Convergence analysis of stochastic higher-order majorization–minimization algorithms

Daniela Lupu\textsuperscript{a} and Ion Necoara\textsuperscript{b}

\textsuperscript{a}Automatic Control and Systems Engineering Department, University Politehnica Bucharest, Bucharest, Romania; \textsuperscript{b}Gheorghe Mihoc-Caius Iacob Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy, Bucharest, Romania

ABSTRACT

Majorization–minimization schemes are a broad class of iterative methods targeting general optimization problems, including nonconvex, nonsmooth and stochastic. These algorithms minimize successively a sequence of upper bounds of the objective function so that along the iterations the objective value decreases. We present a stochastic higher-order algorithmic framework for minimizing the average of a very large number of sufficiently smooth functions. Our stochastic framework is based on the notion of stochastic higher-order upper bound approximations of the finite-sum objective function and minibatching. We derive convergence results for nonconvex and convex optimization problems when the higher-order approximation of the objective function yields an error that is \( p \) times differentiable and has Lipschitz continuous \( p \) derivative. More precisely, for general nonconvex problems we present asymptotic stationary point guarantees and under Kurdyka–Lojasiewicz property we derive local convergence rates ranging from sublinear to linear. For convex problems with uniformly convex objective function, we derive local (super)linear convergence results for our algorithm. Compared to existing stochastic (first-order) methods, our algorithm adapts to the problem’s curvature and allows using any batch size. Preliminary numerical tests support the effectiveness of our algorithmic framework.

1. Introduction

The empirical risk minimization (also called finite-sum) problems appear in various applications such as statistics and machine learning \([2,13]\) or distributed control and signal processing \([14,25]\). Usually, these problems are nonconvex and nonsmooth due to the presence of regularization terms and constraints. Another difficulty when dealing with these problems is the large number of terms in the finite-sum and the large number of variables. Hence, stochastic methods are the most appropriate to solve finite-sum optimization problems. Among these methods, those based on stochastic first-order oracle are widely used, e.g. stochastic gradient descent \([2,19,20,24,27]\) or stochastic proximal point \([24,31]\).
In these works, for (strongly) convex objective functions and diminishing step sizes, sublinear convergence results were derived. Although stochastic first-order methods work well in applications, their convergence speed is known to slow down close to saddle points or in ill-conditioned landscapes [23]. On the other hand, stochastic higher-order methods (including, e.g. cubic Newton and third-order methods) use the curvature of the objective function and thus they converge faster to a (local) minimum [1,16,17].

**Related work.** Most of the optimization methods can be interpreted from the majorization–minimization point of view, which consist of successively minimizing some upper bounds of the objective function [20,26]. More specifically, in these methods at each iteration one constructs and optimizes a local upper bound model (majorizer) of the objective function using first- or higher-order derivatives (e.g. a Taylor approximation with an additional penalty term that depends on how well the model approximates the real objective). In this work, we also focus on stochastic high-order majorization–minimization methods to optimize a finite-sum objective function. Deterministic higher-order majorization–minimization methods that construct a local Taylor model of the objective in each iteration with an additional regularization term has gained a lot of attention in the last decade due to their fast convergence [3,7,21,22,26,28]. For example, first-order methods achieve convergence rates of order \( O(1/k) \) for smooth convex optimization [23], while higher-order methods of order \( p > 1 \) have converge rates \( O(1/k^p) \) [21,22,26], where \( k \) is the iteration counter. Accelerated variants of higher-order methods were also developed recently, see e.g. [12,22]. Yet, these better convergence results assume a deterministic setting where access to exact evaluations of the higher-order derivatives is needed and they do not directly translate to the stochastic case.

To the best of our knowledge, the use of high-order derivatives in a stochastic setting has received little consideration in the literature so far. For example, [1,17] (see also [4,32,33]) present stochastic optimization methods that use third- (second-) order Taylor approximations with a fourth- (cubic-) order regularizations to find local minima of smooth and nonconvex finite-sum objective functions. Usually, these methods use sub-sampled derivatives instead of computing them exactly and are able to find fast (first-/second-order) critical points. However, these methods require large sub-samples, e.g. the batch size must be proportional to \( O(1/\epsilon^2) \), where \( \epsilon \) is the desired accuracy for solving the problem. In general, the theoretical bounds for the batch size usually exceed the number of functions from the finite-sum, leading basically to deterministic algorithms. Furthermore, quasi-Newton type methods that perform updates in a deterministic (cyclic) fashion are given, e.g. in [18]. The papers most related to our work are [16,20], where stochastic variants of the majorization–minimization approach are derived based on first- or second-order Taylor approximations with a proper regularization term. For these methods, (local) linear rates are derived when the finite-sum objective function is sufficiently smooth and strongly convex. However, there is no complete convergence analysis for general stochastic higher-order majorization–minimization algorithms for solving general finite-sum problems (including, e.g. nonconvex problems or problems having nonsmooth regularization terms).

**Contributions.** This paper presents a stochastic higher-order algorithmic framework for minimizing finite-sum (possibly nonconvex and nonsmooth) optimization problems. Our framework is based on stochastic higher-order upper bound approximations of the (non)convex and/or (non)smooth finite-sum objective function, leading to a minibatch stochastic higher-order majorization–minimization algorithm, which we call SHOM. We
derive convergence guarantees for the SHOM algorithm for general optimization problems when the upper bounds approximate the finite-sum objective function up to an error that is $p \geq 1$ times differentiable and has a Lipschitz continuous $p$ derivative; we call such upper bounds stochastic higher-order surrogate functions. The main challenge in analysing convergence of SHOM, especially in the nonconvex setting, is the fact that even in expectation the cost cannot be used as a Lyapunov function, unless a proper secondary sequence is constructed from the algorithm. For general nonconvex problems, we prove that SHOM is a descent method in expectation, derive asymptotic stationary point guarantees and under Kurdyka–Lojasiewicz (KL) property we establish the first local linear or sublinear convergence rates (depending on the KL parameter) for stochastic higher-order type algorithms under such an assumption. For convex problems, we derive local superlinear or linear convergence results, provided that the objective function is uniformly convex.

To prove convergence, stochastic higher-order methods usually require the evaluation of a large number of higher derivatives (e.g. gradients, Hessians, etc.). On the other hand, our algorithmic framework does not have this drawback. In particular, one variant of our method requires at each iteration the random selection of a single function from the finite-sum and the computation of only its higher-order derivatives. Moreover, unlike most existing stochastic higher-order methods (e.g. second- and third-order methods), ours have faster local convergence than first-order methods, as they adapt to the objective function’s curvature. Besides providing a general framework for the design and analysis of stochastic higher-order methods, in special cases, where complexity bounds are known for some particular algorithms, our convergence results recover the existing bounds (see Remark 5.1). In particular, for strongly convex functions and $p = 2$ we recover the local linear rate from [16]. For the deterministic case, i.e. the batch size is equal to the number of functions in the finite-sum, we obtain local superlinear convergence as in [10]. Finally, for $p = 1$ SHOM coincides with the popular MISO algorithm [20] and a byproduct of our convergence analysis leads to new convergence results for MISO in the nonsmooth and nonconvex settings under the KL assumption. Numerical simulations also confirm the efficiency of our algorithm, i.e. increasing the approximation order can have beneficial effects in terms of convergence.

Content. The paper is organized as follows. In Section 2, we introduce notations and some generalities. Then, in Section 3 we formulate the optimization problem and our algorithm. In Section 4, we provide a convergence analysis for the nonconvex case. For convex problems, we derive convergence rates in Section 5. Finally, Section 6 presents some preliminary numerical experiments on convex and nonconvex applications.

2. Notations and generalities

We denote a finite-dimensional real vector space with $E$ and by $E^*$ its dual space composed by linear functions on $E$. For such a function $s \in E^*$, denoted by $\langle s, x \rangle$ its value at $x \in E$. Utilizing a self-adjoint positive-definite operator $B : E \to E^*$ (notation $B = B^* > 0$), we can endow these spaces with conjugate Euclidean norms:

$$\|x\| = \langle Bx, x \rangle^{1/2} \quad \forall x \in E, \quad \|y\|^* = \langle y, B^{-1}y \rangle^{1/2} \quad \forall y \in E^*.$$  

Under these settings, we have the following expression for the dual norm: $\|y\|^* = \max_{x: \|x\| \leq 1} \langle y, x \rangle$. For a smooth function $\psi : \text{dom} \psi \to \mathbb{R}$, $p \geq 1$ times continuously
differentiable on the convex and open domain \( \text{dom}\psi \subseteq E \), denote by \( \nabla \psi(x) \) its gradient and \( \nabla^2 \psi(x) \) its Hessian at the point \( x \in \text{dom}\psi \). Note that \( \nabla \psi(x) \in E^* \) and \( \nabla^2 \psi(x) h \in E^* \) for all \( h \in E \). In what follows, we often work with directional derivatives: \( D^p \psi(x)[h_1, \ldots, h_p] \) denotes the \( p \) directional derivative of function \( \psi \) at \( x \in \text{dom}\psi \) along directions \( h_1, \ldots, h_p \in E \) (see also [21] for a similar exposition). Note that \( D^p \psi(x)[\cdot] \) is a symmetric \( p \) multilinear form on \( E \). For example, for a smooth function \( \psi \) one has for any \( x \in \text{dom}\psi \) and \( h, \bar{h} \in E \) that \( D^2 \psi(x)[h] = \langle \nabla \psi(x), h \rangle \) and \( D^2 \psi(x)[h, \bar{h}] = \langle \nabla^2 \psi(x) h, \bar{h} \rangle \). On the other hand, in this paper we consider for a nonsmooth function \( \psi \) the directional derivative classically defined as \( D^p \psi(x)[h] = \lim_{t \downarrow 0}(\psi(x + th) - \psi(x))/t \) (note that this definition is valid not only for convex functions and, obviously, for differentiable functions, but also for a broader class of functions, so-called locally convex). The abbreviation \( D^p \psi(x)[h]^p \) is used when all directions are the same, i.e. \( h_1 = \cdots = h_p = h \) for some \( h \in E \). The norm of \( D^p \psi(x) \) is defined in the standard way (see [21]):

\[
\|D^p \psi(x)\| := \max_{\|h_1\|, \ldots, \|h_p\| \leq 1} |D^p \psi(x)[h_1, \ldots, h_p]| = \max_{\|h\| \leq 1} |D^p \psi(x)[h]^p|.
\]

Note that for any fixed \( x, y \in \text{dom}\psi \) the form \( D^p \psi(x)[\cdot] - D^p \psi(y)[\cdot] \) is also \( p \) multilinear and symmetric. Then we define the following class of smooth functions:

**Definition 2.1:** Let \( \psi : E \to \mathbb{R} \) be \( p \geq 1 \) times continuously differentiable. Then the \( p \) derivative of \( \psi \) is Lipschitz continuous if there exist \( L^\psi_p > 0 \) for which the following relation holds:

\[
\|D^p \psi(x) - D^p \psi(y)\| \leq L^\psi_p \|x - y\| \quad \forall x, y \in \text{dom}\psi.
\]

We denote the Taylor approximation of \( \psi \) around \( x \in \text{dom}\psi \) of order \( p \) by

\[
T^\psi_p(y; x) = \psi(x) + \sum_{i=1}^{p} \frac{1}{i!} D^i \psi(x)[y - x]^i \quad \forall y \in E.
\]

It is known that if (1) holds, then by the standard integration arguments the residual between function value and its Taylor approximation can be bounded [21]:

\[
|\psi(y) - T^\psi_p(y; x)| \leq \frac{L^\psi_p}{(p + 1)!} \|y - x\|^{p+1} \quad \forall x, y \in \text{dom}\psi.
\]

Applying the same reasoning for the functions \( \langle \nabla \psi(x), h \rangle \) and \( \langle \nabla^2 \psi(x) h, h \rangle \), with direction \( h \in E \) being fixed, we also get the following inequalities valid for all \( x, y \in \text{dom}\psi \) and \( p \geq 2 \), see [21]:

\[
\|\nabla \psi(y) - \nabla T^\psi_p(y; x)\|_* \leq \frac{L^\psi_p}{p!} \|y - x\|^p,
\]

\[
\|\nabla^2 \psi(y) - \nabla^2 T^\psi_p(y; x)\| \leq \frac{L^\psi_p}{(p - 1)!} \|y - x\|^{p-1}.
\]

For the Hessian, we consider the spectral norm of self-adjoint linear operators (maximal module of all eigenvalues computed with respect to operator \( B \)). Below are some examples of functions having known Lipschitz continuous derivatives.
Example 2.2: Given $x_0 \in \mathbb{R}^n$ and a matrix $B > 0$, defining the norm $\|x\| = \langle Bx, x \rangle^{1/2}$ for all $x \in \mathbb{R}^n$, then the power of the Euclidean norm $\psi(x) = \|x - x_0\|^{p+1}$ with $p \geq 1$ satisfies the Lipschitz continuous condition (1) with $L_\psi^p = (p + 1)!$ (see Theorem 7.1 in [30] for more details).

Example 2.3: For given $(a_i)_{i=1}^m \in E^*$, consider the log-sum-exp function:

$$
\psi(x) = \log \left( \sum_{i=1}^m e^{\langle a_i, x \rangle} \right) \quad \forall x \in E.
$$

For $B := \sum_{i=1}^m a_i a_i^*$ (assuming $B > 0$, otherwise we can reduce dimensionality of the problem) we define the norm $\|x\| = \langle Bx, x \rangle^{1/2}$ for all $x \in E$ and then the Lipschitz continuous condition (1) holds for $p = 1, 2$ and $3$ with $L_1^\psi = 1, L_2^\psi = 2$ and $L_3^\psi = 4$, respectively (see Lemma 4 in [11] for details). Note that for $m = 2$ and $a_1 = 0$, we recover the logistic regression function, widely used in machine learning [13].

Example 2.4: If the $p + 1$ derivative of a function $\psi$ is bounded, then the $p$ derivative of $\psi$ is Lipschitz continuous. Moreover, any polynomial of degree $p$ (e.g. the $p$ Taylor approximation of $\psi$, denoted $T^\psi_p$) has the $p$ derivative Lipschitz with the Lipschitz constant zero (in particular, $L^p_{T^\psi_p} = 0$).

Further, let us introduce the class of uniformly convex functions.

Definition 2.5: A function $\psi(\cdot)$ is uniformly convex of degree $q \geq 2$ with the uniform constant $\sigma_q > 0$, if for any $x, y \in \text{dom}\psi$ we have

$$
\psi(y) \geq \psi(x) + \langle \psi_x, y - x \rangle + \frac{\sigma_q}{q} \|x - y\|^q,
$$

where $\psi_x \in E^*$ is an arbitrary vector (subgradient) from the subdifferential $\partial \psi(x)$.

Note that $q = 2$ corresponds to the strongly convex functions. Next, we provide a simple example of a uniformly convex function (see, e.g. [22] for more details).

Example 2.6: Given $x_0 \in \mathbb{R}^n$, $B = I_n$ and $q \geq 2$, then the power of the Euclidean norm $\psi(x) = \frac{1}{q} \|x - x_0\|^q$ is uniformly convex of degree $q$ with $\sigma_q = (\frac{1}{2})^{q-2}$.

3. Stochastic higher-order majorization–minimization algorithms

We consider the finite-sum (possibly nonsmooth and nonconvex) problem:

$$
\min_{x \in \mathcal{X}} f(x) = \frac{1}{N} \sum_{i=1}^N f_i(x),
$$

where each function $f_i : \text{dom}f \rightarrow \mathbb{R}$, having $\text{dom}f \subseteq E$ open set, is lower semi-continuous and $\mathcal{X} \subseteq \text{dom}f$ is a given closed convex set. Our approach consists of associating to $f$ a stochastic $p$ higher-order surrogate function $g$. Let us define this notion:
Definition 3.1: Given $f$ as in (6) and the set of points $\{x_1, \ldots, x_N\} \subset X$, we denote $\hat{x} = [x_1^T \cdots x_N^T]^T$. Then the function $g(y; \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} g_i(y; x_i)$, with $g_i(\cdot; x_i) : \text{dom} g_i \to \mathbb{R}$ having $X \subseteq \text{dom} g_i$, is called a stochastic $p$ higher-order surrogate function at $\hat{x}$ over $X$ if the following relations hold for all $i = 1:N$:

(i) each $g_i$ is a majorizer for the function $f_i$ on $X$, i.e. $g_i(y; x_i) \geq f_i(y)$ $\forall y \in X$.

(ii) The error function $h_i(y; x_i) = g_i(y; x_i) - f_i(y)$ with $\text{dom} h_i = E$ has the $p$ derivative Lipschitz continuous with Lipschitz constant $L_{hi}$.

(iii) The $j$ derivatives satisfy $\nabla^j h_i(x_i; x_i) = 0$ $\forall j = 0:p$, where $j = 0$ means $h_i(x_i; x_i) = 0$.

Hence, the previous three properties yield the following relations for the finite-sum objective function $f$:

(1) $g$ is a majorizer for the finite-sum objective function $f$, i.e. $g(y; \hat{x}) \geq f(y)$ $\forall y \in X$.

(2) The error function $h(y; \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} h_i(y; x_i)$ has the $p$ derivative smooth with Lipschitz constant $L_{h}^p = \frac{1}{N} \sum_{i=1}^{N} L_{hi}^p$.

See also [20] for a similar definition of a stochastic first-order surrogate function (i.e. for the particular case $p = 1$) and [26] for the definition of a higher-order surrogate function in the deterministic settings. Note that our formulation of the surrogate extends both definitions given in [20] and in [26]. Next, we give several nontrivial examples of stochastic higher-order surrogate functions satisfying requirements of Definition 3.1.

Example 3.2 (Lipschitz $p$ derivative stochastic surrogate): Assume that each function $f_i : E \to \mathbb{R}$ has the $p$ derivative smooth with the Lipschitz constant $L_{fi}^p$. Then, we can define the following stochastic $p$ higher-order surrogate over $E$:

$$g(y; \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} \left[ T_{fi}^p(y; x_i) + \frac{M_{fi}^p}{(p+1)!} \|y - x_i\|^{p+1} \right],$$

where $M_{fi}^p \geq L_{fi}^p$ and $x_i \in X$ for all $i = 1:N$. Then, the error function $h(y; \hat{x}) = g(y; \hat{x}) - f(y)$ has the $p$ derivative smooth with the Lipschitz constant $L_{h}^p = \frac{1}{N} \sum_{i=1}^{N} M_{fi}^p + L_{fi}^p$. Let us note that when each function $f_i$ is convex the stochastic $p$ higher-order surrogate function $g$ is also convex in the first argument for any $\hat{x}$, provided that $M_{fi}^p \geq pL_{fi}^p$. This is a consequence of the following lemma.

Lemma 3.3 ([21]): Let $f$ be convex function with the $p > 2$ derivative Lipschitz continuous with constant $L_{fi}^p$. Then, for $M_p \geq pL_{fi}^p$ and any $x \in E$ the function:

$$g(y; x) = T_{fi}^p(y; x) + \frac{M_p}{(p+1)!} \|y - x\|^{p+1}$$

is convex in the first argument.
**Example 3.4 (composite functions):** Assume that \( f(x) = \frac{1}{N} \sum_{i}^{N} \phi_{i}(x) + \psi(x) \), where each (possibly nonconvex) function \( \phi_{i} \) has the \( p \) derivative smooth with the Lipschitz constant \( L_{p}^{\phi_{i}} \) and \( \psi \) is a simple function (possibly nonsmooth and nonconvex). Assume also that \( X \subseteq \text{dom} \psi \) and \( \psi \) is proper lower-semicontinuous function. Define \( f_{i} = \phi_{i} + \psi \). Then, we have the following stochastic \( p \) higher-order surrogate over \( X \):

\[
g(y; \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} \left[ T_{p}^{\phi_{i}}(y; x_{i}) + \frac{M_{p}^{i}}{(p+1)!} \|y - x_{i}\|^{p+1} + \psi(y) \right],
\]

where \( M_{p}^{i} \geq L_{p}^{\phi_{i}} \). Further, the error function \( h(y; \hat{x}) = g(y; \hat{x}) - f(y) \) has the \( p \) derivative smooth with the Lipschitz constant \( L_{p}^{h} = \frac{1}{N} \sum_{i=1}^{N} M_{p}^{i} + L_{p}^{\phi_{i}} \). From Lemma 3.3, it follows that the surrogate \( g \) is convex in the first argument provided that the functions \( \phi_{i} \)'s and \( \psi \) are convex and the constants \( M_{p}^{i} \)'s are sufficiently large.

**Remark 3.1:** In Examples 3.2 and 3.4, if additionally, each function \( f_{i} \) and \( \phi_{i} \) satisfies the condition \( f_{i}(y) \geq T_{p}^{f_{i}}(y; x_{i}) \) and \( \phi_{i}(y) \geq T_{p}^{\phi_{i}}(y; x_{i}) \) for all \( y \) and \( x_{i} \), respectively, then the Lipschitz constant for the error function \( h \) can be improved to \( L_{p}^{h} = \frac{1}{N} \sum_{i=1}^{N} M_{p}^{i} \), see [26].

**Example 3.5 (bounded derivative functions):** Assume \( p \) odd and each function \( f_{i} \) has the \( p + 1 \) derivative bounded by a symmetric multilinear form \( H_{i} \), i.e. \( \langle D_{p+1}^{f_{i}}(x) [h]^{p}, h \rangle \leq \langle H_{i} |[h]^{p}, h \rangle \) for all \( h \in \mathbb{E} \) and \( x \in X \). Then, each \( f_{i} \) has the \( p \) derivative Lipschitz with constant \( L_{p}^{f_{i}} \leq \|H_{i}\| \) (see Example 2.4) and for any constants \( M_{p}^{i} \geq 1 \) we have the following stochastic \( p \) higher-order surrogate over \( X \):

\[
g(y; \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} \left[ T_{p}^{f_{i}}(y; x_{i}) + \frac{M_{p}^{i}}{(p+1)!} \langle H_{i} [y - x_{i}]^{p}, y - x_{i} \rangle \right].
\]

Moreover, the error function \( h(y; \hat{x}) = g(y; \hat{x}) - f(y) \) has the \( p \) derivative smooth with the Lipschitz constant \( L_{p}^{h} = \frac{1}{N} \sum_{i=1}^{N} \|H_{i}\| + L_{p}^{f_{i}} \). From Lemma 3.3, it follows that the surrogate \( g \) is convex in the first argument provided that the functions \( f_{i} \)'s are convex and the constants \( M_{p}^{i} \)'s are sufficiently large.

**Example 3.6 (high-order proximal functions):** Let \( f_{i} \), with \( i = 1:N \), be general proper lower-semicontinuous functions (possibly nonsmooth and nonconvex). Then, for any \( p \geq 1 \) we have the following stochastic \( p \) higher-order surrogate over \( X \):

\[
g(y; \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} \left[ f_{i}(y) + \frac{M_{p}^{i}}{(p+1)!} \|y - x_{i}\|^{p+1} \right],
\]

for parameters \( M_{p}^{i} > 0 \). Moreover, the error function \( h(y; \hat{x}) = g(y; \hat{x}) - f(y) \) has the \( p \) derivative smooth with the Lipschitz constant \( L_{p}^{h} = \frac{1}{N} \sum_{i=1}^{N} M_{p}^{i} \) (according to Example 2.2). Note that if \( f_{i} \)'s are convex functions, then the corresponding surrogates \( g_{i} \)'s
become uniformly convex; if $f_i$’s are weakly convex functions and $p = 1$, then the surrogate functions $g_i$’s become strongly convex for appropriate choices of $M_p$’s; even for nonconvex functions $f_i$’s and $p > 1$, the $p + 1$ regularization can have beneficial effects in convexifying $f_i$’s. The reader can find other examples of stochastic surrogates depending on the application at hand.

For solving the general (possibly nonsmooth and nonconvex) finite-sum problem (6), we propose the following minibatch Stochastic Higher-Order Majorization–minimization (SHOM) algorithm:

\begin{center}
\textbf{SHOM algorithm}
\end{center}

Given $\tau \in [1, N], p \geq 1$ and $x_0 \in \mathcal{X}$, define $x_0^j = x_0 \quad \forall j = 1 : N$. For $k \geq 0$ do:

1. Choose uniformly random a subset $S_k \subseteq \{1, \ldots, N\}$ of size $\tau$.
2. For each $i_k \in S_k$ compute $p$ higher-order surrogate $g_{i_k}(y; x_{i_k}^j)$ of $f_{i_k}$ near $x_{i_k}^j$; otherwise keep the previous surrogates.
3. Compute $x_{k+1}$ as a stationary point of subproblem:
   \[
   \min_{y \in \mathcal{X}} g(y; \hat{x}_k) := \frac{1}{N} \sum_{j=1}^{N} g_j(y; x_{i_k}^j),
   \]
   where $\hat{x}_k = [x_{i_k}^j]_{j=1:N}$ and $x_{i_k}^j = \begin{cases} x_k, & j \in S_k \\ x_{i_k}^{j-1}, & j \notin S_k. \end{cases}$

Note that SHOM is a variance reduced method and for $p = 1$ and $\tau = 1$ it coincides with MISO algorithm given in [20]. In our analysis below for the nonconvex case we consider $x_{k+1}$ to be a stationary point of the subproblem (7) satisfying additionally the following descend property:

\[
    g(x_{k+1}; \hat{x}_k) \leq g(x_k; \hat{x}_k) \quad \forall k \geq 0.
\]

Hence, in the nonconvex case SHOM does not require the computation of a (local) minimum of the surrogate function $g$ at $\hat{x}_k$ over the convex set $\mathcal{X}$, we just need $x_{k+1}$ to be a stationary point of (7) satisfying the descent (8). Moreover, in the convex case this decrease always holds, since in this case $x_{k+1}$ is computed as the global optimum of (7). Let us recall the definition of finite-sum error function (see Definition 3.1):

\[
    h(y; \hat{x}_k) = \frac{1}{N} \sum_{j=1}^{N} h_j(y; x_{i_k}^j) = g(y; \hat{x}_k) - f(y),
\]

where $h_j(y; x_{i_k}^j) = g_j(y; x_{i_k}^j) - f_j(y)$. Note that we can update very efficiently the model approximation $g$ in SHOM algorithm (see also Section 6.1):

\[
    g(y; \hat{x}_k) = g(y; \hat{x}_{k-1}) + \frac{1}{N} \sum_{j \in S_k} g_j(y; x_k) - g_j(y; x_{k-1}^{j-1}).
\]
Moreover, since the subset of indices $S_k$ of size $\tau$ are chosen uniformly random, then for scalars $\theta_i$, with $i = 1:N$, the following holds in expectation:

$$\mathbb{E} \left[ \sum_{i \in S_k} \theta_i \right] = \frac{\tau}{N} \sum_{i=1}^{N} \theta_i.$$  (10)

We denote the conditional expectation w.r.t. the filtration defined as the sigma algebra generated by the history of the random index set $S$, $\mathcal{F}_k = \sigma (\{S_0, \ldots, S_{k-1}\})$, by $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_k]$. Finally, we denote the directional derivative of a function $f$ at any $x \in \text{dom} f$ (assumed open set) along $y-x$ by $\nabla f(x)[y-x]$. Note that if $f$ is differentiable, the directional derivative is $\nabla f(x)[y-x] = \langle \nabla f(x), y-x \rangle$.

## 4. Nonconvex convergence analysis

In this section, we assume a general nonconvex finite-sum (possibly nonsmooth) objective function $f$ which has directional derivative at each point in its open domain $\text{dom} f$ (see Examples 3.2, 3.4, 3.5 and 3.6). Under these general assumptions, we define next the notion of stationary points for problem (6) (see also [20] for a similar definition).

**Definition 4.1 (asymptotic stationary points):** Assume that the directional derivative of the function $f$ at any $x \in \mathcal{X} \subseteq \text{dom} f$ along $y-x$ exists for all $y \in \mathcal{E}$. Then, a sequence $(x_k)_{k \geq 0}$ satisfies the asymptotic stationary point condition for the nonconvex problem (6) if the following holds:

$$\liminf_{k \to +\infty} \inf_{x \in \mathcal{X}} \frac{Df(x_k)[x-x_k]}{\|x-x_k\|} \geq 0.$$

Now we are ready to analyze the convergence behaviour of SHOM under these general settings (nonconvexity and nonsmoothness). First, we prove that the sequence generated by SHOM satisfies the asymptotic stationary point condition and the sequence of function values decreases in expectation. Hence, SHOM algorithm is a descent method (in expectation). These results will help us later to derive explicit convergence rates for nonconvex and convex cases.

**Lemma 4.2:** Assume that the objective function $f$ is nonconvex, bounded below and has directional derivatives over $\mathcal{X}$. Moreover, assume that $f$ admits a stochastic $p \geq 1$ higher-order surrogate function $g$ over $\mathcal{X}$ as given in Definition 3.1 and the sequence $(x_k)_{k \geq 0}$ generated by SHOM is bounded. Then the sequence of function values in expectation $(\mathbb{E}[f(x_k)])_{k \geq 0}$ monotonically decreases, $(f(x_k))_{k \geq 0}$ converges almost surely and $(x_k)_{k \geq 0}$ satisfies the asymptotic stationary point condition with probability one.
Proof: From (8), the definition of the surrogate function $g$ and the update rule of SHOM algorithm, we get
\[ g(x_{k+1}; \hat{x}_k) \leq g(x_k; \hat{x}_k) = g(x_k; \hat{x}_{k-1}) + \frac{\sum_{j \in S_k} g_{j}(x_k; x_k) - g_{j}(x_k; \hat{x}_{k-1})}{N}. \] (11)
Further, from the definition of the surrogate function, we have that for $\forall j \in S_k$:
\[ g_{j}(x_k; x_k) = g_{j}(x_k; x_k) = f_{j}(x_k) \leq g_{j}(x_k; x_{k-1}). \]
Moreover, for $j \notin S_k$, we have
\[ g_{j}(x_k; x'_{k}) = g_{j}(x_k; x_{k-1}). \]
Using these relations and (8), we obtain the following descent:
\[ f(x_{k+1}) \leq g(x_{k+1}; \hat{x}_k) \leq g(x_k; \hat{x}_k) \leq g(x_k; \hat{x}_{k-1}), \] (12)
where the first inequality follows from the fact that $g$ majorizes $f$. Thus the sequence $(g(x_{k+1}; \hat{x}_k))_{k \geq 0}$ is monotonically decreasing and bounded below with probability 1, since $g$ by definition is majorizing $f$, which by our assumption is bounded from below. Taking expectation w.r.t. $\mathcal{F}_{k+1}$, we obtain that the sequence $(\mathbb{E}[g(x_{k+1}; \hat{x}_k)])_{k \geq 0}$ converges. Further, using basic properties of conditional expectation in (12), we have
\[ \mathbb{E}[g(x_{k+1}; \hat{x}_k)] = \mathbb{E}[g(x_k; \hat{x}_k)] = \mathbb{E}[\mathbb{E}_k[f(x_k)]] = \mathbb{E}[f(x_k)]. \] (13)
Taking expectation w.r.t. the whole set of minibatches $\mathcal{F}_{k+1}$ in (12) and combining with (13), we also get $\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k)]$, i.e. SHOM algorithm is a descent method in expectation. Note also that the nonnegative quantity $\mathbb{E}[g_{i}(x_{k+1}; x_k) - f_{i}(x_{k+1})]$ is the summand of a converging sum and we have
\[ 0 \leq \mathbb{E}\left[ \sum_{k=0}^{\infty} \sum_{i \in S_{k+1}} g_{i}(x_{k+1}; x_k) - f_{i}(x_{k+1}) \right] = \sum_{k=0}^{\infty} \mathbb{E}\left[ \sum_{i \in S_{k+1}} g_{i}(x_{k+1}; x_k) - f_{i}(x_{k+1}) \right] \]
\[ = \sum_{k=0}^{\infty} \mathbb{E}_{k+1}\left[ \sum_{i \in S_{k+1}} g_{i}(x_{k+1}; x_k) - f_{i}(x_{k+1}) \right] \]
\[ \leq \tau \sum_{k=0}^{\infty} \mathbb{E}\left[ g \left( x_{k+1}; \hat{x}_k \right) - f(x_{k+1}) \right], \]
where we used the Beppo–Lèvy theorem to interchange the expectation and the sum in front of nonnegative quantities (see also [20] for a similar derivation in the case $\tau = 1$). Moreover, considering (13), we obtain
\[ \mathbb{E}\left[ \sum_{k=0}^{+\infty} \sum_{i \in S_{k+1}} g_{i}(x_{k+1}; x_k) - f_{i}(x_{k+1}) \right] \leq \tau \sum_{k=0}^{+\infty} \mathbb{E}\left[ f(x_k) - f(x_{k+1}) \right] \]
\[ \leq \tau \left( f(x_0) - f^* \right) < \infty, \]
where for the last inequality we used that $f$ is assumed lower bounded, e.g. by $f^* > -\infty$. Hence, nonnegative sequence $(g(x_{k+1}; \hat{x}_k) - f(x_{k+1}))_{k \geq 0}$ converges to 0 almost surely, and thus sequence of function values $(f(x_k))_{k \geq 0}$ also converges almost surely.
For the second part of the theorem, we have from the optimality of \( x_{k+1} \) that
\[
\nabla g(x_{k+1}; \hat{x}_k)[y - x_{k+1}] \geq 0 \quad \forall y \in \mathcal{X}.
\]

Thus, from the definition of the error function \( h \) and the fact that \( f \) is assumed to have directional derivatives, we have using basic calculus rules that
\[
\nabla f(x_{k+1})[y - x_{k+1}] = \nabla g(x_{k+1}; \hat{x}_k)[y - x_{k+1}] - \nabla h(x_{k+1}; \hat{x}_k)[y - x_{k+1}]
\geq -\langle \nabla h(x_{k+1}; \hat{x}_k), y - x_{k+1} \rangle \quad \forall y \in \mathcal{X}.
\]

Using the Cauchy–Schwarz inequality in the previous relation, we get
\[
\nabla f(x_{k+1})[y - x_{k+1}] \geq -\| \nabla h(x_{k+1}; \hat{x}_k) \| \| y - x_{k+1} \| \quad \forall y \in \mathcal{X}.
\]

For simplicity of the notation, let us define the function \( \tilde{h}(y) = h(y; \hat{x}_k) \), which, according to the Definition 3.1, has the \( p \) derivative smooth with Lipschitz constant \( L_p^h \). Further, let us consider the following auxiliary point:
\[
y_{k+1} = \arg \min_{y \in \mathbb{R}} \tilde{T}_p^h(y; x_{k+1}) + \frac{M_p^h}{(p + 1)!} \| y - x_{k+1} \|^{p+1},
\]
where \( M_p^h > L_p^h \). Let us consider first the nontrivial case \( p \geq 2 \). Using the (global) optimality of \( y^{k+1} \), we have
\[
\tilde{T}_p^h(y_{k+1}; x_{k+1}) + \frac{M_p^h}{(p + 1)!} \| y_{k+1} - x_{k+1} \|^{p+1}
\leq \tilde{T}_p^h(x_{k+1}; x_{k+1}) + \frac{M_p^h}{(p + 1)!} \| x_{k+1} - x_{k+1} \|^{p+1}
= h(x_{k+1}; \hat{x}_k) + \sum_{i=1}^{p} \frac{1}{i!} D^i h(x_{k+1}; \hat{x}_k)[x_{k+1} - x_{k+1}]^i = h(x_{k+1}; \hat{x}_k).
\]

Moreover, developing the left term of the previous inequality, we get
\[
h(x_{k+1}; \hat{x}_k) + \sum_{i=1}^{p} \frac{1}{i!} D^i h(x_{k+1}; \hat{x}_k)[y_{k+1} - x_{k+1}]^i + \frac{M_p^h}{(p + 1)!} \| y_{k+1} - x_{k+1} \|^{p+1}
\leq h(x_{k+1}; \hat{x}_k).
\]

In conclusion, we obtain
\[
\sum_{i=1}^{p} \frac{1}{i!} D^i h(x_{k+1}; \hat{x}_k)[y_{k+1} - x_{k+1}]^i + \frac{M_p^h}{(p + 1)!} \| y_{k+1} - x_{k+1} \|^{p+1} \leq 0. \quad (15)
\]

From relation (2), we also have
\[
\tilde{h}(y) = h(y; \hat{x}_k) \leq \tilde{T}_p^h(y; x_{k+1}) + \frac{L_p^h}{(p + 1)!} \| y - x_{k+1} \|^{p+1} \quad \forall y \in \mathbb{R}.
\]
We rewrite this relation for our chosen auxiliary point \( y_{k+1} \) and get

\[
h(y_{k+1}; \hat{x}_k) \leq T^h_p(y_{k+1}; x_{k+1}) + \frac{L^h_p}{(p+1)!} \|y_{k+1} - x_{k+1}\|^{p+1}
\]

\[
= h(x_{k+1}; \hat{x}_k) + \sum_{i=1}^{p} \frac{1}{i!} D^i h(x_{k+1}; \hat{x}_k) [y_{k+1} - x_{k+1}]^i + \frac{L^h_p}{(p+1)!} \|y_{k+1} - x_{k+1}\|^{p+1}
\]

\[
\leq h(x_{k+1}; \hat{x}_k) - \frac{M^h_p - L^h_p}{(p+1)!} \|y_{k+1} - x_{k+1}\|^{p+1}.
\]

Recalling that \( h \) is nonnegative, then it follows that

\[
0 \leq h(x_{k+1}; \hat{x}_k) - \frac{M^h_p - L^h_p}{(p+1)!} \|y_{k+1} - x_{k+1}\|^{p+1}.
\]

This leads to

\[
h(x_{k+1}; \hat{x}_k) \geq \frac{M^h_p - L^h_p}{(p+1)!} \|y_{k+1} - x_{k+1}\|^{p+1} \geq 0. \tag{16}
\]

Since \( h(x_{k+1}; \hat{x}_k) = g(x_{k+1}; \hat{x}_k) - f(x_{k+1}) \), using the first part of the proof we get that the sequence \( (h(x_{k+1}; \hat{x}_k))_{k \geq 0} \) converges to 0 almost sure. Hence, the sequence \( (y_{k+1} - x_{k+1})_{k \geq 0} \) converges to 0 almost sure. Moreover, from the optimality conditions for \( y_{k+1} \), we have

\[
\nabla h(x_{k+1}; \hat{x}_k) + H_{k+1} [y_{k+1} - x_{k+1}] = 0, \tag{17}
\]

where we denote the matrix

\[
H_{k+1} = \nabla^2 h(x_{k+1}; \hat{x}_k) + \sum_{i=3}^{p} D^i h(x_{k+1}; \hat{x}_k) [y_{k+1} - x_{k+1}]^{i-2} + \frac{M^h_p}{p!} \|y_{k+1} - x_{k+1}\|^{p-1} B.
\]

Since the sequence \((x_k)_{k \geq 0}\) generated by SHOM is assumed bounded and \( h \) is \( p \) times continuously differentiable, then \( \nabla^i h(x_{k+1}; \hat{x}_k) \) are bounded for all \( i = 2:p \). Moreover, since \( y_{k+1} - x_{k+1} \to 0 \) as \( k \to \infty \), it follows that \( y_{k+1} \) is also bounded (due to boundedness of \( x_k \)). Therefore, for \( p \geq 2 \) we have proved that \( H_{k+1} \) is bounded and consequently:

\[
\nabla h(x_{k+1}; \hat{x}_k) \to 0 \ \text{as} \ k \to \infty \ \text{almost sure}.
\]

For \( p = 1 \) we can just take \( y_{k+1} = x_{k+1} - 1/L^h_1 \nabla h(x_{k+1}; \hat{x}_k) \). Then, using that \( h(\cdot; \hat{x}_k) \) has gradient Lipschitz with constant \( L^h_1 \), from (2) for \( h \) and \( p = 1 \), we obtain

\[
0 \leq h(y_{k+1}; \hat{x}_k) \leq h(x_{k+1}; \hat{x}_k) - \frac{1}{2L^h_1} \|\nabla h(x_{k+1}; \hat{x}_k)\|_*^2,
\]

which further yields

\[
\frac{1}{2L^h_1} \|\nabla h(x_{k+1}; \hat{x}_k)\|_*^2 \leq h(x_{k+1}; \hat{x}_k) \to 0 \ \text{as} \ k \to \infty,
\]
as \( h(x_{k+1}; \hat{x}_k) \) converges to zero almost surely. Therefore, also in the case \( p = 1 \) we have \( \nabla h(x_{k+1}; \hat{x}_k) \to 0 \). Using this result in (14), minimizing over \( y \in X \) and taking the infimum limit, we get that the sequence \((x_k)_{k \geq 0}\) satisfies the asymptotic stationary point condition for the nonconvex problem (6). This concludes our statements.

Note that the main difficulty in the previous proof is to handle \( h \) having \( p \geq 2 \) derivatives smooth. We overcome this difficulty by introducing a new sequence \((y_k)_{k \geq 0}\), proving that it has similar properties as the sequence \((x_k)_{k \geq 0}\) generated by SHOM and then using the optimality conditions for \( y_k \) instead of \( x_k \). Note that in practice we do not need to compute \( y_k \). Moreover, if \( f \) is differentiable and \( X = E \), the previous theorem states that the sequence of gradients \((\nabla f(x_k))_{k \geq 0}\) converges to 0 almost surely.

4.1. Convergence rates under Kurdyka–Lojasiewicz property

Recent works demonstrate that special geometric properties of the objective function, such as the Kurdyka–Lojasiewicz (KL) property, enable fast convergence rates for some first- or higher-order algorithms, see, e.g. [5, 26]. However, from our knowledge, this is the first work deriving convergence rates for stochastic higher-order majorization–minimization algorithms under the KL property. The main challenge when analysing the convergence of a stochastic algorithm under the KL is the fact that it is not trivial to combine the local property of the KL with the stochasticity of the iterates. We overcome this problem using Egorov’s theorem, which allows us to pass from almost sure convergence to uniform convergence. In this section, we derive convergence rates for SHOM under the KL condition.

For simplicity of the exposition, we assume in this section \( X = E \) and the functions \( f \) and \( g \) are subdifferentiable on \( E \). Denote

\[ S_f(x) = \text{dist}(0, \partial f(x)) := \inf_{f^* \in \partial f(x)} \| f^* \|_* \quad \forall x \in E. \]

If \( \partial f(x) = \emptyset \), we set \( S_f(x) = \infty \). We also assume that the optimization problem (6) has finite optimal value. Let \( \Omega \) be the set of limit points of the sequence \((x_k)_{k \geq 0}\) generated by SHOM. From Lemma 4.2, we note that any limit point \( x^* \in \Omega \) satisfies the first-order stationary point condition for the original problem (6). Let us prove that the objective function \( f \) is constant on the set of limit points of the sequence \((x_k)_{k \geq 0}\), i.e. \( \Omega \). Indeed, from Lemma 4.2 we have that \((f(x_k))_{k \geq 0}\) converges almost surely to, e.g. \( f_* > -\infty \). Let \( x_* \) be a limit point of \((x_k)_{k \geq 0}\). Then there exists a subsequence \((x_{k_\ell})_{\ell \geq 0}\) such that \( x_{k_\ell} \to x_* \), i.e. \( x_* \in \Omega \). In this section only, we assume that \( f \) is continuous. Then we obtain \( \lim_{\ell \to \infty} f(x_{k_\ell}) = f(x_*) = f_* \text{ a.s.} \), i.e. \( f(\Omega) = f_* \). Based on these properties of \((x_k)_{k \geq 0}\), in the rest of this section we assume that \( f \) satisfies additionally the following KL inequality on a neighbourhood \( U \) of \( \Omega \) and for some \( \varepsilon > 0 \), i.e. [5]:

\[ f(x) - f_* \leq \sigma_q S_f(x)^q \quad \forall x \in U, \quad f_* < f(x) < f_* + \varepsilon, \tag{18} \]

where the KL parameter \( q > 1 \) and \( \sigma_q > 0 \). Note that the KL property holds for a large class of functions including semi-algebraic functions and uniformly convex functions, see [5] for more examples. The next lemma will be useful later in this section. Denote by \( 1_{\Gamma} \) the indicator function of a set \( \Gamma \).
Lemma 4.3: Given a bounded random variable $0 \leq X \leq \Delta$ and a measurable set $\Gamma_{\delta}$ such that $\mathbb{P}(\Gamma_{\delta}) \geq 1 - \delta$, the following inequality holds:

$$\mathbb{E}[X] - \Delta \sqrt{\delta} \leq \mathbb{E}[X1_{\Gamma_{\delta}}] \leq \mathbb{E}[X].$$

Proof: The inequality from the right-hand side is obvious, since $X \geq 0$. For the other inequality, we can write $X = X1_{\Gamma_{\delta}} + X1_{\Gamma_{\delta}^C}$, where $\Gamma_{\delta}^C$ is the complement of the set $\Gamma_{\delta}$. Then, using the Cauchy–Schwartz inequality and the boundedness of $X$, we have

$$\mathbb{E}[X] - \mathbb{E}[X1_{\Gamma_{\delta}}] = \mathbb{E}[X] \leq (\mathbb{E}[X^2] \mathbb{E}[1_{\Gamma_{\delta}^C}^2])^{1/2} \leq (\Delta^2 \mathbb{P}(\Gamma_{\delta}^C))^{1/2} \leq (\Delta^2)^{1/2}.$$}

Note that since $f(x_k) \to f_*$ a.s., then there exists $\Delta < \infty$ such that $|f(x_k) - f_*| \leq \Delta$ for all $k \geq 0$. We get the following convergence rates for SHOM under the KL.

Theorem 4.4: Let the objective function $f$ be proper and continuous (possibly nonconvex), satisfy the KL property (18) for some $q > 1$ and $\varepsilon > 0$, admit a stochastic $p \geq 1$ higher-order surrogate $g$ as in Definition 3.1 and $M_p^h > L_p^h$. Additionally, assume that the sequence $(x_k)_{k \geq 0}$ generated by SHOM satisfies $0 \leq f(x_k) - f_* \leq \Delta$ for all $k \geq 0$. Then, for any $\delta > 0$ exists $k_{\delta,\varepsilon}$ such that with probability at least $1 - \delta$:

1. If $q = p + 1$, then locally $f(x_k)$ converges linearly to a neighbourhood of $f_*$:

$$\mathbb{E}[f(x_k)] - f_* \leq \left(\frac{\sigma q^q q!}{M_p^h - L_p^h + \sigma q^q q!}\right)^{k-k_{\delta,\varepsilon}} \mathbb{E}[f(x_{k_{\delta,\varepsilon}})] - f_* + \Delta \sqrt{\delta},$$

where $c = \max_{k \geq 0} \|H_{k+1}\|$ and $k > k_{\delta,\varepsilon}$.

2. If $q < p + 1$, then locally $f(x_k)$ converges sublinearly to a neighbourhood of $f_*$:

$$\mathbb{E}[f(x_k)] - f_* \leq \frac{\mathbb{E}[f(x_{k_{\delta,\varepsilon}})] - f_* - \Delta \sqrt{\delta}}{(1 + \alpha(k - k_{\delta,\varepsilon}))^{p+q}},$$

where $\alpha > 0$ is some appropriate constant and $k > k_{\delta,\varepsilon}$.

Proof: From the proof of Lemma 4.2 (i.e. combining (16) with $h(x_{k+1}; \hat{x}_k) = g(x_{k+1}; \hat{x}_k) - f(x_{k+1})$ and then taking expectation and the inequality (13)), we have

$$(M_p^h - L_p^h)((p + 1)!)^{-1} \mathbb{E} \left[ \|y_{k+1} - x_{k+1}\|^{p+1} \right] \leq \mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})] \quad \forall k \geq 0. \quad (19)$$

Therefore, in expectation the cost function in combination with the sequences $x_k$ and $y_k$ can be used as a Lyapunov function. From the optimality condition for $x_{k+1}$, we have $0 \in \partial g(x_{k+1}; \hat{x}_k)$ and from $h = g - f$ we get $-\nabla h(x_{k+1}; \hat{x}_k) \in \partial f(x_{k+1})$ the (limiting) subdifferential of $f$ at $x_{k+1}$. Then, using (17) we obtain

$$S_f(x_{k+1}) = \text{dist}(0, \partial f(x_{k+1})) \leq \|\nabla h(x_{k+1}; \hat{x}_k)\|_* = \|H_{k+1}(y_{k+1} - x_{k+1})\|_* \leq c \cdot \|y_{k+1} - x_{k+1}\| \quad \forall k \geq 0,$$

where $c = \max_{k \geq 0} \|H_{k+1}\| < \infty$ according to Lemma 4.2. Further, from Lemma 4.2 we have that $f(x_k) \to f_*$ a.s., which means that there exists some measurable set $\Gamma$ such
that \( \mathbb{P}[\Gamma] = 1 \) and for \( \varepsilon \) and \( \gamma \in \Gamma \) there exists \( k_{\varepsilon}(\gamma) \) such that for any \( k > k_{\varepsilon}(\gamma) \) we have \( f(x_k(\gamma)) - f_\ast \leq \sigma_q S_f(x_k(\gamma)) \). Further, invoking measure theoretic arguments to pass from almost sure convergence to almost uniform convergence thanks to Egorov’s theorem, we can prove that for any \( \delta > 0 \) there exists a measurable set \( \Gamma_{\delta} \subseteq \Gamma \), such that \( \mathbb{P}([\Gamma_{\delta}]) \geq 1 - \delta \), and for any \( \varepsilon > 0 \) there exists \( k_{\delta,\varepsilon} > 0 \) such that for all \( \gamma \in \Gamma_{\delta} \) and \( k \geq k_{\delta,\varepsilon} \) we have \( f(x_k(\gamma)) - f_\ast \leq \sigma_q S_f(x_k(\gamma)) \). Hence, with probability at least \( 1 - \delta \) the sequence \( (x_k)_{k \geq 0} \) satisfies KL on \( \Gamma_{\delta} \) for all \( k \geq k_{\delta,\varepsilon} \), i.e. \( (f(x_k) - f_\ast) 1_{\Gamma_{\delta}} \leq \sigma_q S_f(x_k) 1_{\Gamma_{\delta}} \) for all \( k \geq k_{\delta,\varepsilon} \). Then, with probability at least \( 1 - \delta \):

\[
(f(x_{k+1}) - f_\ast) 1_{\Gamma_{\delta}} \leq \sigma_q S_f(x_{k+1}) \leq \sigma_q S_f(x_k) \leq \sigma_q \|y_{k+1} - x_{k+1}\|^q.
\]

Denote \( \Delta_k = \mathbb{E}[f(x_k)] - f_\ast \) and \( C = \sigma_q c^d \frac{(p + 1)!}{M_p^q - L_p^q} \frac{\sqrt{p+1}}{\sqrt{p+1}} \). If \( q = p + 1 \), using Lemma 4.3, taking expectation in (21) and combining with (19), we get

\[
\Delta_{k+1} - \Delta \leq C(\Delta_k - \Delta_{k+1}) \quad \forall k \geq k_{\delta,\varepsilon},
\]

which proves the first statement. If \( q < p + 1 \), using Lemma 4.3, taking expectation in (21) and combining with (19), we obtain the following:

\[
\mathbb{E}[f(x_{k+1})] - f_\ast - \Delta \leq \sigma_q c^d \mathbb{E}[S_f(x_{k+1})] \leq \sigma_q c^d \mathbb{E}[\|y_{k+1} - x_{k+1}\|^p] \leq C \mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})],
\]

where in the third inequality we used concavity of the function \( u \mapsto (u)^{\frac{q}{p+1}} \) on \( \mathbb{R}_+ \) for \( q < p + 1 \). Hence, in this case we obtain the following recurrence:

\[
\Delta_{k+1} - \Delta \leq C(\Delta_k - \Delta_{k+1}) \quad \forall k \geq k_{\delta,\varepsilon}.
\]

If we denote \( \Delta_k - \Delta \leq C \frac{p+1}{p+1+q} \Delta_k \), we further get the recurrence:

\[
\tilde{\Delta}_{k+1} \leq \tilde{\Delta}_k - \tilde{\Delta}_{k+1} \quad \text{with} \quad \frac{p + 1 - q}{q} > 0,
\]

and then using similar arguments as in Lemma 11 in [22], we obtain the following sublinear rate:

\[
\tilde{\Delta}_k \leq \frac{\tilde{\Delta}_{k_{\delta,\varepsilon}}}{(1 + \tilde{\alpha}(k - k_{\delta,\varepsilon}))^{\frac{q}{p+1+q}}}, \quad \forall k > k_{\delta,\varepsilon},
\]

for some appropriate constant \( \tilde{\alpha} > 0 \), which proves the second statement.

Previous theorem establishes linear or sublinear convergence results when the objective function satisfies the KL property without imposing convexity requirements on it. For \( p = 1 \) and \( q = 2 \), we also get new convergence rates for MISO algorithm from [20] in the general nonsmooth and nonconvex settings. In the next section, we derive convergence rates for convex objective functions.
5. Convex convergence analysis

Throughout this section, we assume that the objective function $f$ is convex and $x_{k+1}$ is the global minimum of subproblem (7). Note that if the functions $f_i$’s are all convex, then for appropriate choices of constants $M_p$’s (see Lemma 3.3), the subproblem (7) remains convex and then the global minimum $x_{k+1}$ can be computed using convex optimization algorithms. We analyse further the global convergence of the algorithm SHOM in the convex settings. Let $f^*$ and $x^*$ be the optimal value and a minimizer of the convex optimization problem (6). As usually assumed in convex optimization (see also, e.g. [21,23]), we consider that there exists $R > 0$ such that:

$$\|x_k - x^*\| \leq R \quad \forall k \geq 0. \quad (25)$$

**Theorem 5.1:** Let the objective function $f$ be convex and admit a stochastic $p \geq 1$ higher-order surrogate function $g$ as given in Definition 3.1 and $\bar{L}_p = \max_{i=1:N} L_i^{h_i}$. Moreover, let the sequence $(x_k)_{k \geq 0}$ generated by SHOM with $\tau = 1$ satisfy (25) and $x_{k+1}$ be the global minimum of subproblem (7). Then $f(x_k)$ converges globally in expectation to $f^*$ at a sublinear rate:

$$E[f(x_k)] - f^* \leq \frac{\bar{L}_p R^{p+1}}{p!} \left( \frac{p + 1}{k} \right)^p.$$  

**Proof:** Using Definition 3.1 and $x_{k+1}$ the global minimum of subproblem (7), we have

$$f(x_{k+1}) \leq g(x_{k+1}; \hat{x}_k) \leq g(x; \hat{x}_k) \quad \forall x \in \mathcal{X}.$$  

Furthermore, taking expectation and then the minimum over $\mathcal{X}$, we obtain

$$E \left[ f(x_{k+1}) \right] \leq \min_{x \in \mathcal{X}} E \left[ g(x; \hat{x}_k) \right] = \min_{x \in \mathcal{X}} E \left[ E_k \left[ g_i(x; x_k) \right] \right]$$

$$\leq \min_{x \in \mathcal{X}} E \left[ f_i(x) + \frac{L_i^{h_i}}{(p + 1)!} \|x - x_k\|^{p+1} \right]$$

$$\leq \min_{x \in \mathcal{X}} E \left[ f(x) + \frac{\bar{L}_p^{h_i}}{(p + 1)!} \|x - x_k\|^{p+1} \right],$$  

where $\bar{L}_p = \max_{i=1:N} L_i^{h_i}$. Further, our proof follows a similar reasoning as in [21] (Theorem 2). For $k = 1$, using the above inequality, we get

$$E \left[ f(x_1) \right] \leq \min_{x \in \mathcal{X}} E \left[ f(x) + \frac{\bar{L}_p}{(p + 1)!} \|x - x_0\|^{p+1} \right] \leq f^* + \frac{\bar{L}_p}{(p + 1)!} R^{p+1},$$

and denoting $\Delta_k = E[f(x_k)] - f^*$, we obtain

$$\Delta_1 \leq \frac{\bar{L}_p}{(p + 1)!} R^{p+1}. \quad (27)$$
Taking \( x = \alpha x^* + (1 - \alpha)x_k \) and using the convexity of \( f \) in (26), we get for \( k \geq 1 \):

\[
\mathbb{E} \left[ f(x_{k+1}) \right] \leq \min_{\alpha \in [0,1]} \alpha f^* + (1 - \alpha)\mathbb{E}[f(x_k)] + \frac{\tilde{L}_p}{(p+1)!} \alpha^{p+1} R^{p+1}.
\]

The optimal \( \alpha \) in the previous problem has the expression:

\[
0 < \alpha^* = \left( \frac{\Delta_k p!}{R^{p+1} \tilde{L}_p} \right)^\frac{1}{p} \leq \left( \frac{\Delta_1 p!}{R^{p+1} \tilde{L}_p} \right)^\frac{1}{p} \leq \left( \frac{1}{p+1} \right)^\frac{1}{p} < 1 \quad \forall k \geq 1.
\] (3)

Thus we conclude that

\[
\mathbb{E} \left[ f(x_{k+1}) \right] \leq \mathbb{E} \left[ f(x_k) \right] - \frac{p}{p+1} \alpha^* \Delta_k \Leftrightarrow \Delta_{k+1} \leq \left( 1 - \frac{p}{p+1} \alpha^* \right) \Delta_k
\]

\[
\Leftrightarrow \Delta_k - \Delta_{k+1} \geq \frac{p}{p+1} \alpha^* \Delta_k \Leftrightarrow \frac{p}{p+1} \left( \frac{p!}{R^{p+1} \tilde{L}_p} \right)^\frac{1}{p} \Delta_k^{\frac{p}{p+1}}.
\]

Using further the same arguments as in Theorem 2 in [21] for the previous recurrence, we obtain the statement.

Note that in the previous theorem, we do not require each individual function \( f_i \), with \( i = 1:N \), to be convex, only the finite sum function \( f \) needs to be convex. In the sequel we prove that SHOM can achieve locally faster rates when the finite-sum objective function \( f \) is uniformly convex. More precisely, we get local (super)linear convergence rates for SHOM in several optimality criteria: function values and distance of the iterates to the optimal point. For this we first need some auxiliary results. Let \( x^* \) be the unique solution of (6) and assume, for simplicity, that \( L_{h_j} = L_h \) for all \( j = 1:N \).

**Lemma 5.2:** Let \( f \) be uniformly convex of degree \( q \geq 2 \) and constant \( \sigma_q \) and admit a stochastic \( p \geq 1 \) higher-order surrogate function \( g \) over \( X \) as given in Definition 3.1. Then the sequence \( (x_k)_{k \geq 0} \) generated by SHOM satisfies

\[
\frac{\sigma_q}{q} \| x_{k+1} - x^* \|^q \leq f(x_{k+1}) - f(x^*) \leq \frac{L_h}{N(p+1)!} \sum_{j=1}^N \| x_j^i - x^* \|^{p+1}.
\] (29)

**Proof:** Since \( h_j(y;x_k^i) = g_j(y;x_k^i) - f_j(y) \) has the \( p \) derivative smooth with Lipschitz constant \( L_p \), then from (2) we have

\[
|h_j(y;x_k^i) - T_p^h_j(y;x_k^i)| \leq \frac{L_h}{(p+1)!} \| y - x_k^i \|^{p+1} \quad \forall y.
\]

Note that from property 3 of Definition 3.1, we have

\[
T_p^h_j(y;x_k^i) = h_j(x_k^i;x_k^i) + \sum_{i=1}^p \frac{1}{i!} D^i h_j(x_k^i;x_k^i) [y - x_k^i]^i = 0, \quad \forall j = 1:N.
\]
Therefore, summing over \( j \) in the previous inequality, we obtain
\[
\frac{1}{N} \sum_{j=1}^{N} |h_j(y;x_k^j)| \leq \frac{1}{N} \frac{L_p^h}{(p+1)!} \sum_{j=1}^{N} \|y - x_k^j\|^{p+1}.
\]

Further, since the surrogate function \( h_j(\cdot;x_k^j) \geq 0 \), we get
\[
h(y;\hat{x}_k) \leq \frac{L_p^h}{N(p+1)!} \sum_{j=1}^{N} \|y - x_k^j\|^{p+1}.
\]

On the other hand, from the global optimality of \( x_{k+1} \), we have
\[
f(x_{k+1}) \leq g(x_{k+1};\hat{x}_k) \leq g(y;\hat{x}_k) = f(y) + h(y;\hat{x}_k) \quad \forall y \in X.
\]

Using (30) in the above inequality, we obtain
\[
f(x_{k+1}) \leq f(y) + \frac{L_p^h}{N(p+1)!} \sum_{j=1}^{N} \|y - x_k^j\|^{p+1} \quad \forall y \in X.
\]

Choosing \( y = x^\ast \) we obtain the right side of (29). On the other hand, from the uniform convexity relation (5) for the function \( f \) at \( x = x^\ast \), we have
\[
f(x_{k+1}) \geq f(x^\ast) + \frac{\sigma_q q}{q} \|x^\ast - x_{k+1}\|^q,
\]
which yields the left side of relation (29). This proves our statements. \( \square \)

### 5.1. Local (super)linear convergence in function values

In this section, we prove local (super)linear convergence in function values for SHOM, provided that the objective function is uniformly convex. Note that in Lemma 4.2, we have proved already that the objective function decreases along the iterates of SHOM. From Lemma 5.2, we obtain the following relation:

**Lemma 5.3:** Let \( f \) be uniformly convex of degree \( q \geq 2 \) with constant \( \sigma_q \) and admit a stochastic \( p \geq 1 \) higher-order surrogate function \( g \) over \( X \) as given in Definition 3.1. Consider the following Lyapunov function along the sequence \( \hat{x}_k \) generated by SHOM:
\[
\Phi_k = \frac{1}{N} \sum_{j=1}^{N} (f(x_k^j) - f(x^\ast))^\frac{p+1}{q}.
\]

Then, we have
\[
f(x_{k+1}) - f(x^\ast) \leq \frac{L_p^h}{(p+1)!} \left( \frac{q}{\sigma_q} \right)^\frac{p+1}{q} \Phi_k.
\]

(31)
Proof: From the left-hand side of relation (29), we have for all $j = 1:N$ that
\[
\left[ \frac{q}{\sigma_q} (f(x_j^k) - f(x^*)) \right]^{\frac{1}{q}} \geq \|x^* - x_j^k\|. \quad (32)
\]
Returning with this relation into the right side of (29), we obtain
\[
f(x_{k+1}) - f(x^*) \leq \frac{L_p}{N(p+1)!} \left( \frac{q}{\sigma_q} \right)^{p+1} q N \sum_{j=1}^{N} (f(x_j^k) - f(x^*))^{\frac{p+1}{q}}. \quad (33)
\]
Thus, proving the relation (31).

Note that a similar Lyapunov function for $p = 2$ and $q = 2$ was considered in [16].

Lemma 5.4: Let $f$ be uniformly convex of degree $q \geq 2$ with constant $\sigma_q$ and admit a stochastic $p \geq 1$ higher-order surrogate function $g$ over $X$ as given in Definition 3.1. Furthermore, let $\alpha \in (0,1)$ and $C = \frac{L_p}{(p+1)!} \left( \frac{q}{\sigma_q} \right)^{p+1}$. If $q < p + 1$, then the following implication holds with probability 1:
\[
f(x_0) - f(x^*) \leq \frac{\alpha^{\frac{q}{p+1-q}}}{C^{p+1-q}} \Rightarrow \Phi_k \leq \frac{\alpha^{\frac{q}{p+1-q}}}{C^{p+1-q}} \forall k \geq 0. \quad (34)
\]
Proof: The statement follows as a corollary of a stronger result, that is for any $j = 1:N$ and $k \geq 0$ we have with probability 1 that
\[
(f(x_j^k) - f(x^*))^{\frac{p+1}{q}} \leq \frac{\alpha^{\frac{q}{p+1-q}}}{C^{p+1-q}} \forall j = 1:N.
\]
Let us show this by induction in $k$. For $k = 0$ it is satisfied by our assumption. Let us assume that it holds for $k$. If $j \notin S_{k+1}$, then $x_{k+1}^j = x_k^j$ and from the induction step $k$ we also get that it is valid for $k + 1$. If $j \in S_{k+1}$, we have
\[
f(x_{k+1}^j) - f(x^*) = f(x_{k+1}^j) - f(x^*) \leq \frac{C}{N} \sum_{j=1}^{N} (f(x_j^k) - f(x^*))^{\frac{p+1}{q}} \leq \frac{C}{N} \sum_{j=1}^{N} (f(x_j^k) - f(x^*))^{\frac{p+1}{q}} \leq \frac{\alpha^{\frac{q}{p+1-q}}}{C^{p+1-q}},
\]
as $\alpha \in (0,1)$ and $q < p + 1$. This shows that
\[
(f(x_{k+1}^j) - f(x^*))^{\frac{p+1}{q}} \leq \frac{\alpha^{\frac{q}{p+1-q}}}{C^{p+1-q}} \forall j = 1:N.
\]
This concludes our induction and the statement of the lemma.
Theorem 5.5: Let $f$ be uniformly convex of degree $q \geq 2$ with constant $\sigma_q$ and admit a stochastic $p \geq 1$ higher-order surrogate function $g$ over $\mathcal{X}$ as given in Definition 3.1. Then the Lyapunov function along the sequence $\hat{x}_k$ generated by SHOM, $\Phi_k = \frac{1}{N} \sum_{j=1}^{N} (f(x_j^k) - f(x^*))^{\frac{p+1}{q}}$, satisfies the recursion:

$$\mathbb{E}_k [\Phi_{k+1}] \leq \left[ 1 - \frac{\tau}{N} + \frac{\tau}{N} \left( \frac{\frac{p+1}{q} \frac{L_h}{\sigma_q^q} (p+1)!}{\sigma_q^q} \right)^{\frac{p+1}{q}} \Phi_k \right] \Phi_k.$$

Furthermore, let $\alpha \in (0, 1)$ and $C = \frac{L_h}{(p+1)!} \left( \frac{q+1}{q} \right)^{\frac{p+1}{q}}$. If $q < p + 1$ and $f(x_0) - f(x^*) \leq \alpha \frac{\sigma^2}{q(q+1)^2 (p+1)!} C \frac{1}{q^{p+q}}$, then we have the following linear rate:

$$\mathbb{E}_k [\Phi_{k+1}] \leq \left( 1 - \left( \frac{1-\alpha}{N} \right) \right) \Phi_k.$$

Proof: Let us note that for all $j = 1:N$, the points $x_{k+1}^j$ are drawn from the following conditional probability distribution:

$$\mathbb{P} \left( x_{k+1}^j = x_{k+1} \mid \mathcal{F}_k \right) = \delta \text{ and } \mathbb{P} \left( x_{k+1}^j = x_k^j \mid \mathcal{F}_k \right) = 1 - \delta,$$

where $\delta := \tau/N$ and $x_0^j = x_0$ for all $j = 1:N$. Thus we have for all $k$:

$$\mathbb{E}_k [\Phi_{k+1}] = \delta \mathbb{E}_k \left[ (f(x_{k+1}) - f(x^*))^{\frac{p+1}{q}} \right] + (1 - \delta) \Phi_k \leq \delta \mathbb{E}_k \left[ (C \Phi_k)^{\frac{p+1}{q}} \right] + (1 - \delta) \Phi_k$$

$$= \delta \left( C \Phi_k \right)^{\frac{p+1}{q}} + (1 - \delta) \Phi_k$$

$$= \left( 1 - \delta + \delta C \frac{\Phi_k}{q^{\frac{p+1-q}{q}}} \right) \Phi_k \lesssim (1 - (1-\alpha)\delta) \Phi_k.$$ 

This proves our statements. ■

Remark 5.1: Although for stochastic first-order variance reduction methods, such as SAGA [9], SVRG [15] or SARAH [29], one can also obtain linear convergence rates, they depend on the condition number of the problem and consequently they converge slow on ill-conditioned problems. On the other hand, the convergence rates from Theorem 5.5 are independent of the conditioning, thus, SHOM provably adapts to the curvature of the objective function. For example, for the particular case $p = q = 2$ and $\alpha = \frac{1}{2}$ we get a very fast linear rate depending on the minibatch size $\tau$:

$$\mathbb{E}_k [\Phi_{k+1}] \leq \left( 1 - \frac{\tau}{2N} \right) \Phi_k.$$ 

A similar convergence rate as in (36) was derived in [16] for an algorithm similar to SHOM, but for the particular surrogate function of Example 3.2 with $p = 2$ (i.e. second-order Taylor expansion with cubic regularization) and strongly convex objective (i.e.
\( q = 2 \). Our results cover more general problems (including composite objective, see Examples 3.4, 3.5, 3.6) and uniform convex cost.

Further, from Lemma 5.3 and Theorem 5.5 we get that \( \mathbb{E}[f(x_k)] - f(x^*) \) converges linearly to 0, locally. Moreover, when \( \tau = N = 1 \) SHOM becomes deterministic and for such algorithm we get from the first statement of Theorem 5.5 local superlinear convergence in function values, i.e. for \( q < p + 1 \) we get

\[
f(x_{k+1}) - f(x^*) \leq C(f(x_k) - f(x^*))^{\frac{p+1}{q}}.
\]

A similar convergence result has been obtained in [10] for a deterministic algorithm derived from the particular surrogate function of Example 3.4 with \( N = 1 \) (Taylor expansion with regularization). However, our convergence analysis of SHOM is derived for general surrogates. Note that there is a major difference between the Taylor expansion and the model approximation based on general surrogates (majorization–minimization framework). Taylor expansion is unique. The idea of the majorization–minimization approach is the opposite. A given objective function may admit many upper bound models and every model leads to a different optimization algorithm.

### 5.2. Local (super)linear convergence in the iterates

In this section, we prove local linear convergence in the iterates of SHOM, provided that the objective function is uniformly convex. From Lemma 5.2, we obtain immediately the following relation:

**Lemma 5.6:** Let \( f \) be uniformly convex of degree \( q \geq 2 \) with constant \( \sigma_q \) and admit a stochastic \( p \geq 1 \) higher-order surrogate function \( g \) over \( \mathcal{X} \) as in Definition 3.1. If we consider the following Lyapunov function along the sequence \( \hat{x}_k \) generated by SHOM:

\[
W_k = \frac{1}{N} \sum_{j=1}^{N} \|x_j^k - x^*\|^{p+1}.
\]

then, we have

\[
\|x_{k+1} - x^*\|^q \leq \frac{t_{p,q}^k}{\sigma_q(p+1)!} W_k.
\]  

(37)

**Lemma 5.7:** Let \( f \) be uniformly convex of degree \( q \geq 2 \) with constant \( \sigma_q \) and admit a stochastic \( p \geq 1 \) higher-order surrogate function \( g \) over \( \mathcal{X} \) as given in Definition 3.1. Furthermore, let \( \alpha \in (0, 1) \) and \( \hat{C} = \frac{t_{p,q}^k}{\sigma_q(p+1)!} \). If \( q < p + 1 \), then the following implication holds with probability 1:

\[
\|x_0 - x^*\| \leq \alpha^{\frac{q}{p+1-q(p+1)}} \hat{C}^{\frac{1}{p+1-q}} \Rightarrow W_k \leq \frac{\alpha^{\frac{q}{p+1-q}}}{\hat{C}^{\frac{1}{p+1-q}}} \forall k \geq 0.
\]  

(38)
**Proof:** The statement follows as a corollary of a stronger statement, that is for any \( j = 1:N \) and \( k \geq 0 \) we have with probability 1 that

\[
\|x_k^j - x^*\|^{p+1} \leq \frac{\alpha^{\frac{q}{p+1-q}}}{C^{\frac{p+1-q}{p+1}}}. \tag{39}
\]

We will prove this relation by induction over \( k \). Clearly, (39) holds for \( k = 0 \) from our assumption. Assume it holds for \( k \) and let us show that it holds for \( k + 1 \). If \( j \not\in S_{k+1} \), then \( x_{k+1}^j = x_k^j \) and (39) holds from our induction assumption. On the other hand, for \( j \in S_{k+1} \), we have

\[
\|x_{k+1}^j - x^*\|^{p+1} = \|x_{k+1} - x^*\|^{p+1} \leq \left( \frac{\tilde{C}}{N} \sum_{j=1}^N \|x^* - x_k^j\|^{p+1} \right)^{\frac{p+1}{q}} \leq \frac{\alpha^{\frac{q}{p+1-q}}}{C^{\frac{p+1-q}{p+1}}},
\]

as \( \alpha \in (0, 1) \) and \( p + 1 > q \). Thus relation (38) holds.

**Theorem 5.8:** Let \( f \) be uniformly convex of degree \( q \geq 2 \) with constant \( \sigma_q \) and admit a stochastic \( p \geq 1 \) higher-order surrogate function \( g \) over \( X \) as given in Definition 3.1. Then the Lyapunov function along the sequence \( \hat{x}_k \) generated by SHOM, \( \mathcal{W}_k = \frac{1}{N} \sum_{j=1}^N \|x^* - x_k^j\|^{p+1} \), satisfies the recursion:

\[
\mathbb{E}_k [\mathcal{W}_{k+1}] \leq \left[ 1 - \frac{\tau}{N} + \frac{\tau}{N} \left( \frac{L_h q}{\sigma_q(p+1)!} \right)^{\frac{p+1}{q}} \mathcal{W}_k^{\frac{p+1-q}{q}} \right] \mathcal{W}_k.
\]

Furthermore, let \( \alpha \in (0, 1) \) and \( \tilde{C} = \frac{L_h q}{\sigma_q(p+1)!} \). If \( q < p + 1 \) and \( \|x_0 - x^*\| \leq \frac{\alpha (p+1-q)(p+1)}{\tilde{C}^{p+1-q}} \), then we have the following linear rate:

\[
\mathbb{E}_k [\mathcal{W}_{k+1}] \leq \left( 1 - \frac{(1 - \alpha)\tau}{N} \right) \mathcal{W}_k.
\]

**Proof:** From (35), we have for any \( k \) that

\[
\mathbb{E}_k [\mathcal{W}_{k+1}] = \delta \mathbb{E}_k \left[ \|x_{k+1} - x^*\|^{p+1} \right] + (1 - \delta) \mathcal{W}_k
\]

\[
= \delta \mathbb{E}_k \left[ \left( \|x_{k+1} - x^*\|^{q} \right)^{\frac{p+1}{q}} \right] + (1 - \delta) \mathcal{W}_k
\]

\[
\leq \delta \mathbb{E}_k \left[ \left( \tilde{C} \mathcal{W}_k \right)^{\frac{p+1}{q}} \right] + (1 - \delta) \mathcal{W}_k
\]

\[
= \delta \left( \tilde{C} \mathcal{W}_k \right)^{\frac{p+1}{q}} + (1 - \delta) \mathcal{W}_k = \left( 1 - \delta + \delta \tilde{C}^{\frac{p+1-q}{q}} \mathcal{W}_k^{\frac{p+1-q}{q}} \right) \mathcal{W}_k
\]
\[ \frac{38}{38} \leq \left( 1 - \frac{(1 - \alpha) \tau}{N} \right) W_k. \]

This proves our statements. \[ \square \]

Note that for any particular choice of \( \alpha \) and \( \tau \) we obtain from Theorem 5.8 linear rate independent of the conditioning, thus the method provably adapts to the curvature of the function \( f \). For example, \( p = q = 2 \) and \( \alpha = \frac{1}{4} \) yields

\[ \mathbb{E}_k [W_{k+1}] \leq \left( 1 - \frac{3\tau}{4N} \right) W_k. \]

Moreover, from Lemma 5.6 and Theorem 5.8 we get that \( \mathbb{E}[\|x_k - x^*\|^q] \) converges linearly to 0, locally. Furthermore, when \( \tau = N = 1 \) SHOM becomes deterministic and for such algorithm we obtain from the first statement of Theorem 5.8 local superlinear convergence in the iterates, i.e. for \( q < p + 1 \) we get

\[ \|x_{k+1} - x^*\| \leq \tilde{C}^\frac{1}{q} \|x_k - x^*\|^{\frac{p+1}{q}}. \]

6. Numerical tests

In this section, we consider two applications, a classification problem based on the logistic loss, which can be formulated as a convex finite-sum problem, and the independent component analysis problem, formulated as a nonconvex finite-sum problem. In all our numerical experiments, we use the metric \( B = I_n \) (the identity matrix).

6.1. Implementation details for SHOM

In this section, we show that for structured finite-sum problems, such as composite problems with structure (see Example 3.4 and also the applications below), SHOM subproblem can be solved efficiently. More specifically, for given vectors \( a_i \)'s, let us consider each term in the finite-sum of the form \( f_i(x) = \phi_i(a_i^T x) + \psi(x) \), where typically \( \phi_i \) are sufficiently smooth (possibly nonconvex) functions (see Example 2.3 and the applications below) and \( \psi \) is a simple function such as a regularizer or indicator function of some (possibly nonconvex) set (see the problems (43) or (44)). In this case, we can easily compute higher-order directional derivatives along a direction \( h \), e.g. \( D\phi_i(a_i^T x)[h] = \phi_i'(a_i^T x) \cdot a_i^T h \), \( D^2\phi_i(a_i^T x)[h] = \phi_i''(a_i^T x)a_i^T h \cdot a_i \) and \( D^3\phi_i(a_i^T x)[h]^2 = \phi_i'''(a_i^T x)(a_i^T h)^2 \cdot a_i \). Hence, in these settings, with \( p \in \{1, 2, 3\} \) and \( M_p = M_p \) for all \( j = 1:N \), the surrogate of each \( f_j \) is

\[ g_j(y; x_k^j) = T_p^\phi (y; x_k^j) + \frac{M_p}{(p + 1)!} \|y - x_k^j\|^{p+1} + \psi(y). \]

Note that we can update very efficiently the model approximation \( g \) in SHOM algorithm using (9). Then the iteration of SHOM requires solving a subproblem of the following form:

\[ x_{k+1} = \arg \min_y g(y; \hat{x}_k) = \frac{1}{N} \sum_{j=1}^N \left( T_p^\phi (y; x_k^j) + \frac{M_p}{(p + 1)!} \|y - x_k^j\|^{p+1} \right) + \psi(y). \]
Moreover, subproblem (40) can be solved typically in closed form for \( p = 1 \). For \( p \in \{2, 3\} \), one can use the powerful tools of convex optimization when the subproblem is (uniformly) convex (see, e.g. [9,21,22]), IPOPT [35] or can use, e.g. ADMM, which is known to converge also fast [6], when the subproblem is nonconvex. Below, we briefly sketch the main steps of ADMM. More precisely, after a change of variables \( z = y - x_k \), subproblem (40) can be written equivalently as

\[
\min_z \frac{1}{\ell!} \Delta_{\ell} [z]^{\ell} + \frac{\tau M_p}{N(p + 1)!} \|z\|^{p+1} + \bar{\psi}(z) + \frac{1}{N} \sum_{j \notin S_k} \frac{M_p}{(p + 1)!} \|z + x_k - x^j_k\|^{p+1},
\]

where \( \Delta_{\ell} [z]^{\ell} \) are computed from higher-order directional derivatives of \( \phi_j \)'s along directions \( x^j_k \) and/or \( z \) as explained above. Hence, we need to keep track of only scalar values, no need to memorize matrices or tensors. Further, we can reformulate (41) as the following problem:

\[
\min_{z, z^j} \sum_{\ell=1}^{p} \frac{1}{\ell!} \Delta_{\ell} [z]^{\ell} + \frac{\tau M_p}{N(p + 1)!} \|z\|^{p+1} + \bar{\psi}(z) + \frac{1}{N} \sum_{j \notin S_k} \frac{M_p}{(p + 1)!} \|z^j\|^{p+1}
\]

s.t. \( z^j = z + x_k - x^j_k \quad \forall j \notin S_k \). (42)

At each iteration of ADMM, one needs to solve a subproblem in \( z \) of the form:

\[
\min_z \frac{1}{\ell!} \Delta_{\ell} [z]^{\ell} + \frac{\tau M_p}{N(p + 1)!} \|z\|^{p+1} + \bar{\psi}(z),
\]

for which there are efficient algorithms when \( p \in \{2, 3\} \) (see, e.g. [22,28]); and \( N - \tau \) subproblems in \( z^j \) of the form (here \( \Lambda_j \in \mathbb{R}^n \) and \( \rho > 0 \)):

\[
\min_{z^j} (\Lambda_j, z^j) + \frac{\rho}{2} \|z^j\|^2 + \frac{M_p}{(p + 1)!} \|z^j\|^{p+1},
\]

which can be solved in parallel and in closed form. More specifically, the optimality condition yields \( \Lambda_j + \rho z^j + \frac{M_p}{(p)!} \|z^j\|^{p-1} z^j = 0 \). One can note that if \( \|z^j\|^{p-1} \) is known, then one can easily compute the optimal \( z^j \) from the previous equality. Hence, taking the norm in the optimality condition we get an equation in \( \|z^j\| \) of the form \( \|\Lambda_j\| = \rho \|z^j\| + \frac{M_p}{(p)!} \|z^j\|^p \), whose roots can be found explicitly for \( p \in \{2, 3\} \). Finally, the Lagrange multipliers are updated appropriately, see [6]. We can stop the algorithm when the feasibility constraints in (42) are below some desired accuracy or after a maximum number of ADMM steps.

### 6.2. Classification problem

The problem of fitting a generalized linear model with \( l_2 \) regularization can be formulated as [13]

\[
\min_x f(x) = \frac{1}{N} \sum_{i=1}^{N} \phi(a_i^T x) + \frac{\lambda}{2} \|x\|^2,
\]

(43)
where \( a_1, \ldots, a_N \in \mathbb{R}^n \) are given vectors, \( \lambda > 0 \) is a fixed scalar, and \( \phi: \mathbb{R} \to \mathbb{R} \) are convex functions, at least three times differentiable and sufficiently smooth. One such example is the logistic function \( \phi(t) = \log(1 + e^t) \), which appears often in machine learning applications, e.g. classification or regression tasks [13]. We also consider this function for numerical experiments and two datasets taken from LIBSVM [8]: madelon (\( N = 2000 \) data with \( n = 500 \) features) and a8a (the first \( N = 5000 \) data with \( n = 122 \) features). The regularization parameter is \( \lambda = 10^{-3} \) and the other parameters are: \( L_1^f = \frac{1}{4}\|a_i\|^2 \) (\( M_1 \equiv M_1^i = \max_{i=1:N} L_1^f \)), \( L_2^f = \frac{1}{3}\|a_i\|^3 \) (\( M_2 \equiv M_2^i = \max_{i=1:N} 2L_2^f \)) and \( L_3^f = \frac{2}{3}\sum_{i=1}^N \|a_i\|^4 \) (\( M_3 \equiv M_3^i = \max_{i=1:N} 3L_3^f \)), for all \( i = 1:N \), respectively.

We first analyse the behaviour of algorithm SHOM for different values of \( p \in \{1,2,3\} \). We also compare the SHOM variants with SAGA, a first-order variance reduction algorithm [9]. In this experiment, we consider the minibatch size \( \tau = 300 \) for SHOM and SAGA. The results are given in Figure 1, which shows that for a relatively small minibatch size the increasing of the approximation order \( p \) in SHOM algorithm has beneficial effects in terms of convergence rates (less number of epochs, i.e. number of passing through data) and in terms of achieving very high accuracies. Moreover, SHOM algorithm with \( p > 1 \) is performing better than SAGA or SHOM with \( p = 1 \), i.e. much less number of epochs and less CPU time (see Table 1).

Further, we analyse the behaviour of algorithm SHOM for different values of the minibatch size \( \tau \) ranging from 50 to 1000 on a8a dataset. The results are given in Figure 2 for \( p = 2 \) (top) and \( p = 3 \) (bottom). We observe that there are two optimal regimes for the minibatch size \( \tau^* \) and they coincide for the two variants of SHOM: \( p = 2 \) and \( p = 3 \), respectively. However, the number of epochs is smaller for \( p = 3 \) than for \( p = 2 \). One can also observe that large minibatch sizes do not necessarily improve convergence.

### Table 1. CPU time in seconds for algorithms SHOM (\( p = 1, 2, 3 \)) and SAGA on dataset a8a.

| Alg.   | SHOM \( p = 3 \) | SHOM \( p = 2 \) | SHOM \( p = 1 \) | SAGA |
|--------|------------------|------------------|------------------|-------|
| CPU time | 268              | 283              | 305              | 296   |

---

**Figure 1.** Behaviour of SHOM (\( p = 1, 2, 3 \)) and SAGA: left – a8a, right – madelon.
6.3. Independent component analysis problem

The independent component analysis (ICA) problem consists of finding the sources $V = [v_1, v_2, \ldots, v_r] \in \mathbb{R}^{r \times N}$ from some observed linear mixed set of signals $\hat{V} = [\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_N] \in \mathbb{R}^{m \times N}$, i.e. $\hat{V} = AV$, where $A$ is the mixing matrix, which is assumed unknown [14]. The basic assumption is that the signals are statistically independent, so that a distribution of sum of independent signals with arbitrary distributions tends toward a Gaussian distribution. In other words, the purpose of the ICA is to find a demixing matrix $X = A^*$ that separates the observed signals into a set of sources which are statistically independent. Several techniques for measuring independence (non-Gaussianity) were proposed in the literature and posed as the following optimization problem, after some simplifications (see, e.g. [14] for more details):

$$\max_{\|x\|=1} \frac{1}{N} \sum_{i=1:N} \phi(x^T \hat{v}_i) =: f_i(x).$$  \hspace{1cm} (44)

One standard measure for independence is the kurtosis (the fourth-order moment): $\bar{\phi}(t) = t^4$. Although this measure of non-Gaussianity is theoretically good, it is sensitive to outliers. Therefore, a more practical choice is the approximation of negentropy that uses non-quadratic functions. Two classical choices are [14]: $\bar{\phi}(t) = \frac{1}{\alpha} \ln \cosh(\alpha t)$, with $1 \leq \alpha \leq 2$ and $\bar{\phi}(t) = -e^{-t^2}$. Note that in all these cases the optimization problem (44) is nonconvex, but for $p \in (1, 2)$ the objective function $\bar{\phi}$ is $p$ times differentiable and with the $p$th derivative Lipschitz over the bounded feasible set $\mathcal{X} = \{x : \|x\| = 1\}$. We apply ICA framework to reduce the spectral dimension of a hyperspectral image. A hyperspectral image $\hat{V}$ is a three-dimensional hyperspectral data cube $n_1 \times n_2 \times m$, having $n_1 \times n_2$
pixels and $m$ is the number of spectral channels. Denote $N = n_1 \cdot n_2$ the number of pixels in each spectral channel. Then the hyperspectral image can be represented as a matrix:

$$
\hat{V} = \begin{bmatrix}
\hat{v}_{11} & \hat{v}_{12} & \cdots & \hat{v}_{1N} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{v}_{m1} & \hat{v}_{m2} & \cdots & \hat{v}_{mN}
\end{bmatrix} \in \mathbb{R}^{m \times N},
$$

where $\hat{v}_{ij}$ is the pixel $j$ on band $i$. We aim at reducing the number of bands in the hyperspectral image from $m$ to $r \ll m$ using the negentropy-based ICA framework, replacing the expectation in the nonconvex problem (44) with the empirical risk using the samples from $\hat{V}$. In our tests, we use Salinas scene that has the spatial dimension $512 \times 217$ (i.e. the number of functions in the finite-sum objective is $N = 111,104$) and 204 bands, see [34]. We compare SHOM with $p = 1, 2$ and $\tau = 150$ and the state of the art algorithm FastICA developed specially for the ICA problem in [14]. Note that FastICA is a full batch gradient method with a scalar stepsize depending on the second derivatives of $\bar{\phi}$, see [14] for details. We choose $\bar{\phi}(t) = -e^{-t^2/2}$ in (44), $r = 1$, $L_1^f = (1 + \|\hat{v}_i\|^2)\|\hat{v}_i\|^2$ ($M_1 \equiv M_1^i = 1.1 \max_{i=1:N} L_1^f$) and $L_2^f = (3\|\hat{v}_i\| + \|\hat{v}_i\|^3)\|\hat{v}_i\|^3$ ($M_2 \equiv M_2^i = 1.1 \max_{i=1:N} L_2^f$), for all $i = 1:N$, respectively, and we solve the subproblem (40) using ADMM described in Section 6.1. From Figure 3 (left), one can observe that SHOM for both $p = 1$ and 2 is superior to FastICA algorithm, being able to achieve even high accuracy in a very small number of epochs. Finally, Figure 3 (right) displays the sample band of Salinas dataset found with FastICA algorithm and SHOM. We observe that the two images, corresponding to a single band, are similar.

7. Conclusion

In this paper, we have designed a minibatch stochastic higher-order algorithm for minimizing finite-sum optimization problems. Our method is based on the minimization of a higher-order upper bound approximation of the finite-sum objective function. We have presented convergence guarantees for nonconvex and convex optimization when the higher-order upper bounds approximate the objective function up to an error whose derivative of a certain order is Lipschitz continuous. More precisely, we have derived asymptotic stationary point guarantees for nonconvex problems, and for objective functions having the KL property or uniformly convex ones we have established
local (super/sub)linear convergence rates. Moreover, unlike other higher-order algorithms, our method works with any batch size. Numerical simulations have also confirmed the efficiency of our algorithm.

**Note**

1. By Lyapunov function along the sequence \( \hat{x}_k \) generated by SHOM we mean that we evaluate the Lyapunov function

\[
\Phi(\hat{x}) = \frac{1}{N} \sum_{j=1}^{N} (f(x^j) - f(x^*))^{p+1} q
\]

at \( \hat{x}_k \).

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Funding**

The research leading to these results has received funding from the NO Grants 2014–2021 RO-NO-2019-0184, under project ELO-Hyp, contract no. 24/2020; UEFISCDI PN-III-P4-PCE-2021-0720, under project L2O-MOC, nr. 70/2022.

**Notes on contributors**

*Daniela Lupu* is a Ph.D. student at University Politehnica Bucharest, Romania. Her research interests lies in the area of continuous optimization and machine learning with applications in control. She holds B.Sc. (2017) and M.Sc. degrees (2019) from University Politehnica Bucharest.

*Ion Necoara* received an M.Sc. degree (2002) in Optimization and Statistics from University of Bucharest, Romania, and a PhD degree (2006) in Applied Sciences (cum laude) from Delft University of Technology, Netherlands. From 2007 to 2009, he was a research fellow at Katholieke Universiteit Leuven, Belgium. Since 2009 he is with University Politehnica Bucharest, Romania, where he is now full professor. Since 2021 he is also a senior researcher at the Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy. His research interests cover various topics in developing new optimization algorithms with a focus on mathematical guarantees about their performance.

**References**

[1] A. Agafonov, D. Kamzolov, P. Dvurechensky, and A. Gasnikov, *Inexact tensor methods and their application to stochastic convex optimization*, arXiv preprint: 2012.15636, 2020.

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

[3] E.G. Birgin, J.L. Gardenghi, J.M. Martinez, S.A. Santos, and P.H.L. Toint, *Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models*, Math. Program. 163(1–2) (2017), pp. 359–368.

[4] R. Bollapragada, R.H. Byrd, and J. Nocedal, *Exact and inexact subsampled Newton methods for optimization*, IMA J. Numer. Anal. 39(2) (2018), pp. 545–578.

[5] J. Bolte, A. Daniilidis, and A. Lewis, *The Lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems*, SIAM J. Optim. 17(4) (2007), pp. 1205–1223.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn. 3(1) (2011), pp. 1–122.
[7] C. Cartis, N. IM. Gould, and P.L. Toint, A concise second-order complexity analysis for unconstrained optimization using high-order regularized models, Optim. Methods Softw. 35(2) (2020), pp. 243–256.

[8] C.C. Chang and C.J. Lin, LIBSVM: a library for support vector machines, ACM. Trans. Intell. Syst. Technol. 2(3) (2011), pp. 1–27.

[9] A. Defazio, F. Bach, and S. Lacoste-Julien, SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives, Adv. Neural. Inf. Process. Syst. 27 (2014), pp. 1646–1654.

[10] N. Doikov and Y.U. Nesterov, Local convergence of tensor methods, Math. Program. 193(1) (2022), pp. 315–336.

[11] N. Doikov and Y.U. Nesterov, Inexact tensor methods with dynamic accuracies, in International Conference on Machine Learning, 2020, pp. 2577–2586.

[12] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, C.A. Uribe, B. Jiang, H. Wang, S. Zhang, S. Bubeck, and Q. Jiang, Near optimal methods for minimizing convex functions with Lipschitz 4th derivatives, in Conference on Learning Theory, 2019, pp. 1392–1393.

[13] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.

[14] A. Hyvarinen, J. Karhunen, and E. Oja, Independent Component Analysis, Wiley, 2001.

[15] R. Johnson and T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction, in Advances in Neural Information Processing Systems, pp. 315–323, 2013.

[16] D. Kovalev, K. Mishchenko, and P. Richtarik, Stochastic Newton and cubic Newton methods with simple local linear-quadratic rates, in Advances in Neural Information Processing Systems, 2019.

[17] A. Lucchi and J. Kohler, A Stochastic tensor method for non-convex optimization, arXiv preprint: 1911.10367, 2019.

[18] A. Mokhtari, M. Eisen, and A. Ribeiro, IQN: an incremental quasi-Newton method with local superlinear convergence rate, SIAM J. Optim. 28(2) (2018), pp. 1670–1698.

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

[20] J. Mairal, Incremental majorization–minimization optimization with application to large-scale machine learning, SIAM J. Optim. 25(2) (2015), pp. 829–855.

[21] Y.U. Nesterov, Implementable tensor methods in unconstrained convex optimization, Math. Program. 186(1–2) (2021), pp. 157–183.

[22] Y.U. Nesterov, Inexact basic tensor methods for some classes of convex optimization problems, Optim. Methods Softw. 37(3) (2022), pp. 878–906. doi:10.1080/10556788.2020.1854252

[23] Y.U. Nesterov, Gradient methods for minimizing composite functions, Math. Program. 140(1) (2013), pp. 125–161.

[24] I. Necoara, General convergence analysis of stochastic first order methods for composite optimization, J. Optim. Theory. Appl. 189(1) (2021), pp. 66–95.

[25] I. Necoara, V. Nedelcu, and I. Dumitrache, Parallel and distributed optimization methods for estimation and control in networks, J. Process. Control. 21(5) (2011), pp. 756–766.

[26] I. Necoara and D. Lupu, General higher-order majorization–minimization algorithms for (non)convex optimization, arXiv preprint: 2010.13893, 2020.

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

[28] Y.U. Nesterov and B.T. Polyak, Cubic regularization of Newton method and its global performance, Math. Program. 108(1) (2006), pp. 177–205.

[29] L.M. Nguyen, J. Liu, K. Scheinberg, and M. Takac, SARAH: A novel method for machine learning problems using stochastic recursive gradient, in International Conference on Machine Learning, 2017.

[30] A. Rodomanov and Y.U. Nesterov, Smoothness parameter of power of Euclidean norm, J. Optim. Theory. Appl. 185(2) (2020), pp. 303–326.

[31] L. Rosasco, S. Villa, and B.C. Vu, Convergence of stochastic proximal gradient algorithm, Appl. Math. Optim. 82(3) (2020), pp. 891–917.

[32] N. Tripuraneni, M. Stern, C. Jin, J. Regier, and M.I. Jordan, Stochastic cubic regularization for fast nonconvex optimization, Adv. Neural. Inf. Process. Syst. 31 (2018), pp. 2899–2908.
[33] D. Zhou, P. Xu, and Q. Gu, *Stochastic variance-reduced cubic regularized Newton method*, in *International Conference on Machine Learning*, 2018, pp. 5985–5994.

[34] http://www.ehu.eus/ccwintco/index.php/Hyperspectral-Remote-Sensing-Scenes.

[35] A. Wächter and L.T. Biegler, *On the implementation of a primal–dual interior point filter line search algorithm for large-Scale nonlinear programming*, Math. Program. 106(1) (2006), pp. 25–57.