A Projected Stochastic Gradient Method for Finite-Sum Problems with Linear Equality Constraints

Nataša Krklec Jerinkić[1], Benedetta Morini[2], Mahsa Yousefi[2]

[1]Department of Mathematics and Informatics, University of Novi Sad, Trg Dositeja Obradovića 4, Novi Sad, 21000, Serbia.
[2]Department of Industrial Engineering, University of Florence, Viale G.B. Morgagni 40, Florence, 50134, Italy.

Contributing authors: natasa.krklec@dmi.uns.ac.rs; benedetta.morini@unifi.it; mahsa.yousefi@unifi.it;

Abstract

A stochastic gradient method for finite-sum minimization subject to deterministic linear constraints is proposed and analyzed. The procedure presented adapts the projected gradient method on a convex set to the use of both a stochastic gradient and a possibly inexact projection map. Under standard assumptions in the field of stochastic gradient methods, we provide theoretical results in agreement with the theory for unconstrained problems. Numerical results are presented to show the practical behavior of the procedure.

Keywords: Constrained finite-sum minimization; stochastic gradient; exact and inexact projection.

MSC Classification: 90C30 , 90C06 , 90C53 , 90C90 , 65K05

1 Introduction

We consider the finite-sum optimization problem with linear equality constraints

$$\min_{x\in S} f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x), \quad S = \{x \in \mathbb{R}^n \mid Ax = b\}, \tag{1}$$

where the functions $f_i : \mathbb{R}^n \to \mathbb{R}, i = 1, ..., N$, are continuously-differentiable, $A \in \mathbb{R}^{m \times n}$, $m < n$, is a full row-rank matrix and b is a vector in $\mathbb{R}^m$.

Problems of minimizing finite-sums are often encountered in applications, such as least-squares approximation and Machine Learning within training phase where parameters of a model function are optimized. This usually assumes a large number of data which corresponds to a large N in (1). The so called Big Data setup motivates stochastic optimization approach since evaluating the whole function and/or its derivatives is too expensive to be treated by classical, deterministic methods. This raised a number of first-order stochastic strategies proposing different function and/or gradient approximations [1]. Moreover, analogously to deterministic optimization, second-order information can speed up convergence, and this yielded a new direction in stochastic optimization research based on appropriate Hessian. A special case of this approach leans on spectral methods and stochastic Barzilai-Borwein stepsizes, see e.g., [2–5].

Subsampling is a main-stream procedure for obtaining approximations to function and/or gradient evaluations; subsampling strategies range from the mini-batch approach where the sample size is usually small and fixed (e.g., [6]), to increasing sample size strategies where the full sample is eventually reached (e.g., [7]), with many variations in between ([3, 8, 9] to name just a few). Determining a suitable stepsize sequence is also an actuator in stochastic optimization field. In addition to prefixed constant stepsize sequences and diminishing stepsizes, globalization strategies such as line search and trust region can be adapted to the stochastic framework (see e.g. [10–12]) but the majority of such approaches requires at least approximate function evaluations.

Finite-sum problems can also come with constraints incorporating prior knowledge and physical meaning [13–16]. In this work, we propose a stochastic projected gradient method for problem (1) based on the projected gradient method. Function evaluations are not required while a mini-batch approach is employed to compute stochastic gradients. Regarding the mini-batch strategy, the set of indices in the sum (1) is divided into chunks of data which can be redefined, possibly at each iteration. At any iteration k, each chunk can be sorted with the same probability; the corresponding gradient estimate is calculated on the sampled portion and by using an appropriate scaling that provides an unbiased estimate of the full gradient. In order to handle constraints, we allow the use of inexact projections, especially suited in the presence of a large number of constraints (e.g., [17]); specifically, inexact but controlled projections provide a nonmonotone decay of the infeasibility measure. The proposed algorithm is accompanied with a theoretical analysis that establishes convergence results, distinguishing constant and diminishing step stepsize sequences, in line with the existing literature. Convexity of the objective function is not required. The numerical behavior of the method is shown considering some stepsize selections in agreement with the theory.

Outline of the paper. In the following section we state the method and provide some insights and preliminary results. Section 3 is devoted to the theoretical analysis of the method, while Section 4 provides a comparison with relevant papers in the literature. Some numerical results are presented in Section 5, while the last section draws the main conclusions.

Notations. The symbol $\|\cdot\|$ indicates the Euclidean norm. $Pr(\cdot)$ and $\mathbb{E}[\cdot]$ represent the probability function and expected value, respectively.

2 The Method

In this section, we introduce our Projected Stochastic Gradient method for Linear Equality COnstrained problems, named the PSG_LECO method, to solve the problem (1). We start by describing two main tasks in the algorithm: the construction of a stochastic gradient and the computation of the projection of a point in $\mathbb{R}^n$ onto S.

At k-th iteration, given $x_k \in \mathbb{R}^n$, a stochastic gradient g_k is computed as a mini-batch gradient by using the following strategy. Let us define a partition $\{\mathcal{N}_k^i\}_{i=1}^r$ of $\mathcal{N} = \{1, \ldots, n\}$ into r disjoint mini-batches at iteration k, namely

$$\mathcal{N} = \bigcup_{i=1}^{r} \mathcal{N}_k^i, \quad \text{where} \quad \mathcal{N}_k^i \cap \mathcal{N}_k^j = \emptyset, \quad \text{for all } i \neq j. \tag{2}$$

This partition can be either fixed or varying along the iterations. Then we let $g_k^{(i)}$ be the eligible mini-batch gradients associated with the partition in (2), i.e.,

$$g_k^{(i)} = \frac{r}{N} \sum_{j \in \mathcal{N}_k^i} \nabla f_j(x_k), \quad i = 1, \ldots, r, \tag{3}$$

and g_k be our stochastic gradient that corresponds to a uniformly and randomly selected mini-batch $\mathcal{N}_k^i$ from the partition. By

$$Pr\left(g_k = g_k^{(i)} \mid \mathcal{F}_k\right) = \frac{1}{r}, \quad i = 1, \ldots, r, \tag{4}$$

where $Pr(\cdot)$ represents the probability of the outcomes, and $\mathcal{F}_k$ is a σ-algebra generated by $x_0, \ldots, x_k$, i.e., by $g_0, \ldots, g_{k-1}$, we have

$$\mathbb{E}[g_k \,|\mathcal{F}_k] = \nabla f(x_k). \tag{5}$$

Regarding the projection map onto S, since $A \in \mathbb{R}^{m \times n}$ has full rank, the orthogonal projection $\pi_S(y)$ of any given point $y \in \mathbb{R}^n$ onto S takes the form

$$\begin{aligned} \pi_S(y) &= y - A^T \lambda(y), \\ \lambda(y) &= (AA^T)^{-1}(Ay - b). \end{aligned} \tag{6}$$

The computation of $\pi_S(y)$ is viable for moderate values of m as it depends on the solution of the linear system

$$AA^T \widetilde{\lambda}(y) = Ay - b. \tag{7}$$

Algorithm 1 PSG_LECO

1: Choose an initial iterate $x_0 \in \mathbb{R}^n$, a positive sequence $\{\alpha_k\}$, nonnegative sequences $\{\eta_k\}$ and $\{\mu_k\}$ with $\eta_k \in [0, \bar{\eta})$, $\bar{\eta} < 1$, positive scalars $\delta_\ell < \delta_u < \infty$, and an initial stepsize $\Delta_0 \in [\alpha_0\delta_\ell, \alpha_0\delta_u]$.
2: **for** $k = 0, 1, \ldots$ **do**
3: Select uniformly at random a mini-batch $\mathcal{N}_k^i$ from the partition $\{\mathcal{N}_k^1, \ldots, \mathcal{N}_k^r\}$ as in (2), and set $g_k = g_k^{(i)}$ as in (3).
4: Set $y_k = x_k - \Delta_k g_k$.
5: Compute $x_{k+1} = \widetilde{\pi}_S(y_k)$ as in (7), where $r(y_k)$ satisfies

$$\|r(y_k)\| \leq \eta_k \|Ax_k - b\| + \mu_k. \tag{9}$$

6: Choose a scalar δ_k.
7: Compute $\Delta_{k+1} = \alpha_{k+1}\pi_\delta(\delta_k) = \alpha_{k+1} \max\{\delta_\ell, \min\{\delta_k, \delta_u\}\}$.
8: Set $k = k + 1$.
9: **end for**

More generally, it is advisable to allow for inexact projections of the form

$$\begin{aligned} \widetilde{\pi}_S(y) &= y - A^T\widetilde{\lambda}(y), \\ \widetilde{\lambda}(y) &= (AA^T)^{-1}(Ay - b + r(y)), \end{aligned} \tag{8}$$

and $r(y)$ denotes the residual vector in the solution of (7).

Algorithm 1 sketches the k-th iteration of our procedure. Step 1 indicates the hyper-parameters required for execution: two nonnegative sequences $\{\eta_k\}$ and $\{\mu_k\}$ that control inexactness of the projection onto S, a positive stepsize related sequence $\{\alpha_k\}$, two positive scalars δ_ℓ, δ_u that define the projection $\pi_\delta(\cdot) = \max\{\delta_\ell, \min\{\cdot, \delta_u\}\}$ employed in the computation of the stepsize.

Step 3 refers to the construction of the stochastic gradient g_k as described above. Steps 4 and 5 concern the computation of the iterate x_{k+1}. Specifically, first the vector y_k is formed using the step length Δ_k fixed in the previous iteration, then y_k is projected onto S and this gives rise to the new iterate x_{k+1}. Inequality (9) rules the accuracy in the calculation of the projection of y_k by means of scalars η_k and μ_k. We observe that if $\eta_k = \mu_k = 0$, $\forall k$, then the iterates x_k, $k \geq 1$, are feasible irrespective of x_0.

Steps 6 and 7 are devoted to the computation of the step length to be used at the successive iteration. The choice of a (positive) scalar δ_k offers a variety of options and we discuss some possible adaptive choices in Section 5. At the end of iteration k, Δ_{k+1} is formed by means of α_{k+1} and $\pi_\delta(\delta_k)$. Consequently, Δ_{k+1} is deterministic conditioning on x_{k+1} and this feature is crucial in the analysis of the procedure. The predetermined sequences $\{\alpha_k\}$ and $\{\mu_k\}$ affect the convergence properties, as shown in the subsequent theoretical analysis.

3 Theoretical Analysis

In this section, we analyze the theoretical properties of the PSG_LECO method. We make the following standard assumptions on the objective function.

Assumption 1. *The objective function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable on $\mathbb{R}^n$. The gradient of f is Lipschitz continuous with constant $L > 0$.*

Assumption 2. *The objective function $f : \mathbb{R}^n \to \mathbb{R}$ is bounded from below by f^*,*

$$f(x) \geq f^*, \qquad \forall\, x \in \mathbb{R}^n. \tag{10}$$

The following result introduces an optimality measure $d(x)$ for problem (1) which will be used in our subsequent analysis.

Theorem 1. *Assume that f is continuously differentiable in an open set containing S and let*

$$d(x) = \pi_S(x - \nabla f(x)) - x \tag{11}$$

Then, it holds $d(\bar{x}) = 0$, $\bar{x} \in S$, if and only if $\bar{x}$ is a stationary point for the problem (1).

Proof. See [18, Lemma 2.1]. □

Under suitable assumptions, our analysis will provide the following results: each limit point of the sequence $\{x_k\}$ is feasible; if $\alpha_k = \alpha$, $\forall k$, then the expected sum of average-squared norms of $d(x_k)$ is bounded and decreases with α; if $\{\alpha_k\}$ is a diminishing sequence, then $\|d(x_k)\|$ cannot stay bounded away from zero almost surely. Therefore we establish results in accordance with the corresponding algorithms for unconstrained optimization.

The first step in our analysis is to characterize the infeasibility of the iterates generated by Algorithm 1. To this end, we introduce the measures

$$\widetilde{e}(y) = Ay - b, \qquad e(y) = \|\widetilde{e}(y)\|, \tag{12}$$

and make the following assumption on the sequences $\{\eta_k\}$ and $\{\mu_k\}$.

Assumption 3. *The nonnegative sequence $\{\eta_k\}$ satisfies $\eta_k \leq \bar{\eta} < 1$, $\forall k \geq 0$. The nonnegative sequence $\{\mu_k\}$ converges to zero R-linearly.*

Lemma 1. *Suppose that Assumptions 1 and 3 hold. Then,*

a) For $y \in \mathbb{R}^n$, it holds

$$\widetilde{e}(\widetilde{\pi}_S(y)) = -r(y). \tag{13}$$

b) The iterate x_{k+1} satisfies

$$\widetilde{e}(x_{k+1}) = -r(y_k), \quad \text{for all} \quad k \geq 0. \tag{14}$$

c) It holds

$$e(x_k) \leq \kappa_e q^k, \quad \text{for all} \quad k \geq 0, \tag{15}$$

for some positive constant κ_e *and some* $q \in (\bar{\eta}, 1)$ *and every limit point of the sequence* $\{x_k\}$ *is feasible.*

d) It holds

$$e(x_k) \leq \kappa_e, \quad \textit{for all} \quad k \geq 0. \tag{16}$$

Proof. a) Equations (8), (7) and (12) give

$$\widetilde{e}(\widetilde{\pi}_S(y)) = A\widetilde{\pi}_S(y) - b = A\left(y - A^T\tilde{\lambda}(y)\right) - b = -(AA^T\tilde{\lambda}(y) - Ay + b)$$

which corresponds to (13).

b) Equality (14) follows from (13) and Step 4 of Algorithm 1.

c) Equations (14), (9), (7) and Assumption 3 give

$$e(x_{k+1}) \leq \bar{\eta} e(x_k) + \mu_k.$$

Therefore, for all k, there holds

$$e(x_k) \leq \bar{\eta}^k e(x_0) + v_k, \qquad v_k = \sum_{i=1}^{k} \bar{\eta}^{i-1}\mu_{k-i}. \tag{17}$$

Since $\{\mu_k\}$ converges to zero R-linearly and $\bar{\eta} < 1$, $\{v_k\}$ converges to zero R-linearly which implies (15), see e.g., [19, Lemma 4.2]. Thus, $\lim_{k\to\infty} e(x_k) = 0$.

d) The claim trivially follows from (15). □

Now we proceed with an intermediate result on the step d_k taken at iteration k

$$d_k = x_{k+1} - x_k. \tag{18}$$

We denote

$$D = A^T(AA^T)^{-1}, \qquad P_A = I - DA, \tag{19}$$

$\kappa_D = \|D\|$ and introduce the following assumptions.

Assumption 4. *The sequence* $\{x_k\}$ *is either feasible or such that*

$$\|\nabla f(x_k)\| \leq \kappa_\nabla,$$

for some positive constant κ_∇ *and for all* $k \geq 0$.

Note that we assume bounded gradients only in the case where the iterates are infeasible.

Assumption 5. *There exist a positive constant* $\nu > 0$ *such that*

$$\mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2 \,|\mathcal{F}_k\right] \leq \nu. \tag{20}$$

This assumption states that the conditional variance of the sampled gradient is uniformly bounded.

Lemma 2. *Let x_k be generated by Algorithm* PSG_LECO. *Suppose that Assumption 1 holds.*

a) The step d_k defined in (18) satisfies

$$\mathbb{E}\big[d_k\,|\mathcal{F}_k\big] = \Delta_k d(x_k) - (1-\Delta_k)D\widetilde{e}(x_k) - D\mathbb{E}\big[r(y_k)\,|\mathcal{F}_k\big], \tag{21}$$

where D is defined in (19).
b) Suppose further that Assumptions 3 and 4 hold. Then, $d(x_k)$ defined in (11) satisfies

$$d(x_k)^T\nabla f(x_k) \leq -\|d(x_k)\|^2 + \kappa_1 e(x_k), \tag{22}$$

for some positive constant κ_1.
c) Suppose further that Assumptions 3, 4 and 5 hold, and that $(1-\Delta_k)^2 \leq \Delta^*$, for all k and some positive Δ^*. Then,

$$\mathbb{E}\big[\|d_k\|^2\,|\mathcal{F}_k\big] \leq \kappa_2\Delta_k^2 + \kappa_3 q^{2k} + \kappa_4\mu_k + 2\Delta_k^2\|d(x_k)\|^2, \tag{23}$$

for some positive constants κ_2, κ_3, κ_4.

Proof. a) First, we note that (18), (19), (8) and (7) give

$$\begin{aligned} d_k &= \widetilde{\pi}_S(x_k - \Delta_k g_k) - x_k \\ &= x_k - \Delta_k g_k - D(Ax_k - \Delta_k A g_k - b + r(y_k)) - x_k \\ &= -\Delta_k P_A g_k - D\widetilde{e}(x_k) - Dr(y_k), \end{aligned} \tag{24}$$

and that (11), (6) give

$$d(x_k) = -P_A\nabla f(x_k) - D\widetilde{e}(x_k). \tag{25}$$

Hence, using (5) and that Δ_k is $\mathcal{F}_k$ -measurable, we have

$$\begin{aligned} \mathbb{E}\big[d_k\,|\mathcal{F}_k\big] &= -\Delta_k P_A\mathbb{E}\big[g_k\,|\mathcal{F}_k\big] - D\widetilde{e}(x_k) - D\mathbb{E}\big[r(y_k)\,|\mathcal{F}_k\big] \\ &= -\Delta_k P_A\nabla f(x_k) - D\widetilde{e}(x_k) - D\mathbb{E}\big[r(y_k)\,|\mathcal{F}_k\big], \end{aligned}$$

and (25) concludes the proof.

b) Using (25), $P_A D = 0$, $P_A^T = P_A$ and $P_A^2 = P_A$, it follows

$$\begin{aligned} d(x_k)^T d(x_k) &= \nabla f(x_k)^T P_A^2\nabla f(x_k) + \|D\widetilde{e}(x_k)\|^2 \\ &= \nabla f(x_k)^T P_A\nabla f(x_k) + \|D\widetilde{e}(x_k)\|^2 \\ &= -\nabla f(x_k)^T d(x_k) - \nabla f(x_k)^T D\widetilde{e}(x_k) + \|D\widetilde{e}(x_k)\|^2. \end{aligned}$$

Hence, using (12), (16) we obtain

$$\begin{aligned}\nabla f(x_k)^T d(x_k) &\leq -\|d(x_k)\|^2 + \|\nabla f(x_k)\|\,\kappa_D e(x_k) + \kappa_D^2 e(x_k)^2 \\ &\leq -\|d(x_k)\|^2 + \big(\kappa_\nabla \kappa_D + \kappa_D^2 \kappa_e\big) e(x_k).\end{aligned}$$

Thus, (22) holds with $\kappa_1 = \kappa_\nabla \kappa_D + \kappa_D^2 \kappa_e$.

c) Using (24), (25), $\|P_A\| = 1$, and (15) we obtain

$$\begin{aligned}\|d_k - \Delta_k d(x_k)\|^2 &= \| - \Delta_k P_A(g_k - \nabla f(x_k)) - (1-\Delta_k) D\widetilde{e}(x_k) - Dr(y_k)\|^2 \\ &\leq 2\Delta_k^2 \|g_k - \nabla f(x_k)\|^2 + 2\|(1-\Delta_k) D\widetilde{e}(x_k) + Dr(y_k)\|^2 \\ &\leq 2\Delta_k^2 \|g_k - \nabla f(x_k)\|^2 + 4\|(1-\Delta_k) D\widetilde{e}(x_k)\|^2 + 4\|Dr(y_k)\|^2 \\ &\leq 2\Delta_k^2 \|g_k - \nabla f(x_k)\|^2 + 4\kappa_D^2 \Delta^* e(x_k)^2 + 4\kappa_D^2 (\eta_k e(x_k) + \mu_k)^2 \\ &\leq 2\Delta_k^2 \|g_k - \nabla f(x_k)\|^2 + 4\kappa_D^2 (\Delta^* + \bar{\eta}^2)\kappa_e^2 q^{2k} + 4\kappa_D^2 (2\bar{\eta}\kappa_e q^k + \mu_k)\mu_k.\end{aligned}$$

Now, using this last inequality, Assumption 5 and the equation

$$\|d_k\|^2 = \|d_k - \Delta_k d(x_k) + \Delta_k d(x_k)\|^2 \leq 2\|d_k - \Delta_k d(x_k)\|^2 + 2\Delta_k^2 \|d(x_k)\|^2,$$

we obtain

$$\begin{aligned}\mathbb{E}\big[\|d_k\|^2 \,|\mathcal{F}_k\big] &\leq 2\mathbb{E}\big[\|d_k - \Delta_k d(x_k)\|^2 \,|\mathcal{F}_k\big] + 2\Delta_k^2 \mathbb{E}\big[\|d(x_k)\|^2 \,|\mathcal{F}_k\big] \\ &\leq 4\Delta_k^2 \mathbb{E}\big[\|g_k - \nabla f(x_k)\|^2 \,|\mathcal{F}_k\big] + 8\kappa_D^2 (\Delta^* + \bar{\eta}^2)\kappa_e^2 q^{2k} \\ &\quad + 8\kappa_D^2 (2\bar{\eta}\kappa_e q^k + \mu_k)\mu_k + 2\Delta_k^2 \|d(x_k)\|^2. \qquad (26)\end{aligned}$$

Thus, the claim holds with $\kappa_2 = 4\nu$, $\kappa_3 = 8\kappa_D^2(\Delta^* + \bar{\eta}^2)\kappa_e^2$, $\kappa_4 = 8\kappa_D^2(2\bar{\eta}\kappa_e + \mu_{\max})$ where $\mu_{\max} = \max_k \mu_k$. □

For sake of completeness, the following lemma rephrases the results above when x_0 is feasible and the exact projection π_S is used.

Corollary 1. *Let x_k be generated by Algorithm* PSG_LECO. *Suppose that Assumption 1 holds, $x_0 \in S$, $\eta_k = \mu_k = 0$, $\forall k \geq 0$.*
a) The sequence $\{x_k\}$ is feasible.
b) The step d_k defined in (18) satisfies

$$\mathbb{E}\big[d_k \,|\mathcal{F}_k\big] = \Delta_k d(x_k). \qquad (27)$$

c) The direction $d(x_k)$ defined in (11) satisfies

$$d(x_k)^T \nabla f(x_k) \leq -\|d(x_k)\|^2. \qquad (28)$$

d) Suppose further that Assumption 5 holds and that $(1-\Delta_k)^2 \leq \Delta^*$ for all k and some positive Δ^*. Then,

$$\mathbb{E}\big[\|d_k\|^2 \,|\mathcal{F}_k\big] \leq \kappa_2 \Delta_k^2 + 2\Delta_k^2 \|d(x_k)\|^2. \qquad (29)$$

for some positive κ_2.

Proof. Item (a) follows directly from (17). Items (b)–(d) follow straightforwardly from Lemma 2. □

Now we show the behavior of the PSG_LECO method in case the sequence $\{\alpha_k\}$ is diminishing.

Theorem 2. *Suppose that Assumptions 1–5 hold, and that*

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{30}$$

Then,

$$\liminf_{k\to\infty} \|d(x_k)\| = 0, \tag{31}$$

almost surely and the sequence of function values $\{f(x_k)\}$ *converges almost surely.*

Proof. Conditions (30) imply $(1-\Delta_k)^2 \leq \Delta^*$, $\forall k$, and some positive Δ^*. The Assumption 1 implies that

$$f(x) \leq f(y) + \nabla f(y)^T(x-y) + \frac{L}{2}\|x-y\|^2 \ \ \forall\ x, y \in \mathbb{R}^n. \tag{32}$$

Hence, by (18) we obtain

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T d_k + \frac{L}{2}\|d_k\|^2.$$

By taking the conditional expectation of both sides with respect to the σ-algebra $\mathcal{F}_k$, we obtain

$$\mathbb{E}\big[f(x_{k+1})\,|\mathcal{F}_k\big] \leq f(x_k) + \nabla f(x_k)^T \mathbb{E}\big[d_k\,|\mathcal{F}_k\big] + \frac{L}{2}\,\mathbb{E}\big[\|d_k\|^2\,|\mathcal{F}_k\big].$$

Now, by (21), (22) and (23) we have

$$\begin{aligned}
\mathbb{E}\big[f(x_{k+1})\,|\mathcal{F}_k\big] &\leq f(x_k) - \Delta_k\|d(x_k)\|^2 + \kappa_1\Delta_k e(x_k) - (1-\Delta_k)\nabla f(x_k)^T D\widetilde{e}(x_k) \\
&\quad - \mathbb{E}\big[\nabla f(x_k)^T Dr(y_k)\,|\mathcal{F}_k\big] + \frac{L}{2}(\kappa_2\Delta_k^2 + \kappa_3 q^{2k} + \kappa_4\mu_k + 2\Delta_k^2\|d(x_k)\|^2) \\
&\leq f(x_k) - \Delta_k\|d(x_k)\|^2 + \kappa_1\Delta_k e(x_k) + \sqrt{\Delta^*}\|\nabla f(x_k)\|\kappa_D e(x_k) \\
&\quad + \|\nabla f(x_k)\|\kappa_D\mathbb{E}\big[\|r(y_k)\|\,|\mathcal{F}_k\big] \\
&\quad + \frac{L}{2}(\kappa_2\Delta_k^2 + \kappa_3 q^{2k} + \kappa_4\mu_k + 2\Delta_k^2\|d(x_k)\|^2).
\end{aligned} \tag{33}$$

If the sequence $\{x_k\}$ is feasible, then $r(y_k) = 0$. Otherwise, the upper bound $\bar{\eta}e(x_k) + \mu_k$ on $\|r(y_k)\|$ given in (9) is $\mathcal{F}_k$-measurable, and we obtain the following inequality that includes exact and inexact projections,

$$\begin{aligned}
\mathbb{E}\big[f(x_{k+1})\,|\mathcal{F}_k\big] &\leq f(x_k) - \Delta_k\|d(x_k)\|^2 + \big((1+\sqrt{\Delta^*})\kappa_1 + \kappa_\nabla\kappa_D(\sqrt{\Delta^*}+\bar{\eta})\big)e(x_k) \\
&\quad + \kappa_\nabla\kappa_D\mu_k + \frac{L}{2}(\kappa_2\Delta_k^2 + \kappa_3 q^{2k} + \kappa_4\mu_k + 2\Delta_k^2\|d(x_k)\|^2) \\
&\leq f(x_k) - \Delta_k(1 - L\Delta_k)\|d(x_k)\|^2 + \kappa_5 q^k + \kappa_6\mu_k + \frac{L}{2}\kappa_2\Delta_k^2,
\end{aligned} \tag{34}$$

with $\kappa_5 = \Big(\kappa_1(1+\sqrt{\Delta^*}) + \kappa_\nabla\kappa_D(\sqrt{\Delta^*}+\bar{\eta})\Big)\kappa_e + \frac{L}{2}\kappa_3$, $\kappa_6 = \Big(\kappa_\nabla\kappa_D + \frac{L}{2}\kappa_4\Big)$ and the last inequality obtained using (15).

By construction Δ_k satisfies

$$\alpha_k\delta_\ell \leq \Delta_k \leq \alpha_k\delta_u, \tag{35}$$

hence

$$\mathbb{E}\big[u_{k+1}\,|\mathcal{F}_k\big] \leq u_k - \alpha_k\big(\delta_\ell - L\alpha_k\delta_u^2\ \big)\|d(x_k)\|^2 + \kappa_5 q^k + \kappa_6\mu_k + \frac{L}{2}\kappa_2\alpha_k^2\delta_u^2, \tag{36}$$

where $u_k = f(x_k) - f^*$ from (10). Since $\{\alpha_k\}$ is diminishing, let $\bar{k}$ be the index such that $\big(\delta_\ell - L\alpha_k\delta_u^2\big) \geq \frac{\delta_\ell}{2}$ for all $k \geq \bar{k}$. Hence, for $k \geq \bar{k}$

$$\mathbb{E}\big[u_{k+1} \mid \mathcal{F}_k\big] \leq u_k - \alpha_k\frac{\delta_\ell}{2}\|d(x_k)\|^2 + \kappa_5 q^k + \kappa_6\mu_k + \frac{L}{2}\kappa_2\alpha_k^2\delta_u^2.$$

Since $\sum_{k=0}^{\infty}\alpha_k^2$, $\sum_{k=0}^{\infty} q^k$ and $\sum_{k=0}^{\infty}\mu_k$ are summable, the Robbins–Siegmund supermartingale convergence Theorem [20] gives that

$$\sum_{k=\bar{k}+1}^{\infty}\alpha_k\|d(x_k)\|^2 < \infty,$$

almost surely. Now, assume by contradiction that $\|d(x_k)\| \geq c > 0$ for all $k > \bar{k}$. Thus

$$\sum_{k=\bar{k}+1}^{\infty}\alpha_k\|d(x_k)\|^2 \geq c^2\sum_{k=\bar{k}+1}^{\infty}\alpha_k,$$

which contradicts (30). Therefore, (31) holds almost surely. Finally, from the Robbins–Siegmund supermartingale convergence theorem, we also conclude that $\lim_{k\to\infty} u_k$ exists and is finite almost surely. □

The theorem above implies that almost surely there exists a subsequence $\{\|d(x_k)\|\}_{k\in K}$ of $\{\|d(x_k)\|\}$ convergent to zero; if $\{x_k\}_{k\in K}$ admits limit points, then such points are stationary.

Now we analyze the convergence of the PSG_LECO method using constant steplengths and show that the expected sum of average-squared norms of $d(x_k)$ is bounded and decreases with α.

Theorem 3. *Suppose that Assumptions 1–5 hold, $\eta_k \leq \bar{\eta} < 1, \forall k$. If $\alpha_k = \alpha$, $\forall k \geq 0$, with α such that*

$$0 < \alpha \leq \frac{\delta_\ell}{2L\delta_u^2}, \tag{37}$$

then the iterates generated by PSG_LECO *method satisfy*

$$\lim_{K\to\infty}\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K}\|d(x_k)\|^2\right] \leq \frac{\alpha L\kappa_2\delta_u^2}{\delta_\ell}, \tag{38}$$

where κ_2 is scalar in (23).

Proof. Condition (37) implies $(1-\Delta_k)^2 \leq \Delta^*$, $\forall k$, and some positive Δ^*. Applying (35) and (37) in (34) and taking total expectation, we obtain the following

$$\mathbb{E}\big[f(x_{k+1})\big] \leq \mathbb{E}\big[f(x_k)\big] - \frac{1}{2}\alpha\delta_\ell\mathbb{E}\big[\|d(x_k)\|^2\big] + \kappa_5 q^k + \kappa_6\mu_k + \frac{L}{2}\kappa_2\alpha^2\delta_u^2,$$

and consequently

$$\mathbb{E}\big[\|d(x_k)\|^2\big] \leq \frac{2}{\alpha\delta_\ell}\left(\mathbb{E}\big[f(x_k)\big] - \mathbb{E}\big[f(x_{k+1})\big]\right) + \frac{2\kappa_5}{\alpha\delta_\ell}q^k + \frac{2\kappa_6}{\alpha\delta_\ell}\mu_k + \frac{\alpha L\kappa_2\delta_u^2}{\delta_\ell}.$$

Summing both sides of this inequality for $k = 0, \ldots, K$, and dividing by K, we have

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K}\|d(x_k)\|^2\right] \leq \frac{2}{\alpha\delta_\ell K}\left(f(x_0) - \mathbb{E}\big[f(x_{K+1})\big]\right) + \frac{2\kappa_5}{\alpha\delta_\ell K}\sum_{k=0}^{K}q^k$$

$$+ \frac{2\kappa_6}{\alpha\delta_\ell K}\sum_{k=0}^{K}\mu_k + \frac{\alpha L\kappa_2\delta_u^2}{\delta_\ell}$$

$$\leq \frac{2}{\alpha\delta_\ell K}(f(x_0) - f^*) + \frac{2\kappa_5}{\alpha\delta_\ell K}\sum_{k=0}^{K} q^k$$

$$+ \frac{2\kappa_6}{\alpha\delta_\ell K}\sum_{k=0}^{K}\mu_k + \frac{\alpha L\kappa_2\delta_u^2}{\delta_\ell}.$$

Since $\sum_{k=0}^{\infty} q^k$ and $\sum_{k=0}^{\infty}\mu_k$ are summable, the claim follows. □

4 Related work

The extension of methods with random models for the unconstrained setting to the setting of deterministic equality and inequality constrained problems is a recent area of research which is drawing much interest, see [13–17, 21–23]. Focusing on papers [13, 15–17] for deterministic equality constrained optimization, we sketch their main features.

The work [17] introduces a projected gradient method for finite-sum minimization where iterates are allowed to be infeasible in a controlled way whose form is different from the one proposed here. Functions and gradients are approximated by subsampling and the stepsize is determined by a nonmonotone line-search rule. With respect to [17], PSG_LECO Algorithm does not require function evaluation nor an acceptance test of the trial iterate. Further, PSG_LECO Algorithm does not require adjusting the size of the batches used for function and gradient approximations.

Papers [13, 15, 16] present objective function-free procedures in the class of Sequential Quadratic Programming (SQP) algorithms. The procedures in [15, 16] employ stochastic gradients of the objective function and prescribed approximations of the Hessian of the objective function and/or a Lagrangian function; the stepsize selection is adaptive and is based on estimated Lipschitz constants. In particular, in [15] the search direction results from the use of a merit function with ℓ_1-norm penalty function; it is computed solving a quadratic optimization problem based on a local quadratic minimizer of the objective function and a local affine model of the constraint. The choice of the stepsize is inspired by a line search strategy and is based on a rule using Lipschitz constant estimates. The hyper-parameters of the algorithm are: a sequence of Lipschitz constant estimates of the objective function, a sequence of Lipschitz constant estimates of the constraint function, and a sequence to control the stepsize. Assuming that the stochastic gradient is unbiased and condition (20), the theoretical analysis shows that the generated sequence provides stationarity and feasibility in expectation. In our linearly constrained case, the algorithm generates feasible iterates.

In [16] the search direction is decomposed into a normal step and a tangential step. Exact projections are supposed to be computable and consequently the normal step has a closed form. On the other hand, the tangential step solves a trust-region problem which employs a basis for the null space of the Jacobian of the constraints. The adaptive trust-region radius is computed using estimates of the Lipschitz constants of both the objective function and the constraint functions.The hyper-parameters of the algorithm are: a sequence of Lipschitz constant estimates of the objective function, a sequence of Lipschitz constant estimates of the constraint function, and two sequences of positive scalars to control the trust-region radius. Assuming that the stochastic gradient is unbiased and condition (20), the theoretical analysis shows that KKT residuals converge to zero almost surely.

Finally, the very recent paper [13] extends stochastic momentum methods for unconstrained optimization to the stochastic SQP setting and presents two algorithms: a projected

stochastic heavy-ball SQP algorithm and a projected stochastic Adam SQP algorithm. The projection map is supposed to be computed exactly; momentum terms are implemented with projected gradient estimates and with two prefixed stepsize related sequences. Assuming that the stochastic projected gradient is unbiased, condition (20) holds and $\{\|\nabla f(x_k)\|\}$ is bounded from above, the convergence analysis shows that the theoretical behavior of the projected stochastic momentum methods is analogous to the behavior of the corresponding procedures for the unconstrained setting. With respect to [13], PSG_LECO Algorithm allows inexact projections.

5 Numerical experiments

In this section we illustrate the numerical behavior of PSG_LECO method both with the exact projection π_S in (6) and the inexact projection $\widetilde{\pi}_S$ in (8). We applied PSG_LECO in the solution of the logistic regression model with three binary datasets and in the solution of problems employed from the CUTEst collection.

All runs were performed in MATLAB R2025a on a Rocky Linux 8.10 (64-bit) server with 256 GiB RAM.

5.1 Data driven equality–constrained logistic regression

We used MUSHROOMS, MNIST, and DIABETES datasets from LIBSVM [24]. For MUSHROOMS, we mapped labels $1, 2$ to $+1, -1$ respectively, and used the feature matrix as returned. For the multi-class dataset MNIST, we restricted the dataset to a binary one with digits $0, 8$ and mapped 0 to $+1$ and 8 to -1; all features were scaled to $[0, 1]$. For DIABETES, we remapped labels $0, 1$ to $-1, +1$ respectively, and scaled the features to $[-1, 1]$.[1]

Given $\{(z_i, y_i)\}_{i=1}^N$ with $z_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$, in (1) we let[2]

$$f(x) \;=\; \frac{1}{N}\sum_{i=1}^{N} \log\bigl(1 + \mathrm{e}^{-y_i z_i^\top x}\bigr). \tag{39}$$

The constraint matrix $A \in \mathbb{R}^{m\times n}$ and the vector $b \in \mathbb{R}^m$ were generated using a fixed random seed (`rng(0)`); the dimension m was set equal to $\lfloor 0.5n \rfloor$.

As for the dimensions n and N, MUSHROOMS dataset has $n = 112$ and $N = 8124$, MNIST dataset has $n = 780$ and $N = 11774$, and DIABETES dataset has $n = 8$ and $N = 768$. The mini-batch size was set to 256 for MNIST and MUSHROOMS, and to 64 for DIABETES.

We used the initial iterate $x_0 = A^T(AA^T)^{-1}b$ for the runs with exact projection and $x_0 = 0$ for runs with inexact projection.

5.2 CUTEst-derived equality–constrained problems

We considered the problems HUESTIS, DTOC1L, and HS50 subject to linear constraints; the dimension n and the numbers of linear constraints m are: $n = 10$ and $m = 2$ for HUESTIS, $n = 58$ and $m = 36$ for DTOC1L, and $n = 5$ and $m = 3$ for HS50. Matrix A and the initial feasible point x_0 are given in the collection.

Following [17], we let $\widetilde{f}(x) : \mathbb{R}^n \to \mathbb{R}$ denote and objective function in the CUTEst collection, and defined a finite-sum objective function for (1) as

$$f(x) = \widetilde{f}(x) + \sum_{i=1}^{N} \xi_i^2 \,\|x\|_2^2, \tag{40}$$

[1] We used `Mnist.mat` and `diabetes.txt` from the LIBSVM Library.
[2] In our runs we used the stable formulation of the logistic regression.

Table 1: Strategies for computing the step length Δ_{k+1}.

Strategy	α_{k+1}	δ_k
S1	$\alpha > 0$	δ_k in (41)
S2	α_{k+1} in (42)-(43)	δ_k in (41)
S3	α_{k+1} in (42)–(43)	$\delta_k = 1$

where each ξ_i is an independent random sample drawn from a Gaussian distribution, i.e., $\xi_i \sim \mathcal{N}(0, \sigma^2)$,with $\sigma = 0.1$ and fixed random seed (`rng(0)`). In our runs, we set $N = 10000$ and batch size equal to 256.

5.3 Experimental configuration

Regarding the given parameters in PSG_LECO, we set $\delta_\ell = 10^{-3}$, $\delta_u = 10^2$ and initial stepsize $\Delta_0 = \alpha_0 \delta_\ell$. Each run consisted of $k_{\max} = 10^4$ iterations. Regarding the partition $\{\mathcal{N}_k^i\}_{i=1}^r$ in (2), it was kept fixed along the iterations. We built ten independent random partitions, each of which created by using an independent random seed. Each problem was solved using such partitions, thus giving rise to ten solves per problem.

Regarding the selection of the stepsizes, we tested three strategies denoted **S1**–**S3**. They differ in the choice of the scaling parameter α_{k+1} and/or δ_k. In **S1** we considered a fixed sequence $\{\alpha_k\}$, $\alpha_k = \alpha$, $\forall k \geq 0$, while the scalar δ_k was computed as a Barzilai-Borwein step-length [5, 25]. Our requirement that Δ_k is fully determined at x_k motivated the use of the retarded Barzilai-Borwein (BB) steps [25, 26], defined as $\delta_k = \left| \frac{d_{k-\widetilde{q}}^T d_{k-\widetilde{q}}}{d_{k-\widetilde{q}}^T z_{k-\widetilde{q}}} \right|$, where $\widetilde{q} = \min\{q, k-1\}$, $q \geq 1$, and $d_{t-1} = x_t - x_{t-1}$, $z_{t-1} = g_t - g_{t-1}$ $t \geq 1$, see [25, Eq. (2.3)]. Setting $q = 1$ yields

$$\delta_k = \left| \frac{d_{k-1}^T d_{k-1}}{d_{k-1}^T z_{k-1}} \right|. \tag{41}$$

Following guidelines on stochastic Barzilai-Borwein steplengths, z_t has to be computed using stochastic gradients with the same mini batch; since this would require an additional gradient evaluation at each iteration, we updated δ_k using (41) every 20 iterations.

In **S2** we considered δ_k as in (41), and the diminishing parameter α_{k+1} of the form

$$\alpha_{k+1} = \frac{a}{a+k} c_k(\gamma_0, \gamma_1), \tag{42}$$

$$c_k(\gamma_0, \gamma_1) = \gamma_1 + 0.5(\gamma_0 - \gamma_1)\Big(1 + \cos\Big(\frac{k\pi}{k_{\max}}\Big)\Big), \tag{43}$$

with $\gamma_0, \gamma_1 > 0$ [27, 28]. The cosine-decay step rule (43) implies $c_0(\gamma_0, \gamma_1) = \gamma_0$, $c_{k_{\max}}(\gamma_0, \gamma_1) = \gamma_1$. The strategy **S3** depends only on $\{\alpha_k\}$ as we fixed $\delta_k = 1, \forall k$. In **S3**, we let α_{k+1} as in (42). Table 1 summarizes the three strategies above.

In our experiments we set $a = 1000$ in (42) and $\gamma_1 = 10^{-5}$ in (43). The parameter α in **S1** and the parameter γ_0 in **S2** and **S3** were varied. We tuned α and γ_0 exploring a set of reasonable values; the results obtained are satisfactory by further performance improvements may be achieved through more extensive hyper-parameter tuning. Our results report on the statistics over ten runs. The x-axis of each figure is iteration count and y-axis is in base-10 logarithmic scale.

5.4 Numerical results

First, we tested PSG_LECO Algorithm using the projection π_S in (6) evaluated via Cholesky factorization of AA^T. We performed our experiments letting α and γ_0 vary in the range $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$.

The inexact version of PSG_LECO Algorithm was applied in the solution of MNIST and MUSHROOMS tests, due to the larger values of m compared to the other datasets. The evaluation of $\widetilde{\pi}_S$ in (8) was performed applying the Conjugate Gradient Method (Matlab built-in function `pcg`) to (7) with stopping criterion (9). The solver `pcg` stops upon meeting the relative residual condition $\|r(y_k)\|/\|Ay_k - b\| < \tau$ or exhausting the maximum iteration limit (ITER). We set ITER $= m$ and let

$$\tau = \max\left\{10^{-10}, \min\left\{\frac{\eta_k\|Ax_k - b\| + \mu_k}{\|Ay_k - b\|}, 10^{-3}\right\}\right\}.$$

In practice, we used two thresholds to avoid too small and too large values of $\|r(y_k)\|$; to save computations, we skipped projection when $\|Ay_k - b\| \leq 10^{-12}$. The experiments were performed setting $\eta_k = 0.5, \forall k$, $\mu_k = \mu_0\rho^k, \forall k$, $\mu_0 = 0.1$ and $\rho = 0.95$. As for the parameters of **S1**–**S3** we tested the same values as in the previous subsection.

In Figures 1 we report selected results, obtained with exact projection π_S, that exhibited relatively better performance across the tested hyper-parameters and show the optimality measure $\|d(x_k)\|$ through the iterations. In Figures 2 we report selected results obtained with the inexact projection $\widetilde{\pi}_S$ and show both the optimality measure $\|d(x_k)\|$ and the infeasibility measure $\|Ax_k - b\|$ through the iterations. The minimum value achieved by $\|d(x_k)\|$, and by $\|Ax_k - b\|$ in case of inexact projection, are reported in the legends for each tested strategy. The runs refer to the hyper-parameters α and γ_0 declared in the figures.

Our experiments indicate that Algorithm PSG_LECO coupled with **S1**, **S2** and **S3** strategies generally shows good progress to optimality in the initial phase of the execution and is not erratic at the final stage. Focusing on Figure 1, we remark that results for HUESTIS dataset is displayed for the first 360 iterates as the rate of convergence is fast and small values $\|d(x_k)\|$ are rapidly achieved; if more iterates are performed, the behavior of Algorithm PSG_LECO with **S2** and **S3** is steady while using **S1** gives rise to some peaks which are recovered thereafter. The strategy **S1** appears to be the most effective in terms of accuracy in the solution of MUSHROOMS test and HUESTIS test. The strategies **S2** and **S3** based on diminishing stepsizes are similar in the decay rate of $\|d(x_k)\|$ except for MUSHROOM test where **S2** outperforms **S3**. Overall, using the stochastic counterpart of the Barzilai-Borwein step appears to be beneficial and the scheme **S2** is preferable to **S1**.

The results concerning the inexact projection $\widetilde{\pi}_S$ displayed in Figure 2 are coherent with the previous observations. The hyper-parameters α and γ_0 are the same as in Figure 2. First, we note that the decrease of the infeasibility is fast and that the decrease of $\|d(x_k)\|$ is not affected by the inexact projection.

Summarizing the results obtained, Algorithm PSG_LECO works well on our test problems. Since it is based on a stochastic gradient methodology, the performance depends on the choice of the step length. Strategies **S1**–**S3** appeared to be effective. In general the adaptive choice of the step length in **S2** yielded lower values of the optimality measure quickly and provided good accuracy.

6 Conclusions

In this work, we proposed a projected stochastic gradient method for minimizing a function subject to deterministic linear constraints. Our method involves a projection map that can

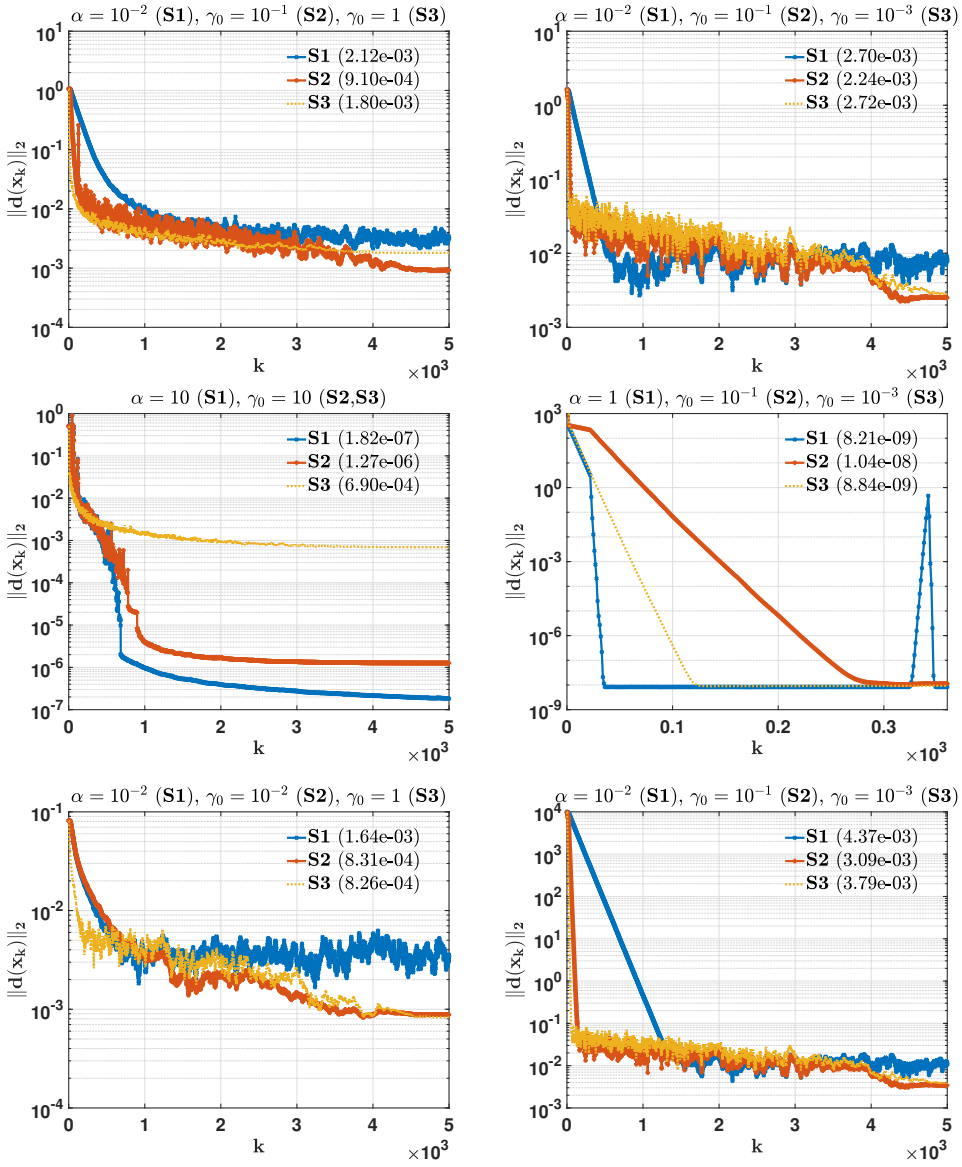

Fig. 1: The average values of $\|d(x_k)\|$ over ten runs using strategies **S1**–**S3**; left column: MNIST (top), MUSHROOMS (middle), DIABETES (bottom), and right column: DTOC1L (top), HUESTIS (middle), and HS50 (bottom)

be evaluated either exactly or inexactly. Theoretical properties depend on the choice of a stepsize related sequence and are equivalent to those established in the unconstrained setting. Numerical illustration of our procedure was presented to show its effectiveness.

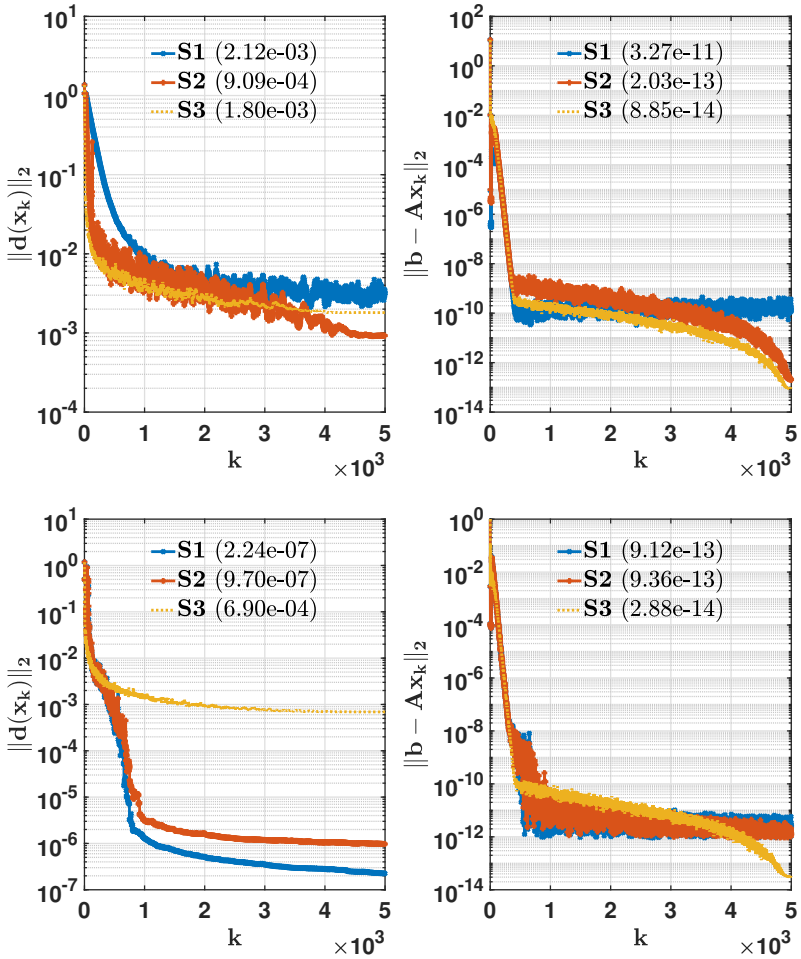

Fig. 2: The average values of $\|d(x_k)\|$ and $\|Ax_k - b\|$ over ten runs using strategies S1–S3 on MNIST (top) and MUSHROOMS (bottom)

Acknowledgments

Natasa Krklec Jerinkić was supported by the Science Fund of the Republic of Serbia, GRANT No 7359, Project LASCADO. The author also gratefully acknowledge the financial support of the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Grants No. 451-03-33/2026-03/ 200125 & 451-03-34/2026-03/ 200125)
Benedetta Morini and Mahsa Yousefi are members of the INdAM Research Group GNCS. The research that led to the present paper was partially supported by INDAM-GNCS through Progetti di Ricerca 2025 and by PNRR - Missione 4 Istruzione e Ricerca - Componente C2 Investimento 1.1, Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN) funded by the European Commission under the NextGeneration EU programme, project "Advanced optimization METhods for automated central

veIn Sign detection in multiple sclerosis from magneTic resonAnce imaging (AMETISTA)", code: P2022J9SNP, MUR D.D. financing decree n. 1379 of 1st September 2023 (CUP E53D23017980001), project "Numerical Optimization with Adaptive Accuracy and Applications to Machine Learning", code: 2022N3ZNAX MUR D.D. financing decree n. 973 of 30th June 2023 (CUP B53D23012670006).

Data availability

The datasets utilized in this research are publicly accessible and commonly employed benchmarks in the field of machine learning and numerical optimization, see [24, 29].

Declarations

Conflict of interest. The authors have no relevant financial or non-financial interests to disclose.

References

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

[2] Tan, C., Ma, S., Dai, Y.-H., Qian, Y.: Barzilai-borwein step size for stochastic gradient descent. Advances in Neural Information Processing Systems **29**, 685–693 (2016)

[3] Krklec Jerinkić, N., Ruggiero, V., Trombini, I.: Spectral stochastic gradient method with additional sampling for finite and infinite sums. Computational Optimization and Applications **91(2)**, 717–758 (2025)

[4] Bellavia, S., Krejić, N., N., K.J., Raydan, M.: Slises: Subsampled line search spectral gradient method for finite sums. Optimization Methods and Software **91(2)**, 1–26 (2024)

[5] Bellavia, S., Morini, B., Yousefi, M.: Fully stochastic trust-region methods with Barzilai-Borwein steplengths. Journal of Computational and Applied Mathematics **476**, 117059 (2026)

[6] Robbins, H., Monro, S.: A stochastic approximation method. SIAM J. Optim **21**, 1109–1140 (1951)

[7] Friedlander, M.P., Schmidt, M.: Hybrid deterministic-stochastic methods for data fitting. SIAM Journal on Scientific Computing **34(3)**, 1380–1405 (2012)

[8] Franchini, G., Porta, F., Ruggiero, V., Trombini, I., Zanni, L.: A stochastic gradient method with variance control and variable learning rate for deep learning. Journal of Computational and Applied Mathematics **451**, 116083 (2024)

[9] Bastin, F., Cirillo, C., Toint, P.L.: An adaptive Monte Carlo algorithm for computing mixed logit estimators. Computational Management Science **3(1)**, 55–79 (2006)

[10] Bellavia, S., Krejić, N., Morini, B., Rebegoldi, S.: A stochastic first-order trust-region method with inexact restoration for finite-sum minimization. Computational Optimization and Applications **84**, 53–84 (2023)

[11] Paquette, C., Scheinberg, K.: A stochastic line search method with expected complexity analysis. SIAM Journal on Optimization **30(1)**, 349–376 (2020)

[12] Curtis, F.E., Scheinberg, K., Shi, R.: A stochastic trust region algorithm based on careful step normalization. Informs Journal on Optimization **1(3)**, 200–220 (2019)

[13] Wang, Q., Piermarini, C., Zhu, Y., Curtis, F.E.: Projected stochastic momentum methods for nonlinear equality-constrained optimization for machine learning. arXiv preprint arXiv:2601.11795 (2026)

[14] Curtis, F., Robinson, D., Zhou, B.: Sequential quadratic optimization for stochastic optimization with deterministic nonlinear inequality and equality constraints. SIAM Journal on Optimization **34**, 3592–3622 (2024)

[15] Berahas, A.S., Curtis, F.E., Robinson, D., Zhou, B.: Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. SIAM Journal on Optimization **31**(3), 1352–1379 (2021)

[16] Fang, Y., Na, S., Mahoney, M.W., Kolar, M.: Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. SIAM Journal on Optimization **34**(2), 2000–2037 (2024)

[17] Krejić, N., Krklec Jerinkić, N., Rapajić, S., Rutešić, L.: IPAS: An adaptive sample size method for weighted finite sum problems with linear equality constraints. arXiv preprint arXiv:2504.19629 (2025)

[18] Birgin, E.G., Martínez, J.M., Raydan, M.: Nonmonotone spectral projected gradient methods on convex sets. SIAM Journal on Optimization **10**(4), 1196–1211 (2000)

[19] Krejić, N., Krklec Jerinkić, N.: Non-monotone line search methods with variable sample size. Numerical Algorithms **68**, 711–739 (2015)

[20] Robbins, H., Siegmund, D.: A convergence theorem for non negative almost supermartingales and some applications. In: Optimizing Methods in Statistics, pp. 233–257. Elsevier, New York (1971)

[21] Bellavia, S., Gratton, S., Morini, B., Toint, P.L.: Fast stochastic second-order Adagrad for nonconvex bound-constrained optimization. arXiv:2505.06374 (2025)

[22] Krejić, N., Krklec Jerinkić, N., Ostojić, T., Vučićević, N.: AS-BOX: Additional sampling method for weighted sum problems with box constraints. arXiv preprint arXiv:2509.00547 (2025)

[23] Krejić, N., Krklec Jerinkić, N., Ostojić, T., Vučićević, N.: Aspen: An additional sampling penalty method for finite-sum optimization problems with nonlinear equality constraints. arXiv preprint arXiv:2508.02299 (2025)

[24] Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST) **2**(3), 1–27 (2011)

[25] Dai, Y.H., Fletcher, R.: Projected Barziali-Borwein methods for large box-constrained quadratic programming. Numer. Math. **100**, 21–47 (2005)

[26] Friedlander, A., Martiínez, J.M., Molina, B., Raydan, M.: Gradient methods with retard and generalizations. SIAM J. Numer. Anal. **36**, 275–289 (1999)

[27] Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)

[28] MathWorks: Cosine learning rate schedule. Deep Learning Toolbox Documentation, available at: https://it.mathworks.com/help/deeplearning/ref/cosinelearnrate.html (2024)

[29] Gratton, S., Toint, P.L.: S2MPJ and CUTEst optimization problems for Matlab, Python and Julia. Optimization Methods and Software, 1–33 (2025)