ASPEN: An Additional Sampling Penalty Method for Finite-Sum Optimization Problems with Nonlinear Equality Constraints

Nataša Krejić[*], Nataša Krklec Jerinkić[*], Tijana Ostojić[†], Nemanja Vučićević[‡§]

August 4, 2025

Abstract

We propose a novel algorithm for solving non-convex, nonlinear equality-constrained finite-sum optimization problems. The proposed algorithm incorporates an additional sampling strategy for sample size update into the well-known framework of quadratic penalty methods. Thus, depending on the problem at hand, the resulting method may exhibit a sample size strategy ranging from a mini-batch on one end, to increasing sample size that achieves the full sample eventually, on the other end of the spectrum. A non-monotone line search is used for the step size update, while the penalty parameter is also adaptive. The proposed algorithm avoids costly projections, which, together with the sample size update, may yield significant computational cost savings. Also, the proposed method can be viewed as a transition of an additional sampling approach for unconstrained and linear constrained problems, to a more general class with non-linear constraints. The almost sure convergence is proved under a standard set of assumptions for this framework, while numerical experiments on both academic and real-data based machine learning problems demonstrate the effectiveness of the proposed approach.

Key words: Constrained optimization, Sample Average Approximation, non-monotone line search, additional sampling, penalty functions, almost sure convergence, KKT point.
MSC Classification: 90C15, 90C30, 65K05, 65K10.

[*]Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia. Emails: natasak@uns.ac.rs, natasa.krklec@dmi.uns.ac.rs

[†]Department of Fundamental Sciences, Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia. Email: tijana.ostojic@uns.ac.rs

[‡]Department of Mathematics and Informatics, Faculty of Sciences, University of Kragujevac, Radoja Domanovića 12, 34000 Kragujevac, Serbia. Email: nemanja.vucicevic@pmf.kg.ac.rs.

[§]Corresponding author.

1 Introduction

We consider a nonlinear equality-constrained optimization problem of the form

$$\min_x f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x), \quad \text{subject to} \quad h(x) = 0, \tag{1.1}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, N$ and $h : \mathbb{R}^n \to \mathbb{R}^m$ are continuously differentiable. Optimization problems of this structure naturally arise in various real-world contexts, particularly in areas such as machine learning, deep learning, and network system optimization. They are commonly encountered when training models such as logistic regression and deep neural networks, where the goal is to minimize a cumulative loss across a dataset while satisfying specific constraints. Owing to both their practical relevance and mathematical complexity, these problems continue to attract significant attention in contemporary research.

A wide range of deterministic algorithms has been developed for solving equality-constrained nonlinear optimization problems. One of the frequently used approaches is based on penalty methods and augmented Lagrangian techniques [13]. In these methods one reformulates the constrained problem by adding a penalty term to the objective function to penalize constraints violation, governed by a penalty parameter, thereby transforming the problem into an unconstrained one. Penalty methods can also be effectively applied in stochastic environments. In the stochastic case, when either the objective function or the constraints are defined in terms of expectations, several methods are proposed for solving such problems by penalty-based techniques [19, 24, 33, 34].

Stochastic approaches that use stochastic approximations of the gradient have been proposed to reduce computational cost per iteration, leading to various stochastic penalty and projection methods [28, 31, 32]. Also, stochastic gradient methods have found wide application in classification problems, especially in recent years, with variable metric methods emerging as one of the approaches for preconditioning the stochastic gradient, as presented in the work [7].

Another popular class of algorithms for the considered problems is based on Sequential Quadratic Programming (SQP). In the stochastic setting, the search direction is typically generated by solving a quadratic subproblem by using a stochastic gradient estimate [4], and the convergence complexity of the method has been studied in [9]. Various extensions have been proposed, including adaptive sampling techniques [2], variance reduction [5], and structural improvements [3, 8, 29, 30]. In [11], the TR-StoSQP algorithm is introduced, which combines a trust region approach with adaptive trust region radii and allows indefinite Hessians in the subproblem.

In large-scale finite-sum minimization problems, exact evaluation of the objective and its derivatives is computationally expensive. Hence, approximation via subsampling is widely used. Adaptive subsampling - adjusting sample size during iterations - has proven to be an efficient strategy when coupled with line search or trust region mechanisms [1, 17, 18, 21, 23]. Some methods employ additional sampling to decide whether to increase the sample size. Moreover, additional sampling is also used to govern the acceptance of a candidate point, and this approach proved to be very efficient for unconstrained finite-sum minimization [10, 21, 26, 27].

In [22], a stochastic first-order method with variable sample size was proposed for minimizing a weighted finite sum under linear equality constraints, using adaptive sampling and approximate projections. The method proposed here generalizes this approach to nonlinear equality constraints, but does not rely on the approximate projection strategy used in [22].

The non-monotone line search we propose is presented in [14] and is used in many methods relax the strict Armijo line search conditions and enhance convergence speed. In the stochastic framework, non-monotone strategies have also proved their efficiency. Paper [20] proposed a class of algorithms with adaptive sample size combined with non-monotone line search. Non-monotonicity has also been employed in stochastic optimization settings in [21, 22, 25].

Solving nonlinear equality-constrained problems with finite-sum structure is challenging due to the high cost of evaluating the objective and its derivatives at each iteration. While penalty methods and SQP-type algorithms have been successfully applied in this setting, most of the existing methods either assume access to full gradients or rely on expensive projection steps. Motivated by these limitations, we designed an efficient first-order algorithm that uses subsampling, avoids expensive projections and the strict decrease imposed by the standard Armijo line search.

To summarize, our main contributions are as follows:

- We propose a novel first-order algorithm (ASPEN) for nonlinear equality-constrained finite-sum problems that combines adaptive sampling, non-monotone line search, and adaptive penalty parameter update;
- ASPEN extends the adaptive sampling framework of [22] to a more general class of problems with nonlinear constraints, while avoiding the (approximate) projections;
- The theoretical analysis of the proposed method is presented - almost sure convergence of ASPEN is proved under some standard assumptions for the considered framework;
- The practical efficiency of ASPEN is demonstrated by comparing it to the relevant state-of-the-art methods and by illustrating its adaptive nature.

Paper organization. The paper is organized as follows. In Section 2, we present the proposed algorithm. Section 3 contains theoretical analysis and convergence results. Section 4 presents numerical experiments. We conclude the paper in Section 5 with a summary and directions for future research.

Notation. Throughout the paper, we use the following notation: $\mathbb{R}_+$ denotes the set of non-negative real numbers. Depending on the argument, the symbol $\|\cdot\|$ represents the Euclidean vector norm and the spectral matrix norm. We use "a.s." to abbreviate "almost sure"/"almost surely". $\mathbb{E}(\cdot)$ and $\mathbb{E}(\cdot \mid \mathcal{F})$ denote mathematical expectation and conditional expectation with respect to the σ-algebra $\mathcal{F}$, respectively. Finally, for a finite set N, $|N|$ denotes its cardinality.

2 The algorithm

The method that will be proposed within this section - ASPEN - belongs to a class of first-order methods, i.e., we assume that only the relevant functions and their first-order derivatives are attainable. Given that the full sample size N (the number of data points in machine learning problems) is typically and the data sets often include some data redundancy, we employ a subsampling strategy to reduce the computational load. More precisely, we use the standard Sample Average Approximations (SAA) to form the objective function approximations. Furthermore, we use the same sample to calculate the gradient approximation, i.e., we take

$$f_{\mathcal{N}_k}(x) = \frac{1}{N_k} \sum_{i \in \mathcal{N}_k} f_i(x), \;\; \nabla f_{\mathcal{N}_k}(x) = \frac{1}{N_k} \sum_{i \in \mathcal{N}_k} \nabla f_i(x), \tag{2.1}$$

where $\mathcal{N}_k \subseteq \mathcal{N} := \{1, 2, ..., N\}$ and $N_k = |\mathcal{N}_k|$. We emphasize that the choice of sample $\mathcal{N}_k$ is arbitrary, i.e., we do not make any restrictions on the sampling strategy for $\mathcal{N}_k$.

ASPEN fits into the framework of penalty methods. In particular, we use the quadratic penalty function, which takes the following form when considering the original problem (1.1)

$$F(x, \mu) = f(x) + \frac{\mu}{2} ||h(x)||^2 =: f(x) + \mu q(x).$$

Here, the parameter μ is the penalty parameter, and by q we denote the constraints violation function defined by the square norm of h. We assume that the constraints are non-linear in general, but computable under reasonable costs as well as their first-order derivatives. Therefore, we form an approximation of the penalty function as follows

$$F_{\mathcal{N}_k}(x, \mu_k) = f_{\mathcal{N}_k}(x) + \frac{\mu_k}{2} ||h(x)||^2,$$

where μ_k represents the penalty parameter used at iteration k. We define the gradient[1] of the approximate (SAA) penalty function at iteration k accordingly, i.e.,

$$g_k := \nabla F_{\mathcal{N}_k}(x_k, \mu_k) = \nabla f_{\mathcal{N}_k}(x_k) + \mu_k \nabla^T h(x_k) h(x_k). \tag{2.2}$$

Since we work with approximate functions, applying a monotone line search may be too restrictive and ineffective. Therefore, we apply a nonmonotone Armijo-type line search similar to one used in [22], for instance. Setting the search direction to $p_k = -g_k$, we perform the backtracking line search with respect to the approximate penalty function. More precisely, we determine the step size α_k that satisfies the following condition

$$F_{\mathcal{N}_k}(x_k + \alpha_k p_k, \mu_k) \leq F_{\mathcal{N}_k}(x_k, \mu_k) + \eta \alpha_k g_k^T p_k + \epsilon_k, \tag{2.3}$$

where $\{\epsilon_k\}_{k \in \mathbb{N}}$ is assumed to be a sequence of summable positive numbers, i.e., it satisfies

$$\sum_{k=0}^{\infty} \epsilon_k < \infty.$$

After the line search, we set $\bar{x}_k = x_k + \alpha_k p_k$ to be the candidate point for the subsequent iterate.

As mentioned before, depending on the problem, ASPEN may use subsampling during the whole optimization process (i.e., we can have $N_k < N$ for all $k \in \mathbb{N}$), or it can reach the full sample size (i.e., $N_k = N$ for all k large enough) and switch to the deterministic mode. We will refer to the first scenario as mini-batch (MB), while the latter one will be referred to as the full sample (FS) scenario. We assume that, if $N_k = N$, then $\mathcal{N}_k = \mathcal{N}$, i.e., we use the whole set of local cost functions f_i (i.e., the whole set of data points). If ASPEN is in the FS mode at iteration k (i.e., if $N_k = N$), then the candidate point is unconditionally accepted, and we set $x_{k+1} = \bar{x}_k$. Moreover, since the sample size sequence is nondecreasing, the method behaves like a deterministic[2] sequential programming penalty method - when a vicinity of a stationary point of the penalty function $F(\cdot, \mu_k)$ is reached, the penalty parameter is increased to μ_{k+1} and an approximate solution $\min_{x \in \mathbb{R}^n} F(x, \mu_{k+1})$ is computed. The novelty of ASPEN lies in the MB phase that we describe in detail as follows. If ASPEN is in the MB phase at iteration k ($N_k < N$), then an additional check is performed to decide whether to accept the candidate point or to reject the step completely. This check is based on additional sampling [10, 21, 27], which also guides the sample size update. Namely, we use an independent, arbitrarily small subsample

[1] The gradient will always be taken with respect to variable x, i.e., $\nabla F_{\mathcal{N}_k}(x, \mu_k) := \nabla_x F_{\mathcal{N}_k}(x, \mu_k)$. We drop x from the subscript in order to simplify the notation.

[2] The method still yields a stochastic sequence of iterates since in the FS scenario there exists a finite, but random iteration $\bar{k}$ such that $N_k = N$ for all $k \geq \bar{k}$.

$\mathcal{D}_k \subseteq \mathcal{N}$, which is selected randomly, without replacement, and check if the candidate point is "good enough" for the function

$$F_{\mathcal{D}_k}(x_k, \mu_k) := \frac{1}{D_k} \sum_{i \in \mathcal{D}_k} f_i(x_k) + \frac{\mu_k}{2} ||h(x_k)||^2,$$

where $D_k = |\mathcal{D}_k|$. More precisely, we check if

$$F_{\mathcal{D}_k}(\overline{x}_k, \mu_k) \leq F_{\mathcal{D}_k}(x_k, \mu_k) - c\|\nabla F_{\mathcal{D}_k}(x_k, \mu_k)\|^2 + C\epsilon_k, \tag{2.4}$$

where c and C are positive constants, arbitrarily small and arbitrarily large, respectively. This is an important step because it allows for evaluating the algorithm's progress on independent data, with low computational cost (in the experiments we use $D_k = 1$, but any other choice such that $D_k \leq N_k$ is eligible). If the condition (2.4) is satisfied, we accept the candidate point; otherwise, we set $x_{k+1} = x_k$.

The additional check also serves as a test of the similarity of local cost functions (see e.g. [21] for more details). If the condition (2.4) is met, then we decide that the current approximation is sufficient to describe the true objective function and keep the same sample size for the subsequent iteration. On the contrary, if the condition (2.3) does not hold, we try to increase the level of precision and obtain a better approximation of the objective function by increasing the sample size to $N_{k+1} \in \{N_k + 1, \ldots, N\}$. This increase is arbitrary, which makes ASPEN very flexible and adaptive to various types of problems.

The penalty parameter in the MB phase is updated according to the constraint violation measure $\|h(x_k)\|$. If the current iterate is relatively far away from the feasible set, i.e., if $\|h(x_k)\| > \epsilon_k$, we increase the penalty parameter in order to encourage feasibility improvement. Otherwise, we keep it at the same level.

The algorithm is stated as follows.

Algorithm 1 ASPEN:
Additional **S**ampling **P**enalty method for **E**quality **N**onlinear constraints

Input: $x_0 \in \mathbb{R}^n$, $N_0 \in \mathbb{N}$, $c, \beta \in (0,1)$, C, γ, $\mu_0 > 0$, $\{\epsilon_k\}_{k=0}^{\infty}$.
For $k = 0, 1, 2, ...$
if $N_k < N$ **then**
    choose $\mathcal{N}_k \subseteq \mathcal{N}$
else
    set $\mathcal{N}_k = \mathcal{N}$.
end if
Calculate $p_k = -g_k$ via (2.2).
Find the smallest $j \in N_0$ such that $\alpha_k = \beta^j$ satisfies (2.3).
Set $\overline{x}_k = x_k + \alpha_k p_k$.
if $N_k = N$ **then**
    Set $x_{k+1} = \overline{x}_k$
    if $\|\nabla F_{\mathcal{N}_k}(x_k, \mu_k)\| < \frac{1}{\mu_k}$ **then**
        Set $\mu_{k+1} = \gamma \mu_k$.
    else
        Set $\mu_{k+1} = \mu_k$.
    end if
else
    Choose $\mathcal{D}_k \subseteq \mathcal{N}$ randomly and uniformly, without replacement.
    if (2.4) holds **then**
        Set $x_{k+1} = \overline{x}_k$ and $N_{k+1} = N_k$.
    else
        Set $x_{k+1} = x_k$, and choose $N_{k+1} \in \{N_k + 1, \ldots, N\}$.
    end if
    if $\|h(x_k)\| > \epsilon_k$ **then**
        Set $\mu_{k+1} = \gamma \mu_k$.
    else
        Set $\mu_{k+1} = \mu_k$.
    end if
end if

3 Convergence analysis

Within this section, we prove almost sure convergence results for ASPEN. The analysis is conducted by observing two possible scenarios (MB and FS) separately, and integrating them at the end. We start by stating the following assumption.

Assumption A 1. *The functions f_i, $i = 1, \dots, N$ are bounded from below and continuously differentiable with L-Lipschitz continuous gradients. Moreover, the constraints violation function q is continuously differentiable with L-Lipschitz continuous gradient.*

Notice that this assumption implies that all the sampled functions $f_{\mathcal{N}_k}$ are also bounded from below and continuously differentiable with L-Lipschitz continuous gradients. Notice that the penalty functions $F_{\mathcal{N}_k}$ are also bounded from below and continuously differentiable. Moreover, the gradients $\nabla F_{\mathcal{N}_k}$ are also Lipschitz-continuous with the Lipschitz constant $(1 + \mu_k)L$. Since the line search is performed with the negative gradient direction of the approximate penalty function, it can be shown that

$$\alpha_k \geq \min\left\{1, \frac{2\beta(1-\eta)}{(1+\mu_k)L}\right\} =: \bar{\alpha}_{\mu_k}. \tag{3.1}$$

Now, let us consider the MB phase of ASPEN. As usual for an additional sampling-based method, we define the following sets needed for the analysis. By $\mathcal{D}_k^+$ we denote the set of all possible outcomes of $\mathcal{D}_k$ at iteration k that satisfy (2.4), i.e.,

$$\mathcal{D}_k^+ = \{\mathcal{D}_k \subseteq \mathcal{N} \mid F_{\mathcal{D}_k}(\overline{x}_k, \mu_k) \leq F_{\mathcal{D}_k}(x_k, \mu_k) - c\|\nabla F_{\mathcal{D}_k}(x_k, \mu_k)\|^2 + C\epsilon_k\},$$

and by $\mathcal{D}_k^-$ be the complementary subset, i.e.,

$$\mathcal{D}_k^- = \{\mathcal{D}_k \subseteq \mathcal{N} \mid F_{\mathcal{D}_k}(\overline{x}_k, \mu_k) > F_{\mathcal{D}_k}(x_k, \mu_k) - c\|\nabla F_{\mathcal{D}_k}(x_k, \mu_k)\|^2 + C\epsilon_k\}.$$

The following lemma shows that, if ASPEN stays in the MB phase during the whole optimization process, then the condition (2.4) is satisfied for all possible choices of $\mathcal{D}_k$. The proof of this lemma is essentially the same as the proof of Lemma 1 [21], but we state it here for completeness.

Lemma 3.1. *Suppose that Assumption A1 holds. If $N_k < N$ for all $k \in \mathbb{N}$, then a.s. there exists $k_1 \in \mathbb{N}$ such that $\mathcal{D}_k^- = \emptyset$ for all $k \geq k_1$.*

Proof. Let us assume, aiming for contradiction, that the lemma does not hold. Then there exists an infinite set of indices $K \subseteq \mathbb{N}$ such that $\mathcal{D}_k^- = \emptyset$ for all $k \in K$. Since the sample size sequence is non-decreasing and satisfies $N_k < N$ for all $k \in \mathbb{N}$ by the assumption of this lemma, there must exist some $k_1 \in \mathbb{N}$ such that $N_k = \bar{N} < N$ for all $k \geq k_1$. This implies that the

number of elements in the subset $\mathcal{D}_k$ satisfies $D_k \leq N_k \leq \bar{N} \leq N-1$. Since $\mathcal{D}_k$ is selected randomly and uniformly from a finite collection, there exists a constant $q > 0$ such that

$$\mathbb{P}(\mathcal{D}_k \in D_k^-) \geq q \quad \text{for all } k \in K.$$

Moreover, without loss of generality, we may assume that $K \subseteq \{k \in \mathbb{N} : k \geq k_1\}$, i.e., every $k \in K$ satisfies $k \geq k_1$. Therefore, choosing $\mathcal{D}_k \in D_k^+$ for all $k \in K$ is a.s. impossible since

$$\mathbb{P}(\mathcal{D}_k \in \mathcal{D}_k^+, k \in K) \leq \prod_{k \in K} (1-q) = 0.$$

This means that a.s. there exists an iteration $\tilde{k} \in K$ such that

$$\mathcal{D}_{\tilde{k}} \in \mathcal{D}_{\tilde{k}}^-.$$

According to the algorithm, this implies that $N_{\tilde{k}+1} > N_{\tilde{k}} = \bar{N}$, contradicting the assumption that $N_k = \bar{N}$ for all $k \geq k_1$. Hence, the statement holds. □

Again, following the concept of additional sampling, one can show that the Armijo-like condition related to the full sample penalty function is satisfied for all k large enough, even in the MB scenario.

Lemma 3.2. *Suppose that the assumptions of Lemma 3.1 hold. Then, a.s. the following holds for all $k \geq k_1$*

$$F(x_{k+1}, \mu_k) \leq F(x_k, \mu_k) - c\|\nabla F(x_k, \mu_k)\|^2 + C\epsilon_k.$$

Proof. First, recall that Lemma 3.1 implies that a.s. for all $k \geq k_1$ we have $D_k^- = \emptyset$ and thus the condition (2.4) is satisfied. Therefore, the candidate point is accepted for all k large enough, i.e., $x_{k+1} = \bar{x}_k$ for all $k \geq k_1$. Moreover, notice that $D_k^- = \emptyset$ also implies[3] that for each $i \in \mathcal{N}$ we have

$$F_i(x_{k+1}, \mu_k) \leq F_i(x_k, \mu_k) - c\|\nabla F_i(x_k, \mu_k)\|^2 + C\epsilon_k, \tag{3.2}$$

where

$$F_i(x, \mu_k) := f_i(x_k) + \frac{\mu_k}{2}||h(x)||^2.$$

Thus, considering (3.2), by summing up and dividing by N we obtain that

$$F(x_{k+1}, \mu_k) \leq F(x_k, \mu_k) - c\sum_{i=1}^{N} \|\nabla F_i(x_k, \mu_k)\|^2 + C\epsilon_k \tag{3.3}$$

[3]Since $\mathcal{D}_k$ is chosen uniformly, with replacement, one possible choice for $\mathcal{D}_k$ is $\{i, \ldots, i\}$. If there exists i that violates the condition (3.2), then the condition (2.4) would be violated for $\mathcal{D}_k = \{i, ..., i\}$ which would imply that $D_k^- \neq \emptyset$.

for all $k \geq k_1$. Finally, using the convexity of the norm, we obtain

$$\|\nabla F(x_k, \mu_k)\|^2 = \|\frac{1}{N}\sum_{i=1}^{N} \nabla F_i(x_k, \mu_k)\|^2 \leq \frac{1}{N}\sum_{i=1}^{N} \|\nabla F_i(x_k, \mu_k)\|^2.$$

Combining this with (3.3), we conclude the proof. □

Before stating the next theorem, let us denote by $\mathbb{E}_{MB}(\cdot)$ and $\mathbb{E}_{FS}(\cdot)$ the conditional expectation concerning all the sample paths of ASPEN that fall into the MB and FS scenario, respectively. Similarly, we define the corresponding conditional probabilities $\mathbb{P}_{MB}(\cdot)$ and $\mathbb{P}_{FS}(\cdot)$. The proof is essentially the same as the proof of Theorem 1 in [27] e.g., but we state its short version for completeness.

Theorem 3.3. *Suppose that Assumption A1 holds and that the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *generated by ASPEN is bounded. Then*

$$\mathbb{P}_{MB}(\lim_{k\to\infty} \nabla F(x_k, \mu_k) = 0) = 1.$$

Proof. Since we observe only the MB scenario sample paths, Lemma 3.2 implies that a.s. the following holds for each $l \in \mathbb{N}$

$$F(x_{k_1+l}, \mu_{k_1+l}) \leq F(x_{k_1}, \mu_{k_1}) - c\sum_{j=0}^{l-1} \|\nabla F(x_{k_1+j}, \mu_{k_1+j})\|^2 + C\sum_{j=0}^{l-1} \epsilon_{k_1+j}.$$

Due to summability of ϵ_k and boundedness of iterates, applying the conditional expectation $\mathbb{E}_{MB}(\cdot)$ and letting $l \to \infty$ we obtain

$$\sum_{j=0}^{\infty} \mathbb{E}_{MB}(\|\nabla F(x_{k_1+j}, \mu_{k_1+j})\|^2) < \infty$$

and the result follows from the extended form of Markov's inequality and the Borel-Cantelli lemma. □

Next, we prove a.s. convergence towards a KKT point in the MB case by considering different scenarios with respect to the penalty parameter sequence. The proof is conducted under the Linear Independence Constraint Qualification (LICQ) assumption. The second part of the proof, considering unbounded μ_k follows the same steps as the deterministic quadratic penalty method analysis, but we state it here for completeness.

Theorem 3.4. *Let assumptions of Theorem 3.3 hold and assume that* $N_k < N$ *for every* $k \in \mathbb{N}$. *Then, a.s., every accumulation point of the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *at which LICQ holds is a KKT point of problem* (1.1).

Proof. Recall that Theorem 3.3 implies that $\lim_{k\to\infty} \nabla F(x_k, \mu_k) = 0$ a.s. Let x^* be an arbitrary accumulation point of the considered sequence, i.e., let

$$\lim_{k\in K} x_k = x^*.$$

Under the MB scenario, we distinguish two possible cases regarding the penalty parameter - bounded and unbounded μ_k.

First, let us assume that μ_k is bounded. Since the sequence of penalty parameters is non-decreasing, this further implies the existence of $\bar{\mu}$ such that $\mu_k = \bar{\mu}$ for all k large enough. According to lines 25-29 of ASPEN, for all k large enough we have $\mu_{k+1} = \mu_k$ and $\|h(x_k)\| \leq \epsilon_k$. Since $\{\epsilon_k\}_{k\in\mathbb{N}}$ is assumed to be summable, we know that $\lim_{k\to\infty} \epsilon_k = 0$, which further implies

$$\lim_{k\to\infty} h(x_k) = 0.$$

Thus, each accumulation point of $\{x_k\}_{k\in\mathbb{N}}$ is feasible and we have $h(x^*) = 0$ and therefore

$$0 = \lim_{k\in K} \nabla F(x_k, \mu_k) = \lim_{k\in K} (\nabla f(x_k) + \mu_k \nabla^T h(x_k) h(x_k)) = \nabla f(x^*),$$

holds a.s. and the statement is proved.

Now, let us consider the case where $\lim_{k\to\infty} \mu_k = \infty$. Then, according to (2.2) we have

$$\|\nabla^T h(x_k) h(x_k)\| \leq \frac{1}{\mu_k} (\|\nabla F(x_k, \mu_k)\| + \|\nabla f(x_k)\|)$$

and taking the limit over K we obtain

$$\|\nabla^T h(x^*) h(x^*)\| = 0$$

and the LICQ condition implies $h(x^*) = 0$. Moreover, LICQ also implies that $\nabla h(x^*) \nabla^T h(x^*)$ is non-singular and, due to continuity, $\nabla h(x_k) \nabla^T h(x_k)$ is also non-singular for each $k \in K$ large enough. Thus, considering (2.2) again and defining $\lambda_k := \mu_k h(x_k)$ we obtain

$$\lambda_k = (\nabla h(x_k) \nabla^T h(x_k))^{-1} \nabla h(x_k) (\nabla F(x_k, \mu_k) - \nabla f(x_k)) \tag{3.4}$$

for all k large enough and a.s.

$$\lim_{k\in K} \lambda_k = -(\nabla h(x^*) \nabla^T h(x^*))^{-1} \nabla h(x^*) \nabla f(x^*) =: \lambda^*. \tag{3.5}$$

Thus, a.s.,

$$0 = \lim_{k\in K} \nabla F(x_k, \mu_k) = \nabla f(x^*) + \nabla^T h(x^*) \lambda^*, \tag{3.6}$$

which together with $h(x^*) = 0$ implies that x^* is a KKT point. □

Next, we prove a.s. convergence for the FS scenario.

Theorem 3.5. *Suppose that Assumption A1 holds and that the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *generated by ASPEN is bounded. Moreover, suppose that there exists* $\tilde{k}$ *such that* $N_k = N$ *for all* $k \geq \tilde{k}$*. Then, a.s., there exists an accumulation point of the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *which is a KKT point of problem* (1.1) *provided that LICQ holds at that point.*

Proof. Let us consider iterations $k \geq \tilde{k}$. Then, $\mathcal{N}_k = \mathcal{N}$ and the line search implies

$$F(x_{k+1}, \mu_k) \leq F(x_k, \mu_k) - \eta\alpha_k \|\nabla F(x_k, \mu_k)\|^2 + \epsilon_k.$$

Notice that in the FS phase, the penalty parameter μ_k is increased only if $\|\nabla F(x_k, \mu_k)\| < 1/\mu_k$. On the other hand, if μ_k is kept fixed on some $\bar{\mu}$, from (3.1) and the line search we obtain

$$F(x_{k+1}, \bar{\mu}) \leq F(x_k, \bar{\mu}) - \eta\bar{\alpha}_{\bar{\mu}} \|\nabla F(x_k, \bar{\mu})\|^2 + \epsilon_k$$

and, due to boundedness of iterates and summability of ϵ_k, we conclude that $\|\nabla F(x_k, \bar{\mu})\|^2$ tends to zero. This further implies that after a finite number of iterations we will have $\|\nabla F(x_k, \bar{\mu})\| < 1/\bar{\mu}$. Therefore, we conclude that the penalty parameter cannot be fixed for an infinite number of iterations. In fact, it must be increased infinitely many times, and thus $\lim_{k\to\infty} \mu_k = \infty$ holds. Moreover, we conclude that there exists a subset of iterations $\tilde{K}$ such that $\|\nabla F(x_k, \mu_k)\| < 1/\mu_k$ for all $k \in \tilde{K}$ which further implies that

$$\lim_{k\in\tilde{K}} \nabla F(x_k, \mu_k) = 0.$$

Since the sequence of iterates is bounded, there exist x^* and $\tilde{K}_1 \subseteq \tilde{K}$ such that $\lim_{k\in\tilde{K}_1} x_k = x^*$ and following the steps of the second part of the proof of Theorem 3.4 we obtain the result. □

Finally, considering both possible scenarios (FS and MB), we obtain the main result for the ASPEN method.

Theorem 3.6. *Suppose that Assumption A1 holds and that the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *generated by ASPEN is bounded. Then, a.s., there exists an accumulation point of the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *which is a KKT point of problem* (1.1) *provided that LICQ holds.*

4 Numerical results

In this section, we evaluate the performance of the proposed method on real-data machine learning tasks, considering several binary classification datasets from LIBSVM collection [6] listed in Table 1. Moreover, to provide further insights into the behavior of the proposed method, we also consider

an academic problem (HS24) from the CUTEst collection [12] and modify it by adding different levels of noise. Other benchmark datasets frequently referenced in the literature include those from the UCI Machine Learning Repository [15] and the MNIST handwritten digit database [16].

We begin our numerical study by demonstrating that ASPEN offers benefits compared to the deterministic approach, where $N_k = N$ for all $k \in \mathbb{N}$. We call this method "Full" since it uses the full sample during the whole optimization process. More precisely, Full follows the standard penalty approach by solving each subproblem approximately - until $\|\nabla F_{\mathcal{N}_k}(x_k, \mu_k)\| < 1/\mu_k$ is satisfied, and then it increases the penalty parameter by $\mu_{k+1} = \gamma\mu_k$. Furthermore, we also compare ASPEN to a heuristic method named "Heur" which starts with a subsample of size N_0 and increases it by $N_{k+1} = \min\{\lceil 1.1N_k \rceil, N\})$ whenever $\|\nabla F_{\mathcal{N}_k}(x_k, \mu_k)\| < 1/\mu_k$ happens. The penalty parameter is updated as in the Full method, and the same non-monotone line search is applied for solving the subproblems for all the methods mentioned above.

The parameters of ASPEN are the following. The starting penalty parameter is $\mu_0 = 1$, while the initial sample size is $N_0 = \lceil 0.01N \rceil$. The additional sampling is applied with $D_k = 1$, $c = 10^{-4}$ and $C = 1$, while the line search is performed with $\beta = 0.1$, $\eta = 10^{-4}$ and $\epsilon_k = k^{-1.1}$. When needed, the sample size of ASPEN is increased by one, i.e., $N_{k+1} = N_k + 1$ in line 23 of ASPEN. Heur uses the same starting values for the penalty parameter and the sample size. We set $\gamma = 1.1$ for all the considered methods. Starting points x_0 are equal for all the considered methods and are obtained by normalizing random vectors with a Gaussian distribution.

Table 1: Binary classification data set details [6].

Dataset	**Dimension** (n)	**Datapoints** (N)
a9a	123	32,561
Australian	14	690
Heart	13	270
Mushrooms	112	8,124
Splice	60	1,000
MNIST	784	60000

We consider constrained logistic regression binary classification problems as in [5], commonly used in machine learning applications. The form is the following

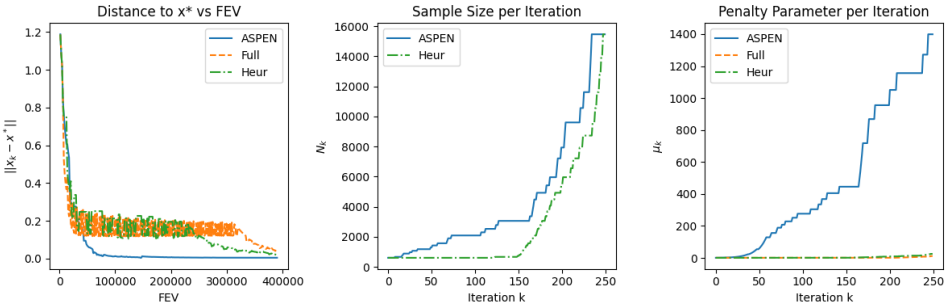
$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + e^{-b_i a_i^\top x}\right) \quad \text{s.t.} \quad \|x\|_2^2 = 1, \tag{4.1}$$

where $a_i \in \mathbb{R}^n$ and $b_i \in \{-1, 1\}$ represent the attributes and the corresponding label for the i-th data point, respectively. We model the computational

cost by FEV_k - the number of scalar products required by the specified method to compute x_k, starting from the initial point x_0.

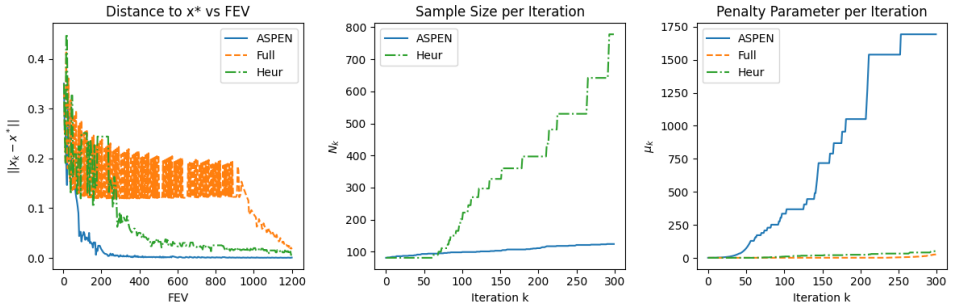
To evaluate the performance of the considered methods, we show:
1) the distance between x_k and the solution x^* of the considered problem, i.e., $||x_k - x^*||$, against computational cost measure FEV_k in graphs a);
2) the sample size behavior across iterations, graphs b);
3) the penalty parameter update across iterations, graphs c).
The results for all the datasets from Table 1 in Figures 1-6 are presented.

The results show that ASPEN manages to outperform Full and Heur on most of the datasets, especially in light of achieving a better vicinity of the solution with significantly lower computational costs. The results also confirm the adaptive nature of ASPEN, especially when the sample size is considered. The graphs b) reveal that the sample size is increased according to the problem at hand, therefore highlighting ASPEN's data-driven adaptivity. For instance, the MNIST dataset (Fig. 1) requires faster sample size growth to cope with the diversity of the data, while the Mushrooms dataset (Fig. 2) obviously contains more similar data points, which allows good approximate solutions even under modest sample sizes which practically fall into a mini-batch framework. Interestingly, the full sample is not reached in any of the considered problems. Furthermore, as expected, the penalty parameter is increased more rapidly for ASPEN than for the other two methods (graphs c), but according to the optimality gap, it seems to be beneficial - it allows the algorithm to progressively enforce feasibility while maintaining efficiency.



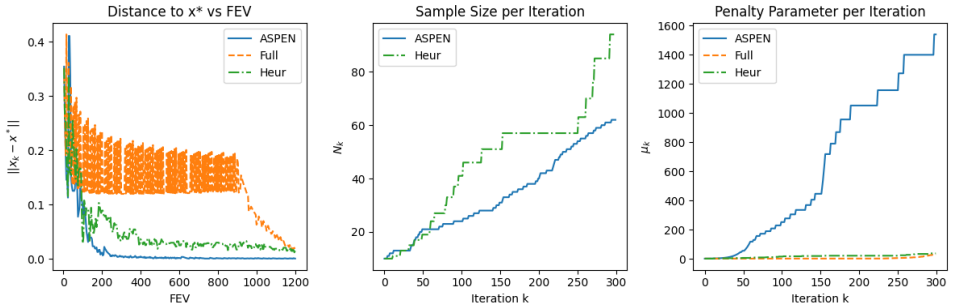
a) b) c)

Figure 1: *MNIST* dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).



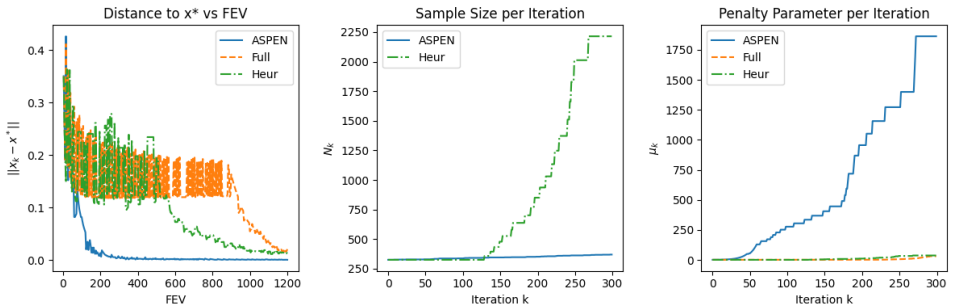
a) b) c)

Figure 2: *Mushrooms* dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).



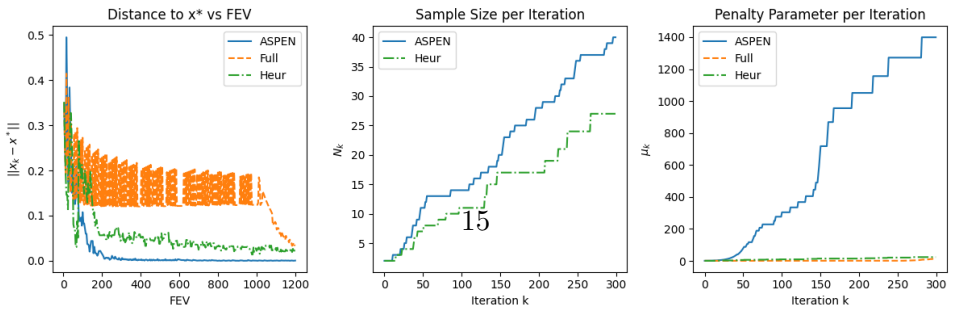
a) b) c)

Figure 3: *Splice* dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).



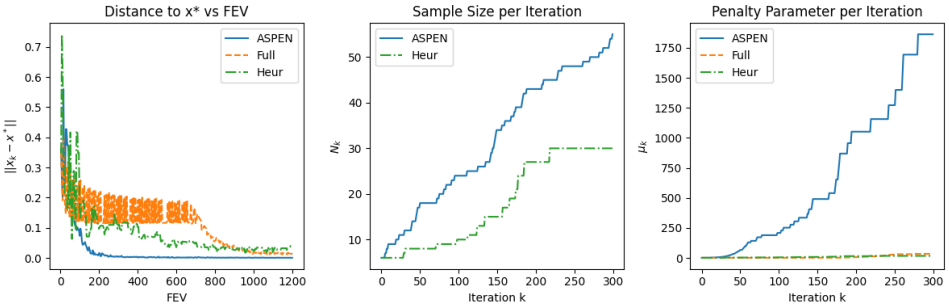
a) b) c)

Figure 4: *a9a* dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).



a) b) c)

Figure 5: *Heart* dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).



a) b) c)

Figure 6: *Australian* dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).

In the sequel, we compare ASPEN to the state-of-the-art stochastic optimization methods of the relevant framework: Sto-SQP[4] and SVR-STO[5]. While the ASPEN keeps the same parameter settings as in the previous experiments, the selected parameters for the competing methods are based on empirical tuning and the recommendations from the literature. After validation, the values were fixed and consistently applied across all experiments, as was done in [5]. In detail, the setup is the following: Sto-SQP uses $\theta = 10^4$, $\tilde{\tau}_{-1} = 0.1$, $\epsilon_\tau = 10^{-6}$, $\tilde{\xi}_{-1} = 0.1$, $\epsilon_\xi = 10^{-2}$, $\sigma_{Sto-SQP} = 0.5$; SVR-STO is employed with $\sigma_{SVR-STO} = 0.5$, $\theta = 10^4$, $\tilde{\tau}_{-1,0} = 0.1$, $\epsilon_\tau = 10^{-6}$, $\alpha_u = 10^6$, $\beta_{SVR-STO} = 1$. Considering Figure 7, ASPEN shows to be competitive with the state-of-the-art methods, outperforming them on most of the considered datasets. The detailed analysis is provided below.

Figure 7 a) – *Mushrooms dataset.* ASPEN's optimality gap drops down sharply after only a few FEVs and continues to decline throughout the interval. STO-SQP shows a stable, monotone decrease but with a much smaller slope, indicating a higher computational cost for the same accuracy. SVR-SQP with $b = 16$ converges more slowly than STO-SQP, whereas the SVR-SQP with $b = 128$ yields only negligible improvement within the available FEV budget.

Figure 7 b) – *a9a dataset.* ASPEN is the only method achieving a pronounced error reduction; after a brief plateau, it resumes decreasing, underscoring the effectiveness of its adaptive scheme. STO-SQP steadily lowers the distance, albeit at a moderate rate. Both SVR-SQP variants display limited convergence, suggesting that this dataset would require more iterations or a different batch-size schedule.

Figure 7 c) – *Australian dataset.* ASPEN again attains the largest error reduction and maintains the lowest optimality gap across the entire FEV horizon. STO-SQP provides a uniform but slower decrease. SVR-SQP with $b = 16$ accelerates in the later phase and approaches STO-SQP's accuracy, highlighting the stochastic variant's sensitivity to the data's statistical prop-

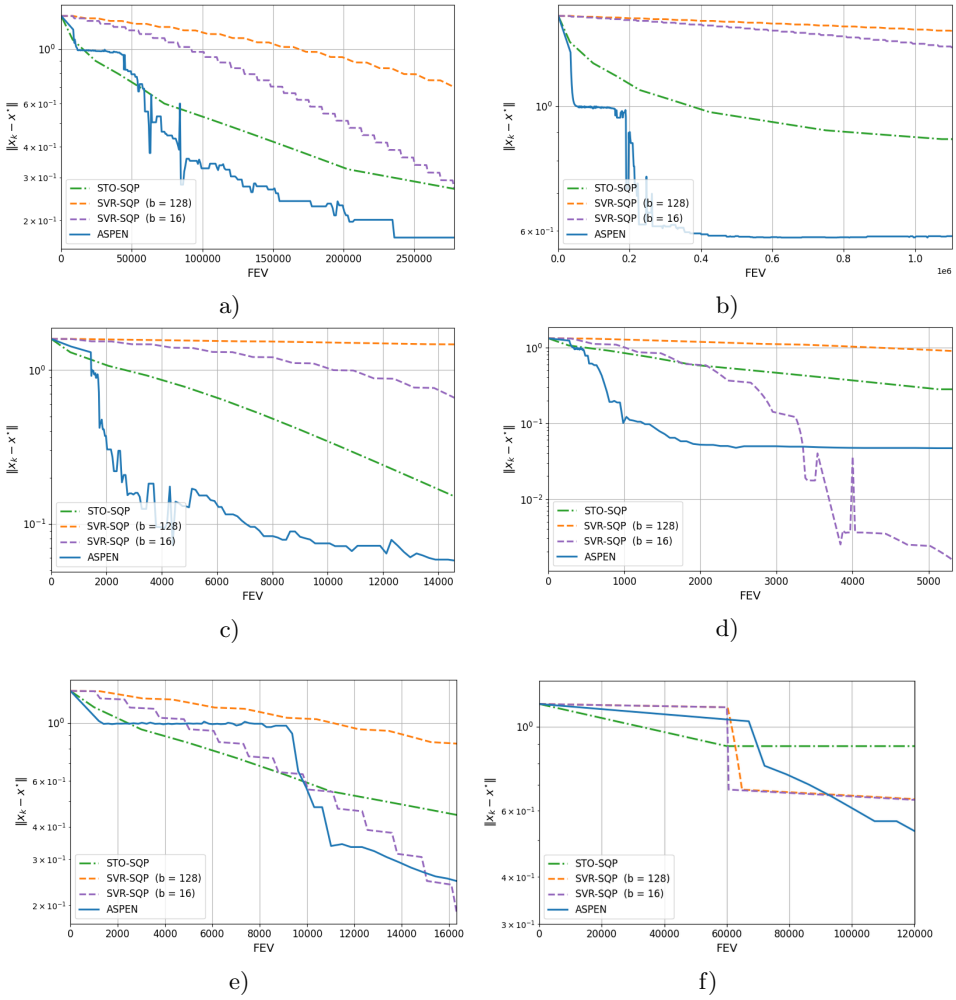


Figure 7: ASPEN vs. STO-SQP and two variants of SVR-SQP. Distance to the optimal solution versus the number of scalar products (FEV): a) *Mushrooms dataset*; b) *a9a dataset*; c) *Australian dataset*; d) *Heart dataset*; e) *Splice dataset*; f) *MNIST dataset*.

erties. The $b = 128$ variant shows the smallest improvement, confirming that large batch sizes can hamper stochastic acceleration under a fixed evaluation budget.

Figure 7 d) – *Heart dataset*. ASPEN demonstrates the fastest initial convergence, rapidly reducing the distance to the optimal solution $||x_k - x^*||$ below 10^{-1} within just a few hundred FEVs. However, after this sharp drop, its progress stagnates, maintaining a nearly constant level. Interestingly, SVR-SQP ($b = 16$) continues to steadily decrease over time and eventually surpasses ASPEN in terms of final accuracy, achieving the lowest distance to the optimum. STO-SQP and SVR-SQP ($b = 128$) exhibit slower and more

gradual convergence, remaining less competitive throughout.

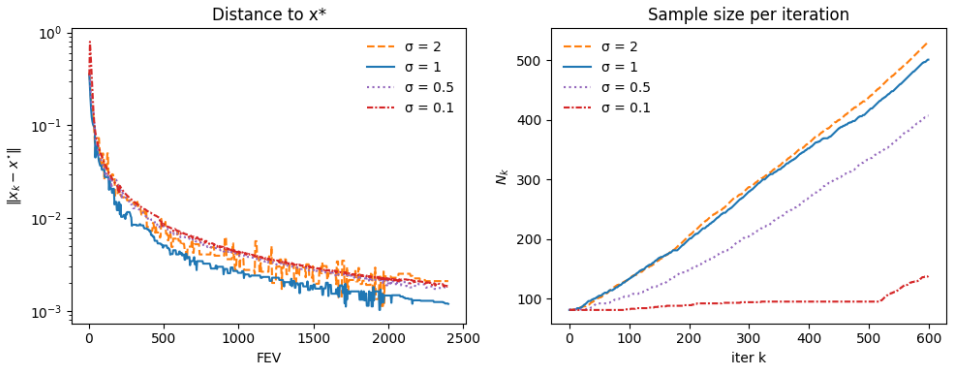
Figure 7 e) – *Splice dataset*. On this dataset, SVR-SQP ($b = 16$) eventually outperforms ASPEN in terms of final accuracy, although ASPEN demonstrates a significantly faster initial convergence. ASPEN again achieves strong early convergence but experiences prolonged stagnation between 2000 and 9000 FEV, only improving afterward. In contrast, SVR-SQP ($b = 16$) exhibits consistent and monotonic progress and ultimately outperforms all other methods by achieving the best final solution. STO-SQP remains moderately effective, while SVR-SQP ($b = 128$) shows the slowest rate of improvement. These results highlight that while ASPEN excels in fast early convergence, SVR-SQP ($b = 16$) demonstrates superior long-term accuracy on both datasets.

Figure 7 f) – MNIST dataset. On this dataset, for the binary classification problem, we can conclude that the ASPEN algorithm is the most successful in terms of solution accuracy, although it requires slightly more function evaluations to outperform the other algorithms, which are also reliable in terms of convergence but somewhat slower. The SVR-SQP algorithms (with batch sizes $b = 16$ and $b = 128$) produced better results than STO-SQP in scenarios with a larger number of FEV, even though STO-SQP initially shows the best performance.

We end this section by providing some more insights on the sample size behavior of the proposed method. To this end, we consider an academic problem (HS24) from the CUTEst collection [12], modified by introducing a Gaussian noise. More precisely, in order to simulate a stochastic environment, we consider a perturbed problem

$$\min_{x\in\mathbb{R}^n} f(x) := \frac{1}{N}\sum_{i=1}^{N}(\tilde{f}(x) + \varepsilon_i^2||x||^2) \quad \text{s.t.} \quad \|x\|_2^2 = 1, \tag{4.2}$$

where $\tilde{f}(x) = (x_1-2)^4+(x_1-2x_2)^2$ is the objective function of problem HS24 in CUTEst collection, and ε_i values are drawn from Gaussian distribution $\mathcal{N}(0,\sigma^2)$. Different levels of noise, i.e., variance, are employed to model different levels of similarity of local cost functions, where higher level of noise indicates more heterogeneous data.



a) b)

Figure 8: Influence of dissimilarity of local cost functions (modeled by $\sigma \in \{0.1, 0.5, 1, 2\}$) on the behavior of ASPEN algorithm applied on problem (4.2): a) optimality gap; b) sample size increase.

The results presented in Figure 8 show the robustness of ASPEN with respect to the optimality gap and illustrate the adaptive nature of the proposed method. As expected, more heterogeneous data require larger mini-batch sizes to mimic the original objective function properly, and these results clearly indicate that ASPEN is adaptable with respect to the sample size increase.

5 Conclusions

We introduced a novel first-order adaptive sampling algorithm for finite-sum minimization (ASPEN) that extends the work of [22] to a more general class of problems with nonlinear equality constraints. The method combines an additional sampling technique with non-monotone line search and puts it into the framework of quadratic penalty methods. The resulting method may behave like a mini-batch or an increasing sample method, depending on the problem at hand. Besides the sample size, the penalty parameter is also updated in an adaptive manner. We proved almost sure convergence of ASPEN under some standard assumptions for the considered framework, thus providing theoretical support for the proposed method. Numerical results conducted on real-world binary classification problems show that ASPEN is competitive with other state-of-the-art methods. Moreover, a numerical study on an academic problem reveals ASPEN's capability of adapting to different data structures. Future work will include potential extensions to problems with nonlinear inequality constraints.

Funding. N. Krejić and N. Krklec Jerinkić are supported by the Science Fund of the Republic of Serbia, Grant no. 7359, Project LASCADO. T. Ostojić is partially supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Grants No. 451-03-136/2025-03/200156) and by the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2025" (No. 01-50/295). N. Vučićević is partially supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Grant no. 451-03-137/2025-03/200122)

References

[1] S. Bellavia, G. Gurioli, B. Morini, & P. L. Toint, (2022). Trust-region algorithms: Probabilistic complexity and intrinsic noise with applications to subsampling techniques. *EURO Journal on Computational Optimization, 10, 100043.*

[2] A. S. Berahas, R. Bollapragada, & B. Zhou, (2022). An adaptive sampling sequential quadratic programming method for equality constrained stochastic optimization. Available at: *arXiv:2206.00712.*

[3] A. S. Berahas, F. E. Curtis, M. J. O'Neill, & D. P. Robinson, (2024). A stochastic sequential quadratic optimization algorithm for nonlinear-equality-constrained optimization with rank-deficient Jacobians. *Mathematics of Operations Research, 49(4), 2212-2248.*

[4] A. S. Berahas, F. E. Curtis, D. Robinson, & B. Zhou, (2021). Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization, 31(2), 1352-1379.*

[5] A. S. Berahas, J. Shi, Z. Yi, B. Zhou, (2023). Accelerating stochastic sequential quadratic programming for equality constrained optimization using predictive variance reduction, *Computational Optimization and Applications, 86(1), 79-116.*

[6] C. C. Chang, C. J. Lin, (2011). LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27.*

[7] P. Cascarano, G. Franchini, E. Kobler, F. Porta, & A. Sebastiani (2024). A variable metric proximal stochastic gradient method: An application to classification problems. *EURO Journal on Computational Optimization, 12, 100088-100088.*

[8] F. E. Curtis, D. P. Robinson, & B. Zhou, (2021). Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraintsi. Available at: *arXiv: 2107.03512.*

[9] F. E. Curtis, M. J. O'Neill, & D. P. Robinson, (2024). Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization. *Mathematical Programming, 205(1), 431-483.*

[10] D. Di Serafino, N. Krejić, N. Krklec Jerinkić, M. Viola, (2023). LSOS: Line-search Second-Order Stochastic optimization methods for nonconvex finite sums, *Mathematics of Computation, 92(341), 1273-1299.*

[11] Y. Fang, S. Na, M. W. Mahoney, & M. Kolar, (2024). Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. *SIAM Journal on Optimization, 34(2), 2007-2037.*

[12] N. I. M. Gould, D. Orban, & P. L. Toint, (2015). CUTEst: a Constrained and Unconstrained Testing Environment with safe threads for mathematical optimization. *Computational Optimization and Applications, 60, 545–557.*

[13] W. Huyer, & A. Neumaier, (2003). A new exact penalty function. *SIAM Journal on Optimization, 13(4), 1141-1158.*

[14] D. H. Li, & M. Fukushima, (2000). A derivative-free line search and global convergence of Broyden-like method for nonlinear equations. *Optimization Methods and Software, 13(3), 181–201.*

[15] M. Lichman, (2013). UCI Machine Learning Repository. Available at: *https://archive.ics.uci.edu/ml/index.php*

[16] Y. LeCun, C. Cortes, & C. J. C. Burges, (1998). The MNIST database of handwritten digits. Available at: *http://yann.lecun.com/exdb/mnist/*

[17] A. N. Iusem, A. Jofré, R. I. Oliveira, & P. Thompson, (2019). Variance-based extragradient methods with line search for stochastic variational inequalities. *SIAM Journal on Optimization, 29(1), 175-206.*

[18] A. N. Iusem, A. Jofré, R. I. Oliveira, & P. Thompson, (2017). Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization, 27(2), 686-724.*

[19] N. Krejić, N. Krklec Jerinkić, & A. Rožnjik, (2018). Variable sample size method for equality constrained optimization problems. *Optimization Letters, 12, 485-497.*

[20] N. Krejić & N. Krklec Jerinkić, (2015). Nonmonotone line search methods with variable sample size. *Numerical Algorithms, 68(4), 711-739.*

[21] N. Krejić, N. Krklec Jerinkić, A. Martinez, & M. Yousefi, (2024). A non-monotone trust-region method with noisy oracles and additional sampling. *Computational Optimization and Applications, 89(1), 247-278.*

[22] N. Krejić, N. Krklec Jerinkić, S. Rapajić, & L. Rutešić, (2025). IPAS: An Adaptive Sample Size Method for Weighted Finite Sum Problems with Linear Equality Constraints. Available at: *arXiv: 2504.19629.*

[23] N. Krejić, Z. Lužanin, Z. Ovcin, & I. Stojkovska, (2015). Descent direction method with line search for unconstrained optimization in noisy environment. *Optimization Methods and Software, 30(6), 1164-1184.*

[24] N. Krklec Jerinkić, & A. Rožnjik, (2020). Penalty variable sample size method for solving optimization problems with equality constraints in a form of mathematical expectation. *Numerical Algorithms, 83(2), 701-718.*

[25] N. Krklec Jerinkić, & T. Ostojić, (2024). AN-SPS: adaptive sample size nonmonotone line search spectral projected subgradient method for convex constrained optimization problems, *Optimization Methods and Software, 39(5), 1143-1167.*

[26] N. Krklec Jerinkić, F. Porta, V. Ruggiero, & I. Trombini, (2025). Variable metric proximal stochastic gradient methods with additional sampling, *Computational Optimization and Applications, https://doi.org/10.1007/s10589-025-00720-w.*

[27] N. Krklec Jerinkić, V. Ruggiero, & I. Trombini, (2025). Spectral Stochastic Gradient Method with Additional Sampling for Finite and Infinite Sums, *Computational Optimization and Applications, 91 (2), 717–758.*

[28] G. Lan, (2020). First-order and stochastic optimization methods for machine learning, *Cham: Springer International Publishing, (Vol. 1), 21-51.*

[29] S. Na, M. Anitescu, & M. Kolar, (2023). An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Mathematical Programming, 199(1), 721-791.*

[30] S. Na, M. Anitescu, & M. Kolar, (2023). Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *Mathematical Programming, 202(1), 279-353.*

[31] Y. Nandwani, A. Pathak, & P. Singla,(2019). A primal dual formulation for deep learning with constraints. *Advances in neural information processing systems, 32, 12157-12168.*

[32] G. Négiar, G. Dresdner, A. Tsai, L. El Ghaoui, F. Locatello, R. Freund, & F. Pedregosa, (2020, November). Stochastic Frank-Wolfe for constrained finite-sum minimization. *In international conference on machine learning, PMLR, 7253-7262.*

[33] E. Polak, & J. O. Royset, (2008). Efficient sample sizes in stochastic nonlinear programming. *Journal of Computational and Applied Mathematics, 217(2), 301-310.*

[34] X. Wang, S. Ma, & Y. X. Yuan, (2017). Penalty methods with stochastic approximation for stochastic nonlinear programming. *Mathematics of Computation, 86(306), 1793-1820.*