Journal of Machine Learning Research 23 (2022) 1-44 Submitted 9/21; Revised 5/22; Published 7/22

Stochastic DCA with Variance Reduction and Applications in Machine Learning

Hoai An Le Thi HOAI-AN.LE-THI@UNIV-LORRAINE.FR
Université de Lorraine, LGIPM, Département IA, F-57000 Metz, France
Institut Universitaire de France (IUF)

Hoang Phuc Hau Luu HOANG-PHUC-HAU.LUU@UNIV-LORRAINE.FR

Hoai Minh Le MINH.LE@UNIV-LORRAINE.FR
Université de Lorraine, LGIPM, Département IA, F-57000 Metz, France

Tao Pham Dinh PHAM@INSA-ROUEN.FR
Laboratory of Mathematics, INSA-Rouen, University of Normandie
76801 Saint-Étienne-du-Rouvray Cedex, France

Editor: Ohad Shamir

Abstract

We design stochastic Difference-of-Convex-functions Algorithms (DCA) for solving a class of structured Difference-of-Convex-functions (DC) problems. As the standard DCA requires the full information of (sub)gradients which could be expensive in large-scale settings, stochastic approaches rely upon stochastic information instead. However, stochastic estimations generate additional variance terms making stochastic algorithms unstable. Therefore, we integrate some novel variance reduction techniques including SVRG and SAGA into our design. The almost sure convergence to critical points of the proposed algorithms is established and the algorithms' complexities are analyzed. To study the efficiency of our algorithms, we apply them to three important problems in machine learning: nonnegative principal component analysis, group variable selection in multiclass logistic regression, and sparse linear regression. Numerical experiments have shown the merits of our proposed algorithms in comparison with other state-of-the-art stochastic methods for solving nonconvex large-sum problems.

Keywords: DC programming, DCA, DCA-SVRG, DCA-SAGA, variance reduction technique

1. Introduction

We are concerned with the following nonconvex optimization problem

$$(P) \quad \min\left\{F(x) = G(x) - H(x) + r_1(x) - r_2(x) : x \in \mathbb{R}^n\right\},$$

where $H := \frac{1}{N}\sum_{i=1}^{N} h_i$, $G, h_i : \mathbb{R}^n \to \mathbb{R}$ and $r_1, r_2 : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are convex, lower semicontinuous functions while each h_i has Lipschitz continuous gradient (with a common constant L) on the effective domain of r_1.

This type of problems is often encountered in many areas, in particular in machine learning, where $G - H$ represents the data-fitting term (the loss function) while $r_1 - r_2$ represents the regularization term to encourage some desired properties on the found solutions (some low dimensional structures such as low-rank or sparsity - for instance) or to model constraints on x. Moreover, in the era of big data, optimization models are expected to take into account the large-sum structure in order to capture the high volume nature of data. Therefore, the loss function often has a (nonconvex) large-sum structure. Here, we consider a more general structure of the loss function including the large-sum case. More precisely, we assume the large-sum structure of H but not G, because that, as will be seen later, our approach treats impartially whether G is a large-sum function or not. As for the regularization term, we consider two cases of r_2: r_2 is convex (the regularizer term is a DC function), and r_2 is a composite function defined by $r_2(x) = \sum_{i=1}^m l_i(p_i(x))$ with $l_i : \mathbb{R} \to \mathbb{R}$ being convex, decreasing and $p_i : \mathbb{R}^n \to \mathbb{R}$ being convex. In the second case, r_2 is no longer convex in general. It is well-known that most nonconvex sparsity-promoting regularizers (usually used to approximate the ℓ_0 norm) have the above composite form of r_2 (Le Thi et al., 2015; Ong and Le Thi, 2013), for instance, log-sum penalty (Candes et al., 2008), smoothly clipped absolute deviate (SCAD) (Fan and Li, 2001), capped ℓ_1 (Zhang, 2010b), minimax concave penalty (Zhang, 2010a), (nonconcave) piecewise linear function (Le Thi, 2012).

The problem (P) attracts our special attention, thanks to its important role in machine learning and big data analytics via two prominent features: the (nonconvex) DC structure of the objective (on both loss and regularizer) and the large-sum structure of H (more precisely, the very large value of N). In fact, since many newly developed machine learning models are complicated, nonconvex optimization is indispensable to model them. Here, the nonconvexity is represented by the DC structure, which covers a large class of nonconvex optimization problems (Le Thi and Pham Dinh, 2018). Moreover, as mentioned above, the large-sum structure is one of the most popular forms encountered in practice to model big data-driven problems. For example, the large-sum structure arises naturally in empirical risk minimization (ERM) in stochastic programming. That is, for instance, the stochastic problem

$$\min_{x \in D} \mathbb{E}(f(x,\xi)),$$

where ξ is a random variable, can be approximated by the following regularized empirical risk minimization problem

$$\text{(ERM)} \quad \min_{x \in D} \frac{1}{N} \sum_{i=1}^{N} f(x,\xi_i) + \lambda\Omega(x),$$

where $\{\xi_1, \xi_2, \ldots \xi_N\}$ is a set of i.i.d. realizations of ξ, and Ω is the regularization term. In particular, when Ω is a convex function, D is a convex set, and $f(\cdot, \xi_i)$ has L-Lipschitz gradients for all i, the problem (ERM) falls into the spectrum of the problem (P) with $G(x) = (L/2)\|x\|^2$, $h_i(x) = (L/2)\|x\|^2 - f(x,\xi_i)$, $i = 1, 2, \ldots, N$, and $r_1(x) = \lambda\Omega(x) + \chi_D(x), r_2(x) = 0$. Many important applications such as LASSO, Principal component analysis, logistic regression, etc., can be expressed in this form.

Related works and our motivation. The two mentioned features of the problem (P) also create a huge challenge. On one hand, there are very few efficient and scalable algorithms dealing with nonconvex problems, and, on another hand, the large-sum structure is difficult to be handled by deterministic algorithms. Therefore, an efficient approach could be using stochastic methods for dealing with the large-sum structure combined with power algorithms for nonconvex programming. Intuitively, a stochastic algorithm usually employs some stochastic approximation techniques based on a framework of a deterministic algorithm. This deterministic frame plays an important role in the overall quality of a stochastic algorithm and should therefore be highly suited to the problem's structure. And the stochastic approximation should be determined in an inexpensive way while having a small noise.

In the convex setting, the literature on stochastic optimization is vast. In particular, the problem (ERM) with $f(\cdot, \xi)$ and Ω being convex, $D = \mathbb{R}^n$, has been studied intensively. By the classical idea of Stochastic gradient descent, which is traced back to the seminal work (Robbins and Monro, 1951), many variants have been developed, to name but a few, stochastic dual coordinate descent (Shalev-Shwartz and Zhang, 2013), stochastic average gradient (SAG) (Schmidt et al., 2017), stochastic variance reduced gradient (SVRG) (Johnson and Zhang, 2013), SAGA (Defazio et al., 2014), StochAstic Recursive grAdient algoritHm (SARAH) (Nguyen et al., 2017), etc. Some other works considered the (strongly) convex sum function but relaxed the convexity of each component function $f(\cdot, \xi)$ (Allen-Zhu and Yuan, 2016; Shalev-Shwartz, 2016).

Works for nonsmooth, nonconvex large-sum problems remain rare. The (ERM) problem with L-smooth $f(\cdot, \xi)$ and (possibly nonsmooth) convex regularizer Ω (which is a special case of (P)) is probably the most common model that has been studied via the proximal-based approach (J. Reddi et al., 2016; Pham et al., 2020). Among stochastic algorithms for solving the (ERM), the two algorithms prox-SVRG and prox-SAGA (J. Reddi et al., 2016) are the most related to our work since they adopt the SVRG and SAGA estimators, respectively. Another work based on the Majorize-Minimization (MM) method was proposed in (Mairal, 2015) which considered a large-sum of nonconvex functions where each function admit surrogates with L-smooth error (i.e. the difference between a function and its surrogates is L-smooth).

For nonsmooth DC large-sum problems, the current body of research is even more limited. There were some recent works (Le Thi et al., 2017, 2020; Xu et al., 2019b) that developed stochastic methods based on DCA (DC Algorithm). The first stochastic DCA was proposed in (Le Thi et al., 2017) for a large-sum of (nonconvex) L-smooth functions with DC regularizer, which was further extended to a more general problem where the L-smooth assumption is relaxed (Le Thi et al., 2020). In these works, the stochastic DCA has been developed based on the SAG estimator (Schmidt et al., 2017) that is a variance reduction estimator. This research direction is promising and should be continued for a very large class to cover various problems arising in practice.

Another approach was studied in (Xu et al., 2019b) where the data-fitting term is large-sum DC and the regularizer is nonconvex, nonsmooth whose proximal operator can be efficiently computed. By using the Moreau envelope (which is a DC function) of the regularizer, the authors approximated the original problem by a DC program and proposed DCA schemes for solving it. In general, the proposed algorithms can be regarded as the standard DCA in which the (large-sum) convex subproblems are solved by stochastic algorithms (e.g., Adagrad (Duchi et al., 2011), SVRG (Johnson and Zhang, 2013)) up to a certain level of accuracy.

Starting from the above observations, we are motivated to develop new stochastic algorithms based on DCA for solving (P). Also, the virtue of stochastic variance-reduced gradients given by SVRG and SAGA has been theoretically and practically justified by many researchers. Hence, combining them with the DCA could be efficient approaches.
In (deterministic) nonconvex optimization, DC programming and DCA were well known to be powerful tools, as they established a nice philosophy allowing researchers to travel from the convex world to the land of nonconvexity, where passengers are able to carry with them useful instruments of convex analysis/programming which have been developed for decades. The original key idea of DCA relies on the DC structure of the objective function. DCA consists in iteratively approximating the considered DC program by a sequence of convex ones. DC programming and DCA, constitute the backbone of nonconvex programming and global optimization, were introduced in 1985 by Pham Dinh Tao and have been extensively developed by Le Thi Hoai An and Pham Dinh Tao since 1994, to become now classic and increasingly popular (see e.g. (Pham Dinh and Le Thi, 1997; Le Thi and Pham Dinh, 2005, 2018)). Their popularity is due to their robustness and efficiency compared to existing methods, their adaptation to the structures of treated problems and their ability to solve large-scale real world nonconvex programs (see a comprehensive survey on thirty years of developments of DC programming and DCA in (Le Thi and Pham Dinh, 2018)). It was analyzed in (Le Thi and Pham Dinh, 2018) that most algorithms (classical as well as recent algorithms) in nonconvex programming framework can be seen as versions of DCA with suitable DC decompositions / DC formulations. DCA has been successfully applied for solving numerous problems in divers areas, especially in large-scale settings: transport logistic, finance, data mining and machine learning, computational chemistry, computational biology, robotics and computer vision, combinatorial optimization, cryptology, inverse problems and ill-posed problems, etc., see e.g. (Le Thi and Pham Dinh, 2005; Pham Dinh and Le Thi, 1997, 1998; Le Thi, 2005; Le Thi and Pham Dinh, 2018) and references therein.
Given the well-established strength in deterministic optimization, DCA is an appropriate frame for designing nonconvex stochastic algorithms.

Paper's contribution. Aiming to tackle the two above mentioned challenges of (P), namely, the nonconvexity and the large-sum structure of the objective, we investigate variance reduction techniques in stochastic algorithms to deal with the large-sume structure, and develop DCA to solve nonconvex programs, that result in the so-called stochastic DCA. We propose two stochastic DCA schemes integrated SVRG and SAGA, namely DCA-SVRG and DCA-SAGA. In our design, we consider two commonly-used sample strategies: sample with replacement and sample without replacement. Sampling is called *with replacement* if one sample selected at random from the population is returned to the population, then the next sample is selected in the same way; therefore, it allows repetition. Meanwhile, sample *without replacement* strategy sequentially samples from the population where each chosen unit is not placed back in the population. In the literature, algorithms for solving large-sum problems mainly adopt the former sampling strategy due to its convenient property: each sample is independent of the others. However, in practice, the latter one is a very natural strategy to employ and sometimes leads to remarkably better results (which is illustrated by our numerical experiments in Section 5). Therefore, we provide the analysis for both sampling strategies. Our algorithms follow the new trend in the development of DCA: design novel efficient algorithms based on DCA for nonconvex optimization which improve standard DCA in some aspects.

The convergence of the DCA-SVRG and DCA-SAGA has been studied rigorously. The almost sure convergence to critical points of the proposed algorithms is established. We show that, each accumulated point of the sequence generated by DCA-SVRG and DCA-SAGA is a DC critical point of $F = (G + r_1) - (H + r_2)$. Furthermore, if r_2 is L-smooth, we obtain the $\mathcal{O}(k^{-1/2})$ convergence rate with respect to the measure of proximity to criticality, where k is the number of iterations. As a consequence, to find ϵ-criticality, DCA-SVRG and DCA-SAGA have the complexity of $\mathcal{O}(N^{2/3}/\epsilon^2)$ and $\mathcal{O}(N + N^{3/4}/\epsilon^2)$ (respectively) in terms of gradient evaluations. Another important contribution is novel arguments used in the analysis of DCA-SAGA. We introduce an elegant idea of how to escape the non-independence (which causes difficulties in the analysis) and give good quantitative links between $\mathbb{E}\|x^t - \alpha_i^t\|^2$ and $\mathbb{E}\|x^t - x^{t-1}\|^2, \mathbb{E}\|x^{t-1} - x^{t-2}\|^2$, etc., where $\{\alpha_i^t\}_{i=1}^N$ are local auxiliary variables introduced by the SAGA technique, which are used for building variance reduction terms and are updated progressively. These new arguments are expected to be useful in analyzing future SAGA-type algorithms.

Furthermore, we give additional convergence analysis to handle the composite structure of r_2. By introducing new optimization variables, $z_i = p_i(x)$ with $i = \overline{1, m}$, one can reformulate this problem as a DC program in form of (P) with respect to a couple of old and new variables. However, our aim here is to provide a more practical convergence result where we do not simply treat this couple of variables as a whole. Instead, we still focus on the main role of the primitive variable x in our convergence results. This special treatment offers us more flexible results to be employed in practice and avoids an undesirable situation when straightforwardly applying the original convergence analysis of the proposed algorithms (discussed in Section 4).

Finally, we conduct numerical experiments to study carefully the proposed algorithms' behaviors. The real-life problems being considered are the nonnegative principal component analysis, the group variable selection in multiclass logistic regression, and sparse linear regression.

Comparison with related works. Comparing with (Le Thi et al., 2017, 2020), we use SVRG and SAGA estimators in the construction of convex subproblems, while (Le Thi et al., 2017, 2020) employed the SAG estimator. Intuitively, the SAG is quite "conservative" due to its averaging feature: at each iteration, it computes the average of b new gradients and $N - b$ kept gradients (b is the minibatch size and N is the number of functions of the large sum) to form a stochastic direction used to construct a subproblem. Consequently, if b is relatively small in comparison with N, the old information dominates the new information, which makes this stochastic direction change slowly from iteration to iteration. In the convex case, numerical experiments in the paper (Defazio et al., 2014) have also shown some merits of SVRG and SAGA over SAG. Hence, SVRG and SAGA are expected to continue exhibiting these advantages in the context of DC programming and DCA.

As for (Xu et al., 2019b), our proposed algorithms are quite different: their algorithms are in fact deterministic DCA in which stochastic convex solvers are used to solve convex subproblems, while our algorithms stochastically construct convex subproblems. Moreover, in their work, the authors employ the SVRG algorithm for solving their subproblem, which results in the so-named SSDC-SVRG algorithm. Our DCA-SVRG differs from the SSDC-SVRG by the fact that in the DCA-SVRG, the SVRG estimator is employed as an outer method to compute stochastic gradients; whereas, the SVRG is used as an inner convex solver in the SSDC-SVRG.

Regarding the proximal-SVRG and the prox-SAGA in (J. Reddi et al., 2016) for solving the (ERM) problem where each $f(\cdot, \xi_i)$ is L-smooth and Ω is convex, which is a special case of (P): the DCA-SVRG via the option *with replacement* applies on this particular problem with G, h_i, r_1, r_2 defined above recovers the prox-SVRG. Meanwhile, the DCA-SAGA via the option *with replacement* do not correspond to the prox-SAGA, but a closely related algorithm, where the "stochastic variance-reduced gradient" of prox-SAGA, $\tilde{\nabla}_{\text{prox-SAGA}}$, is drifted by another unbiased variance-reduction term: $\tilde{\nabla}_{\text{DCA-SAGA}} = \tilde{\nabla}_{\text{prox-SAGA}} - \text{drift}$, where

$$\text{drift} = \frac{L}{N}\sum_{i=1}^{N} \alpha_i^t - \frac{L}{b}\sum_{i \in I} \alpha_i^t.$$

Furthermore, in the prox-SAGA step of updating the table of gradients, the authors do not update directly on I but drawing another independent set of indexes J, that mainly aims to facilitate their theoretical analysis. This procedure additionally consumes - in the worst case - b gradients computation, where b is the minibatch size. By contrast, the DCA-SAGA does not need to use another set J, rather, we update the table of gradients based on I directly, which omits these additional computations.

2. Background

The optimal value α of the problem (P) is assumed to be finite, i.e., $\alpha > -\infty$, which implies $\operatorname{dom} r_1 \subset \operatorname{dom} r_2$. Moreover, we assume that $\operatorname{dom} r_1 \subset \operatorname{dom} \partial r_2$.

The space $X := \mathbb{R}^n$ is equipped with the canonical inner product $\langle \cdot, \cdot \rangle$. Its dual space Y is identified with X itself. The effective domain of a function $f : X \to \mathbb{R} \cup \{+\infty\}$, denoted by $\operatorname{dom} f$, is $\operatorname{dom} f := \{x \in X : f(x) < +\infty\}$. It is called proper if $\operatorname{dom} f \neq \emptyset$. We adopt the convention that $+\infty - (+\infty) = +\infty$.

A function f is called ρ-convex for some $\rho \geq 0$, if for all $x, y \in \mathbb{R}^n, \lambda \in [0, 1]$, one has

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) - \frac{\rho}{2}\lambda(1-\lambda)\|x - y\|^2.$$

The supremum of all $\rho \geq 0$ such that the above inequality holds is called convex modulus of f, denoted by $\rho(f)$ or ρ_f. Let $x \in \operatorname{dom} f$, a vector $z \in X$ is called a subgradient of f at x if

$$f(y) - f(x) \geq \langle z, y - x \rangle, \quad \forall y \in X.$$

The set of all subgradients of f at x is called the subdifferential of f at x, denoted by $\partial f(x)$. And, f is said to be subdifferentiable at x if $\partial f(x) \neq \emptyset$. Also, by definition,

$$\operatorname{dom} \partial f = \{x \in X : \partial f(x) \neq \emptyset\}.$$

Let f be differentiable on an open set U, and let $V \subset U$. The function f is called L-smooth (L-Lipschitz derivative) on V if

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \ \forall x, y \in V.$$

Let $\Gamma_0(X)$ denote the convex cone of all lower semicontinuous proper convex functions on X. The vector space of DC functions is denoted by $DC(X) := \Gamma_0(X) - \Gamma_0(X)$. The standard DC program takes the form

$$(P_{dc}) \qquad \alpha := \inf\{f(x) := g(x) - h(x) : x \in X\}$$

where $g, h \in \Gamma_0(X)$. Such a function f is called DC, $g - h$ is DC decomposition while g and h are DC components of f. It is worth noting that a DC function f has infinitely many DC decompositions.

A DC program with closed convex constraint $x \in C$ can be equivalently written as a standard DC program by adding the indicator function χ_C to the first component g,

$$\inf\{f(x) := g(x) - h(x) : x \in C\} = \inf\{\chi_C(x) + g(x) - h(x) : x \in X\}.$$

Without loss of generality, we can assume that g and h are strongly convex since f can be reformulated as a difference of two strongly convex functions as

$$f = \left(g + \frac{\rho}{2}\|\cdot\|^2\right) - \left(h + \frac{\rho}{2}\|\cdot\|^2\right), \quad \text{with } \rho > 0.$$

A point x^* is called a critical point of (P_{dc}) (or $f = g - h$), iff $\partial g(x^*) \cap \partial h(x^*) \neq \emptyset$, or equivalently $0 \in \partial g(x^*) - \partial h(x^*)$. Critical points are generalized KKT points for DC programs. They coincide, under technical assumptions, with zero of Clarke subdifferential (or Clarke's stationarity) of f (Le Thi and Pham Dinh, 2018; Le Thi et al., 2018). Also x^* is called a strongly critical point of (P_{dc}) (or $f = g - h$), iff $\emptyset \neq \partial h(x^*) \subset \partial g(x^*)$. This inclusion expresses the d(irectional)- stationarity of x^* for f. Finally, note that strong criticality is a necessary (and sufficient with additional assumptions) condition for local optimality in DC programming (Le Thi et al., 2018; Le Thi and Pham Dinh, 2018).

In practice, the notion of ϵ-criticality has also been used (Xu et al., 2019b). A point x^* is called an ϵ-critical point of $f = g - h$ if $\text{dist}(\partial g(x^*), \partial h(x^*)) \leq \epsilon$, where $\text{dist}(A, B)$ denotes the distance between two sets A and B.

Philosophy of DCA: DCA is based on local optimality conditions and duality in DC programming, which introduces the nice and elegant concept of approximating a DC program by a sequence of convex ones: at each iteration k, DCA approximates the second DC component h by its affine minorization $h_k(x) = h(x^k) + \langle x - x^k, y^k \rangle$, with $y^k \in \partial h(x^k)$, and then solves the resulting convex subprogram to get x^{k+1}. The standard DCA is formally described as follows.

```
*Standard DCA.*
**Initialization:** Let $x^0 \in \operatorname{dom} \partial h$ and $k = 0$.
**repeat**
    Step 1: Compute the subgradient $y^k \in \partial h(x^k)$.
    Step 2: Solve the convex program $x^{k+1} \in \arg\min\{g(x) - h_k(x) : x \in X\}$.
```

Step 3: $k = k + 1$.
until Stopping criterion.

Convergences properties of the standard DCA and its complete theoretical foundation in the DC programming framework can be found in (Le Thi and Pham Dinh, 2005; Pham Dinh and Le Thi, 1997, 1998). For instance, it is especially worth mentioning that the sequence $\{x^k\}$ generated by DCA has the following properties:

1. The sequence $\{(g - h)(x^k)\}$ is decreasing.

2. If $(g-h)(x^{k+1}) = (g-h)(x^k)$, then x^k and x^{k+1} are critical points of (P_{dc}) and DCA terminates at k-th iteration.

3. If $\rho(g) + \rho(h) > 0$ then the series $\sum_{k=1}^{\infty} \|x^{k+1} - x^k\|^2$ converges.

4. If the optimal value α of the problem (P_{dc}) is finite and the sequences $\{x^k\}$ and $\{y^k\}$ are bounded, then every limit point $\tilde{x}$ of $\{x^k\}$ is a critical point of $g - h$.

Remark 1 *In general, the convergence of the whole sequence $\{x^k\}$ generated by DCA does not hold. Proving this property is difficult and requires more sophisticated tools. Hence, there were a very few results about it. The first known positive result has been stated for the class of DC programs with subanalytic data, i.e., their DC objective functions and convex* constraints are subanalytic *(see (Le Thi et al., 2018)). Its proof has been based on the famous Lojasiewics inequality.*

3. Stochastic DCA with Variance Reduction

Throughout this section, we study the problem (P) with r_2 being a convex function.

3.1 The First Stochastic DCA: DCA-SVRG

In this subsection, we develop the first stochastic DCA, called DCA-SVRG, for solving the problem (P). The algorithm can be considered as a combination of the deterministic DCA and the SVRG-style of gradient update. That is, based on the deterministic DCA, we replace the gradient of H by the "stochastic variance reduced gradient" of H. The algorithm is epoch-based, where the full gradient of H is computed at the beginning of each epoch and used as a variance-reduction term inside that epoch.

To construct the stochastic variance reduction gradient of H, we can choose one of two options of sampling: with or without replacement. If the option is sample with replacement, at each iteration, a set of indexes $I_b = \{i_1, i_2, \ldots, i_b\}$ is randomly chosen from $\{1, 2, \ldots, N\}$ where each i_j is independent of the others. Otherwise, the repetition in I_b is not allowed.

On the other hand, since a critical point x^* of $F = (G + r_1) - (H + r_2)$ satisfies $\operatorname{dist}(\partial(G + r_1)(x^*), \partial(H + r_2)(x^*)) = 0$, we define the measure of proximity to criticality as follows

$$d_K = \min_{k=0,1,\ldots,K-1, j=0,1,\ldots,M-1} \mathbb{E}\operatorname{dist}(\partial(H + r_2)(x_{j+1}^{k+1}), \partial(G + r_1)(x_{j+1}^{k+1})).$$

Algorithm 1 DCA-SVRG

```
**Initialization:** $\tilde{x}^0 \in \operatorname{dom} r_1$, inner-loop length $M$, minibatch size $b$, $k = 0$, option (either *with replacement* or *without replacement*).
**repeat**
    Compute the full gradient $\tilde{\nu}^k = \frac{1}{N}\sum_{i=1}^N \nabla h_i(\tilde{x}^k)$ and set $x_0^{k+1} = \tilde{x}^k$.
    **for** $j = 0 : M-1$ **do**
        **if** option is *with replacement* **then**
            Randomly choose with replacement the set $I_b$ of $b$ elements of $[N]$.
        **else**
            Randomly choose without replacement the set $I_b$ of $b$ elements of $[N]$.
        **end if**
        Compute the "stochastic variance reduced gradient" $t_j^{k+1}$ by
            $$t_j^{k+1} = \frac{1}{b}\sum_{i\in I_b} \nabla h_i(x_j^{k+1}) + \tilde{\nu}^k - \frac{1}{b}\sum_{i\in I_b}\nabla h_i(\tilde{x}^k).$$
        Compute $y_j^{k+1} \in \partial r_2(x_j^{k+1})$ and let $z_j^{k+1} = t_j^{k+1} + y_j^{k+1}$.
        Solve the convex problem $x_{j+1}^{k+1} \in \arg\min\{G(x) + r_1(x) - \langle z_j^{k+1}, x\rangle : x \in \mathbb{R}^n\}$.
    **end for**
    Set $\tilde{x}^{k+1} = x_M^{k+1}$, and $k = k+1$.
**until** Stopping criterion.
```

For brevity, we denote $\bar{x}^k = \{x_0^k, x_1^k, \dots, x_{M-1}^k\}$ and $\bar{y}^k = \{y_0^k, y_1^k, \dots, y_{M-1}^k\}$ for all $k \in \mathbb{N}^*$. Let $\mathcal{P}_j^{k+1}$ be the sigma algebra defined as

$$\mathcal{P}_j^{k+1} = \sigma(\bar{x}^1, \dots, \bar{x}^k, x_0^{k+1}, x_1^{k+1}, \dots, x_j^{k+1}, \bar{y}^1, \dots, \bar{y}^k, y_0^{k+1}, y_1^{k+1}, \dots, y_j^{k+1}).$$

With the suitable selection of the minibatch size b and the inner-loop length M, we derive the following convergence results.

Theorem 2 *If the minibatch size b and the inner-loop length M satisfy*

$$\frac{M}{\sqrt{b}} \le \frac{1}{4\sqrt{e-1}}\frac{\rho_{G+r_1} + \rho_H + \rho_{r_2}}{L},$$

then

1. *The sequence $\{F(\tilde{x}^k)\}$ converges almost surely.*
2. $\sum_{k=0}^\infty \sum_{j=0}^{M-1} \mathbb{E}\|x_{j+1}^{k+1} - x_j^{k+1}\|^2 < +\infty.$
3. *Suppose the sequence $\{y_j^k\}$ is bounded almost surely, then every limit point of $\{x_j^k\}$ is a critical point of $F = (G + r_1) - (H + r_2)$ almost surely.*
4. *If r_2 has Lipschitz continuous gradient over* $\operatorname{dom} r_1$, *then $d_K = \mathcal{O}(1/\sqrt{K})$. Moreover, by choosing $b = \lfloor N^{2/3}\rfloor$, $M = \lfloor \mu\sqrt{b}\rfloor$, where*

$$\mu = \frac{1}{4\sqrt{e-1}}\frac{\rho_{G+r_1} + \rho_H + \rho_{r_2}}{L},$$

the complexity in terms of the number of gradient evaluations to obtain ϵ-criticality in expectation is $\mathcal{O}(N^{2/3}/\epsilon^2)$; Meanwhile, the complexity in terms of the number of convex subproblems being solved is $\mathcal{O}(1/\epsilon^2)$.

We will prove Theorem 2 simultaneously for both options with replacement and without replacement since the proofs for two options are essentially overlap. In the following proof, we will highlight parts that need to be analyzed differently for each option. Before giving the proof for Theorem 2, we first introduce the following lemma for the option of sample without replacement.

Lemma 3 *For the option of sample without replacement, the following inequality holds*

$$\mathbb{E}(\|t_j^{k+1} - \nabla H(x_j^{k+1})\|^2|\mathcal{P}_j^{k+1}) \leq \frac{L^2}{b}\left(1 - \frac{b-1}{N-1}\right)\|x_j^{k+1} - \tilde{x}^k\|^2 \leq \frac{L^2}{b}\|x_j^{k+1} - \tilde{x}^k\|^2.$$

The proof of this lemma is given in the appendix.

Proof [Proof of Theorem 2] 1. Since H is ρ_H-convex and by the definition of x_{j+1}^{k+1},

$$H(x_{j+1}^{k+1}) \geq H(x_j^{k+1}) + \langle \nabla H(x_j^{k+1}), x_{j+1}^{k+1} - x_j^{k+1}\rangle + \frac{\rho_H}{2}\|x_{j+1}^{k+1} - x_j^{k+1}\|^2,$$
$$r_2(x_{j+1}^{k+1}) \geq r_2(x_j^{k+1}) + \langle y_j^{k+1}, x_{j+1}^{k+1} - x_j^{k+1}\rangle + \frac{\rho_{r_2}}{2}\|x_{j+1}^{k+1} - x_j^{k+1}\|^2.$$

It follows from the definition of x_{j+1}^{k+1} that

$$(G + r_1)(x_j^{k+1}) \geq (G + r_1)(x_{j+1}^{k+1}) + \langle z_j^{k+1}, x_j^{k+1} - x_{j+1}^{k+1}\rangle + \frac{\rho_{G+r_1}}{2}\|x_j^{k+1} - x_{j+1}^{k+1}\|^2.$$

These inequalities imply

$$F(x_{j+1}^{k+1}) \leq F(x_j^{k+1}) + \langle x_{j+1}^{k+1} - x_j^{k+1}, t_j^{k+1} - \nabla H(x_j^{k+1})\rangle - \frac{\rho}{2}\|x_{j+1}^{k+1} - x_j^{k+1}\|^2,$$

where $\rho = \rho_H + \rho_{r_2} + \rho_{G+r_1}$. Let $\gamma > 0$ that will be determined later. By applying Schwartz inequality and AM-GM inequality,

$$F(x_{j+1}^{k+1}) \leq F(x_j^{k+1}) + \frac{1}{2\gamma}\|t_j^{k+1} - \nabla H(x_j^{k+1})\|^2 - \frac{\rho-\gamma}{2}\|x_{j+1}^{k+1} - x_j^{k+1}\|^2.$$

By taking conditional expectation with respect to $\mathcal{P}_j^{k+1}$,

$$\begin{aligned}\mathbb{E}(F(x_{j+1}^{k+1})|\mathcal{P}_j^{k+1}) \leq F(x_j^{k+1}) - \frac{\rho-\gamma}{2}\mathbb{E}(\|x_{j+1}^{k+1} - x_j^{k+1}\|^2|\mathcal{P}_j^{k+1})& \\ + \frac{1}{2\gamma}\mathbb{E}(\|t_j^{k+1} - \nabla H(x_j^{k+1})\|^2|\mathcal{P}_j^{k+1}).&\end{aligned}$$

If option is sample with replacement, the set I_b consists of independent indexes, we therefore evaluate $\mathbb{E}(\|t_j^{k+1} - \nabla H(x_j^{k+1})\|^2 | \mathcal{P}_j^{k+1})$ as follows,

$$\begin{aligned}
&\mathbb{E}(\|t_j^{k+1} - \nabla H(x_j^{k+1})\|^2 | \mathcal{P}_j^{k+1}) \\
&= \mathbb{E}_{I_b}\left(\left\|\frac{1}{b}\sum_{i\in I_b} \nabla h_i(x_j^{k+1}) - \frac{1}{b}\sum_{i\in I_b}\nabla h_i(\tilde{x}^k) + \nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1})\right\|^2\right) \\
&= \frac{1}{b}\mathbb{E}_i\left(\left\|\nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k) + \nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1})\right\|^2\right), \text{ where } i \overset{\text{uni}}{\sim} [N] \\
&\leq \frac{1}{b}\mathbb{E}_i\|\nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k)\|^2 \leq \frac{L^2}{b}\|x_j^{k+1} - \tilde{x}^k\|^2. \qquad (1)
\end{aligned}$$

Together with Lemma 3, the following inequality holds for two options

$$\mathbb{E}(\|t_j^{k+1} - \nabla H(x_j^{k+1})\|^2 | \mathcal{P}_j^{k+1}) \leq \frac{L^2}{b}\|x_j^{k+1} - \tilde{x}^k\|^2. \qquad (2)$$

Therefore,

$$\mathbb{E}(F(x_{j+1}^{k+1})|\mathcal{P}_j^{k+1}) \leq F(x_j^{k+1}) + \frac{L^2}{2b\gamma}\|x_j^{k+1} - \tilde{x}^k\|^2 - \frac{\rho-\gamma}{2}\mathbb{E}(\|x_{j+1}^{k+1} - x_j^{k+1}\|^2|\mathcal{P}_j^{k+1}).$$

Consider the sequence of Lyapunov functions $V_j^{k+1} = F(x_j^{k+1}) + c_j\|x_j^{k+1} - \tilde{x}^k\|^2$, where $\{c_j\}$ are non-negative numbers determined later (the idea of such sequence of Lyapunov functions is adopted from (J. Reddi et al., 2016)). We have

$$\begin{aligned}
\mathbb{E}(V_{j+1}^{k+1}|\mathcal{P}_j^{k+1}) &= \mathbb{E}(F(x_{j+1}^{k+1}) + c_{j+1}\|x_{j+1}^{k+1} - \tilde{x}^k\|^2|\mathcal{P}_j^{k+1}) \\
&\leq \mathbb{E}(F(x_{j+1}^{k+1}) + c_{j+1}(1+\beta)\|x_{j+1}^{k+1} - x_j^{k+1}\|^2 \\
&\quad + c_{j+1}(1+1/\beta)\|x_j^{k+1} - \tilde{x}^k\|^2|\mathcal{P}_j^{k+1}), \text{ where } \beta > 0 \\
&\leq F(x_j^{k+1}) + \left(\frac{L^2}{2b\gamma} + c_{j+1}\left(1+\frac{1}{\beta}\right)\right)\|x_j^{k+1} - \tilde{x}^k\|^2 \\
&\quad + \left(c_{j+1}(1+\beta) - \frac{\rho-\gamma}{2}\right)\mathbb{E}(\|x_{j+1}^{k+1} - x_j^{k+1}\|^2|\mathcal{P}_j^{k+1}). \qquad (3)
\end{aligned}$$

To obtain V_j^{k+1} in the right-hand side of (3), we choose the sequence $\{c_j\}$ in such a way $c_M = 0$ and $c_j = \frac{L^2}{2b\gamma} + c_{j+1}\left(1+\frac{1}{\beta}\right)$ if $j < M$. These relations yield $c_j = \frac{\beta L^2}{2b\gamma}\left(\left(1+\frac{1}{\beta}\right)^{M-j} - 1\right)$. Next, to achieve the descent property on the Lyapunov sequence, i.e. $\mathbb{E}(V_{j+1}^{k+1}|\mathcal{P}_j^{k+1}) < V_j^{k+1}$, we want $c_{j+1}(1+\beta) + \frac{\gamma}{2} \leq \frac{\rho}{4}, \forall j = \overline{0, M-1}$, or equivalently,

$$(1+\beta)\frac{\beta L^2}{2b\gamma}\left(\left(1+\frac{1}{\beta}\right)^{M-j-1} - 1\right) + \frac{\gamma}{2} \leq \frac{\rho}{4}, \forall j = \overline{0, M-1},$$

which is further equivalent to

$$(1+\beta)\frac{\beta L^2}{2b\gamma}\left(\left(1+\frac{1}{\beta}\right)^{M-1} - 1\right) + \frac{\gamma}{2} \leq \frac{\rho}{4}. \qquad (4)$$

By choosing $\beta = M-1$, and noticing that $\left(1+\frac{1}{M-1}\right)^{M-1} \leq e$ and $(1+\beta)\beta < (1+\beta)^2 = M^2$, (4) holds if the following stronger inequality holds

$$\frac{M^2L^2}{2b\gamma}(e-1) + \frac{\gamma}{2} \leq \frac{\rho}{4}. \tag{5}$$

Now by choosing $\gamma = \frac{ML\sqrt{e-1}}{\sqrt{b}}$ to optimize the LHS of (5), the inequality (5) becomes $\frac{M}{\sqrt{b}} \leq \frac{\rho}{4L\sqrt{e-1}}$. Consequently, if the minibatch size b and the inner-loop length M satisfy this inequality, together with (3) we get

$$\mathbb{E}(V_{j+1}^{k+1}|\mathcal{P}_j^{k+1}) \leq V_j^{k+1} - \frac{\rho}{4}\mathbb{E}(\|x_{j+1}^{k+1} - x_j^{k+1}\|^2|\mathcal{P}_j^{k+1}). \tag{6}$$

Note that $\{V_j^{k+1}\}$ is an adapted process with respect to the filtration $\{\mathcal{P}_j^{k+1}\}$, and $V_j^{k+1} \geq F(x_j^{k+1}) \geq \alpha, \forall k, j$. It follows from supermartingale convergence theorem (Bertsekas et al., 2003) that there exists a random variable V^∞ such that the sequence $\{V_j^{k+1}\}$ converges to V^∞ almost surely. As a consequence, the sequence $\{F(\tilde{x}^k)\}$ converges to V^∞ almost surely since $V_0^{k+1} = F(x_0^{k+1}) = F(\tilde{x}^k)$.

2. From (6), we obtain $\frac{\rho}{4}\mathbb{E}\|x_{j+1}^{k+1}-x_j^{k+1}\|^2+\mathbb{E}(V_{j+1}^{k+1}) \leq \mathbb{E}(V_j^{k+1})$. From this inequality and the fact that $\mathbb{E}(V_0^1) = \mathbb{E}(F(\tilde{x}^0)) = F(\tilde{x}^0) < +\infty$, by induction, one obtains $\mathbb{E}(V_j^{k+1}) < +\infty$, for all k, j. By telescoping with j and k,

$$\frac{\rho}{4}\sum_{k=0}^{\infty}\sum_{j=0}^{M-1}\mathbb{E}\|x_{j+1}^{k+1} - x_j^{k+1}\|^2 \leq \mathbb{E}(V_0^1) - \alpha < +\infty. \tag{7}$$

3. From the evaluation (2), we obtain

$$\mathbb{E}\left\|t_j^{k+1} - \nabla H(x_j^{k+1})\right\|^2 \leq \frac{L^2(M-1)}{b}\left(\mathbb{E}\|x_1^{k+1} - x_0^{k+1}\|^2 + \ldots + \mathbb{E}\|x_{M-1}^{k+1} - x_{M-2}^{k+1}\|^2\right),$$

which implies

$$\sum_{k=0}^{\infty}\sum_{j=0}^{M-1}\mathbb{E}\left\|t_j^{k+1} - \nabla H(x_j^{k+1})\right\|^2 < +\infty. \tag{8}$$

Let $S_1 = \left(t_j^{k+1} - \nabla H(x_j^{k+1}) \to 0\right)$. It follows from (8) that $\mathbb{P}(S_1) = 1$. On the other hand, let $S_2 = \left(x_{j+1}^{k+1} - x_j^{k+1} \to 0\right)$, it follows from (7) that $\mathbb{P}(S_2) = 1$. Furthermore, $\mathbb{P}(S_3) = 1$, where $S_3 = (y_j^k$ is bounded).

Now, consider an event in $S_1 \cap S_2 \cap S_3$ which gives rise to a realization $\{x_j^k\}$ (here, $\{x_j^k\}$ is a sequence of real numbers rather than a sequence of random variables). Let x^* be a limit point of $\{x_j^k\}$. Note that the sequence $\{x_j^k\}$ generated by the algorithm is expressed explicitly in the following order

$$x_0^1, x_1^1, \ldots, x_{M-1}^1, x_M^1 \equiv x_0^2, x_1^2, \ldots, x_{M-1}^2, x_M^2 \equiv x_0^3, x_1^3, \ldots$$

We observe that a subsequence of $\{x_j^k\}$ has the following form $\{x_{v(l)}^{u(l)}\}_{l\in\mathbb{N}}$, where $u:\mathbb{N}\to\mathbb{N}^*$ and $v:\mathbb{N}\to\{0,1,\dots,M-1\}$ satisfy the following relation: for every $i<t$ in $\mathbb{N}$, either $u(i)<u(t)$ or $u(i)=u(t)$ and $v(i)<v(t)$. Since x^* is a limit point of $\{x_j^k\}$, there exists a subsequence $\{x_{v(l)}^{u(l)}\}$ converging to x^* as $l\to\infty$, which implies $x_{v(l)+1}^{u(l)}\to x^*$. As a consequence, $t_{v(l)}^{u(l)}\to\nabla H(x^*)$. By passing to a subsequence if necessary, we assume that $y_{v(l)}^{u(l)}$ converges to y^*. Since $y_{v(l)}^{u(l)}\in\partial r_2\left(x_{v(l)}^{u(l)}\right)$, we obtain $y^*\in\partial r_2(x^*)$ thanks to the closedness of the graph of the subdifferential operator.

By the definition of $x_{v(l)+1}^{u(l)}$, we obtain $z_{v(l)}^{u(l)}\in\partial(G+r_1)\left(x_{v(l)+1}^{u(l)}\right)$. Since the graph of the subdifferential operator is closed, by letting $l\to\infty$ we derive $\nabla H(x^*)+y^*\in\partial(G+r_1)(x^*)$, hence x^* is a critical point of $F=(G+r_1)-(H+r_2)$.

4. We first evaluate $\mathbb{E}\operatorname{dist}(\nabla H(x_{j+1}^{k+1})+\nabla r_2(x_{j+1}^{k+1}),\partial(G+r_1)(x_{j+1}^{k+1}))$. Since $z_j^{k+1}\in\partial(G+r_1)(x_{j+1}^{k+1})$, we obtain

$$\begin{aligned}
&\mathbb{E}\operatorname{dist}(\nabla H(x_{j+1}^{k+1})+\nabla r_2(x_{j+1}^{k+1}),\partial(G+r_1)(x_{j+1}^{k+1}))\\
&\le\mathbb{E}\|\nabla H(x_{j+1}^{k+1})-t_j^{k+1}\|+\mathbb{E}\|\nabla r_2(x_{j+1}^{k+1})-\nabla r_2(x_j^{k+1})\|\\
&\le\mathbb{E}\|\nabla H(x_{j+1}^{k+1})-\nabla H(x_j^{k+1})\|+\mathbb{E}\|\nabla H(x_j^{k+1})-t_j^{k+1}\|+\mathbb{E}\|\nabla r_2(x_{j+1}^{k+1})-\nabla r_2(x_j^{k+1})\|\\
&\le(L+L_{r_2})\mathbb{E}\|x_{j+1}^{k+1}-x_j^{k+1}\|+\left(\mathbb{E}\|t_j^{k+1}-\nabla H(x_j^{k+1})\|^2\right)^{\frac{1}{2}}\\
&\le(L+L_{r_2})\mathbb{E}\|x_{j+1}^{k+1}-x_j^{k+1}\|+\frac{L}{\sqrt{b}}\left(\mathbb{E}\|x_j^{k+1}-\tilde{x}^k\|^2\right)^{\frac{1}{2}}.
\end{aligned}\tag{9}$$

Furthermore, we have

$$\begin{aligned}
&\|x_j^{k+1}-\tilde{x}^k\|^2=\|x_j^{k+1}-x_0^{k+1}\|^2\\
&\le(M-j+M-j-1+\ldots+M-1)\times\\
&\quad\left(\frac{1}{M-j}\|x_j^{k+1}-x_{j-1}^{k+1}\|^2+\ldots+\frac{1}{M-1}\|x_1^{k+1}-x_0^{k+1}\|^2\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{j=1}^{M-1}\left(\mathbb{E}\|x_j^{k+1}-\tilde{x}^k\|^2\right)^{\frac{1}{2}}\le\sum_{j=1}^{M-1}\left(\sum_{r=1}^{j}(M-r)\right)^{\frac{1}{2}}\left(\sum_{r=1}^{j}\frac{1}{M-r}\mathbb{E}\|x_r^{k+1}-x_{r-1}^{k+1}\|^2\right)^{\frac{1}{2}}\\
&\le\left(\sum_{j=1}^{M-1}\sum_{r=1}^{j}(M-r)\right)^{\frac{1}{2}}\left(\sum_{j=1}^{M-1}\sum_{r=1}^{j}\frac{1}{M-r}\mathbb{E}\|x_r^{k+1}-x_{r-1}^{k+1}\|^2\right)^{\frac{1}{2}}\\
&=\left(\frac{(M-1)M(2M-1)}{6}\right)^{\frac{1}{2}}\left(\sum_{r=1}^{M-1}\mathbb{E}\|x_r^{k+1}-x_{r-1}^{k+1}\|^2\right)^{\frac{1}{2}}.
\end{aligned}\tag{10}$$

It follows from (9) and (10) that

$$\sum_{k=0}^{K-1}\sum_{j=0}^{M-1}\mathbb{E}\operatorname{dist}(\nabla H(x_{j+1}^{k+1})+\nabla r_2(x_{j+1}^{k+1}),\partial(G+r_1)(x_{j+1}^{k+1}))\leq(L+L_{r_2})\sum_{k=0}^{K-1}\sum_{j=0}^{M-1}\mathbb{E}\|x_{j+1}^{k+1}-x_j^{k+1}\|$$

$$+\frac{L}{\sqrt{b}}\left(\frac{(M-1)M(2M-1)}{6}\right)^{\frac{1}{2}}\sum_{k=0}^{K-1}\left(\sum_{r=1}^{M-1}\mathbb{E}\|x_r^{k+1}-x_{r-1}^{k+1}\|^2\right)^{\frac{1}{2}}$$

$$\leq\sqrt{K}\left((L+L_{r_2})\sqrt{M}+\frac{L}{\sqrt{b}}\left(\frac{(M-1)M(2M-1)}{6}\right)^{\frac{1}{2}}\right)\left(\sum_{k=0}^{K-1}\sum_{j=0}^{M-1}\mathbb{E}\|x_{j+1}^{k+1}-x_j^{k+1}\|^2\right)^{\frac{1}{2}}.$$

As a consequence, the iteration convergence rate is given by $d_K=O(1/\sqrt{K})$.

Now we derive the complexity to find ϵ-criticality as follows. After K iterations, the output x_a is chosen uniformly from $\{x_{j+1}^{k+1}\}_{j=0,\dots,M-1}^{k=0,\dots,K-1}$, it holds, for some $C>0$,

$$KM\cdot\mathbb{E}\operatorname{dist}(\nabla H(x_a)+\nabla r_2(x_a),\partial(G+r_1)(x_a))$$
$$\leq\sqrt{K}\left((L+L_{r_2})\sqrt{M}+\frac{L}{\sqrt{b}}\left(\frac{(M-1)M(2M-1)}{6}\right)^{\frac{1}{2}}\right)C.$$

Let's choose $M=\lfloor\mu\sqrt{b}\rfloor$, where

$$\mu=\frac{1}{4\sqrt{e-1}}\frac{\rho_{G+r_1}+\rho_H+\rho_{r_2}}{L}$$

one gets

$$KM\cdot\mathbb{E}\operatorname{dist}(\nabla H(x_a)+\nabla r_2(x_a),\partial(G+r_1)(x_a))\leq C\sqrt{K}\left((L+L_{r_2})\sqrt{M}+\frac{L}{2\sqrt{b}}M^{3/2}\right)$$
$$\leq C\sqrt{K}\left((L+L_{r_2})\sqrt{M}+\frac{L\mu}{2}\sqrt{M}\right)\leq C\sqrt{MK}\left(L+L_{r_2}+\frac{L\mu}{2}\right).$$

Therefore,

$$\mathbb{E}\operatorname{dist}(\nabla H(x_a)+\nabla r_2(x_a),\partial(G+r_1)(x_a))=\mathcal{O}\left(\frac{1}{\sqrt{KM}}\right).$$

By choosing $b=\lfloor N^{2/3}\rfloor$, the algorithm needs $K=\mathcal{O}(\frac{1}{\epsilon^2N^{1/3}})$ outer iterations to attain an ϵ-DC critical point. The total number of gradient evaluations is then:

$$K\left(N+2bM\right)\leq K(N+2N^{2/3}\mu N^{1/3})=KN(1+2\mu)=\mathcal{O}\left(\frac{N^{2/3}}{\epsilon^2}\right).$$

On the other hand, there are KM convex subproblems to be solved, leading to the complexity of $\mathcal{O}(1/\epsilon^2)$ in terms of the number of convex subproblems.

■

Remark 4 *About the constraint between b and M, $\frac{M}{\sqrt{b}} \leq \frac{1}{4\sqrt{e-1}} \frac{\rho}{L}$:*

i. If the DC decomposition admits a large number $\frac{\rho}{L}$, we have great flexibility to choose b and M.

ii. If the DC problem has a very small value of $\frac{\rho}{L}$, we can adjust the DC decomposition to increase $\frac{\rho}{L}$. Indeed, to this end, we add $\frac{\gamma}{2}\|\cdot\|^2$ ($\gamma > 0$) to both DC components, $F = \left(G + r_1 + \frac{\gamma}{2}\|\cdot\|^2\right) - \left(H + r_2 + \frac{\gamma}{2}\|\cdot\|^2\right)$. With this new DC decomposition ($\bar{G} = G, \bar{H} = H, \bar{r_1} = r_1 + \frac{\gamma}{2}\|\cdot\|^2, \bar{r_2} = r_2 + \frac{\gamma}{2}\|\cdot\|^2$), the larger value is obtained since

$$\frac{\bar{\rho}}{\bar{L}} = \frac{\rho + 2\gamma}{L} \to +\infty \text{ as } \gamma \to +\infty.$$

Nevertheless, one should be careful when using the above technique since adding a strongly convex term to both DC components also yields an increase in the "gap" between the second DC component and its linear minorant, which potentially leads to bad approximations.

3.2 The Second Stochastic DCA: DCA-SAGA

In this subsection, we propose another stochastic DCA called DCA-SAGA (algorithm 2). Just as DCA-SVRG, the new scheme is a combination of the deterministic DCA and the SAGA-style of stochastic gradient update. That is, the gradient of H in the deterministic DCA is replaced by the stochastic gradient given by SAGA.

Algorithm 2 DCA-SAGA

```
Initialization: $x^0 \in \operatorname{dom} r_1$, set $\alpha_i^0 = x^0, \forall i = \overline{1, N}, t = 0$, option (either with replacement or without replacement).
Compute the full gradient $\nu^0 = \frac{1}{N}\sum_{i=1}^N \nabla h_i(\alpha_i^0)$.
repeat
    if option is with replacement then
        Randomly choose with replacement the set $I_t$ of $b$ elements of $[N]$.
    else
        Randomly choose without replacement the set $I_t$ of $b$ elements of $[N]$.
    end if
    Compute the "stochastic variance reduced gradient" $v^t$,
        $v^t = \frac{1}{b}\sum_{i \in I_t} \nabla h_i(x^t) + \nu^t - \frac{1}{b}\sum_{i \in I_t} \nabla h_i(\alpha_i^t).$
    Compute $y^t \in \partial r_2(x^t)$, and let $z^t = v^t + y^t$.
    Solve the convex program $x^{t+1} \in \arg\min\{G(x) + r_1(x) - \langle z^t, x\rangle : x \in \mathbb{R}^n\}$.
    Update $\alpha_j^{t+1} = \begin{cases} x^t, & \text{if } j \in I_t \\ \alpha_j^t, & \text{otherwise} \end{cases}$
    Update the "full gradient" $\nu^{t+1} = \nu^t - \frac{1}{N}\sum_{i \in I_t} \nabla h_i(\alpha_i^t) + \frac{1}{N}\sum_{i \in I_t} \nabla h_i(\alpha_i^{t+1})$.
    Update $t = t + 1$.
until Stopping criterion.
```

We denote $d_T = \min_{t=0,1,\ldots,T-1} \mathbb{E}\operatorname{dist}(\partial(H + r_2)(x^{t+1}), \partial(G + r_1)(x^{t+1}))$. For brevity, we denote $\bar{\alpha}^t = \{\alpha_1^t, \alpha_2^t, \ldots, \alpha_N^t\}, \forall t \in \mathbb{N}$. We denote the sequence of increasing sigma algebra $\{\mathcal{P}_t\}$ as $\mathcal{P}_t = \sigma(x^0, x^1, \dots, x^t, y^0, y^1, \dots, y^t, \bar{\alpha}^0, \bar{\alpha}^1, \dots, \bar{\alpha}^t)$. The convergence results of DCA-SAGA is described as follows.

Theorem 5 *If the minibatch size b satisfies $\frac{N\sqrt{N+b}}{b^2} \leq \frac{\rho_H+\rho_{r_2}+\rho_{G+r_1}}{4L}$ for the option of sample with replacement or $\frac{N\sqrt{N+1}}{b^2} \leq \frac{\rho_H+\rho_{r_2}+\rho_{G+r_1}}{4L}$ for the option of sample without replacement, then*

1. *$\sum_{t=0}^{\infty} \mathbb{E}\|x^{t+1} - x^t\|^2 < \infty$.*

2. *$\sum_{t=0}^{\infty} \sum_{i=1}^{N} \mathbb{E}\|x^t - \alpha_i^t\|^2 < \infty$.*

3. *The sequence $\{F(x^t)\}$ converges almost surely.*

4. *Suppose the sequence $\{y^t\}$ is bounded almost surely. Then, every limit point of $\{x^t\}$ is a critical point of $F = (G + r_1) - (H + r_2)$ almost surely.*

5. *If r_2 has Lipschitz continuous gradient over $\operatorname{dom} r_1$, then $d_T = \mathcal{O}(1/\sqrt{T})$. Furthermore, by choosing*

$$b = \left\lceil \frac{\sqrt[4]{2}}{\sqrt{\mu}} N^{3/4} \right\rceil, \text{ where } \mu = \frac{\rho_H + \rho_{r_2} + \rho_{G+r_1}}{4L},$$

the complexity in terms of number of gradient evaluations to obtain ϵ-criticality in expectation is $\mathcal{O}(N + N^{3/4}/\epsilon^2)$; Meanwhile, the complexity in terms of number of convex subproblems solved is $\mathcal{O}(1/\epsilon^2)$.

Proof [Proof of Theorem 5]

Firstly, we will prove the above properties for the option of sample with replacement. Then, the proof for the option of sample without replacement is sketched, where the main different arguments are highlighted.

Consider the option of sample with replacement.

1. Since H is ρ_H-convex, r_2 is ρ_{r_2}-convex, and by the definition of x^{t+1},

$$H(x^{t+1}) \geq H(x^t) + \langle \nabla H(x^t), x^{t+1} - x^t \rangle + \frac{\rho_H}{2}\|x^{t+1} - x^t\|^2.$$
$$r_2(x^{t+1}) \geq r_2(x^t) + \langle y^t, x^{t+1} - x^t \rangle + \frac{\rho_{r_2}}{2}\|x^{t+1} - x^t\|^2,$$
$$(G + r_1)(x^{t+1}) \leq (G + r_1)(x^t) + \langle z^t, x^{t+1} - x^t \rangle - \frac{\rho_{G+r_1}}{2}\|x^{t+1} - x^t\|^2.$$

Combining these inequalities,

$$F(x^{t+1}) \leq F(x^t) + \langle x^{t+1} - x^t, v^t - \nabla H(x^t) \rangle - \frac{\rho}{2}\|x^{t+1} - x^t\|^2, \tag{11}$$

where $\rho = \rho_H + \rho_{r_2} + \rho_{G+r_1}$. By taking conditional expectation with respect to $\mathcal{P}_t$,

$$\mathbb{E}_{I_t}(F(x^{t+1})) \leq F(x^t) + \mathbb{E}_{I_t}\left(\langle x^{t+1} - x^t, v^t - \nabla H(x^t) \rangle\right) - \frac{\rho}{2}\mathbb{E}_{I_t}\|x^{t+1} - x^t\|^2,$$

which implies

$$\mathbb{E}_{I_t}(F(x^{t+1})) \leq F(x^t) + \frac{1}{2\gamma}\mathbb{E}_{I_t}\|v^t - \nabla H(x^t)\|^2 - \frac{\rho-\gamma}{2}\mathbb{E}_{I_t}\|x^{t+1}-x^t\|^2 \tag{12}$$

where $\gamma > 0$ that will be determined later to gain some advantages.

We now evaluate $\mathbb{E}_{I_t}\|v^t - \nabla H(x^t)\|^2$ as follows,

$$\begin{aligned}\mathbb{E}_{I_t}\|v^t - \nabla H(x^t)\|^2 &= \mathbb{E}_{I_t}\left\|\frac{1}{b}\sum_{i\in I_t}\nabla h_i(x^t) + \nu^t - \frac{1}{b}\sum_{i\in I_t}\nabla h_i(\alpha_i^t) - \nabla H(x^t)\right\|^2\\ &= \frac{1}{b}\mathbb{E}_i\left\|\nabla h_i(x^t) - \nabla h_i(\alpha_i^t) + \nu^t - \nabla H(x^t)\right\|^2 \text{ where } i \overset{\text{uni}}{\sim} \{1,2,\dots,N\}\\ &\leq \frac{1}{b}\mathbb{E}_i\|\nabla h_i(x^t) - \nabla h_i(\alpha_i^t)\|^2 \leq \frac{L^2}{b}\mathbb{E}_i\|x^t-\alpha_i^t\|^2 = \frac{L^2}{bN}\sum_{i=1}^N\|x^t-\alpha_i^t\|^2. \end{aligned}\tag{13}$$

Therefore,

$$\mathbb{E}_{I_t}(F(x^{t+1})) \leq F(x^t) - \frac{\rho-\gamma}{2}\mathbb{E}_{I_t}\|x^{t+1}-x^t\|^2 + \frac{L^2}{2\gamma bN}\sum_{i=1}^N\|x^t-\alpha_i^t\|^2.$$

We define the sequence of Lyapunov functions $V^t = F(x^t) + c_t\sum_{i=1}^N\|x^t-\alpha_i^t\|^2$, where the sequence $\{c_t\}$ will be determined later. For all $\beta > 0$,

$$\begin{aligned}&\mathbb{E}_{I_t}(V^{t+1}) = \mathbb{E}_{I_t}\left(F(x^{t+1}) + c_{t+1}\sum_{i=1}^N\|x^{t+1}-\alpha_i^{t+1}\|^2\right)\\ &\leq \mathbb{E}_{I_t}\left(F(x^{t+1}) + c_{t+1}(\beta+1)\sum_{i=1}^N\|x^{t+1}-x^t\|^2 + c_{t+1}(1+1/\beta)\sum_{i=1}^N\|x^t-\alpha_i^{t+1}\|^2\right)\\ &\leq \mathbb{E}_{I_t}\Bigg(F(x^t) + \left(Nc_{t+1}(\beta+1) - \frac{\rho-\gamma}{2}\right)\|x^{t+1}-x^t\|^2\\ &\qquad + \frac{L^2}{2\gamma bN}\sum_{i=1}^N\|x^t-\alpha_i^t\|^2 + c_{t+1}(1+1/\beta)\sum_{i\notin I_t}\|x^t-\alpha_i^t\|^2\Bigg)\end{aligned}$$

$$\begin{aligned}&= F(x^t) + \left(Nc_{t+1}(\beta+1) - \frac{\rho-\gamma}{2}\right)\mathbb{E}_{I_t}\|x^{t+1}-x^t\|^2\\ &\quad + \left(\frac{L^2}{2\gamma bN} + c_{t+1}(1+1/\beta)\right)\sum_{i=1}^N\|x^t-\alpha_i^t\|^2 - c_{t+1}\left(1+\frac{1}{\beta}\right)b\mathbb{E}_i\|x^t-\alpha_i^t\|^2\\ &= F(x^t) + \left(Nc_{t+1}(\beta+1) - \frac{\rho-\gamma}{2}\right)\mathbb{E}_{I_t}\|x^{t+1}-x^t\|^2\\ &\quad + \left(\frac{L^2}{2\gamma bN} + c_{t+1}\left(1+\frac{1}{\beta}\right)\left(1-\frac{b}{N}\right)\right)\sum_{i=1}^N\|x^t-\alpha_i^t\|^2.\end{aligned}$$

Similar to the proof of DCA-SVRG, we set $c_t = \frac{L^2}{2\gamma bN} + c_{t+1}\left(1+\frac{1}{\beta}\right)\left(1-\frac{b}{N}\right)$, or equivalently, $c_{t+1} = \left(c_t - \frac{L^2}{2\gamma bN}\right)\frac{\beta N}{(\beta+1)(N-b)}$. We obtain by recursion that

$$c_t = \left(c_0 - \frac{L^2\beta}{2\gamma b(\beta b + b - N)}\right)\left(\frac{\beta N}{(\beta+1)(N-b)}\right)^t + \frac{L^2\beta}{2\gamma b(\beta b+b-N)}.$$

From the above recursion, if we choose the initial value $c_0 = \frac{L^2\beta}{2\gamma b(\beta b+b-N)}$, we will get $c_0 = c_1 = c_2 = \ldots = c_t = \ldots$ Next, we choose β such that $\beta b + b - N > 0 \Leftrightarrow \beta > \frac{N-b}{b}$ $(\star)$ to make the sequence $\{c_t\}$ nonnegative. With this choice, we obtain $\mathbb{E}_{I_t}(V^{t+1}) \leq V^t + \left(Nc_{t+1}(\beta+1) - \frac{\rho-\gamma}{2}\right)\mathbb{E}_{I_t}\|x^{t+1}-x^t\|^2$.

In order to obtain the "decreasing property" of the sequence $\{V^t\}$, we want the term $Nc_{t+1}(\beta+1) - \frac{\rho-\gamma}{2}$ to be negative. More specifically, we want $Nc_{t+1}(\beta+1) + \frac{\gamma}{2} \leq \frac{\rho}{4}$, or equivalently,

$$N(\beta+1)\frac{L^2\beta}{2\gamma b(\beta b+b-N)} + \frac{\gamma}{2} \leq \frac{\rho}{4}. \tag{14}$$

Now by choosing $\gamma > 0$ to make the LHS(14) as small as possible (AM-GM inequality), $\gamma = L\sqrt{\frac{\beta(\beta+1)N}{b(\beta b+b-N)}}$, we need $\sqrt{\frac{N(\beta+1)L^2\beta}{b(\beta b+b-N)}} \leq \frac{\rho}{4}$, or equivalently

$$\sqrt{\frac{N(\beta+1)\beta}{b(\beta b+b-N)}} \leq \frac{\rho}{4L}. \tag{15}$$

From $(\star)$, we choose $\beta = N/b$. The inequality (15) becomes

$$\frac{N\sqrt{N+b}}{b^2} \leq \frac{\rho}{4L}. \tag{16}$$

Consequently, if b satisfies (16), we obtain $\mathbb{E}_{I_t}(V^{t+1}) \leq V^t - \frac{\rho}{4}\mathbb{E}_{I_t}\|x^{t+1}-x^t\|^2$. Since $c_t \geq 0$, we obtain $V^t \geq F(x^t) \geq \alpha > -\infty$. By applying supermartingale convergence theorem, we obtain: there exits V^∞ such that $V^t \to V^\infty$ a.s. Next, we have $\frac{\rho}{4}\mathbb{E}\|x^{t+1}-x^t\|^2 \leq \mathbb{E}(V^t) - \mathbb{E}(V^{t+1})$. By induction based on this inequality, we obtain $\mathbb{E}(V^t)$ is finite forall t. By telescoping the above inequality, we derive $\frac{\rho}{4}\sum_{t=0}^{\infty}\mathbb{E}\|x^{t+1}-x^t\|^2 \leq \mathbb{E}(V^0) - \alpha < +\infty$.

2. We now evaluate the term $\mathbb{E}\|x^t - \alpha_i^t\|^2$. This is a key for us to establish the convergence to critical points. By the definition, we can write $\alpha_i^t = x^{t-1}\cdot 1_{\{i\in I_{t-1}\}} + \alpha_i^{t-1}\cdot 1_{\{i\notin I_{t-1}\}}$. Similarly, we can track back one further step $\alpha_i^t = x^{t-1}\cdot 1_{\{i\in I_{t-1}\}} + x^{t-2}\cdot 1_{\{i\in \bar{I}_{t-1}\cap I_{t-2}\}} + \alpha_i^{t-2}\cdot 1_{\{i\in\bar{I}_{t-1}\cap\bar{I}_{t-2}\}}$.

By doing so repeatedly,

$$\begin{aligned}\alpha_i^t =& x^{t-1}\cdot 1_{\{i\in I_{t-1}\}} + x^{t-2}\cdot 1_{\{i\in\bar{I}_{t-1}\cap I_{t-2}\}} + x^{t-3}\cdot 1_{\{i\in\bar{I}_{t-1}\cap\bar{I}_{t-2}\cap I_{t-3}\}}\\ &+\ldots + x^0\cdot 1_{\{i\in\bar{I}_{t-1}\cap\bar{I}_{t-2}\cdots\cap\bar{I}_1\}}.\end{aligned}$$

It is worth noting that $\{i \in I_{t-1}\}, \{i \in \bar{I}_{t-1} \cap I_{t-2}\}, \{i \in \bar{I}_{t-1} \cap \bar{I}_{t-2} \cap I_{t-3}\}, \ldots, \{i \in \bar{I}_{t-1} \cap \bar{I}_{t-2} \cdots \cap \bar{I}_1\}$ form a partition of the sample space Ω, and we obtain

$$\alpha_i^t = \begin{cases} x^{t-1} \text{ on } \{i \in I_{t-1}\} \\ x^{t-2} \text{ on } \{i \in \bar{I}_{t-1} \cap I_{t-2}\} \\ x^{t-3} \text{ on } \{i \in \bar{I}_{t-1} \cap \bar{I}_{t-2} \cap I_{t-3}\} \\ \ldots \\ x^0 \text{ on } \{i \in \bar{I}_{t-1} \cap \bar{I}_{t-2} \cdots \cap \bar{I}_1\}. \end{cases}$$

Therefore,

$$\begin{aligned} &\mathbb{E}\|x^t - \alpha_i^t\|^2 = \int_\Omega \|x^t - \alpha_i^t\|^2 d\mathbb{P} \\ &= \int_{\{i \in I_{t-1}\}} \|x^t - x^{t-1}\|^2 d\mathbb{P} + \int_{\{i \in \bar{I}_{t-1} \cap I_{t-2}\}} \|x^t - x^{t-2}\|^2 d\mathbb{P} \\ &+ \int_{\{i \in \bar{I}_{t-1} \cap \bar{I}_{t-2} \cap I_{t-3}\}} \|x^t - x^{t-3}\|^2 d\mathbb{P} + \ldots + \int_{\{i \in \bar{I}_{t-1} \cap \bar{I}_{t-2} \cdots \cap \bar{I}_1\}} \|x^t - x^0\|^2 d\mathbb{P}. \end{aligned} \tag{17}$$

Our main aim is to prove $\sum_{t=1}^\infty \mathbb{E}\|x^t - \alpha_i^t\|^2 < +\infty$.
Here we have several remarks playing as guiding light in our proof:

- We already proved $\sum_{t=0}^\infty \mathbb{E}\|x^{t+1} - x^t\|^2 < +\infty$, therefore, we will find links between $\mathbb{E}\|x^t - \alpha_i^t\|^2$ and $\mathbb{E}\|x^t - x^{t-1}\|^2, \mathbb{E}\|x^{t-1} - x^{t-2}\|^2, \ldots$

- Our main challenge when dealing with the RHS of (17) is that it is very hard to compute explicitly those integrals since, for instance, $\|x^t - x^{t-1}\|^2$ and $\{i \in \bar{I}_{t-1}\}$ are not independent.

- We know that if a random variable X is independent to the set of events B, then we have the property $\int_B X d\mathbb{P} = \mathbb{E}(X)\mathbb{P}(B)$. We will therefore evaluate in such a way that after our evaluations, by grouping integrals with the common function integrated, the combined region is independent to that function. For example, after some evaluations, we obtain something of the form

$$\int_{A_1} X d\mathbb{P} + \int_{A_2} X d\mathbb{P} + \ldots + \int_{A_m} X d\mathbb{P} = \int_{A_1 \cup A_2 \cup \ldots \cup A_m} X d\mathbb{P},$$

  where $A_1, A_2, \ldots, A_m$ are pair-wise disjoint. Here our aim is to have $A_1 \cup A_2 \cup \ldots \cup A_m$ being independent to X, so we can use the mentioned property.

With these meta-ideas kept in mind, we are back to the proof. We denote $\{\epsilon_i\}$ the sequence as $\epsilon_i = \frac{1}{i^2}$. The sequence $\{\epsilon_i\}$ plays a crucial role in our proof. Moreover, we denote $A = \sum_{i=1}^\infty \epsilon_i = \sum_{i=1}^\infty \dfrac{1}{i^2} < +\infty$. We define another sequence $\{\theta_i\}$ as $\theta_i = \frac{\epsilon_i}{A}$, which implies $\sum_{i=1}^{\infty} \theta_i = 1$. We will evaluate $\mathbb{E}\|x^t - \alpha_i^t\|^2, (t \geq 2)$ based on (17) as follows. Cauchy-Schwartz inequality implies

$$\begin{aligned}\int_{\{i\in\bar{I}_{t-1}\cap I_{t-2}\}} \|x^t - x^{t-2}\|^2 d\mathbb{P} \leq & \frac{1}{\theta_1}\int_{\{i\in\bar{I}_{t-1}\cap I_{t-2}\}} \|x^t - x^{t-1}\|^2 d\mathbb{P} \\ & + \frac{1}{1-\theta_1}\int_{\{i\in\bar{I}_{t-1}\cap I_{t-2}\}} \|x^{t-1} - x^{t-2}\|^2 d\mathbb{P},\end{aligned}$$

$$\begin{aligned}&\int_{\{i\in\bar{I}_{t-1}\cap\bar{I}_{t-2}\cap I_{t-3}\}} \|x^t - x^{t-3}\|^2 d\mathbb{P} \leq \frac{1}{\theta_1}\int_{\{i\in\bar{I}_{t-1}\cap\bar{I}_{t-2}\cap I_{t-3}\}} \|x^t - x^{t-1}\|^2 d\mathbb{P} \\ &+ \frac{1}{\theta_2}\int_{\{i\in\bar{I}_{t-1}\cap\bar{I}_{t-2}\cap I_{t-3}\}} \|x^{t-1} - x^{t-2}\|^2 d\mathbb{P} \\ &+ \frac{1}{1-\theta_1-\theta_2}\int_{\{i\in\bar{I}_{t-1}\cap\bar{I}_{t-2}\cap I_{t-3}\}} \|x^{t-2} - x^{t-3}\|^2 d\mathbb{P},\end{aligned}$$

$$\ldots$$

$$\begin{aligned}&\int_{\{i\in\bar{I}_{t-1}\cap\ldots\cap\bar{I}_1\}} \|x^t - x^0\|^2 d\mathbb{P} \leq \frac{1}{\theta_1}\int_{\{i\in\bar{I}_{t-1}\cap\ldots\cap\bar{I}_1\}} \|x^t - x^{t-1}\|^2 d\mathbb{P} \\ &+ \frac{1}{\theta_2}\int_{\{i\in\bar{I}_{t-1}\cap\ldots\cap\bar{I}_1\}} \|x^{t-1} - x^{t-2}\|^2 d\mathbb{P} + \ldots \\ &+ \frac{1}{\theta_{t-1}}\int_{\{i\in\bar{I}_{t-1}\cap\ldots\cap\bar{I}_1\}} \|x^2 - x^1\|^2 d\mathbb{P} \\ &+ \frac{1}{1-\theta_1-\theta_2-\ldots-\theta_{t-1}}\int_{\{i\in\bar{I}_{t-1}\cap\ldots\cap\bar{I}_1\}} \|x^1 - x^0\|^2 d\mathbb{P}.\end{aligned}$$

By adding (17) and these inequalities, then grouping integrals sharing common the function integrated, namely $\|x^t - x^{t-1}\|^2, \|x^{t-2} - x^{t-1}\|^2, \ldots, \|x^1 - x^0\|^2$, and noticing

$$\begin{aligned}&\{i \in \bar{I}_{t-1} \cap I_{t-2}\} \cup \{i \in \bar{I}_{t-1} \cap \bar{I}_{t-2} \cap I_{t-3}\} \ldots \cup \{i \in \bar{I}_{t-1} \cap \ldots \bar{I}_1\} = \{i \in \bar{I}_{t-1}\}, \\ &\{i \in \bar{I}_{t-1} \cap \bar{I}_{t-2} \cap I_{t-3}\} \cup \ldots \cup \{i \in \bar{I}_{t-1} \cap \ldots \bar{I}_1\} = \{i \in \bar{I}_{t-1} \cap \bar{I}_{t-2}\}, \\ &\ldots\end{aligned}$$

we get

$$\begin{aligned}&\mathbb{E}\|x^t - \alpha_i^t\|^2 \leq \int_{\{i\in I_{t-1}\}} \|x^t - x^{t-1}\|^2 d\mathbb{P} + \frac{1}{\theta_1}\int_{\{i\in\bar{I}_{t-1}\}} \|x^t - x^{t-1}\|^2 d\mathbb{P} \\ &+ \frac{1}{1-\theta_1}\int_{\{i\in\bar{I}_{t-1}\cap I_{t-2}\}} \|x^{t-1} - x^{t-2}\|^2 d\mathbb{P} + \frac{1}{\theta_2}\int_{\{i\in\bar{I}_{t-1}\cap\bar{I}_{t-2}\}} \|x^{t-1} - x^{t-2}\|^2 d\mathbb{P} \\ &+ \ldots + \frac{1}{1-\theta_1-\ldots-\theta_{t-2}}\int_{\{i\in\bar{I}_{t-1}\cap\ldots\cap I_1\}} \|x^2 - x^1\|^2 d\mathbb{P} \\ &+ \frac{1}{\theta_{t-1}}\int_{\{i\in\bar{I}_{t-1}\cap\ldots\cap\bar{I}_1\}} \|x^2 - x^1\|^2 d\mathbb{P} \\ &+ \frac{1}{1-\theta_1-\theta_2-\ldots-\theta_{t-1}}\int_{\{i\in\bar{I}_{t-1}\cap\ldots\cap\bar{I}_1\}} \|x^1 - x^0\|^2 d\mathbb{P}.\end{aligned}$$

By observing that $1 \leq \frac{1}{\theta_1}, \quad \frac{1}{1-\theta_1} \leq \frac{1}{\theta_2}, \quad \ldots, \quad \frac{1}{1-\theta_1-\ldots-\theta_{t-2}} \leq \frac{1}{\theta_{t-1}}, \frac{1}{1-\theta_1-\ldots-\theta_{t-1}} \leq \frac{1}{\theta_t}$, we have

$$\begin{aligned}\mathbb{E}\|x^t - \alpha_i^t\|^2 \leq & \frac{1}{\theta_1}\int_{\Omega} \|x^t - x^{t-1}\|^2 d\mathbb{P} + \frac{1}{\theta_2}\int_{\{i\in \bar{I}_{t-1}\}} \|x^{t-1} - x^{t-2}\|^2 d\mathbb{P} \\ & + \ldots + \frac{1}{\theta_{t-1}}\int_{\{i\in \bar{I}_{t-1}\cap\ldots\cap\bar{I}_2\}} \|x^2 - x^1\|^2 d\mathbb{P} + \frac{1}{\theta_t}\int_{\{i\in \bar{I}_{t-1}\cap\ldots\cap\bar{I}_1\}} \|x^1 - x^0\|^2 d\mathbb{P}.\end{aligned}$$

By the independence of each pair $\|x^{t-1} - x^{t-2}\|^2$ and $\{i \in \bar{I}_{t-1}\}$,..., $\|x^2 - x^1\|^2$ and $\{i \in \bar{I}_{t-1} \cap \ldots \cap \bar{I}_2\}$, $\|x^1 - x^0\|^2$ and $\{i \in \bar{I}_{t-1} \cap \ldots \cap \bar{I}_1\}$,

$$\begin{aligned}\mathbb{E}\|x^t - \alpha_i^t\|^2 \leq & \frac{1}{\theta_1}\mathbb{E}\|x^t - x^{t-1}\|^2 + \frac{1}{\theta_2}\mathbb{E}\|x^{t-1} - x^{t-2}\|^2 \mathbb{P}(\{i \in \bar{I}_{t-1}\}) + \ldots \\ & + \frac{1}{\theta_{t-1}}\mathbb{E}\|x^2 - x^1\|^2 \mathbb{P}(\{i \in \bar{I}_{t-1} \cap \ldots \cap \bar{I}_2\}) + \frac{1}{\theta_t}\mathbb{E}\|x^1 - x^0\|^2 \mathbb{P}(\{i \in \bar{I}_{t-1} \cap \ldots \cap \bar{I}_1\}) \\ & = \frac{1}{\theta_1}\mathbb{E}\|x^t - x^{t-1}\|^2 + \frac{1}{\theta_2}\mathbb{E}\|x^{t-1} - x^{t-2}\|^2 \left(\frac{N-1}{N}\right)^b + \ldots \\ & + \frac{1}{\theta_{t-1}}\mathbb{E}\|x^2 - x^1\|^2 \left(\frac{N-1}{N}\right)^{b(t-2)} + \frac{1}{\theta_t}\mathbb{E}\|x^1 - x^0\|^2 \left(\frac{N-1}{N}\right)^{b(t-1)}.\end{aligned}$$

For short, let us denote $\lambda = \left(\frac{N-1}{N}\right)^b < 1$. The inequality can be written as follows

$$\mathbb{E}\|x^t - \alpha_i^t\|^2 \leq \sum_{k=1}^{t} \frac{\lambda^{t-k}}{\theta_{t-k+1}}\mathbb{E}\|x^k - x^{k-1}\|^2.$$

Therefore,

$$\begin{aligned}\sum_{t=1}^{\infty}\mathbb{E}\|x^t - \alpha_i^t\|^2 & \leq \sum_{t=1}^{\infty}\sum_{k=1}^{t} \frac{1}{\theta_{t-k+1}}\lambda^{t-k}\mathbb{E}\|x^k - x^{k-1}\|^2 \\ & = \sum_{k=1}^{\infty}\sum_{t=k}^{\infty} \frac{1}{\theta_{t-k+1}}\lambda^{t-k}\mathbb{E}\|x^k - x^{k-1}\|^2 = \sum_{k=1}^{\infty}\mathbb{E}\|x^k - x^{k-1}\|^2 \sum_{t=k}^{\infty}\frac{\lambda^{t-k}}{\theta_{t-k+1}} \\ & = \sum_{k=1}^{\infty}\mathbb{E}\|x^k - x^{k-1}\|^2 \sum_{t=0}^{\infty}\frac{\lambda^{t}}{\theta_{t+1}} = A\left(\sum_{t=0}^{\infty}\frac{\lambda^t}{\epsilon_{t+1}}\right)\sum_{k=1}^{\infty}\mathbb{E}\|x^k - x^{k-1}\|^2 \\ & = A\left(\sum_{t=0}^{\infty}(t+1)^2\lambda^t\right)\sum_{k=1}^{\infty}\mathbb{E}\|x^k - x^{k-1}\|^2 < +\infty,\end{aligned}$$

since the power series $u(x) = \sum_{t=0}^{\infty}(t+1)^2\lambda^t$ has the convergence radius 1.

3. As a consequence of properties 1,2 above, we derive that $\{F(x^t)\}$ converges to V^{∞} almost surely.

4. If follows from (13) and property *2* that $\mathbb{E}\left(\sum_{k=1}^{\infty}\|v^k - \nabla H(x^k)\|^2\right) < +\infty$. Let

$$S = \left(v^k - \nabla H(x^k) \to 0, x^k - x^{k+1} \to 0, \{y^k\} \text{ is bounded}\right).$$

It is evident that $\mathbb{P}(S) = 1$. We consider a fixed event in S to obtain a realization $\{x^k\}$. Let x^* be a limit point of $\{x^k\}$, there exists a subsequence $\{x^{l_k}\}$ such that $x^{l_k} \to x^*$, implying $x^{l_k+1} \to x^*$. Since H is L-smooth, we obtain $\nabla H(x^{l_k}) \to \nabla H(x^*)$, which implies, $v^{l_k} \to \nabla H(x^*)$. Since $\{y^{l_k}\}$ is bounded, by passing to a subsequence if necessary, we assume that $y^{l_k} \to y^*$, therefore, $y^* \in \partial r_2(x^*)$. On the other hand, we have $z^{l_k} \in \partial(G + r_1)(x^{l_k+1})$. By the closedness of the graph subdifferential operator, we obtain $\nabla H(x^*)+y^* \in \partial(G+r_1)(x^*)$. Therefore, x^* is a critical point of $F = (G + r_1) - (H + r_2)$.

5. Since $z^t \in \partial(G + r_1)(x^{t+1})$,

$$\begin{aligned}
&\operatorname{dist}\left(\nabla H(x^{t+1}) + \nabla r_2(x^{t+1}), \partial(G + r_1)(x^{t+1})\right) \le \|\nabla H(x^{t+1}) + \nabla r_2(x^{t+1}) - v^t - y^t\| \\
&\le \|\nabla H(x^{t+1}) - \nabla H(x^t)\| + \|\nabla H(x^t) - v^t\| + \|\nabla r_2(x^{t+1}) - \nabla r_2(x^t)\| \\
&\le (L + L_{r_2})\|x^{t+1} - x^t\| + \|\nabla H(x^t) - v^t\|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\operatorname{dist}\left(\nabla H(x^{t+1}) + \nabla r_2(x^{t+1}), \partial(G + r_1)(x^{t+1})\right) \\
&\le (L + L_{r_2})\mathbb{E}\|x^{t+1} - x^t\| + \left(\mathbb{E}\|\nabla H(x^t) - v^t\|^2\right)^{\frac{1}{2}} \\
&\le (L + L_{r_2})\left(\mathbb{E}\|x^{t+1} - x^t\|^2\right)^{\frac{1}{2}} + \frac{L}{\sqrt{bN}}\left(\sum_{i=1}^{N}\mathbb{E}\|x^t - \alpha_i^t\|^2\right)^{\frac{1}{2}}.
\end{aligned}$$

Summing these inequalities for $t = 0$ to $T - 1$,

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\operatorname{dist}(\nabla H(x^{t+1}) + \nabla r_2(x^{t+1}), \partial(G + r_1)(x^{t+1})) \\
&\le (L + L_{r_2})\sum_{t=0}^{T-1}\left(\mathbb{E}\|x^{t+1} - x^t\|^2\right)^{\frac{1}{2}} + \frac{L}{\sqrt{bN}}\sum_{t=0}^{T-1}\left(\sum_{i=1}^{N}\mathbb{E}\|x^t - \alpha_i^t\|^2\right)^{\frac{1}{2}} \\
&\le (L + L_{r_2})\sqrt{T}\left(\sum_{t=0}^{T-1}\mathbb{E}\|x^{t+1} - x^t\|^2\right)^{\frac{1}{2}} + \frac{L}{\sqrt{bN}}\sqrt{T}\left(\sum_{t=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\|x^t - \alpha_i^t\|^2\right)^{\frac{1}{2}} \\
&\le (L + L_{r_2})\sqrt{T}C + \frac{L\sqrt{T}C}{\sqrt{b}} \le (2L + L_{r_2})\,C\sqrt{T}, \qquad \text{for some } C > 0.
\end{aligned}$$

It follows that the iteration convergence rate is $d_T = \mathcal{O}(1/\sqrt{T})$.
If the output x_a is chosen randomly from $\{x^t\}_{t=1}^T$,

$$\mathbb{E}\operatorname{dist}(\nabla H(x_a) + \nabla r_2(x_a), \partial(G + r_1)(x_a)) = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

Therefore, to get ϵ-criticality, the algorithm needs $T = \mathcal{O}(1/\epsilon^2)$ iterations. From the condition for b, if we choose

$$b = \left\lceil \frac{\sqrt[4]{2}}{\sqrt{\mu}}N^{3/4} \right\rceil, \text{ where } \mu = \frac{\rho_H + \rho_{r_2} + \rho_{G+r_1}}{4L},$$

one obtains the complexity of $\mathcal{O}(N+N^{3/4}/\epsilon^2)$ in terms of the number of gradient evaluations and the complexity of $\mathcal{O}(1/\epsilon^2)$ in terms of the number of convex subproblems.

Sketched proof for the option of sample without replacement.

1. By following the arguments of the proof above, we obtain the inequality (12). Then, we have the following lemma whose proof is similar to the proof of Lemma 3.

Lemma 6 *The following inequality holds,*

$$\mathbb{E}_{I_t}\|v^t - \nabla H(x^t)\|^2 \leq \frac{L^2}{bN}\left(1 - \frac{b-1}{N-1}\right)\sum_{i=1}^{N}\|x^t - \alpha_i^t\|^2.$$

Similar to the arguments above, by introducing the Lyapunov sequence $V^t = F(x^t) + c_t\sum_{i=1}^N \|x^t - \alpha_i^t\|^2$ with $c_t = \frac{\beta L^2(N-b)}{2\gamma b(N-1)(b-N+\beta b)}$, where $\beta = N/b$, and γ is determined by the AM-GM inequality as above, we want

$$\frac{N\sqrt{N^2-b^2}}{b^2\sqrt{N-1}} \leq \frac{\rho}{4L}. \tag{18}$$

Consequently, if $\frac{N\sqrt{N+1}}{b^2} \leq \dfrac{\rho}{4L}$, (18) holds and we obtain that $\{V^t\}$ converges almost surely and $\sum_{t=0}^{\infty}\mathbb{E}\|x^{t+1} - x^t\|^2 < \infty$.

2. The arguments above can be employed with the following notice: $\mathbb{P}(\{i \in \bar{I}_{t-1}\}) = 1 - b/N, \mathbb{P}(\{i \in \bar{I}_{t-1} \cap \bar{I}_{t-2}\}) = (1-b/N)^2$, etc.

3, 4, 5. These properties are established similar to the option of sample with replacement.

■

Remark 7 *i. We can apply the same technique discussed in Remark 4 to improve the number $\frac{\rho}{L}$ of the problem if necessary.*

ii. The analysis giving links between $\mathbb{E}\|x^t - \alpha_i^t\|^2$ and $\sum_{t=1}^{\infty}\mathbb{E}\|x^{t+1} - x^t\|^2$ is quite novel in the literature, which is expected to be helpful in analyzing SAGA-type algorithms. The analysis provides an elegant idea of how to escape the "non-independence" of the function and the region of an integral.

Remark 8 *Some comparisons with related algorithms:*

i. From the convergence of the standard DCA (Pham Dinh and Le Thi, 1997), one can verify that to find an ϵ-criticality, under the L-smooth assumption of H and r_2, it requires $\mathcal{O}(1/\epsilon^2)$ iterations. As a result, the standard DCA has the complexity of $\mathcal{O}(N/\epsilon^2)$ in terms of gradient evaluations and $\mathcal{O}(1/\epsilon^2)$ in terms of the number of convex subproblems. While the proposed DCA-SVRG and DCA-SAGA enjoy the same complexity as the standard DCA in terms of the number of convex subproblems, they significantly outperform the standard DCA in terms of gradient evaluations, DCA-SVRG: $\mathcal{O}(N^{2/3}/\epsilon^2)$, DCA-SAGA: $\mathcal{O}(N+N^{3/4}/\epsilon^2)$.

ii. (Xu et al., 2019b) considered the following large-sum problem

$$\min_{x\in\mathbb{R}^n} F(x) = G(x) - H(x) + r(x) := \frac{1}{\bar{N}}\sum_{i=1}^{\bar{N}} g_i(x) - \frac{1}{N}\sum_{j=1}^{N} h_j(x) + r(x),$$

where g_i, h_j are convex and the regularizer r is convex. The authors also analyze the case that r is nonconvex (not necessary to be DC) where the proximal mapping can be efficiently computed. For simplicity, we only discuss the case that r is convex (note that their complexity for the case of nonconvex r is obviously worse than the convex case). When each h_j is L-smooth, the best gradient complexity to find (nearly) ϵ-criticality is given in (Xu et al., 2019b, Corollary 9), which is $\mathcal{O}((\bar{N}+N)/\epsilon^2)$. Of course, this complexity takes into account the cost of solving each large-sum convex subproblem by the SVRG. To compare with our algorithms, we need to cast this complexity into the number of gradient evaluations of H and the number of convex subproblems being solved. To obtain a (nearly) ϵ-critical point, they need $\mathcal{O}(1/\epsilon^2)$ stages in total (Xu et al., 2019b, Lemma 3 and Corollary 9). Each stage consumes a full gradient of H and requires solving inexactly a convex subproblem, resulting in the complexity of $\mathcal{O}(N/\epsilon^2)$ in terms of gradient evaluations and $\mathcal{O}(1/\epsilon^2)$ in terms of the number of convex subproblems. Our algorithms enjoy better gradient complexity, while the complexity in terms of convex subproblems is somewhat incomparable since their algorithmic design also takes into account the inexactness of solving each such problem, while we do not quantify this issue here.

4. Additional Convergence Analysis for The Composite Structure of r_2

In this section, we consider $r_2(x) = \sum_{i=1}^m l_i(p_i(x))$ where $l_i : \mathbb{R} \to \mathbb{R}$ is convex and decreasing, $p_i : \mathbb{R}^n \to \mathbb{R}$ is convex and hence continuous, for $i = \overline{1,m}$. The optimization problem (P) becomes

$$(Q) \quad \alpha = \min \left\{ F(x) = G(x) - \frac{1}{N}\sum_{i=1}^{N} h_i(x) + r_1(x) - \sum_{i=1}^{m} l_i(p_i(x)) \right\}.$$

By denoting $\Omega = \{(x,z) \in \mathbb{R}^n \times \mathbb{R}^m : p_i(x) \le z_i, \forall i = \overline{1,m}\}$, we can solve the following problem instead

$$(Q') \quad \alpha = \min \left\{ \varphi(x,z) = G(x) - \frac{1}{N}\sum_{i=1}^{N} h_i(x) + r_1(x) + \chi_\Omega(x,z) - \sum_{i=1}^{m} l_i(z_i) \right\},$$

since if (x^*, z^*) is the optimal solution of (Q'), x^* is the optimal solution of (Q). We observe that the problem (Q') with respect to the optimization variables (x,z) takes the form of (P). Therefore, we can apply the proposed DCA-SVRG and DCA-SAGA to (Q'). However, the convergence analysis of DCA-SVRG and DCA-SAGA should be modified for this case since x and z do not play the same role. We see that, due to the newly introduced optimization variable z, the modulus of convexity of G and H with respect to (x,z) is 0, which is undesirable. By Remark 4, we can add the convex term $\frac{\gamma}{2}\|(x,z)\|^2$ to both DC components. However, this technique has its own risk of resulting in bad approximations. Therefore, we still want to use the modulus of strong convexity of G and H with respect to x only in our analysis, which requires us to adapt the convergence analysis of DCA-SVRG and DCA-SAGA to this case. For convenience of presentation, we give DCA-SVRG scheme applied to (Q') and state the convergence results with its sketched proof. For DCA-SAGA applied to (Q'), we only present convergence results.

The DCA-SVRG scheme applied to (Q') is given in Algorithm 3, where we denote $r_3(z) = \sum_{i=1}^m l_i(z_i)$.

Algorithm 3 DCA-SVRG applied to (Q')

```
**Initialization:** $\tilde{x}^0 \in \operatorname{dom} r_1$, inner-loop length $M$, minibatch size $b$, $k = 0$, option (either *with replacement* or *without replacement*).
**repeat**
    Compute the full gradient $\tilde{\nu}^k = \frac{1}{N}\sum_{i=1}^N \nabla h_i(\tilde{x}^k)$ and set $x_0^{k+1} = \tilde{x}^k$, $z_0^{k+1} = (z_{0,1}^{k+1}, z_{0,2}^{k+1}, \ldots, z_{0,m}^{k+1})^\top$ with $z_{0,r}^{k+1} = p_r(x_0^{k+1})$.
    **for** $j = 0 : M-1$ **do**
        **if** option is *with replacement* **then**
            Randomly choose with replacement the set $I_t$ of $b$ elements of $[N]$.
        **else**
            Randomly choose without replacement the set $I_t$ of $b$ elements of $[N]$.
        **end if**
        Compute the "stochastic variance reduced gradient" $t_j^{k+1}$ by
            $$t_j^{k+1} = \frac{1}{b}\sum_{i \in I_b} \nabla h_i(x_j^{k+1}) + \tilde{\nu}^k - \frac{1}{b}\sum_{i \in I_b} \nabla h_i(\tilde{x}^k).$$
        Compute $y_j^{k+1} \in \partial r_3(z_j^{k+1})$.
        Solve the convex problem
            $$(x_{j+1}^{k+1}, z_{j+1}^{k+1}) = \arg\min\{G(x) + r_1(x) + \chi_\Omega(x,z) - \langle t_j^{k+1}, x\rangle - \langle y_j^{k+1}, z\rangle\}$$
    **end for**
    Set $\tilde{x}^{k+1} = x_M^{k+1}$, and $k = k+1$.
**until** Stopping criterion.
```

Henceforth, we still denote ρ_H, ρ_{G+r_1} the modulus of strong convexity of H and $G + r_1$ and L the Lipschitz constant of ∇h_i with respect to x only. We derive the following convergence results.

Theorem 9 *If the minibatch size b and the inner-loop length M satisfy*

$$\frac{M}{\sqrt{b}} \leq \frac{1}{4\sqrt{e-1}} \frac{\rho_{G+r_1} + \rho_H}{L},$$

then

1. The sequence $\{F(\tilde{x}^k)\}$ converges almost surely.

2. $\sum_{k=0}^\infty \sum_{j=0}^{M-1} \mathbb{E}\|x_{j+1}^{k+1} - x_j^{k+1}\|^2 < +\infty$.

3. Suppose the sequence $\{y_j^k\}$ is bounded almost surely, let x^* be a limit point of $\{x_j^k\}$, then (x^*, z^*) is a critical point of $F = (G + r_1 + \chi_\Omega) - (H + r_3)$ almost surely, where $z^* = (p_1(x^*), p_2(x^*), \ldots, p_m(x^*))^\top$.

Proof [Proof of Theorem 9] 1, 2. The convex subproblem can be solved as follows

$$x_{j+1}^{k+1} = \arg\min \left\{ G(x) + r_1(x) - \langle t_j^{k+1}, x \rangle - \sum_{r=1}^{m} y_{j,r}^{k+1} p_r(x) \right\} \tag{19}$$

$$z_{j+1,r}^{k+1} = p_r(x_{j+1}^{k+1}), \quad \forall r = \overline{1,m}. \tag{20}$$

It follows from (19) that $t_j^{k+1} \in \partial \left(G + r_1 - \sum_{r=1}^{m} y_{j,r}^{k+1} p_r \right) (x_{j+1}^{k+1})$, which implies

$$\begin{aligned} G(x_{j+1}^{k+1}) + r_1(x_{j+1}^{k+1}) \leq & G(x_j^{k+1}) + r_1(x_j^{k+1}) + \langle t_j^{k+1}, x_{j+1}^{k+1} - x_j^{k+1} \rangle \\ & - \sum_{r=1}^{m} y_{j,r}^{k+1} (p_r(x_j^{k+1}) - p_r(x_{j+1}^{k+1})) - \frac{\rho_{G+r_1}}{2} \| x_{j+1}^{k+1} - x_j^{k+1} \|^2. \end{aligned} \tag{21}$$

From the convexity of H and r_3, we obtain

$$H(x_{j+1}^{k+1}) \geq H(x_j^{k+1}) + \langle \nabla H(x_j^{k+1}), x_{j+1}^{k+1} - x_j^{k+1} \rangle + \frac{\rho_H}{2} \| x_{j+1}^{k+1} - x_j^{k+1} \|^2, \tag{22}$$

$$r_3(z_{j+1}^{k+1}) \geq r_3(z_j^{k+1}) + \langle y_j^{k+1}, z_{j+1}^{k+1} - z_j^{k+1} \rangle. \tag{23}$$

From (21), (22) and (23),

$$F(x_{j+1}^{k+1}) \leq F(x_j^{k+1}) + \langle t_j^{k+1} - \nabla H(x_j^{k+1}), x_{j+1}^{k+1} - x_j^{k} \rangle - \frac{\rho_{G+r_1} + \rho_H}{2} \| x_{j+1}^{k+1} - x_j^{k+1} \|^2,$$

by noting that $\varphi(x_{j+1}^{k+1}, z_{j+1}^{k+1}) = F(x_{j+1}^{k+1}), \varphi(x_j^{k+1}, z_j^{k+1}) = F(x_j^{k+1})$.

Here, by considering the Lyapunov sequence $V_j^{k+1} = F(x_j^{k+1}) + c_j \| x_j^{k+1} - \tilde{x}^k \|^2$ similar to the proof of Theorem 2, we arrive at the conclusion 1 and 2.

3. Let $S = \left(t_j^{k+1} - \nabla H(x_j^{k+1}) \to 0, x_{j+1}^{k+1} - x_j^{k+1} \to 0, \{y_j^k\} \text{ is bounded} \right)$. We see that $\mathbb{P}(S) = 1$ and let $\{x_j^k\}$ be a realization with respect to a fixed event in S. Let x^* be a limit point of $\{x_j^k\}$, there exists a subsequent $x_{v(l)}^{u(l)} \to x^*$, which further implies $x_{v(l)+1}^{u(l)} \to x^*$. From (20) and the continuity of p_r, we obtain $z_{v(l)}^{u(l)} \to (p_1(x^*), \ldots, p_m(x^*))^\top$. Without loss of generality, we assume that $y_{v(l)}^{u(l)} \to y^*$, which yields $y^* \in \partial r_3((p_1(x^*), \ldots, p_m(x^*))^\top)$. On the other hand, $t_{v(l)}^{u(l)} \to \nabla H(x^*)$. It follows from

$$t_{v(l)}^{u(l)} \in \partial \left(G + r_1 - \sum_{r=1}^{m} y_{v(l),r}^{u(l)} p_r \right) (x_{v(l)+1}^{u(l)})$$

that $\nabla H(x^*) \in \partial (G + r_1 - \sum_{r=1}^{m} y_r^* p_r)(x^*)$. Therefore,

$$\begin{pmatrix} \nabla H(x^*) \\ y^* \end{pmatrix} \in \partial (G(x) + r_1(x) + \chi_\Omega(x,z))(x^*, z^*),$$

which implies $\partial (G(x) + r_1(x) + \chi_\Omega(x,z))(x^*, z^*) \cap \partial (H(x) + r_3(z))(x^*, z^*) \neq \emptyset$. ■

Similarly, we have the following convergence results when applying DCA-SAGA to minimize the function $\varphi(x, z)$.

Theorem 10 *If the minibatch size b satisfies $\frac{N\sqrt{N+b}}{b^2} \le \frac{\rho_H+\rho_{G+r_1}}{4L}$ for the option of sample with replacement or $\frac{N\sqrt{N+1}}{b^2} \le \frac{\rho_H+\rho_{G+r_1}}{4L}$ for the option of sample without replacement, then*

1. $\sum_{t=0}^{\infty} \mathbb{E}\|x^{t+1} - x^t\|^2 < \infty$.
2. $\sum_{t=0}^{\infty} \sum_{i=1}^{N} \mathbb{E}\|x^t - \alpha_i^t\|^2 < \infty$.
3. *The sequence $\{F(x^t)\}$ converges almost surely.*
4. *Suppose the sequence $\{y^t\}$ is bounded almost surely, let x^* be a limit point of $\{x^t\}$, then (x^*, z^*) is a critical point of $F = (G + r_1 + \chi_\Omega) - (H + r_3)$ almost surely, where $z^* = (p_1(x^*), p_2(x^*), \dots, p_m(x^*))^\top$.*

5. Applications to Nonnegative Principal Component Analysis, Group Variables Selection in Multi-class Logistic Regression, and Sparse Linear Regression

In this section, to study our proposed algorithms' practical behaviors, we apply DCA-SVRG and DCA-SAGA to three important problems in machine learning: nonnegative principal component analysis, group variables selection in multi-class logistic regression, and sparse linear regression.

All numerical experiments in this section are performed on a Processor Intel(R) core(TM) i7-8700, CPU @ 3.20GHz, RAM 16 GB.

5.1 Nonnegative Principal Component Analysis

Principal component analysis is arguably one of the most successful tools for dimensionality reduction. Throughout its history, PCA has numerous success stories in a large spectrum of applied sciences, including neuroscience, medicine, psychology, material science, scientometry, astronomy, geography, and social sciences.

We perform numerical experiments on a special structure of PCA called Non-negative component analysis (NN-PCA). The term "non-negative" indicates that we restrict the search domain to be the non-negative part of the whole search space. This restriction implicitly assumes that we know some additional information on a solution of PCA: the solution belongs to the non-negative orthant. This additional information offers some theoretical advantages as discussed in (Montanari and Richard, 2015). The NN-PCA has the following form

$$\min_{\|x\|\le 1, x\ge 0} -\frac{1}{2N}\sum_{i=1}^{N} \langle x, z_i\rangle^2, \tag{24}$$

where $\{z_i\}_{i=1}^N$ is the given data set.

5.1.1 DCA-SVRG and DCA-SAGA Applied to (24)

We apply DCA-SVRG and DCA-SAGA to (24) with the following DC decomposition $G(x) = \frac{\rho}{2}\|x\|^2$, $h_i(x) = \frac{\rho}{2}\|x\|^2 + \frac{1}{2}\langle x, z_i\rangle^2, r_1(x) = \chi_S(x), r_2(x) = 0$, where $S = \{x \in \mathbb{R}^n : x \ge 0, \|x\| \le 1\}$, and $\rho > 0$. In our experiments, we fix $\rho = 1$.

With this setting, we implement two versions of DCA-SAGA and two versions of DCA-SVRG, namely, DCA-SAGA-v1 (sample with replacement) and DCA-SAGA-v2 (sample without replacement), DCA-SVRG-v1 (sample with replacement) and DCA-SVRG-v2 (sample without replacement), respectively.

5.1.2 Numerical Experiments

Data sets. We use standard machine learning data sets in LIBSVM [1], namely, `a9a` (32561×123), `aloi` (108000×128), `cifar10` (50000×3072), `SensIT Vehicle` (78823×100), `connect-4` (67557×126), `letter` (15000×16), `mnist` (60000×780), `protein` (17766×357), `shuttle` (43500×9), `YearPredictionMSD` (463715×90). Each sample of these data sets is normalized as $\|z_i\| = 1$.

Comparative algorithms. We implement stochastic DCA (denoted by SDCA) (Le Thi et al., 2020). For a more in-depth comparison, we further implement prox-SGD and prox-SAGA (J. Reddi et al., 2016), prox-SARAH (Pham et al., 2020), prox-SpiderBoost (Wang et al., 2019). Here we do not compare with the prox-SVRG since - as discussed earlier - the DCA-SVRG-v1 applied to the above setting recovers a version of the prox-SVRG.

For proximal-type algorithms (prox-SGD, prox-SAGA, prox-SARAH, prox-SpiderBoost), the convex regularizer is set to be $\chi_S(x)$, while each L-smooth component of the large sum is set as $f_i(x) = -\frac{1}{2}\langle x, z_i\rangle^2$ (f_i is 1-smooth).

Experiment setups and results. We study the evolution process of the objective with respect to the computational resource used. This resource is measured by the number of stochastic first-order oracle calls (SFO). That is, every time we access one stochastic gradient (the gradient of one summand) of the large sum, one SFO call is counted. We set the fixed budget of SFO calls to be $15N$. However, we allow all algorithms to use some extra SFO calls to complete their last iteration.

We observe that each function $x \mapsto \frac{1}{2}\langle x, z_i\rangle^2$ is 1-smooth. Based on theoretical convergence results, we choose hyperparameters for the algorithms as follows. The minibatch size b is chosen as $\lfloor N^{\frac{2}{3}} \rfloor$, $\lfloor N^{\frac{2}{3}} \rfloor$, $\lceil 2^{\frac{5}{4}} N^{\frac{3}{4}} \rceil$, $\lceil 2\sqrt{N\sqrt{N+1}} \rceil$ for DCA-SVRG-v1, DCA-SVRG-v2, DCA-SAGA-v1, DCA-SAGA-v2, respectively. We set the inner loop length M for DCA-SVRG-v1 and DCA-SVRG-v2 to be $\lfloor \frac{1}{4\sqrt{e-1}}\sqrt{b} \rfloor$. On the other hand, we use a neutral minibatch size of $10\%N$ which usually yields good results (Le Thi et al., 2020) for the SDCA. For the prox-SGD, $\eta = 1/(2L)$ (Ghadimi et al., 2016) and we choose a neutral minibatch size of 500. For the prox-SAGA, we set $\eta = 1/(5L)$ and the minibatch size $b = \lfloor N^{\frac{2}{3}} \rfloor$ (J. Reddi et al., 2016). The hyperparamters of prox-SARAH are chosen as follows (Pham et al., 2020, Theorem 8): the minibatch size $b = \lfloor \sqrt{n} \rfloor$, the inner loop length $M = \lfloor N/b \rfloor$, the step size η and the averaging parameter γ:

$$\eta = \frac{2\sqrt{\omega M}}{4\sqrt{\omega M}+1}, \gamma = \frac{1}{L\sqrt{\omega M}}, \text{ where } \omega := \frac{3(N-b)}{2b(N-1)}.$$

Lastly, for the prox-SpiderBoost, we choose $\eta = 1/(2L)$ and the inner loop length and the mini-batch size $M = b = \lfloor \sqrt{N} \rfloor$ (Wang et al., 2019, Theorem 1).

[1] The data sets can be downloaded from `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

We report the suboptimality of the objective, $f(x) - f(\hat{x})$, where $\hat{x}$ is the “optimal” solution found by running the deterministic DCA 500 iterations (which means that this deterministic DCA uses $500N$ SFO calls) 10 times with different initial points. The final optimal value $f(\hat{x})$ is the smallest value found by the deterministic DCA and all the competitive algorithms in this experiment. To enhance the visibility, we plot this suboptimality in the logarithmic scale of base 10. Figure 1 shows the results over 10 random runs of all comparative algorithms. The bold lines represent the mean curves, while the shaded regions illustrate the mean absolute deviation (MAD). Note that, in the normal scale, one might plot $x \pm \mathrm{MAD}(x)$; Meanwhile, in the log scale, a reasonable way is to plot $\log_{10}(x) \pm (1/\ln(10))\,\mathrm{MAD}(x)/x \approx \log_{10}(x) \pm 0.434\,\mathrm{MAD}(x)/x$ (see, e.g., (Stuve, 2004)).

Comments. We observe from Figure 1 that two versions of DCA-SVRG and the prox-SARAH, prox-SpiderBoost perform very well and usually take the lead, where the suboptimality is around 10^{-15} in most of the cases. It is followed by the sample without replacement version of DCA-SAGA (DCA-SAGA-v2) who obtains a good suboptimality (usually less than 10^{-10}). While the performance of DCA-SVRG-v1 and DCA-SVRG-v2 are almost identical, DCA-SAGA-v1 is much worse than DCA-SAGA-v2, indicating that sample strategies are of decisive importance on the performance of DCA-SAGA. In this experiment, the prox-SAGA also obtains good results with suboptimality less than 10^{-5} in most data sets. The SDCA performs a little bit worse than the prox-SAGA, meanwhile, the prox-SGD fluctuates around 10^{-5} and can not further decrease the suboptimality.

5.2 Group Variables Selection in Multi-class Logistic Regression

Logistic regression is undoubtedly one of the most successful tools for classification. Logistic regression has a wide range of applications including cancer detection (Kim et al., 2008), social sciences (King and Zeng, 2001), medical (Bagley et al., 2001), etc. On the other hand, for high-dimensional data, feature selection is usually employed in order to select relevant features and encourage weights corresponding to irrelevant features to go to 0.

Let $\{(x_i, y_i) : i = 1, 2, \dots, N\}$ be a training set with feature vectors x_i and the labels $y_i \in \{1, 2, \dots, Q\}$ where Q is the number of classes. Let W be the $d \times Q$ matrix with columns $W_{:,1}, W_{:,2}, \dots, W_{:,Q}$ and $b = (b_1, b_2, \dots, b_Q)$. The conditional probability of class y given the observation $X = x$ can be modeled by

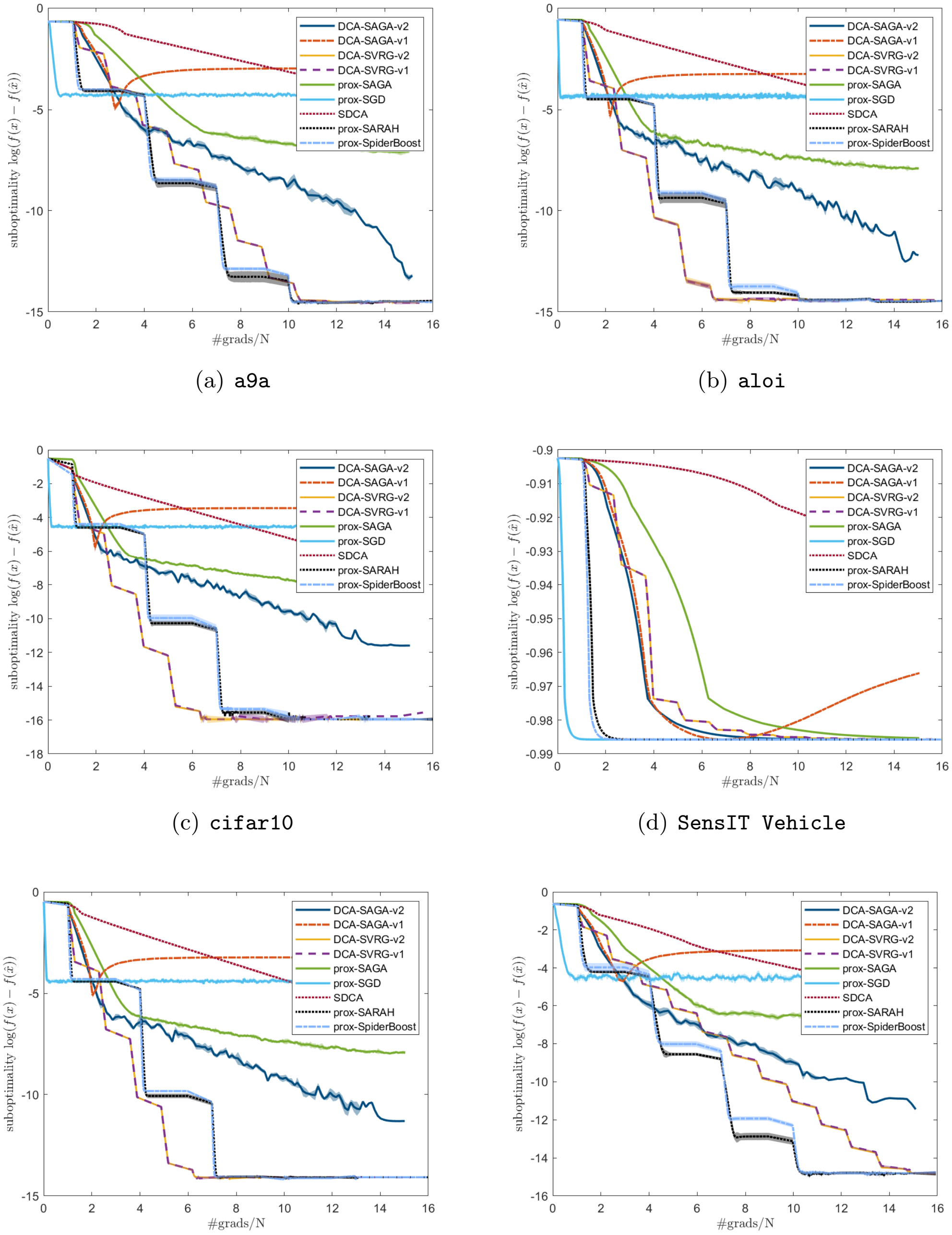
$$p(Y = y|X = x) = \frac{\exp\left(b_y + W_{:,y}^\top x\right)}{\sum_{k=1}^{Q} \exp\left(b_k + W_{:,k}^\top x\right)}.$$

To find (W, b), one minimizes the following negative log-likelihood function over the whole training set

$$\mathcal{L}(W, b) := \frac{1}{N} \sum_{i=1}^{N} \ell(x_i, y_i, W, b)$$

where $\ell(x_i, y_i, W, b) = -\log p(Y = y_i|X = x_i)$. On the other hand, to select relevant features, one employs the following $\ell_{q,0}$-norm of W

$$\|W\|_{q,0} = |\{j \in \{1, 2, \dots, d\} : \|W_{j,:}\|_q \neq 0\}|,$$



(a) `a9a`

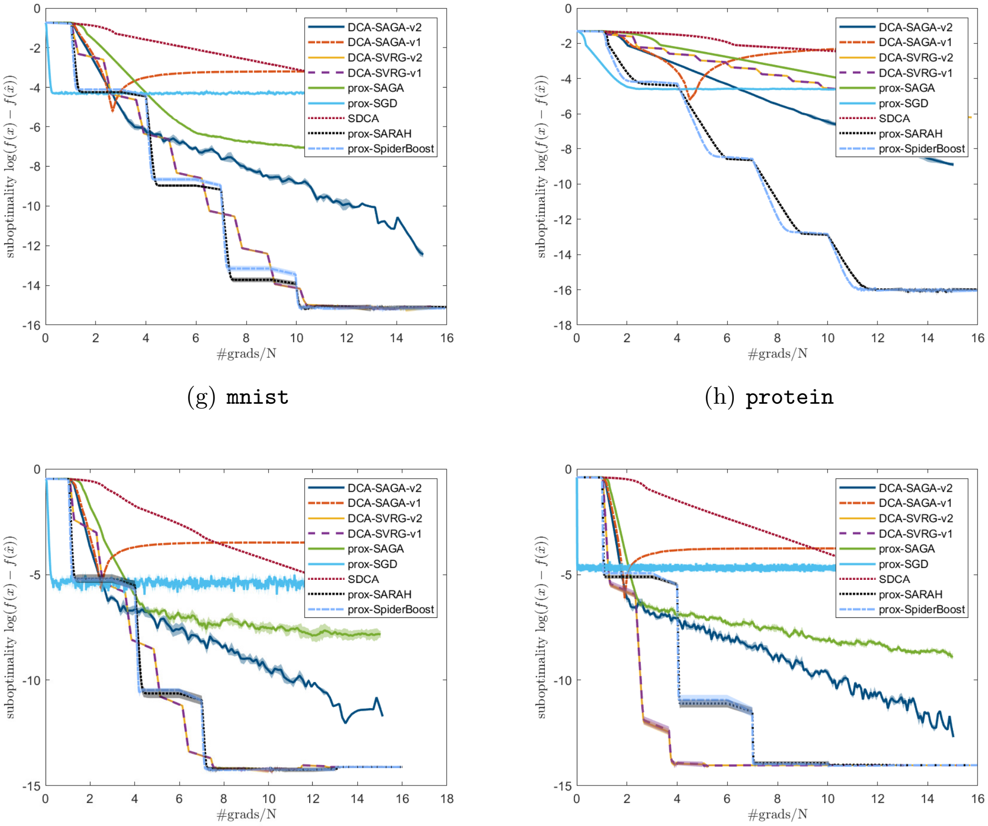
(b) `aloi`

(c) `cifar10`

(d) `SensIT Vehicle`

(e) `connect-4`

(f) `letter`



(g) `mnist`

(h) `protein`

(i) `shuttle`

(j) `YearPredictionMSD`

Figure 1: The suboptimality of all algorithms over 10 data sets

which leads to the following optimization problem

$$\min_{W,b}\left\{\frac{1}{N}\sum_{i=1}^{N}\ell(x_i,y_i,W,b)+\lambda\|W\|_{q,0}\right\}. \tag{25}$$

In practice, the discontinuous $\ell_{q,0}$-norm is usually approximated by continuous functions. In this paper, we approximate $\ell_{q,0}$ by the following nonconvex functions, which have been proven to be efficient in several problems including individual variable selection in SVM (Bradley and Mangasarian, 1998), sparse optimal scoring problem (Le Thi and Phan, 2016), sparse covariance matrix estimation problem (Phan et al., 2017),

$$\begin{aligned}\text{Exponential:}\quad & \eta_\alpha^{\exp}(s)=1-\exp(-\alpha s),\\ \text{Capped}-\ell_1:\quad & \eta_\alpha^{\text{cap}-\ell_1}(s)=\min\{1,\alpha s\}.\end{aligned}$$

The corresponding approximation problem of (25) takes the form

$$\min_{W,b}\left\{\frac{1}{N}\sum_{i=1}^{N}\ell(x_i,y_i,W,b)+\lambda\sum_{j=1}^{d}\eta_\alpha(\|W_{j,:}\|_q)\right\}. \tag{26}$$

5.2.1 DCA-SVRG and DCA-SAGA Applied to (26)

It is worth noting that an L-smooth function φ admits the following DC decomposition

$$\varphi(x)=\frac{\gamma}{2}\|x\|^2-\left(\frac{\gamma}{2}\|x\|^2-\varphi(x)\right) \tag{27}$$

whenever $\gamma\geq L$. We can show that $\ell(x_i,y_i,W,b)$ is $\frac{2\sqrt{2}}{3\sqrt{3}}\sqrt{Q}(\|x_i\|^2+1)$-smooth, so the DC decomposition (27) can be used. On the other hand, η_α is concave and increasing. Therefore, the problem (26) takes the form of (Q) with

$$\begin{aligned}&G(W,b)=\frac{\gamma}{2}\|(W,b)\|^2,\ h_i(W,b)=\frac{\gamma}{2}\|(W,b)\|^2-\ell(x_i,y_i,W,b),\\ &r_1(W,b)=0, l_i(s)=-\lambda\eta_\alpha(s),\ p_i(W,b)=\|W_{i,:}\|_q, \forall i=\overline{1,d},\end{aligned}$$

where $\gamma\geq\frac{2\sqrt{2}}{3\sqrt{3}}\sqrt{Q}(\max_{i=\overline{1,N}}\|x_i\|^2+1)$. We can apply DCA-SVRG and DCA-SAGA to solve (26). In general, the convex subproblem of DCA-SVRG and DCA-SAGA requires the computing of the proximal operator of $r\|\cdot\|_q$ with $q\in\{1,2\}$, where the proximal operator of a function f is defined by

$$\text{prox}_f(x)=\arg\min_y\left\{\frac{1}{2}\|x-y\|^2+f(y)\right\}.$$

It is worth noting that, these proximal operators can be computed explicitly as follows

$$\begin{aligned}&\text{prox}_{r\|\cdot\|_2}(x)=\begin{cases}\left(1-\dfrac{r}{\|x\|_2}\right)x & \text{if } \|x\|_2\geq r,\\ 0 & \text{otherwise}\end{cases}\\ &\text{prox}_{r\|\cdot\|_1}(x)=\max(|x|-r,0)\circ\text{sign}(x).\end{aligned}$$

Therefore, the corresponding DCA-SVRG and DCA-SAGA schemes applied to (26) is explicit, i.e., no convex solver is needed for solving convex subproblems.

In this experiment, we found that DCA-SAGA with replacement performs poorly and is not numerically stable. Hence, we only study the behaviors of three algorithms: DCA-SAGA without replacement, DCA-SVRG-v1 (with replacement), DCA-SVRG-v2 (without replacement).

5.2.2 Numerical Experiments

Data sets. We use standard machine learning data sets obtained from LIBSVM for multi-class classification, namely, `connect-4` (67557×126, 3 classes), `dna` (2000×180, 3 classes), `letter` (15000×16, 26 classes), `SensIT Vehicle` (78823×100, 3 classes), `Sensorless` (58509×48, 11 classes), `shuttle` (43500×9, 7 classes). All data sets are normalized by MinMaxScaler to scale all features into the range $[0, 1]$.

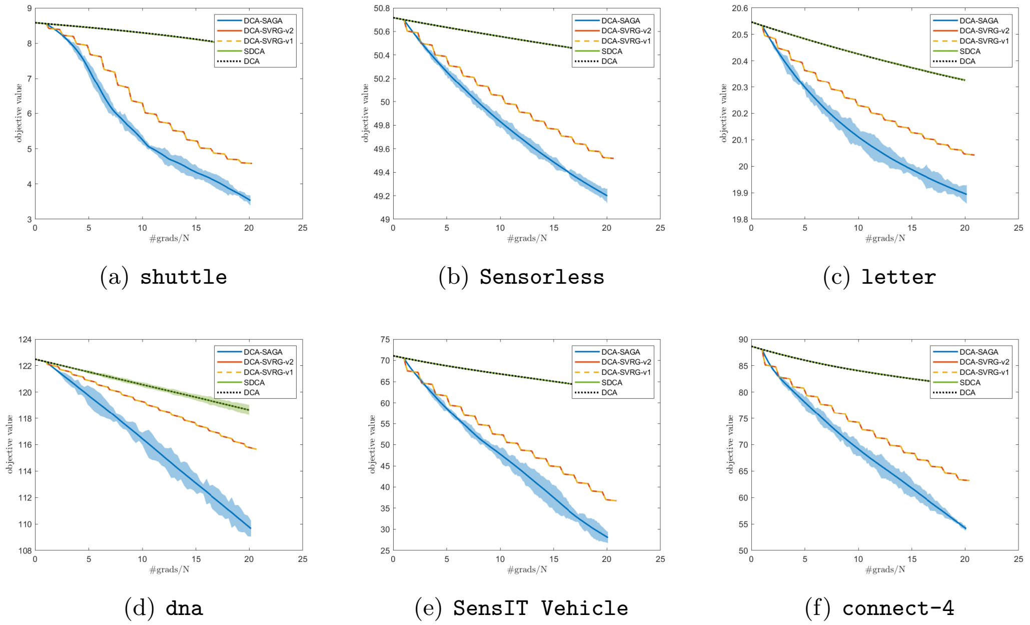
Comparative algorithms. The DCA and the SDCA are chosen as comparative algorithms in this experiment.

Experiment setups and results. We set the budget of SFO calls to be $20N$. Throughout this experiment, regularization parameter λ and the number α that controls the tightness of the $\ell_{q,0}$ approximation are fixed as 1 and 0.5, respectively. As mentioned, it can be shown that each function $\ell(x_i, y_i, W, b)$ is L-smooth, where $L = \frac{2\sqrt{2}}{3\sqrt{3}}\sqrt{Q}(\max_{i=\overline{1,N}} \|x_i\|^2 + 1)$. In order to guarantee the convexity of two DC components, we need to set $\gamma \geq L$. In our experiment, we fix $\gamma = 2L$.

By the convergence analysis above and the numerical experiments in (Le Thi et al., 2020), we choose hyperparameters for algorithms as follows. The minibatch size b is chosen as $\left\lceil 2\sqrt{N\sqrt{N+1}} \right\rceil$, $\lfloor N^{\frac{2}{3}} \rfloor$, $\lfloor N^{\frac{2}{3}} \rfloor$, $\lfloor 10\% N \rfloor$ for DCA-SAGA, DCA-SVRG-v1, DCA-SVRG-v2, SDCA, respectively. The inner loop length M of DCA-SVRG-v1 and DCA-SVRG-v2 is set to be $\lfloor \frac{1}{4\sqrt{e-1}}\sqrt{b} \rfloor$.

Four Figures 2, 3, 4, and 5 report the mean curves of the objective (averaged over 10 runs) and the mean absolute deviation of five algorithms. To enhance visibility, the MAD is scaled up by a factor of 20.

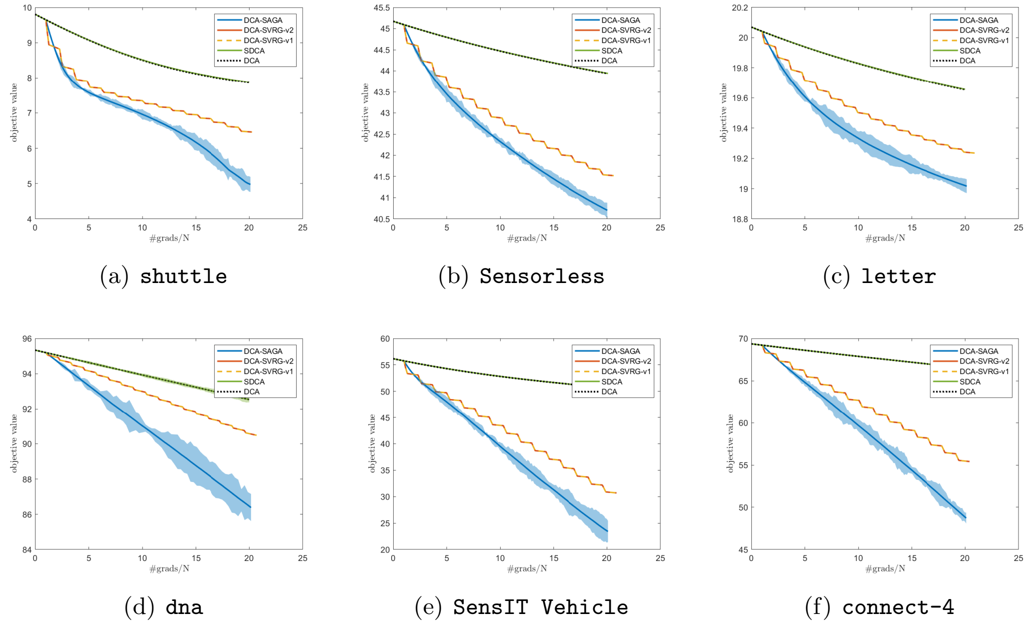
Comments. It can be observed that, in all cases, DCA-SAGA is the best algorithm in both solutions' quality and convergence rate criteria. It is followed by the two versions of DCA-SVRG to be the second-best, where the performances of these two versions are almost identical. These patterns occur consistently over all tested cases. Again, in this experiment, the sample strategies are of paramount importance in the performance of DCA-SAGA: while sample with replacement makes DCA-SAGA numerically unstable, sample without replacement makes DCA-SAGA the best algorithms among all competitors. We see that the SDCA is the worst algorithm in this experiment where it decreases the objective value in a quite slow rate, which partially demonstrates the conservative nature of the averaging feature in its design. The performance of the standard DCA is relatively similar to the SDCA. The gain of DCA-SAGA over the SDCA varies from 0.71% up to 60.15% and the gain of DCA-SVRG-v1 over the SDCA ranges from 0.56% to 54.73%.



(a) `shuttle` (b) `Sensorless` (c) `letter`

(d) `dna` (e) `SensIT Vehicle` (f) `connect-4`

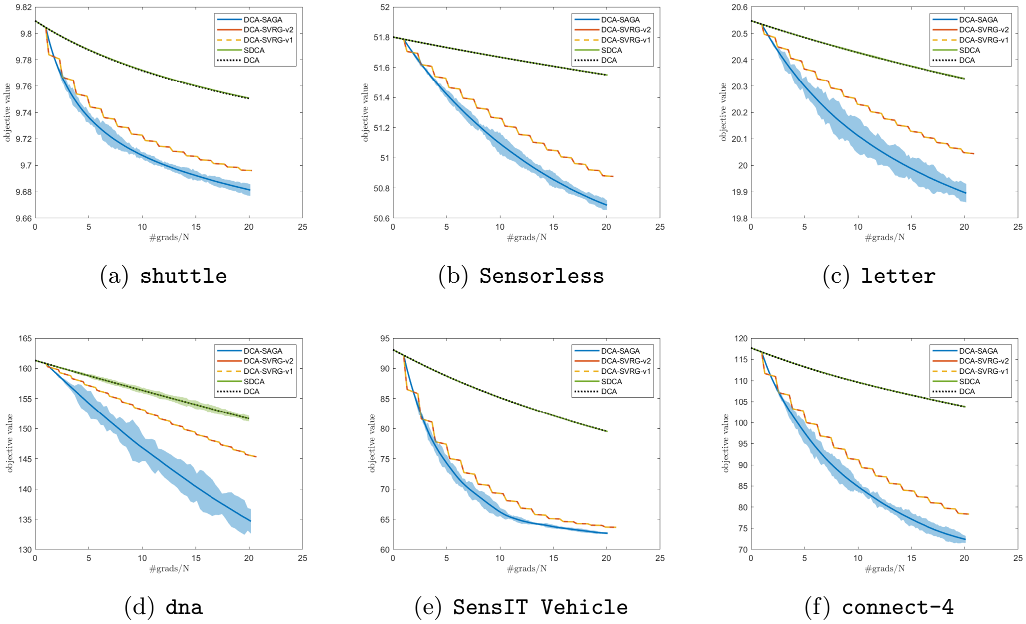
Figure 2: The (averaged) objective value of five algorithms for the case η^{exp} and $q = 1$.



(a) `shuttle` (b) `Sensorless` (c) `letter`

(d) `dna` (e) `SensIT Vehicle` (f) `connect-4`

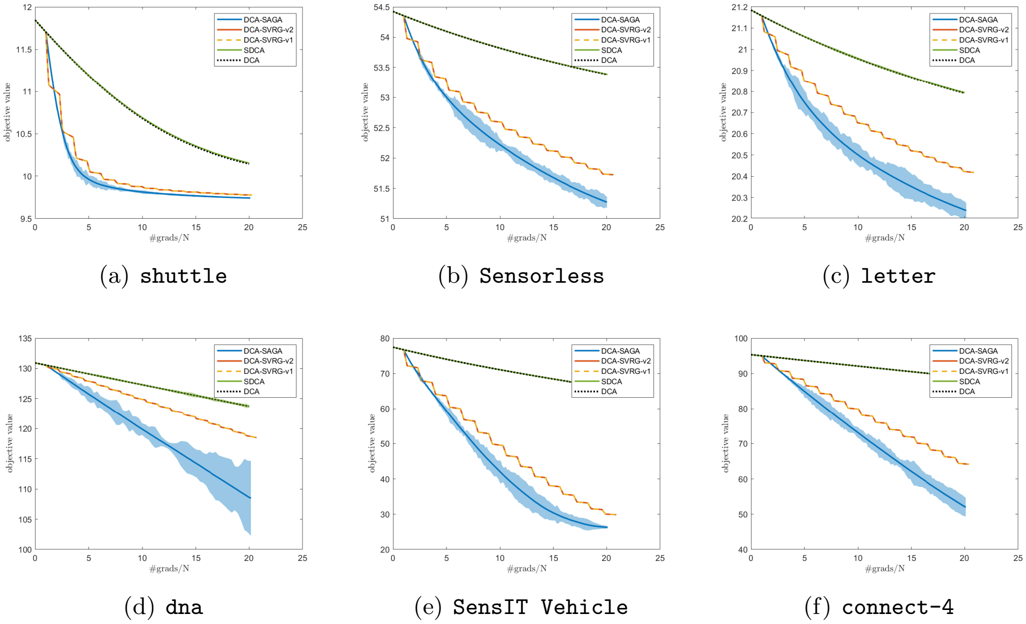
Figure 3: The (averaged) objective value of five algorithms for the case η^{exp} and $q = 2$.



(a) `shuttle` (b) `Sensorless` (c) `letter`

(d) `dna` (e) `SensIT Vehicle` (f) `connect-4`

Figure 4: The (averaged) objective value of five algorithms for the case $\eta^{\text{cap}-\ell_1}$ and $q = 1$.



(a) `shuttle` (b) `Sensorless` (c) `letter`

(d) `dna` (e) `SensIT Vehicle` (f) `connect-4`

Figure 5: The (averaged) objective value of five algorithms for the case $\eta^{\text{cap}-\ell_1}$ and $q = 2$.

5.3 Sparse Linear Regression

Sparse linear regression is a linear regression model that aims to eliminate redundant, noisy, and irrelevant features. Given a data set $\{(x_i, y_i)\}_{i=1}^N$, sparse linear regression tries to fit $\langle \beta, x_i \rangle \approx y_i$ with a sparse solution β in order to select right features, especially in the high-dimensional regime. Formally, the optimization problem associated with sparse linear regression is given by

$$\min_{\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \langle \beta, x_i \rangle)^2 + \lambda \|\beta\|_0, \tag{28}$$

where $\lambda > 0$ is a parameter that makes a trade-off between the data-fitting term and the sparsity-encouraging term.

To make the problem tractable, the discontinuous ℓ_0-norm is usually approximated by a continuous one. In this section, we focus on the Capped-ℓ_1 that is used to approximate the ℓ_0-norm to give rise to the following approximation problem

$$\min_{\beta} F(\beta) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - \langle \beta, x_i \rangle)^2 + \lambda \sum_{j=1}^{m} \min(1, \alpha|\beta_j|). \tag{29}$$

5.3.1 DCA-SVRG and DCA-SAGA Applied to (29)

Firstly, each function $\beta \mapsto \frac{1}{2}(y_i - \langle \beta, x_i \rangle)^2$ is $\|x_i\|^2$-smooth, so it admits the DC decomposition (27). On the other hand, the Capped-ℓ_1 function is DC with the DC decomposition $\min(1, \alpha|\beta|) = \tilde{g}(\beta) - \tilde{h}(\beta)$, where $\tilde{g}(\beta) = 1 + \alpha|\beta|$, and $\tilde{h}(\beta) = \max(1, \alpha|\beta|)$. Therefore, the problem (29) falls into our framework with the following setting

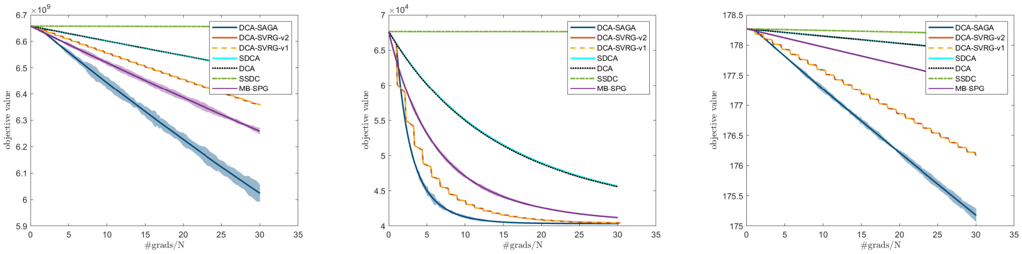
$$\begin{aligned} G(\beta) &= \frac{\gamma}{2}\|\beta\|^2, h_i(\beta) = \frac{\gamma}{2}\|\beta\|^2 - \frac{1}{2}\langle \beta, x_i \rangle^2 + y_i \langle \beta, x_i \rangle, \\ r_1(\beta) &= \lambda\alpha\|\beta\|_1, r_2(\beta) = \lambda \sum_{j=1}^{m} \max(1, \alpha|\beta_j|), \end{aligned}$$

where $\gamma \geq \max_{i=\overline{1,N}} \|x_i\|^2$. When applying DCA-SAGA and DCA-SVRG to this setting, the subproblem is nothing but the proximal operator of ℓ_1, which has a closed-form solution.

On the other hand, similar to the phenomenon happened in the previous experiment on Group Variables Selection in Multi-class Logistic Regression, we have found that DCA-SAGA with replacement performs poorly and numerically unstable in this experiment. Hence, we only consider the following three algorithms: DCA-SAGA without replacement, DCA-SVRG-v1, and DCA-SVRG-v2.

5.3.2 Numerical Experiments

Data sets. We use regression data sets, namely, `cadata` (20640×8), `YearPredictionMSD` (463715×90) from LIBSVM, and SGEMM GPU kernel performance data set, `sgemm_product` (241600×14) (Nugteren and Codreanu, 2015). Each data set is standardized to have mean (μ) of 0 and standard deviation (σ) of 1.



(a) `cadata` (b) `sgemm_product` (c) `YearPredictionMSD`

Figure 6: The performance of seven algorithms on the sparse linear regression problem

Comparative algorithms. We compare our proposed algorithms with the standard DCA, the SDCA, the stochastic stagewise DC (SSDC) proposed in (Xu et al., 2019b), and the mini-batch stochastic proximal gradient (MB-SPG) (Xu et al., 2019a).

Experiment setups and results. The budget of SFO calls is set to be $30N$. The regularization parameter λ and the parameter α of the Capped-ℓ_1 are both set to be 1 throughout this experiment. To guarantee the convexity of DC components, we choose $\gamma = \max_{i=\overline{1,N}} \|x_i\|^2$. On the other hand, the minibatch size b is chosen as $\lfloor N^{\frac{2}{3}} \rfloor$, $\lfloor N^{\frac{2}{3}} \rfloor$, $\lceil 2\sqrt{2}\sqrt{N\sqrt{N+1}} \rceil$, $\lfloor 10\%N \rfloor$ for DCA-SVRG-v1, DCA-SVRG-v2, DCA-SAGA, and SDCA, respectively. For two versions of DCA-SVRG, we set $M = \lfloor \frac{1}{8\sqrt{e-1}} \sqrt{b} \rfloor$. About the SSDC, the nonconvex regularizer Capped-ℓ_1 is replaced by the Moreau envelope r_μ (Xu et al., 2019b, Sect. 4). Moreover, the proximal operator of the Capped-ℓ_1 can be efficiently computed (see, e.g., (Gong et al., 2013)), which allows the SSDC to work. The parameter μ controlling the Moreau envelope is required to match the (square) order of the final error ϵ (Xu et al., 2019b, Theorem 10c), so it is set to be a relatively small number, $\mu = 10^{-6}$. Note that, since the convex subproblem has a closed-form solution and the first DC component is not given as a large sum, no stochastic convex solver is required for the SSDC in this case. On the other hand, as remarked in (Xu et al., 2019a), the use of the Moreau envelope could be a bad idea as it introduces the approximation error while slowing down the convergence. Therefore, we further implement the mini-batch stochastic proximal gradient (MB-SPG) (Xu et al., 2019a) that uses directly the proximal operator of the Capped-ℓ_1. The minibatch size of MB-SPG is of order $\mathcal{O}(1/\epsilon^2)$ where ϵ is the expected final error, so it is proper to set the minibatch size at $\lfloor 10\%N \rfloor$. The constant c (see (Xu et al., 2019a)) is in $(0, 0.5)$, so we set it at 0.25 which is a neutral number between the two extremes; Meanwhile, as discussed earlier, the Lipschitz smoothness parameter can be chosen as $\max_{i=\overline{1,N}} \|x_i\|^2$.

We report the mean curve and the mean absolute deviation (over 10 runs) of the objective value of five comparative algorithms in Figure 6, where the mean absolute deviation is scaled up by a factor of 20 for visibility.

Comments. It is observed that, once again, DCA-SAGA is the best algorithm over all tested cases where it decreases the objective value in a fast rate to obtain good solutions. Meanwhile, the performances of DCA-SVRG-v1 and DCA-SVRG-v2 are almost identical and are the second best on `sgemm_product` and `YearPredictionMSD` data sets. On the other hand, the MB-SPG is the second best on the cadata data set and pushes the DCA-SVRGs to the third place in this case. The gain of DCA-SAGA over DCA-SVRG-v1 are 5.22%, 0.24%, and 0.56% on cadata, sgemm_product, and YearPredictionMSD, respectively. On the other hand, the graphs of DCA and SDCA are also almost the same on all three data sets. The gain of DCA-SVRG-v1 over SDCA on cadata, sgemm_product, and YearPredictionMSD are 2.11%, 11.45%, and 0.98%, respectively. Finally, in this experiment, while sample with replacement makes DCA-SAGA numerically unstable, sample without replacement consistently makes DCA-SAGA the best algorithm over all considered tested cases. Lastly, the performance of the SSDC is relatively bad in this experiment. Although it still manages to decrease the objective function, this decrease is not enough to be visible in the decreasing magnitude of other comparative algorithms. This can be attributed to the slow convergence caused by the use of the Moreau envelope.

6. Conclusion

In this paper, we have proposed two stochastic DCA schemes integrated variance reduction techniques called DCA-SVRG and DCA-SAGA for solving a wide class of nonconvex DC problems including the large-sum case, where both the data-fitting term and the regularization term are given as (nonsmooth) DC functions. The DC structure has been tackled by the DCA framework while the large-sum structure has been handled by the unbiased SVRG, SAGA estimators. The advantages of DC programming and DCA in this work are multiple. DC programming and DCA provide a general look at the considered problems and advise how to flexibly exploit the structural information of these problems. Together, they form a flexible deterministic frame on which stochastic techniques are built to handle the stochastically challenging part of the problems. Our proposed algorithms inherit the virtues as well as the unavoidable limitations of the stochastic estimators used. The DCA-SVRG is an epoch-based algorithm whose advantage is the storage requirement: to build the stochastic variance reduction term, it only needs to keep in the memory one gradient vector. Intuitively, a weakness of this algorithm is that the pivot point $\tilde{x}^k$ is not updated inside its corresponding epoch. Consequently, when the points x_j^{k+1} in an epoch progresses far away from its pivot $\tilde{x}^k$, $\nabla h_i(x_j^{k+1})$ and $\nabla h_i(\tilde{x}^k)$ (i is a random index) are no longer highly correlated, which undermines the effect of the variance reduction term. In contrast, DCA-SAGA can partially address this issue thanks to the progressive update of the variance reduction term through iterations. However, DCA-SAGA has its own limitation that is the storage burden: it needs to store a table of gradients in the memory. Although in some particular problems, one can manage to keep a vector of scalars only, this issue does not go away in general and is the bottleneck of the entire algorithm. We have observed that there is a biased estimator called SARAH (Nguyen et al., 2017) potential to address both limitations of SVRG and SAGA: it progressively updates the variance reduction term inside each epoch and does not need to store a table of gradients. This observation motivates our future works on the combination of DCA and SARAH.

Furthermore, for each proposed algorithm, in the step of choosing a subset of the large sum we have investigated both sampling strategies, namely, sampling with/without replacement. Sampling with replacement enjoys the independence of chosen items, facilitating simpler convergence analysis. However, the latter strategy is a common practice and it avoids some low-probability bad events that happen in sampling with replacement (high repetition that leads to bad approximations). In our numerical experiments on three problems, the results of DCA-SAGA with two options are very different from each other: sampling without replacement makes DCA-SAGA more stable and makes it the best algorithm in two studied problems. Meanwhile, the performance of DCA-SVRG with the two strategies are almost identical.

From the theoretical perspective, we establish the almost sure convergence of the proposed algorithms to DC critical points. Some new techniques and ideas introduced in the analysis of DCA-SAGA are expected to be useful in the analysis of future SAGA-type algorithms. Furthermore, for the class of problem (P) where r_2 in the regularizer $r(x) = r_1(x) - r_2(x)$ has a composite form, we provide additional convergence analysis to facilitate practical choice of parameters. By exploiting the special composite structure of r_2, we reformulate this problem as a DC program. We then give new relations between the proposed algorithms' parameters (minibatch size, inner-loop length) and the problems' parameters (strong convexity, Lipschitz smoothness, data set size) that guarantees the almost sure convergence property to DC critical points of the reformulated problem. The main advantage of the additional analysis is that it allows more relaxing convergence conditions than the ones obtained from straightforwardly applying the entire old convergence analysis to the reformulated problem.

Overall, in three experiments, the proposed algorithms seem to outperform several state-of-the-art stochastic methods for large-sum nonconvex problems. These experiments also intensively justify the merits of the SAGA and SVRG estimators inside the proposed algorithms and illustrate the limitation of the SAG estimator used in SDCA.

Appendix

Proof [Proof of Lemma 3] We have,

$$\begin{aligned}
&\mathbb{E}\left(\left\|\frac{1}{b}\sum_{i\in I_b}\nabla h_i(x_j^{k+1}) - \frac{1}{b}\sum_{i\in I_b}\nabla h_i(\tilde{x}^k) + \nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1})\right\|^2 |\mathcal{P}_j^{k+1}\right)\\
&= \mathbb{E}_{I_b}\left\|\frac{1}{b}\sum_{i\in I_b}\nabla h_i(x_j^{k+1}) - \frac{1}{b}\sum_{i\in I_b}\nabla h_i(\tilde{x}^k) + \nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1})\right\|^2\\
&= \sum_{I\subset[N],|I|=b}\mathbb{P}(I_b = I)\left\|\frac{1}{b}\sum_{i\in I}\nabla h_i(x_j^{k+1}) - \frac{1}{b}\sum_{i\in I}\nabla h_i(\tilde{x}^k) + \nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1})\right\|^2,
\end{aligned}$$

where the probability $\mathbb{P}(I_b = I) = 1/C_N^b$. For a fixed I, we compute

$$\begin{aligned}
R &= \left\| \frac{1}{b}\sum_{i\in I} \nabla h_i(x_j^{k+1}) - \frac{1}{b}\sum_{i\in I} \nabla h_i(\tilde{x}^k) + \nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1}) \right\|^2 \\
&= \frac{1}{b^2}\sum_{i\in I} \|\nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k) + \nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1})\|^2 \\
&+ \frac{1}{b^2}\sum_{i\neq l, i,l\in I} \langle \nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k) + \nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1}), \\
&\qquad\qquad \nabla h_l(x_j^{k+1}) - \nabla h_l(\tilde{x}^k) + \nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1})\rangle := \frac{1}{b^2}(R_1 + R_2).
\end{aligned}$$

$$\begin{aligned}
R_1 &= \sum_{i\in I} \|\nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k)\|^2 + b\|\nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1})\|^2 \\
&+ 2\langle \nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1}), \sum_{i\in I} (\nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k))\rangle, \\
R_2 &= A_b^2 \|\nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1})\|^2 \\
&+ \langle \nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1}), \sum_{i\neq l, i,l\in I} (\nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k) + \nabla h_l(x_j^{k+1}) - \nabla h_l(\tilde{x}^k))\rangle \\
&+ \sum_{i\neq l, i,l\in I} \langle \nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k), \nabla h_l(x_j^{k+1}) - \nabla h_l(\tilde{x}^k)\rangle.
\end{aligned}$$

Therefore,

$$\begin{aligned}
R =& \frac{1}{b^2}\sum_{i\in I} \|\nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k)\|^2 + \frac{2}{b}\langle \nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1}), \sum_{i\in I} (\nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k))\rangle \\
&+ \|\nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1})\|^2 + \frac{1}{b^2}\sum_{i\neq l, i,l\in I} \langle \nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k), \nabla h_l(x_j^{k+1}) - \nabla h_l(\tilde{x}^k)\rangle.
\end{aligned}$$

On the other hand, we have the following computations

$$\begin{aligned}
&\sum_{I\subset[N],|I|=b}\sum_{i\in I} \|\nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k)\|^2 = C_{N-1}^{b-1}\sum_{i=1}^{N} \|\nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k)\|^2, \\
&\sum_{I\subset[N],|I|=b} \langle \nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1}), \sum_{i\in I} (\nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k))\rangle \\
&= -C_{N-1}^{b-1}\cdot N \cdot \|\nabla H(\tilde{x}^k) - \nabla H(x_j^{k+1})\|^2, \\
&\sum_{I\subset[N],|I|=b}\sum_{i\neq l, i,l\in I} \langle \nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k), \nabla h_l(x_j^{k+1}) - \nabla h_l(\tilde{x}^k)\rangle \\
&= C_{N-2}^{b-2}\sum_{i\neq j} \langle \nabla h_i(x_j^{k+1}) - \nabla h_i(\tilde{x}^k), \nabla h_l(x_j^{k+1}) - \nabla h_l(\tilde{x}^k)\rangle.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{I_b}\left\|\frac{1}{b}\sum_{i\in I_b}\nabla h_i(x_j^{k+1})-\frac{1}{b}\sum_{i\in I_b}\nabla h_i(\tilde{x}^k)+\nabla H(\tilde{x}^k)-\nabla H(x_j^{k+1})\right\|^2\\
&=\frac{1}{bN}\sum_{i=1}^{N}\|\nabla h_i(x_j^{k+1})-\nabla h_i(\tilde{x}^k)\|^2-\|\nabla H(\tilde{x}^k)-\nabla H(x_j^{k+1})\|^2\\
&\quad+\frac{b-1}{bN(N-1)}\sum_{i\neq l}\langle\nabla h_i(x_j^{k+1})-\nabla h_i(\tilde{x}^k),\nabla h_l(x_j^{k+1})-\nabla h_l(\tilde{x}^k)\rangle\\
&=\frac{1}{bN}\left(1-\frac{b-1}{N-1}\right)\sum_{i=1}^{N}\|\nabla h_i(x_j^{k+1})-\nabla h_i(\tilde{x}^k)\|^2\\
&\quad+\left(\frac{N(b-1)}{b(N-1)}-1\right)\|\nabla H(\tilde{x}^k)-\nabla H(x_j^{k+1})\|^2\\
&\leq\frac{L^2}{b}\left(1-\frac{b-1}{N-1}\right)\|x_j^{k+1}-\tilde{x}^k\|^2.
\end{aligned}$$

■

References

Zeyuan Allen-Zhu and Yang Yuan. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *International Conference on Machine Learning*, pages 1080–1089. PMLR, 2016.

Steven C Bagley, Halbert White, and Beatrice A Golomb. Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology*, 54(10):979–985, 2001.

Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific optimization and computation series. Athena Scientific, 2003.

Paul S Bradley and Olvi L Mangasarian. Feature selection via concave minimization and support vector machines. In *International Conference on Machine Learning*, volume 98, pages 82–90. Citeseer, 1998.

Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on Machine Learning*, pages 37–45. PMLR, 2013.

Sashank J. Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, volume 29, pages 1145–1153, 2016.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Jinseog Kim, Yuwon Kim, and Yongdai Kim. A gradient-based optimization algorithm for lasso. *Journal of Computational and Graphical Statistics*, 17(4):994–1009, 2008.

Gary King and Langche Zeng. Logistic regression in rare events data. *Political Analysis*, 9 (2):137–163, 2001.

Hoai An Le Thi. DC programming and DCA, 2005. URL `http://www.lita.univ-lorraine.fr/~lethi/index.php/dca.html`.

Hoai An Le Thi. A new approximation for the ℓ_0-norm. *Research Report LITA EA 3097*, 2012.

Hoai An Le Thi and Tao Pham Dinh. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133(1-4):23–46, 2005.

Hoai An Le Thi and Tao Pham Dinh. DC programming and DCA: thirty years of developments. *Mathematical Programming, Special Issue dedicated to : DC Programming - Theory, Algorithms and Applications*, 169(1):5–68, 2018.

Hoai An Le Thi and Duy Nhat Phan. DC programming and DCA for sparse optimal scoring problem. *Neurocomputing*, 186:170–181, 2016.

Hoai An Le Thi, Hoai Minh Le, and Tao Pham Dinh. Feature selection in machine learning: an exact penalty approach using a difference of convex function algorithm. *Machine Learning*, 101(1):163–186, 2015.

Hoai An Le Thi, Hoai Minh Le, Duy Nhat Phan, and Bach Tran. Stochastic DCA for the large-sum of non-convex functions problem and its application to group variable selection in classification. In *International Conference on Machine Learning*, pages 3394–3403. PMLR, 2017.

Hoai An Le Thi, Van Ngai Huynh, and Tao Pham Dinh. Convergence analysis of difference-of-convex algorithm with subanalytic data. *Journal of Optimization Theory and Applications*, 179(1):103–126, 2018.

Hoai An Le Thi, Hoai Minh Le, Duy Nhat Phan, and Bach Tran. Stochastic DCA for minimizing a large sum of dc functions with application to multi-class logistic regression. *Neural Networks*, 132:220–231, 2020.

Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

Andrea Montanari and Emile Richard. Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *IEEE Transactions on Information Theory*, 62(3):1458–1484, 2015.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017.

Cedric Nugteren and Valeriu Codreanu. Cltune: A generic auto-tuner for opencl kernels. In *2015 IEEE 9th International Symposium on Embedded MulticoreMany-core Systems-on-Chip*, pages 195–202. IEEE, 2015.

Cheng Soon Ong and Hoai An Le Thi. Learning sparse classifiers with difference of convex functions algorithms. *Optimization Methods and Software*, 28(4):830–854, 2013.

Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh. Proxsarah: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.

Tao Pham Dinh and Hoai An Le Thi. Convex analysis approach to DC programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 1997.

Tao Pham Dinh and Hoai An Le Thi. A DC optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1998.

Duy Nhat Phan, Hoai An Le Thi, and Tao Pham Dinh. Sparse covariance matrix estimation by dca-based algorithms. *Neural Computation*, 29(11):3040–3077, 2017.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Shai Shalev-Shwartz. SDCA without duality, regularization, and individual convexity. In *International Conference on Machine Learning*, pages 747–754. PMLR, 2016.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(2), 2013.

Eric M Stuve. Estimating and plotting logarithmic error bars, 2004.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yi Xu, Rong Jin, and Tianbao Yang. Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems. In *Advances in Neural Information Processing Systems*, volume 32, 2019a.

Yi Xu, Qi Qi, Qihang Lin, Rong Jin, and Tianbao Yang. Stochastic optimization for dc functions and non-smooth non-convex regularizers with non-asymptotic convergence. *arXiv preprint arXiv:1811.11829v2*, 2019b.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010a.

Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11(3), 2010b.