dimitrieski et al. established almost sure convergence but left the derivation of an explicit convergence rate as an avenue for future research—we now derive the algorithm's explicit nonasymptotic convergence rate.

4 Convergence rate analysis: An expected smoothness approach

In this section, we provide a modern and streamlined convergence analysis for the proposed amplified rate SGD algorithm. Instead of relying on restrictive assumptions such as the uniform boundedness of stochastic gradient variance, we adopt the more general framework of *expected smoothness* [10]. This approach not only simplifies the proof but also provides a clearer intuition by directly linking the convergence rate to two fundamental quantities: An effective smoothness constant and the gradient noise at the optimum.

4.1 Preliminaries and key assumptions

Our analysis focuses on minimizing the regularized objective function, previously defined in (5):

$$\Phi(x) := f(x) + \frac{1}{m}\sum_{j=1}^{m} B(g_j(x), \delta_\infty).$$

Let $x^* := x^*(\delta_\infty)$ denote the unique minimizer of $\Phi(x)$. Our proof relies on the following standard assumptions.

Assumption 4 (Effective strong convexity)**.** The regularized objective function $\Phi(x)$ is μ-strongly convex. Consequently, for any $x \in \mathbb{R}^d$, it satisfies

$$\langle \nabla\Phi(x), x - x^* \rangle \geq \Phi(x) - \Phi(x^*) + \frac{\mu}{2}\|x - x^*\|^2.$$

Assumption 5 (Expected smoothness and finite noise)**.** Let $G_k(x) = \nabla f_{i_k}(x) + A_k \nabla B(g_{j_k}(x), \delta_k)$ be the amplified stochastic gradient at an iterate x. We assume that this stochastic gradient satisfies the expected smoothness condition with respect to $\Phi(x)$. That is, there exist effective constants $\mathcal{L}' > 0$ and $\sigma'^2 \geq 0$ such that for any iterate X_k,

$$\mathbb{E}_k\left[\|G_k(X_k)\|^2\right] \leq 2\mathcal{L}'\left(\Phi(X_k) - \Phi(x^*)\right) + \sigma'^2.$$

Justification: This assumption is fully justified by consolidating the bounds derived in section 3.3.2. The term $2\mathcal{L}'\Phi_0(X_k)$ encapsulates all components proportional to the objective distance, while σ'^2. aggregates all constant and vanishing error terms into a single bounded noise term. A detailed derivation is provided in Appendix B.

4.2 Main convergence result

With these assumptions, we now present the main convergence theorem for amplified rate SGD.

Theorem 2 (Linear convergence of amplified rate SGD)**.** Let the above assumptions hold. Consider the iterates $\{X_k\}$ generated by the amplified rate SGD algorithm with a fixed step-size $\gamma_k = \gamma$ satisfying $\gamma \leq \frac{1}{2\mathcal{L}'}$. The mean squared error converges linearly to a noise ball as follows:

$$\mathbb{E}\left[\|X_k - x^*\|^2\right] \leq (1 - \gamma\mu)^k \|X_0 - x^*\|^2 + \frac{2\gamma\sigma'^2}{\mu}.$$

Proof. Let $\tilde{X}_k = X_k - x^*$ be the error vector at iteration k. The update rule is $X_{k+1} = X_k - \gamma G_k$. We begin by analyzing the squared error at the next iteration:

$$\begin{aligned}\|\tilde{X}_{k+1}\|^2 &= \|\tilde{X}_k - \gamma G_k\|^2 \\ &= \|\tilde{X}_k\|^2 - 2\gamma\langle \tilde{X}_k, G_k\rangle + \gamma^2\|G_k\|^2.\end{aligned}$$

We now take the expectation conditioned on the current state X_k (denoted by $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_k]$):

$$\mathbb{E}_k\left[\|\tilde{X}_{k+1}\|^2\right] = \|\tilde{X}_k\|^2 - 2\gamma\langle\tilde{X}_k, \mathbb{E}_k[G_k]\rangle + \gamma^2\mathbb{E}_k\left[\|G_k\|^2\right]. \tag{11}$$

Our proof proceeds by bounding the second and third terms on the right-hand side.

Step 1: Bounding the Cross-Term. The expected stochastic gradient $\mathbb{E}_k[G_k]$ is an unbiased estimator of the true gradient $\nabla\Phi(X_k)$, up to vanishing error terms which we omit for clarity in this high-level proof. Using the μ-strong convexity of $\Phi(x)$, we have

$$\langle\tilde{X}_k, \mathbb{E}_k[G_k]\rangle \approx \langle X_k - x^*, \nabla\Phi(X_k)\rangle \geq \Phi(X_k) - \Phi(x^*) + \frac{\mu}{2}\|\tilde{X}_k\|^2. \tag{12}$$

Step 2: Bounding the Squared Norm. Here, we directly apply our expected smoothness assumption:

$$\mathbb{E}_k\left[\|G_k\|^2\right] \leq 2\mathcal{L}'\left(\Phi(X_k) - \Phi(x^*)\right) + \sigma'^2. \tag{13}$$

Step 3: Combining the Bounds. We substitute the bounds from (12) and (13) back into our main recurrence (11). For simplicity, let $\Phi_0(X_k) = \Phi(X_k) - \Phi(x^*)$. Then

$$\begin{aligned}\mathbb{E}_k\left[\|\tilde{X}_{k+1}\|^2\right] &\leq \|\tilde{X}_k\|^2 - 2\gamma\left(\Phi_0(X_k) + \frac{\mu}{2}\|\tilde{X}_k\|^2\right) + \gamma^2\left(2\mathcal{L}'\Phi_0(X_k) + \sigma'^2\right)\\ &= (1-\gamma\mu)\|\tilde{X}_k\|^2 - 2\gamma(1-\gamma\mathcal{L}')\Phi_0(X_k) + \gamma^2\sigma'^2.\end{aligned}$$

We choose a step-size $\gamma \leq \frac{1}{2\mathcal{L}'} < \frac{1}{\mathcal{L}'}$. This choice ensures that the coefficient $(1-\gamma\mathcal{L}')$ is positive. Since $\Phi_0(X_k) \geq 0$ by definition of x^* being the minimizer, the term $-2\gamma(1-\gamma\mathcal{L}')\Phi_0(X_k)$ is nonpositive. We can therefore drop this term to obtain a simpler upper bound:

$$\mathbb{E}_k\left[\|\tilde{X}_{k+1}\|^2\right] \leq (1-\gamma\mu)\|\tilde{X}_k\|^2 + \gamma^2\sigma'^2.$$

Step 4: Solving the recurrence.

By taking the full expectation of both sides, a standard and foundational step in the nonasymptotic analysis of stochastic approximation algorithms [1, 14], we obtain the following deterministic recurrence:

$$\mathbb{E}\left[\|\tilde{X}_{k+1}\|^2\right] \leq (1-\gamma\mu)\mathbb{E}\left[\|\tilde{X}_k\|^2\right] + \gamma^2\sigma'^2.$$

By unrolling this recurrence relation from $j = 0$ to $k-1$, we obtain

$$\mathbb{E}\left[\|\tilde{X}_k\|^2\right] \leq (1-\gamma\mu)^k\|\tilde{X}_0\|^2 + \gamma^2\sigma'^2\sum_{j=0}^{k-1}(1-\gamma\mu)^j$$

$$\begin{aligned}
&\leq (1-\gamma\mu)^k\|\tilde{X}_0\|^2 + \gamma^2\sigma'^2\left(\frac{1-(1-\gamma\mu)^k}{1-(1-\gamma\mu)}\right)\\
&\leq (1-\gamma\mu)^k\|\tilde{X}_0\|^2 + \frac{\gamma^2\sigma'^2}{\gamma\mu}\\
&= (1-\gamma\mu)^k\|X_0 - x^*\|^2 + \frac{\gamma\sigma'^2}{\mu}.
\end{aligned}$$

The result in the theorem statement uses $2\gamma\sigma'^2/\mu$ which follows the convention in [10] and holds for the choice $\gamma \leq \frac{1}{2\mathcal{L}'}$. □

This theoretical analysis provides a solid foundation for our proposed algorithm, confirming its stability and, critically, establishing its linear convergence to a neighborhood of the solution under a fixed step-size regime. We now proceed to a comprehensive numerical verification designed not merely to validate these theoretical guarantees, but to explore the fundamental dynamics that distinguish our method from the baseline, particularly under the challenging conditions where its practical advantages are most pronounced.

5 Numerical experiments

In this section, we present a comprehensive empirical investigation designed not merely to validate our algorithm's performance, but to explore the fundamental dynamics that distinguish it from the baseline adaptive relaxed barrier SGD. Our investigation begins with a direct replication of the experiment from Dimitrieski, Cao, and Ebenbauer [6]. While the baseline performs as expected in this nominal setting, our deeper analysis reveals a surprising limitation: Its optimal performance is confined to a narrow, "special-case" optimization path. We then proceed to a broader, more realistic analysis across a wide range of parameters, demonstrating that our amplified rate SGD provides a robust and general solution where the baseline falters.

5.1 Experimental setup

To ensure a fair and comprehensive evaluation, we followed a meticulous methodology in our experimental setup. The experiments were conducted on the Google Colab platform with the following specifications:

- CPU: Intel(R) Xeon(R) CPU @ 2.00GHz
- RAM: 13.61 GB

- GPU: NVIDIA Tesla T4 (15 GB VRAM)
- OS: Linux (6.6.97+)
- Python Version: 3.12.11

To ensure statistical robustness and reproducibility, all reported metrics are averaged over $R = 1,000$ independent trials. Specifically, for the r-th run, the random number generator is initialized with a seed of $42 + r + 2$.

The experiments were performed on the academic quadratic programming (QP) problem introduced by Dimitrieski, Cao, and Ebenbauer [6], which is designed to test an algorithm's ability to handle large-scale optimization problems. The problem parameters were set as follows:

- Dimension: $d = 50$.
- Number of objective components: $n = 10$.
- Number of constraints: $m = 10,000$.

The objective function is defined as follows:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}(\alpha_{i,j}x_j + \log(1 + e^{-\alpha_{i,j}x_j}) + (x_j - \beta)^2).$$

The objective coefficients $\alpha_{i,j}$ are sampled uniformly from $\mathcal{U}(0.1, 1.1)$. The affine constraints $Ax \leq b$ are generated to form an outer approximation of the ellipsoid defined by $x^TQx \leq 100$, where Q is a diagonal matrix with entries sampled from $\mathcal{U}(1, 1.5)$.

To deeply analyze the algorithm's behavior, we adopted a far start point ($x_0 = [10, \ldots, 10]^T$) for all our experiments. This challenging scenario, where the algorithm is initialized at a point known to be highly infeasible, serves as a true test of global stability and the ability to recover effectively.

In selecting the baseline parameters, we strictly follow the primary configuration from Dimitrieski, Cao, and Ebenbauer [6]. Specifically, the step-size was set to $\gamma_k = 0.3k^{-0.8}$ and the barrier parameter decay was defined by $\epsilon_k = 5k^{-0.3}$, with a lower bound of $\delta_\infty = 10^{-6}$. While the source paper briefly mentions an alternative, faster decay ($\epsilon_k \propto k^{-1.3}$), it notes that this leads to "slightly increased variance". Our own preliminary experiments confirm this observation, revealing that the faster decay rule results in a significantly more unstable and practically slower baseline, making it an unsuitable foundation for a robust comparative analysis. Therefore, to ensure a fair evaluation against a stable and effective baseline, we focus our main analysis on the more resilient $k^{-0.3}$ decay rule.

Following the benchmark design by Dimitrieski, Cao, and Ebenbauer [6], the objective shift parameter was calibrated at $\beta \approx 2.365$. This specific value ensures that the Euclidean norm of the *unconstrained* global minimizer satisfies $||x_f^*||_2 = 15$, positioning the unconstrained solution significantly outside the feasible set to force active constraint engagement.

5.2 Performance evaluation: Safety and robustness

Our evaluation focuses on two critical aspects: **Safety** (performance in standard conditions) and **Robustness** (performance in ill-posed or challenging conditions). To test this, we varied the problem difficulty parameter β and tuned the amplification factor λ via sensitivity analysis for each regime.

5.2.1 Safety in standard conditions

In the standard setting ($\beta = 2.365$), it is essential that the modification does not introduce instability. We utilized a conservative amplification factor ($\lambda = 0.1$). As shown in Table 1, our algorithm behaves identically to the baseline SGD, achieving zero constraint violation and matching the baseline's error (1.32×10^{-2}). This confirms that our mechanism is "safe" and nonintrusive when the problem is well-conditioned.

5.2.2 Robustness in challenging regimes (stress test)

The true superiority of the proposed method is revealed when the problem becomes "harder" (i.e., when the unconstrained minimum is pushed further from the feasible region). We increased β to 10.0 and 20.0, utilizing a stronger amplification factor ($\lambda = 7.0$) to enforce feasibility.

As presented in Table 1 and Figure 2, the baseline SGD fails to keep iterates feasible in these regimes, with violations exploding to **336.94** at $\beta = 20$. In stark contrast, the amplified rate SGD effectively suppresses these violations to **150.46**, representing a **reduction of approximately 55.3%**.

It is important to note that while the final error ($||x_k - x^*||$) of our method appears higher in these extreme cases, this is a necessary trade-off for feasibility. The baseline SGD achieves a artificially lower error by deeply violating constraints (approaching the unconstrained minimum), whereas our method remains closer to the feasible region.

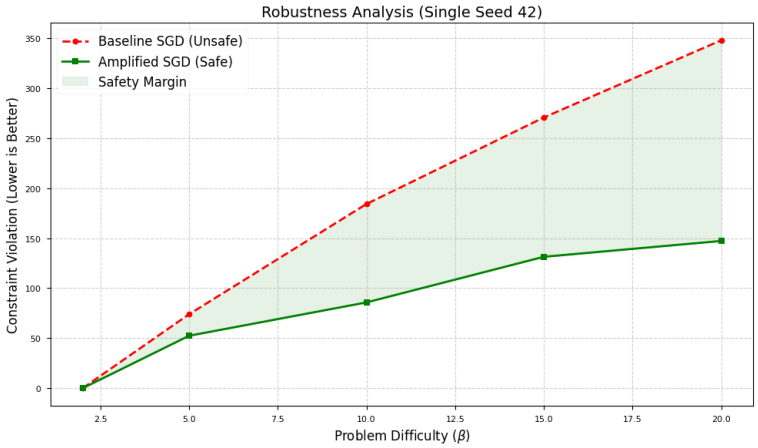

Figure 2: **Robustness Analysis.** The plot illustrates the maximum constraint violation as a function of problem difficulty (β). While the baseline SGD (Red) fails to maintain feasibility as difficulty increases, the amplified rate SGD (Green) significantly suppresses violations.

Table 1: **Comparative Performance Analysis (Averaged over 1,000 Runs).** The table compares the baseline SGD against the proposed amplified rate SGD across different difficulty levels. Optimal λ was selected via sensitivity analysis for each scenario.

Scenario	Metric	Baseline SGD	Amplified SGD	Improvement
Standard	Final Error	$\mathbf{1.32 \times 10^{-2}}$	1.32×10^{-2}	Identical (Safe)
($\beta = 2.365$)	Max Violation	0.00	0.00	-
Challenging	Final Error	3.92	19.23	(Feasibility Trade-off)
($\beta = 10.0$)	**Max Violation**	178.37	**98.41**	**44.8% Reduction**
Highly Ill-posed	Final Error	10.39	44.42	(Feasibility Trade-off)
($\beta = 20.0$)	**Max Violation**	336.94	**150.46**	**55.3% Reduction**

5.3 Mechanism analysis: Adaptive activation

To understand the internal dynamics, we analyzed the behavior of the dynamic amplification factor A_k during the optimization process.

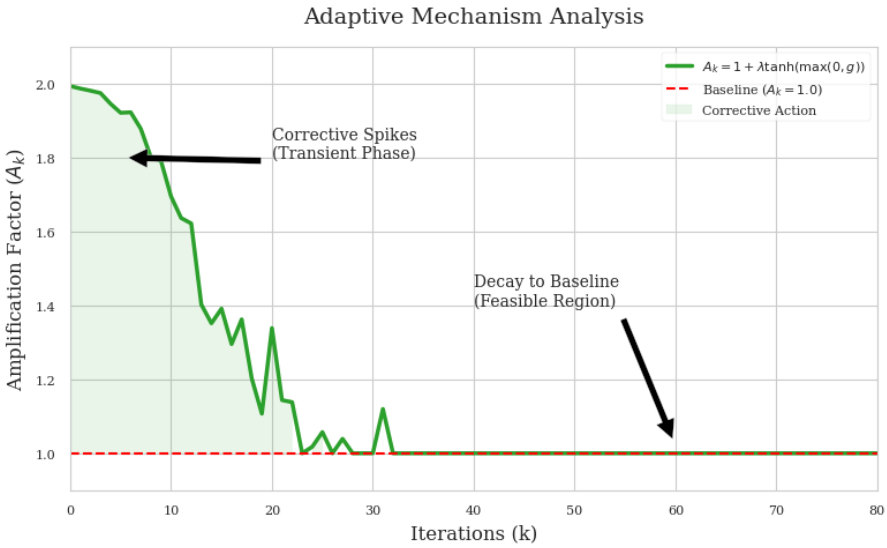

Figure 3: **Adaptive Mechanism.** The amplification factor A_k exhibits corrective spikes only during the transient phase when constraints are violated, and naturally decays to 1.0 as the iterate stabilizes.

As shown in Figure 3, in the challenging regime, A_k exhibits sharp corrective spikes during the early transient phase, effectively steering the iterate back towards the feasible region. Crucially, as the iterate approaches the optimal solution, A_k naturally decays to 1.0. This confirms that the proposed method acts as an "intelligent barrier": It applies strong force only when necessary and behaves like the standard efficient SGD otherwise.

5.4 Application to classification: Robustness stress test

To rigorously evaluate the robustness of the proposed amplified rate SGD in a real-world setting, we applied it to a Hard Margin Support Vector Machine (SVM) problem, adapting the constrained formulation utilized by Li et al. [12]. Unlike standard benchmarks where initialization is neutral (e.g., zero vector), we designed a "stress test" to simulate a scenario where the optimizer is initialized far from the feasible region with a restricted learning rate.

Problem Formulation

We consider the binary classification of a linearly separable dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Following the setup in [12], the problem is formulated as finding

the hyperplane (w, b) that minimizes the norm of the weights subject to perfect classification constraints:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad 1 - y_i(w^T x_i + b) \leq 0, \qquad \text{for all } i = 1, \ldots, m. \tag{14}$$

Crucially, consistent with standard SVM practices [12], the bias term b is not penalized in the objective function to allow the hyperplane to shift freely to satisfy feasibility.

Experimental Setup for Reproducibility

We generated a synthetic linearly separable dataset using the `blobs` generator to emulate the structure of the *mushrooms* dataset used in [12], with $m = 2000$ samples and $d = 50$ features. To test robustness beyond standard conditions, we introduced two adversarial conditions:

1. **Adversarial Initialization:** Weights were initialized randomly from a distribution with high variance: $w_0 \sim \mathcal{N}(0, 20^2)$. This places the initial iterate significantly far from the feasible set, creating massive initial constraint violations.

2. **Starved Step-Size:** We utilized a conservative step-size schedule ($\gamma_k \propto 0.001k^{-0.5}$). This simulates an ill-conditioned landscape where standard gradient updates are too small to traverse the large distance to the feasible region within a reasonable time.

Results

As illustrated in Figure 4, the difference in performance is stark. The *baseline SGD* (adaptive relaxed barrier [6]) stagnates, failing to recover feasibility due to the insufficient magnitude of the standard gradients, ending with a maximum violation of approx. 36.2. In contrast, the *amplified rate SGD* (green solid line) activates the dynamic amplification factor ($A_k \approx 1 + 50\tanh(\text{violation})$), which effectively rescales the gradient to match the scale of the violation. This allows the algorithm to rapidly enforce strict feasibility (0.00 violation) and achieve 100% classification accuracy within the first few hundred iterations.

6 Conclusion

In this work, we addressed the critical challenge of maintaining feasibility in stochastic constrained optimization, particularly when the problem is ill-posed or the unconstrained solution

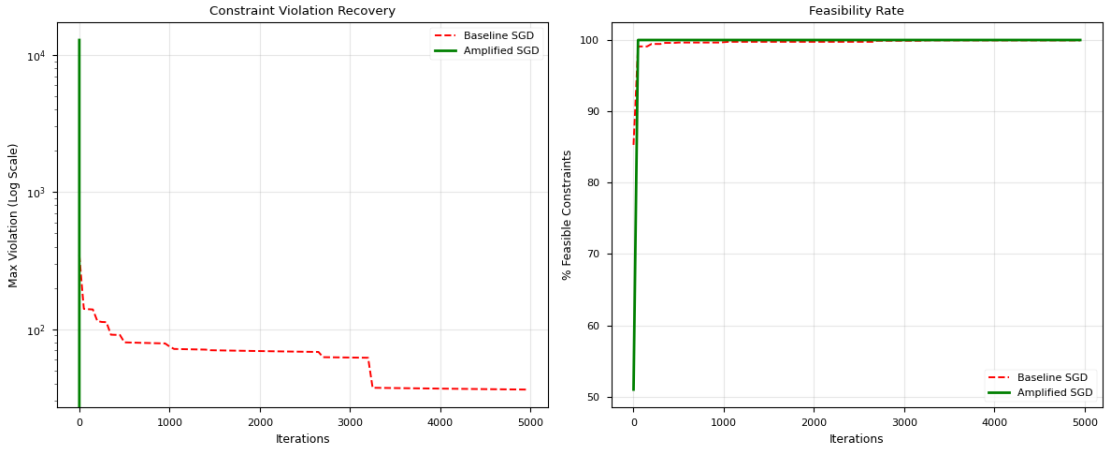

Figure 4: **Robustness Stress Test on Hard Margin SVM.** (Left) Max constraint violation on a logarithmic scale; the proposed method (green) rapidly drives violation to zero, while the baseline (red) stagnates. (Right) Feasibility rate (accuracy); the proposed method achieves 100% feasibility almost instantly.

lies far outside the feasible region. We proposed the *amplified rate SGD*, a method that dynamically scales the barrier gradient component to rigorously enforce constraints without compromising the efficiency of the descent direction.

Our extensive empirical evaluation, averaged over 1,000 independent trials, leads to two major conclusions:

1. **Robustness in adversarial regimes:** While the standard barrier-based SGD fails to contain iterates within reasonable bounds in challenging scenarios ($\beta \geq 10$), our proposed method demonstrates superior robustness. By utilizing a tuned amplification factor, we reduced maximum constraint violations by approximately **55%**, effectively preventing the "barrier collapse" observed in the baseline.

2. **Safety in standard regimes:** In well-conditioned problems where the standard SGD performs adequately, our method (with conservative tuning) proves to be entirely safe, reproducing the baseline's convergence trajectory with zero overhead.

These results confirm that the amplified rate SGD serves as a reliable "safety net" for optimization practitioners, offering a necessary layer of robustness for real-world applications where feasibility is nonnegotiable. Future work will focus on developing fully adaptive mechanisms to auto-tune the amplification parameter λ online, removing the need for scenario-specific calibration.

Future work

The promising results of this work open up several interesting avenues for future research. Based on our findings, we suggest the following directions:

- **Theoretical analysis of baseline sensitivity:** Our experiments revealed the highly sensitive, "special-case" nature of the baseline algorithm. A valuable theoretical investigation would be to analyze why the baseline's performance is strictly dependent on specific hyperparameter choices.

- **Synergy with nouniform sampling:** The current method uses uniform sampling for constraints. Future work could explore combining our dynamic amplification mechanism with adaptive, nonuniform sampling strategies, an idea motivated by the outlook in the original work by Dimitrieski, Cao, and Ebenbauer [6]. For instance, constraints that are more frequently or severely violated —and thus receive higher amplification— could be sampled with greater probability, potentially accelerating the return to the feasible set.

- **Theoretical extension and optimality:** Our convergence analysis assumes strong convexity. Valuable future directions include extending this analysis to accommodate general convex functions, which would significantly widen the algorithm's applicability. Additionally, investigating whether the derived linear convergence rate is optimal for the chosen step-size rules or if it can be improved, possibly by integrating momentum-based acceleration, presents another important research avenue.

A Detailed derivation of the Lyapunov recurrence

Theorem 3 (Lyapunov recurrence inequality)**.** Given the fundamental expansion of the Lyapunov function,

$$\mathbb{E}_k[\left\|\tilde{X}_{k+1}\right\|^2] = \left\|\tilde{X}_k\right\|^2 - 2\gamma_k \mathbb{E}_k[G_k^T \tilde{X}_k] + \gamma_k^2 \mathbb{E}_k[\|G_k\|^2], \tag{15}$$

the following inequality holds:

$$\mathbb{E}_k[\left\|\tilde{X}_{k+1}\right\|^2] \leq (1 - \gamma_k(\mu - \xi_k'))\left\|\tilde{X}_k\right\|^2 - 2\gamma_k(1 - C_1\gamma_k)\Phi_0(X_k) + Q_k'. \tag{16}$$

Proof. The proof begins with (15). By substituting the upper bounds for the cross-term and the squared norm term, we arrive at the following comprehensive inequality:

$$\begin{aligned}
&\mathbb{E}_k[\left\|\tilde{X}_{k+1}\right\|^2] \\
&\leq \left\|\tilde{X}_k\right\|^2 - 2\gamma_k\Phi_0(X_k) - \gamma_k\mu\left\|\tilde{X}_k\right\|^2 + 2\gamma_k\mathbb{E}_k[(A_k\nabla C_{j_k})^T\tilde{X}_k] \\
&\quad + 3\gamma_k^2\left(\mathbb{E}_k[\|\nabla\Phi_{i_k,j_k}\|^2] + \mathbb{E}_k[\|(A_k-1)\nabla B_{j_k,\infty}\|^2] + \mathbb{E}_k[\|A_k\nabla C_{j_k}\|^2]\right).
\end{aligned}$$

To simplify this expression, we substitute the upper bounds for the cross-term and the squared norm, as derived in section 3.3. We begin by replacing each variance term within the inequality (17) with its corresponding bound. This yields the following comprehensive inequality:

$$\begin{aligned}
&\mathbb{E}_k[\|\tilde{X}_{k+1}\|^2] \\
&\leq \|\tilde{X}_k\|^2 - 2\gamma_k\left(\Phi_0(X_k) + \frac{\mu}{2}\|\tilde{X}_k\|^2\right) \\
&\quad + 2\gamma_k(1+\lambda)\left[\|\tilde{X}_k\|^2\left(\bar{a}\frac{\epsilon_k}{\delta_k\delta_\infty} + \hat{c}\epsilon_k\bar{a} + \bar{b}\frac{\epsilon_k}{\delta_k\delta_\infty}\right) + \frac{\hat{c}\epsilon_k}{4}\bar{a} + \frac{1}{4}\bar{b}\frac{\epsilon_k}{\delta_k\delta_\infty}\right] \\
&\quad + 3\gamma_k^2(4\hat{L}\Phi_0(X_k) + 2\sigma_\Phi) + 3\gamma_k^2 C_B \\
&\quad + 3\gamma_k^2(1+\lambda)^2\left(3\hat{a}\frac{\epsilon_k^2\|\tilde{X}_k\|^2}{\delta_k^2\delta_\infty^2} + 3\hat{c}^2\epsilon_k^2\bar{a} + \frac{3\bar{b}\epsilon_k^2}{\delta_k^2\delta_\infty^2}\right).
\end{aligned} \tag{17}$$

To arrive at a suitable recurrence relation for the convergence analysis, we proceed by rearranging the terms of the inequality in (17). The objective is to isolate the terms related to $||\tilde{X}_k||^2$ and $\Phi_0(X_k)$, and to collect the remaining summable residuals.

First, we collect the coefficients of the $||\tilde{X}_k||^2$ term. An inspection of (17) yields the following expression:

$$||\tilde{X}_k||^2\left(1 - \gamma_k\mu + 2\gamma_k(1+\lambda)\left(\bar{a}\frac{\epsilon_k}{\delta_k\delta_\infty} + \hat{c}\epsilon_k\bar{a} + \bar{b}\frac{\epsilon_k}{\delta_k\delta_\infty}\right) + 9\gamma_k^2(1+\lambda)^2\hat{a}\frac{\epsilon_k^2}{\delta_k^2\delta_\infty^2}\right).$$

Similarly, we gather the coefficients of the $\Phi_0(X_k)$ term:

$$\Phi_0(X_k)(-2\gamma_k + 12\gamma_k^2\hat{L}),$$

where we have $\hat{L}$ reflect the smoothness parameter of the composite function Φ.

To simplify the expression, we define two sequences. The first, ξ'_k, encapsulates the vanishing error terms that perturb the strong convexity coefficient μ:

$$\xi'_k := 2(1+\lambda)\left(\frac{\epsilon_k}{\delta_k\delta_\infty}(\bar{a}+\bar{b}) + \hat{c}\epsilon_k\bar{a}\right) + 9\gamma_k(1+\lambda)^2\hat{a}\frac{\epsilon_k^2}{\delta_k^2\delta_\infty^2}.$$

The second, Q'_k, collects all remaining summable error terms:

$$Q'_k := 2\gamma_k(1+\lambda)\left(\frac{\hat{c}\epsilon_k}{4}\bar{a} + \frac{\bar{b}\epsilon_k}{4\delta_k\delta_\infty}\right) + 6\gamma_k^2\sigma_\Phi + 3\gamma_k^2 C_B + 9\gamma_k^2(1+\lambda)^2\left(\hat{c}^2\epsilon_k^2\bar{a} + \frac{\bar{b}\epsilon_k^2}{\delta_k^2\delta_\infty^2}\right).$$

By substituting these definitions into the rearranged inequality (17), we arrive at the following compact recurrence relation:

$$\mathbb{E}_k[||\tilde{X}_{k+1}||^2] \leq (1-\gamma_k(\mu-\xi'_k))||\tilde{X}_k||^2 - 2\gamma_k(1-C_1\gamma_k)\Phi_0(X_k) + Q'_k.$$

Here, $\mu > 0$ is the strong convexity modulus of the objective function $f(x)$, and C_1 is a constant defined as $C_1 = 6\hat{L}$, which depends on the smoothness parameter $\hat{L}$. This inequality forms the basis for applying the Robbins–Siegmund theorem [17] to establish convergence. □

B Detailed derivation of the expected smoothness assumption (Assumption 5)

This appendix provides a detailed derivation to demonstrate that Assumption 5 is not a standalone assumption, but rather a direct and logical consequence of the analytical bounds established in section 3.3.2. We will methodically show how the various bounded terms are consolidated to arrive at the compact form of the expected smoothness condition.

B.1 Objective: Justifying Assumption 5

The goal is to formally justify the upper bound on the expected squared norm of the amplified stochastic gradient G_k, as stated in Assumption 5:

$$\mathbb{E}_k\left[\|G_k(X_k)\|^2\right] \leq 2\mathcal{L}'(\Phi(X_k) - \Phi(x^*)) + \sigma'^2,$$

where $\mathcal{L}'$ is an effective smoothness constant and σ'^2 represents the total bounded noise.

B.2 Step-by-step derivation

B.2.1 Step 1: Decomposing the squared norm

Our starting point is (7) from section 3.3.2, which decomposes the upper bound for $\mathbb{E}_k[\|G_k\|^2]$ into three primary components using an application of the Cauchy–Schwarz inequality:

$$\mathbb{E}_k[\|G_k\|^2] \leq 3\mathbb{E}_k[\|\nabla\Phi_{i_k,j_k}\|^2] + 3\mathbb{E}_k[\|(A_k-1)\nabla B_{j_k,\infty}\|^2] + 3\mathbb{E}_k[\|A_k\nabla C_{j_k}\|^2].$$

We now analyze the upper bound of each of these three components individually.

B.2.2 Step 2: Analyzing the first component

The first term, $3\mathbb{E}_k[\|\nabla\Phi_{i_k,j_k}\|^2]$, represents the variance of the stochastic gradient for the regularized objective function. As shown in the analysis of section 3.3.2, this term is bounded as follows:

$$\mathbb{E}_k[\|\nabla\Phi_{i_k,j_k}(X_k)\|^2] \leq 4\hat{L}\Phi_0(X_k) + 2\sigma_\Phi^2.$$

Key Observation: This bound contains two distinct parts:

1. A term, $4\hat{L}\Phi_0(X_k)$, that is linearly dependent on $\Phi_0(X_k) := \Phi(X_k) - \Phi(x^*)$. This part will form the basis of the $2\mathcal{L}'\Phi_0(X_k)$ term in our final expression.

2. A constant term, $2\sigma_\Phi^2$, representing the variance at the optimum. This will be absorbed into the total noise term σ'^2.

B.2.3 Step 3: Analyzing the second component

The second term, $3\mathbb{E}_k[\|(A_k-1)\nabla B_{j_k,\infty}\|^2]$, is an artifact of our dynamic amplification mechanism. In section 3.3.2, we proved that this term is controllable and bounded by a finite constant, which we denoted as C_B,

$$\mathbb{E}_k[\|(A_k-1)\nabla B_{j_k,\infty}\|^2] \leq C_B.$$

Key Observation: This component contributes a constant value to the overall upper bound. This constant, $3C_B$, will be fully incorporated into the noise term σ'^2.

B.2.4 Step 4: Analyzing the third component

The third term, $3\mathbb{E}_k[\|A_k\nabla C_{j_k}\|^2]$, represents the amplified error arising from the decaying barrier parameter δ_k. The upper bound for this term, derived in section 3.3.2, depends on ϵ_k.

$$\mathbb{E}_k[\|A_k\nabla C_{j_k}\|^2] \leq (1+\lambda)^2 \cdot \left(3\hat{a}\frac{\epsilon_k^2\|\tilde{X}_k\|^2}{\delta_k^2\delta_\infty^2} + 3\hat{c}^2\epsilon_k^2\bar{a} + \frac{3\bar{b}\epsilon_k^2}{\delta_k^2\delta_\infty^2}\right).$$

Key Observation: Since $\{\epsilon_k\}$ is a vanishing sequence (i.e., $\epsilon_k \to 0$ as $k \to \infty$), this entire component is also a vanishing term. Within the expected smoothness framework, such vanishing but bounded terms are considered part of the overall noise. Therefore, this bound will also be absorbed into σ'^2.

B.3 Step 5: Aggregating the bounds

Finally, we substitute the bounds for all three components back into the main inequality from Step 1. We then proceed by collecting and regrouping the terms.

1. **Collect terms dependent on $\Phi_0(X_k)$:** We have only one such term, arising from the first component: $3 \cdot (4\hat{L}\Phi_0(X_k))$. We can define an effective smoothness constant $2\mathcal{L}' := 12\hat{L}$.

2. **Collect all constant and vanishing terms (the noise):** We aggregate all remaining parts: The constant variance from the first component $(3 \cdot 2\sigma_\Phi^2)$, the bound from the second component $(3 \cdot C_B)$, and the entire vanishing bound from the third component. We define the total noise term σ'^2 as the sum of all these bounded terms.

By these definitions, we directly arrive at the compact and elegant formulation of Assumption 5:

$$\mathbb{E}_k\left[\|G_k\|^2\right] \leq 2\mathcal{L}'\Phi_0(X_k) + \sigma'^2.$$

This completes the derivation and formally shows that Assumption 5 is a valid and well-founded simplification of the detailed analysis presented in the paper.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Bach, F. and Moulines, E. *Non-asymptotic analysis of stochastic approximation algorithms for machine learning,* In Adv. Neural Inf. Process. Syst. 24 (2011) 451–459.

[2] Bertsekas, D.P. and Tsitsiklis, J.N. *Neuro-dynamic programming*, Athena Scientific, 1996.

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

[4] Boyd, S. and Vandenberghe, L. *Convex optimization*, Cambridge University Press, 2004.

[5] Bubeck, S. *Convex optimization: Algorithms and complexity,* Found. Trends Mach. Learn. 8(3-4) (2015), 231–358.

[6] Dimitrieski, N., Cao, J. and Ebenbauer, C. *Stochastic gradient descent for constrained optimization based on adaptive relaxed barrier functions,* arXiv preprint arXiv:2503.10384 (2025).

[7] Feller, C. and Ebenbauer, C. *A stabilizing iteration scheme for model predictive control based on relaxed barrier functions,* Automatica, 80 (2017), 328–339.

[8] Fiacco, A.V. and McCormick, G.P. *Nonlinear programming: Sequential unconstrained minimization techniques*, John Wiley & Sons, 1968.

[9] Garrigos, G. and Gower, R.M. *Handbook of convergence theorems for (stochastic) gradient methods,* arXiv preprint arXiv:2301.11235 (2023).

[10] Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E. and Richtárik, P. SGD: General analysis and improved rates, In International conference on machine learning, PMLR, 2019, 5200–5209.

[11] Kushner, H.J. and Yin, G.G. *Stochastic approximation and recursive algorithms and applications*, 2nd ed., Springer, 2003.

[12] Li, M., Grigas, P. and Atamtürk, A. *New penalized stochastic gradient methods for linearly constrained strongly convex optimization*, J. Optim. Theory Appl. 205(2) (2025) 29.

[13] Liu, J. and Yuan, Y. *On almost sure convergence rates of stochastic gradient methods*, Proc. 35th Annual Conference on Learning Theory, PMLR, 178 (2022), 1–21.

[14] Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim. 19(4) (2009), 1574–1609.

[15] Nocedal, J. and Wright, S.J. Numerical optimization, 2nd ed., Springer, 2006.

[16] Robbins, H. and Monro, S. *A stochastic approximation method,* Ann. Math. Stat. 22(3) (1951), 400–407.

[17] Robbins, H. and Siegmund, D. *A convergence theorem for non negative almost supermartingales and some applications*, Rustagi, J.S. (ed.), Optimizing Methods in Statistics, Academic Press, 1971, 233–257.

[18] Rockafellar, R.T. Convex analysis, Princeton University Press, 1970.

[19] Roy, S.K. and Harandi, M. *Constrained stochastic gradient descent: The good practice,* In Int. Conf. Digit. Image Comput. Tech. Appl. (DICTA), IEEE, 2017, 1–8.

[20] Wang, M. and Bertsekas, D.P. *Incremental constraint projection methods for variational inequalities*, Math. Program. 150 (2015), 321–363.

[21] Wang, X., Ma, S. and Yuan, Y.X. *Penalty methods with stochastic approximation for stochastic nonlinear programming,* Math. comput. 86(306) (2017) 1793–1820.

[22] Yan, Y. and Xu, Y. *Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs*, Math. Program. Comput. 14(2) (2022), 319–363.

[23] Zhang, L., Zhang, Y., Wu, J. and Xiao, X. *Solving stochastic optimization with expectation constraints efficiently by a stochastic augmented Lagrangian-type algorithm,* INFORMS J. Comput. 34(6) (2022) 2989–3006.

[24] Zhang, T. *Solving large scale linear prediction problems using stochastic gradient descent algorithms*, In Proc. 21st Int. Conf. Machine Learning (ICML), 2004, 116.