Stochastic Optimization for Non-convex Inf-Projection Problems

Yan Yan 1 Yi Xu 2 Liijun Zhang 3 Xiaoyu Wang 4 Tianbao Yang 1

Abstract
In this paper, we study a family of non-convex and possibly non-smooth inf-projection minimization problems, where the target objective function is equal to minimization of a joint function over another variable. This problem include difference of convex (DC) functions and a family of bi-convex functions as special cases. We develop stochastic algorithms and establish their first-order convergence for finding a (nearly) stationary solution of the target non-convex function under different conditions of the component functions. To the best of our knowledge, this is the first work that comprehensively studies stochastic optimization of non-convex inf-projection minimization problems with provable convergence guarantee. Our algorithms enable efficient stochastic optimization of a family of non-decomposable DC functions and a family of bi-convex functions. To demonstrate the power of the proposed algorithms we consider an important application in variance-based regularization. Experiments verify the effectiveness of our inf-projection based formulation and the proposed stochastic algorithm in comparison with previous stochastic algorithms based on the min-max formulation for achieving the same effect.

1. Introduction
In this paper, we consider a family of non-convex and possibly non-smooth problems with the following structure
\[
\min_{x \in X} F(x) := \{g(x) + \min_{y \in \text{dom}(h)} h(y) - \langle y, \ell(x) \rangle\},
\]
where \(X \subseteq \mathbb{R}^d\) is a closed convex set, \(g : X \to \mathbb{R}\) is lower-semicontinuous, \(h : \text{dom}(h) \to \mathbb{R}\) is uniformly convex, \(\ell : X \to \mathbb{R}^m\) is a lower-semicontinuous differentiable mapping, and \(\langle \cdot, \cdot \rangle\) is the inner product. The requirement of uniform convexity on \(h\) is to ensure the inner minimization problem is well defined and its solution is unique (cf. Section 2). Define \(f(x, y) = g(x) + h(y) - \langle y, \ell(x) \rangle\), the objective function \(F(x)\) is called the inf-projection of \(f(x, y)\) in the literature. When \(g\) is convex, depending on \(\text{dom}(h)\), the two subfamilies of above problem (1) deserve more discussion: difference of convex (DC) and bi-convex functions.

**DC functions.** When \(g\) is convex and \(\text{dom}(h) \subseteq \mathbb{R}_+^m\) and \(\ell\) is convex 1, the inf-projection minimization problem (1) is equivalent to the following DC functions,
\[
\min_{x \in X} F(x) = \{g(x) - h^*(\ell(x))\},
\]
where \(h^*\) denotes the convex conjugate function of \(h\), the convexity of the second component \(h^*(\ell(x))\) is following the composition rule of convexity (Boyd & Vandenberghe, 2004) 2. Minimizing DC functions has wide applications in machine learning and statistics (Nitanda & Suzuki, 2017; Kiryo et al., 2017). Although stochastic algorithms for DC problems have been considered recently (Nitanda & Suzuki, 2017; Xu et al., 2018a; Thi et al., 2017), working with the inf-projection minimization (1) is preferred when \(h^*(\ell(x))\) is non-decomposable such that an unbiased stochastic gradient of \(h^*(\ell(x))\) is not easily accessible as that of \(\langle y, \ell(x) \rangle\) in (1). Inspired by this scenario, let us particularly consider an important instance variance-based regularization. It refers to a learning paradigm that minimizes the empirical loss and its variance simultaneously, by which a better bias-variance trade-off may be achieved (Maurer & Pontil, 2009). To give a condensed understanding of its connection to the inf-projection formulation, we can re-formulate the problem (cf. the details and comparison with a related convex problem of (Namkoong & Duchi, 2017) in Section 5):
\[
\min_{x \in X} \frac{1}{n} \sum_{i=1}^n l_i(x) + \lambda \frac{1}{2n} \sum_{i=1}^n (l_i(x))^2 - \lambda \left(\frac{1}{n} \sum_{i=1}^n l_i(x)\right)^2,
\]
where \(l_i(x) : X \to \mathbb{R}_+\) is the loss function of a model \(x\) on the \(i\)-th example and \(\lambda > 0\) is a regularization parameter. To ensure \(\lambda > 0\), we consider a family of possibly non-smooth problems with the following structure

---

1University of Iowa 2DAMO Academy, Alibaba Group 3Nanjing University 4The Chinese University of Hong Kong (Shenzhen). Correspondence to: Yan Yan <yan-yan-2@uiowa.edu>, Tianbao Yang <tianbao-yang@uiowa.edu>.

Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).
Stochastic Optimization for Non-convex Inf-Projection Problems

Table 1. Summary of results for finding a (nearly) $\epsilon$-stationary solution in this work under different conditions of $g, h$ and $\ell$. SM means smoothness, Lip. means Lipschitz continuous, Diff means differentiable, MO means monotonically increasing or decreasing for $h^\ast$, CVX means convex, and UC means $p$-uniformly convex ($p \geq 2$), $v = 1/(p - 1)$.

| $g$    | $h (h^\ast)$ | $\ell$ | Alg.             | Mini-Batch | Compl.        |
|--------|---------------|--------|------------------|------------|---------------|
| SM     | UC & simple   | SM & Lip | MSPG (Section 3) | Yes        | $O(1/\epsilon^4/v)$ |
| SM & CVX | UC (MO)  | Diff & Lip & CVX | St-SPG (Section 4) | No         | $O(1/\epsilon^4/v)$ |
| Lip & CVX | UC (MO)  | SM & Lip & CVX | St-SPG (Section 4) | No         | $O(1/\epsilon^4/v)$ |

The novelty and significance of our results are (i) this is the first work that comprehensively studies the stochastic optimization of a non-smooth non-convex inf-projection problem; (ii) the application of the inf-projection formulation to variance-based regularization demonstrates much faster convergence of our algorithms comparing with existing algorithms based on a min-max formulation.

A naive idea to tackle (1) is by alternating minimization or block coordinate descent, i.e., alternatively solving the inner minimization problem over $y$ given $x$ and then updating $x$ by certain approaches (e.g., stochastic gradient descent) (Bolte et al., 2014; Davis et al., 2016; Hong et al., 2015; Xu & Yin, 2013; Driggs et al., 2020). However, this approach suffers from two issues: (i) solving the inner minimization might not be a trivial task (e.g., solving the inner minimization problem related to (3) requires passing $n$ examples once); (ii) the target objective function $F(x)$ is not necessarily a smooth function or a convex function, which makes the convergence analysis challenging. Additionally, their convergence analysis focuses on $f(x, y)$ instead of $F(x)$. In this paper, the main question that we tackle is: how to design efficient stochastic algorithms using simple updates for both $x$ and $y$ to enjoy a provable convergence guarantee in terms of finding a stationary point of $F(x)$? Our contributions are summarized below:

- First, we consider the case when $g$ and $\ell$ are smooth but not necessarily convex and $h$ is a simple function whose proximal mapping is easy to compute. Under the condition that $\ell$ is Lipschitz continuous, we prove the convergence of mini-batch stochastic proximal gradient method (MSPG) with increasing mini-batch size that employ parallel stochastic gradient updates for $x$ and $y$, and establish the convergence rate.

- Second, we consider the cases when $g$ and $\ell$ are not necessarily smooth but convex, and $h$ is not necessarily a simple function (corresponding to DC and bi-convex functions). We develop an algorithmic framework that employs a suitable stochastic algorithm for solving strongly convex functions in a stagewise manner. We analyze the convergence rates for finding a (nearly) stationary point when employing the stochastic proximal gradient (SPG) method at each stage, resulting St-SPG. The complexity results of our algorithms under different conditions of $g, h$ and $\ell$ are shown in Table 1.
2. Preliminaries

Let us first present some notations. We let \( \| \cdot \| \) denote the Euclidean norm of a vector and the spectral norm of a matrix. We use \( \xi \) to denote some random variable. Given a function \( g : \mathbb{R}^d \to \mathbb{R} \), we denote the Fréchet subgradients and limiting Fréchet gradients by \( \partial g \) and \( \partial g \) respectively, i.e., at \( x \), \( \partial g(x) = \{ y \in \mathbb{R}^d : \lim_{x_k \to x} \inf_{y \neq y'} \frac{g(x_k) - g(x) - \langle y - y', x - x \rangle}{\| x - x \|} \geq 0 \} \), and \( \partial g(x) = \{ y \in \mathbb{R}^d : \exists x_k^g \to x, v_k \in \partial g(x_k), v_k \to v \} \).

Here \( x_k^g \) represents \( x_k \) and \( g(x_k) \). When the function \( g \) is differentiable, the subgradients (\( \partial g \) and \( \partial g \)) reduce to the standard gradient \( \nabla g \).

We denote by \( \partial_x g(x, y) \) the partial derivative in the direction of \( x \) and \( \partial g(x, y) = (\partial_x g(x, y), \partial_y g(x, y))^T \). In this paper, we will prove the convergence in terms of the limiting gradient. But all results can be extended to the Fréchet subgradients.

Let \( \nabla \ell(x) \in \mathbb{R}^{m \times d} \) denote the Jacobian matrix of the differentiable mapping \( \ell(x) \). \( \ell \) is said \( G_\ell \)-Lipschitz continuous if \( \| \nabla \ell(x) \| \leq G_\ell \). A differentiable function \( f(\cdot) \) has \( (L, v) \)-Hölder continuous gradient if \( \| \nabla f(x_1) - \nabla f(x_2) \| \leq L \| x_1 - x_2 \|^v \) holds for all \( v \in [0, 1] \) and \( L > 0 \). When \( v = 1 \), it is known as \( L \)-smooth function. If \( \nabla f \) is Hölder continuous, then it holds \( f(x_1) - f(x_2) \leq \nabla f(x_2) \cdot (x_1 - x_2) \). A related condition is uniform convexity. A function \( f(\cdot) \) is \( (p, \rho) \)-uniformly convex where \( p \geq 2 \) if \( f(x_1) - f(x_2) \geq \rho \| f(x_2) - f(x_1) \|^2 \). When \( p = 2 \), it is known as strong convexity.

Algorithm 1 MSPG

1. **Input:** initialized \( x_1, y_1 \).
2. **for** \( t = 1, \ldots, T \) **do**
3. Compute mini-batch stochastic partial gradients \( \nabla_{x} f_{0}(t) = \frac{1}{m_t} \sum_{i=1}^{m_t} \nabla_x f_0(x_t, y_t; \xi_i) \) and \( \nabla_{y} f_{0}(t) = \frac{1}{m_t} \sum_{i=1}^{m_t} \nabla_y f_0(x_t, y_t; \xi_i) \).
4. \( x_{t+1} = \Pi_{\mathcal{X}} [x_t - \eta \nabla_{x} f_{0}(t)] \).
5. \( y_{t+1} = P_{\mathcal{Y}} [y_t - \eta \nabla_{y} f_{0}(t)] \).
6. **end for**
7. **Output:** \( w_T = (x_T, y_T) \), where \( T \in \{1, \ldots, T\} \) is randomly sampled.

**Definition 1.** A solution \( x \) satisfying \( \text{dist}(0, \partial F(x)) \leq \epsilon \) is called an \( \epsilon \)-stationary point of \( F \). A solution \( x \) is called a nearly \( \epsilon \)-stationary if there exists \( x \) and a constant \( \epsilon > 0 \) such that \( \| z - x \| \leq \epsilon \) and \( \text{dist}(0, \partial F(z)) \leq \epsilon \).

Particularly, nearly stationarity has been used to measure the convergence for non-smooth non-convex optimization in the literature (Davis & Grimmer, 2017; Davis & Drusvyatskiy, 2018a; Chen et al., 2018; Xu et al., 2018a).

Before ending this section, we state basic assumptions below. For simplicity, here all variance bounds are denoted by \( \sigma^2 \). Additional conditions regarding \( g, h \) and \( \ell \) are presented in individual theorems.

**Assumption 1.** For the problem (1) we assume:
(i) \( h^* \) has \( (L_{h^*}, v) \)-Hölder continuous gradient, and \( \ell \) is continuously differentiable;
(ii) \( \partial g(x; \xi) \) denote a stochastic gradient of \( g(x) \). If \( g(x) \) is smooth, assume \( \mathbb{E}[\| \partial g(x; \xi) - \partial g(x) \|^2] \leq \sigma^2 \), otherwise assume \( \mathbb{E}[\| \partial g(x; \xi) \|^2] \leq \sigma^2 \) for all \( x \in X \);
(iii) Let \( \ell(x; \xi) \) denote a stochastic version of \( \ell(x) \) and assume \( \mathbb{E}[\| \nabla \ell(x; \xi) - \ell(x) \|^2] \leq \sigma^2 \).
(iv) \( \partial h(y; \xi) \) denote a stochastic gradient of \( h(y) \) and assume \( \mathbb{E}[\| \partial h(y; \xi) \|^2] \leq \sigma^2 \) for all \( y \in \text{dom}(h) \);
(v) \( \max_{x \in X, y \in \text{dom}(h)} f(x, y) - \min_{x \in X, y \in \text{dom}(h)} f(x, y) \leq M \).

3. Mini-batch Stochastic Gradient Methods

For Smooth Functions

In this section, we consider the case when \( g \) and \( \ell \) are smooth functions but not necessarily convex. Please note that the target function \( F \) is still not necessarily smooth and is non-convex. We assume \( h \) is simple such that its proximal map-
Solving Inf-Projection Problems

where $h^*$ has $(L_h, v)$-H"older continuous gradient. Then for any $(\tilde{x}, \tilde{y}) \in X \times dom(h)$, we have

$$
\text{dist}(0, \partial F(\tilde{x})) \leq \|\nabla_x f(\tilde{x}, \tilde{y})\|_2 + G_\ell \left(\frac{1 + v}{2v}\right)^v L_h \text{dist}(0, \partial_y f(\tilde{x}, \tilde{y}))^v.
$$

Finally, combining the above results, we can state the main result in this section regarding the convergence of MSPG in terms of the concerned dist$(0, \partial F(x_\tau))$ as follows.

**Theorem 2.** Suppose the same conditions as in Lemma 2 and Assumption 1 hold. Algorithm 1 guarantees that $E[\text{dist}(0, \partial F(x_\tau))^2] \leq O(1/T^\gamma)$. To ensure $E[\text{dist}(0, \partial F(x_\tau))] \leq \epsilon$, we can set $T = O(1/\epsilon^{2/v})$. The total complexity is $O(1/\epsilon^{k/v})$.

### 4. Stochastic Algorithms for Non-Smooth Functions

In this section, we consider the case when $g$ and $\ell$ are not necessarily smooth but are convex. We also assume $h^*$ is monotonic, i.e., $dom(h) \subseteq \mathbb{R}_m$ or $dom(h) \subseteq \mathbb{R}_m$. In the former case, the objective function belongs to DC functions, and in the latter case the objective function belongs to Bi-Convex functions. Please note that the target function $F$ is still not necessarily convex and is non-smooth. The proposed algorithm is inspired by the stagewise stochastic DC algorithm proposed in (Xu et al., 2018a) but with some minor changes. Let us first briefly discuss the main idea and logic behind the proposed algorithm. There are two difficulties that we need to tackle: (i) non-smoothness and non-convexity in terms of $x$, (ii) minimization over $y$.

To tackle the first issue, let us assume the optimal solution $y^*(x) = \arg \min_y h(y) - y^\top \ell(x)$ given $x$ is available. Then the problem regarding $x$ becomes:

$$
\min_{x \in X} g(x) - y^*(x)^\top \ell(x) \quad (5)
$$

When $dom(h) \subseteq \mathbb{R}_m$ (corresponding to a DC function), the above problem is still non-convex. In order to obtain a provable convergence guarantee, we consider the following strongly convex problem from some $\gamma > 0$ and $x_0 \in X$, whose objective function is an upper bound of the function in (5) at $x_0$:

$$
P(x_0) = \arg \min_{x \in X} \left\{ g(x) - y_{x_0}^\top \ell(x_0) + \nabla \ell(x_0)(x - x_0) + \gamma \|x - x_0\|^2 \right\}. \quad (6)
$$
Note \( P(x_0) \) is uniquely defined due to strong convexity. If \( x_0 = P(x_0) \) it can be shown that \( x_0 \) is the critical point of \( F(x) \), i.e., \( 0 \in \partial F(x_0) = \partial g(x_0) - \nabla \ell(x_0)^T y^*(x_0) \). Then we can iteratively solve the fixed-point problem \( x = P(x) \) until it converges.

When \( \text{dom}(h) \subseteq \mathbb{R}^m \) (corresponding to a Bi-convex function), we can simply consider the following strongly convex problem:

\[
P(x_0) = \arg \min_{x \in X} g(x) - y^*(x_0)^T \ell(x) + \frac{\gamma}{2} \| x - x_0 \|^2.
\]

A remaining issue in the above approach is that \( y^*(x_0) \) is assumed available, which is related to the second issue mentioned above. It may not be easy to obtain an exact minimizer \( y^*(x_0) \) given a \( x_0 \). To this end, we can employ an iterative stochastic algorithm to approximate \( \min_y h(y) - y^T \ell(x_0) \) approximately given \( x_0 \), and obtain an inexact solution \( \hat{y}(x_0) \) such that \( h(\hat{y}(x_0)) - \hat{y}(x_0)^T \ell(x_0) - (y^*(x_0)) - y^*(x_0)^T \ell(x_0) \leq \epsilon \) for some approximation error \( \epsilon \). Then, we combine these two pieces together, i.e., replacing \( y^*(x_0) \) in the definition of \( P(x_0) \) with \( \hat{y}(x_0) \), and employing a stochastic algorithm to solve the fixed-point equation by \( x \leftarrow \hat{P}(x) \), where \( \hat{P}(x) \) is an approximation of \( P(x) \). Therefore, we have two sources of approximation error — one from using \( \hat{y} \) instead of \( y^*(x) \) and another one from solving the minimization problem of \( x \) inexact. Our analysis is to show that with well-controlled approximation error, we can still achieve provable convergence guarantee.

For the sake of presentation, let us first introduce some important notations by considering different conditions of DC and bi-convex functions. For the \( k \)-th stage of St-SPG, define

\[
f^k_x(x) = g(x) - y^k_T (\ell(x_k) + \nabla \ell(x_k)(x - x_k)), \quad \text{for dom}(h) \subseteq \mathbb{R}^m,
\]

and

\[
f_x(x) = g(x) - y^T \ell(x), \quad \text{for dom}(h) \subseteq \mathbb{R}^m.
\]

A stochastic gradient of \( f^k_x(x) \) can be computed by \( \partial g(x; \xi_g) - \nabla \ell(x_k; \xi_\ell)^T y_k \) for \( \text{dom}(h) \subseteq \mathbb{R}^m \) or \( \partial g(x; \xi_g) - \nabla \ell(x_k; \xi_\ell)^T y_k \) for \( \text{dom}(h) \subseteq \mathbb{R}^m \). For both conditions, let

\[
f^k_y(y) = h(y) - y^T (x_{k+1})
\]

\[
R^k_x(x) = \frac{\gamma}{2} \| x - x_k \|^2, \quad R^k_y(y) = \frac{\mu}{2} \| y - y_k \|^2.
\]

A stochastic gradient of \( f^k_y(y) \) can be computed by \( \partial h(y; \xi_h) - \ell(x_{k+1}; \xi_\ell) \), where \( \xi_g, \xi_\ell, \xi_h, \xi_\ell \) denote independent random variables.

The proposed algorithm is shown in Algorithm 2 named St-SPG, which employs SPG in Algorithm 3 to solve the subproblems of \( x \) and \( y \) in a stagewise manner. \( x \) and \( y \) share the same update method SPG, so we can summarize it in general notations. To this end, let us consider the convergence of SPG for solving \( H(z) = f(z) + R(z) \), where \( f(z) \) is a convex function and \( R(z) = \frac{\gamma}{2} \| z - z_0 \|^2 \) is a strongly convex function. Its convergence has been considered in many previous works. Here, we adopt the results derived in (Xu et al., 2018a) to establish the convergence of St-SPG under different conditions of \( g \) and \( \ell \) as follows.

**Proposition 2.** Let \( H(z) = f(z) + R(z) \) where \( R(z) = \frac{\gamma}{2} \| z - z_0 \|^2 \) is \( \gamma \)-strongly convex. If \( f(z) \) is \( L \)-smooth and \( E[\| \nabla f(z; \xi) - \nabla f(z) \|^2] \leq \sigma^2 \) and \( \gamma \geq 3L \), then by setting \( \eta_t = 3/(\gamma(t+1)) \) SPG guarantees that

\[
E[H(\hat{z}_T) - H(z_*)] \leq \frac{4\gamma \| z_* - z_1 \|^2}{3T(T+3)} + \frac{6\sigma^2}{(T+3)\gamma}.
\]

If \( f \) is non-smooth with \( E[\| \nabla f(z; \xi) \|^2] \leq \sigma^2 \), then by setting \( \eta_t = 4/(\gamma(t)) \) SPG guarantees that

\[
E[H(\hat{z}_T) - H(z_*)] \leq \frac{\gamma \| z_* - z_1 \|^2}{4T(T+1)} + \frac{17\sigma^2}{\gamma(T+1)},
\]

where \( z_* = \arg \min_{z \in \Omega} H(z) \).

With the above proposition, we can apply the above convergence guarantee of SPG for \( f^k_x(x) + R^k_x(x) \) and \( f^k_y(y) + R^k_y(y) \). Then define \( v_k \) and \( u_k \) as the optimal solutions to the subproblems of \( x \) and \( y \) at the \( k \)-th stage, respectively:

\[
v_k = \arg \min_{x \in X} f^k_x(x) + R^k_x(x),
\]

\[
u_k = \arg \min_{y \in Y} f^k_y(y) + R^k_y(y).
\]

We can establish the following result regarding the convergence of St-SPG related to fixed-point convergence \( (x_{r+1} - x_r) \), and also the minimization error of \( P(x) \),
i.e., $\|x_{\tau+1} - v_{\tau}\|$, for a randomly sampled index $\tau \in \{1, \ldots, K\}$. We have boundedness assumptions on $y$ and $\ell$ below to guarantee the boundedness of the second moment of stochastic gradients, which can be implied by assuming the domain $X$ is a compact set and dom$(h)$ is bounded.

**Theorem 3.** Suppose Assumption 1 holds, and $\max\{\|y_k\|^2, E[\|\ell(x_{k+1}; \xi)\|^2]\} \leq D^2$ for $k \in \{1, \ldots, K\}$. There exists a constant $G = 17 \max\{2\sigma^2 + 2D^2\sigma^2, 2\sigma^2 + 2D^2\}$, and for any constants $\gamma > 0, \mu > 0, \alpha \geq 1$ Algorithm 2 with $T_k^\alpha = k/\gamma + 1, T_k^\alpha = k/\mu + 1$ guarantees that the following inequalities hold:

\[
\frac{1}{2}E[\|x_{\tau+1} - v_{\tau}\|^2] \leq \frac{1}{\gamma K}E[\|x_{\tau+1} - x_{\tau}\|^2] + \frac{A(M + 2G^2)(\alpha + 1)}{\gamma K}, \\
\frac{1}{2}E[\|y_{\tau+1} - u_{\tau}\|^2] \leq \frac{1}{\mu K}E[\|y_{\tau+1} - y_{\tau}\|^2] + \frac{4(M + 2G^2)(\alpha + 1)}{\mu K},
\]

for $\tau$ sampled by $P(\tau = k) = \frac{k^\alpha}{\sum_{i=1}^K i^\alpha}$.

The lemma below connects $\|\nabla F(x_k)\|$ (or dist$(0, F(x_k))$) to the quantities in Theorem 3, by which we can derive the convergence of (nearly) stationary point.

**Lemma 4.** Suppose $g$ is $L_g$-smooth, and $\ell$ is $G_\ell$-Lipschitz continuous. Then for any $k$ we have

\[
\|\nabla F(x_k)\| \leq (\gamma + L_g)\|x_k - v_k\| + G_\ell\|y_k - u_k\| + G_\ell\mu\left(\frac{1 + \nu}{2\nu}\right)^{\frac{n}{2}}L_h\| u_k - y_k \| + G_{\ell + 1}\left(\frac{1 + \nu}{2\nu}\right)^{\frac{n}{2}}L_h\| x_{k+1} - x_k \|.
\]

Suppose $g$ is non-smooth, and $\ell$ is $G_\ell$-Lipschitz continuous and $L_\ell$-smooth and $\max_{y \in \text{dom}(h)} \|y\| \leq D$, then for any $k$ we have

\[
\text{dist}(0, \partial F(v_k)) \leq (\gamma + DL_\ell)\|x_k - v_k\| + G_\ell\|y_k - u_k\| + G_\ell\left(\frac{1 + \nu}{2\nu}\right)^{\frac{n}{2}}L_h\left(\mu\|y_k - u_k\| + G_\ell\|x_{k+1} - z_k\|\right).\]

Combining Lemma 4 and Theorem 3, we have the following corollaries regarding the convergence of St-SPG under different conditions of $g$ and $\ell$.

**Corollary 4.** Suppose $g$ is $L_g$-smooth and $\ell$ is $G_\ell$-Lipschitz continuous and both are convex. Under the same conditions as in Theorem 3, we have $E[\text{dist}(0, \nabla F(x_{\tau}))] \leq \epsilon$ after $K = O(\epsilon^{-\frac{1}{\gamma}})$ stages. Therefore, the total iteration complexity is $\sum_{k=1}^K (T_k^\gamma + T_k^\alpha) = O(\epsilon^{-\frac{1}{\gamma}})$.

**Corollary 5.** Suppose $g$ is non-smooth and convex, $\ell$ is $G_\ell$-Lipschitz continuous and $L_\ell$-smooth and convex, and $\max_{y \in \text{dom}(h)} \|y\| \leq D$. Under the same conditions as in Theorem 3, we have $E[\text{dist}(0, \nabla F(x_{\tau}))] \leq \epsilon$ and $E[\|x_{\tau+1} - v_{\tau}\|] \leq O(\epsilon^{\frac{1}{\mu}})$ after $K = O(\epsilon^{-\frac{1}{\gamma}})$ stages. Therefore, the total iteration complexity is $\sum_{k=1}^K (T_k^\gamma + T_k^\alpha) = O(\epsilon^{\frac{1}{\gamma}})$.

**Remark:** Our algorithms enjoy the same iteration complexity of that in (Xu et al., 2018a) for DC functions when $v$ is unknown or $v = 1$, but we do not assume a stochastic gradient of $h^*(\ell(x))$ is easily computed. It is also notable that St-SPG does not need the knowledge of $v$ to run.

Finally, we would like to mention that the SPG algorithm for solving subproblems in Algorithm 2 can be replaced by other suitable stochastic optimization algorithms for solving a strongly convex problem similar to the developments in (Xu et al., 2018a) for minimizing DC functions. For example, one can use adaptive stochastic gradient methods in order to enjoy an adaptive convergence, and one can use variance reduction methods if the involved functions are smooth and have a finite-sum structure to achieve an improved convergence.

## 5. Application for Variance Regularization

| Datasets | #Examples | #Features | #pos | #neg |
|----------|-----------|-----------|------|------|
| a9a      | 32,561    | 123       | 0.3172:1 |
| covtype  | 581,012   | 54        | 1.0509:1 |
| RCV1     | 697,641   | 47,236    | 1.1033:1 |
| URL      | 2,396,130 | 3,231,961 | 0.4939:1 |

In this section, we consider the application of the proposed algorithms for variance-based regularization in machine learning. Let $l(\theta, z) \in \mathbb{R}^+$ denote a loss of model $\theta \in \Theta$ on a random data $z$. A fundamental task in machine learning is to minimize the expected risk $R(\theta) = E_\theta[l(\theta, z)]$. However, in practice one has to find an approximate model based on sampled data $S_n = \{z_1, \ldots, z_n\}$. An advanced learning theory according to Bennett’s inequality bounds the expected risk by (Maurer & Pontil, 2009):

\[
R(\theta) \leq \frac{1}{n} \sum_{i=1}^n l(\theta, z_n) + c_1\sqrt{\frac{\text{Var}(\ell(\theta, z))}{n}} + \frac{c_2}{n},
\]

where $c_1$ and $c_2$ are constants. This motivates the variance-based regularization approach (Maurer & Pontil, 2009):

\[
\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n l(\theta, z_n) + \lambda\sqrt{\frac{\text{Var}_n(\theta, S_n)}{n}},
\]

where $\text{Var}_n(\theta, S_n) = \frac{1}{n} \sum_{i=1}^n (\ell(\theta, z_i) - \bar{\ell}_n(\theta))^2$ is the empirical variance of loss, $\bar{\ell}_n(\theta)$ is the average of empirical loss, and $\lambda > 0$ is a regularization parameter.

However, the above formulation does not favor efficient stochastic algorithms. To tackle the optimization problem
for variance-based regularization, (Namkoong & Duchi, 2017) proposed a min-max formulation based on distributionally robust optimization, given below and proposed stochastic algorithms for solving the resulting min-max formulation when the loss function is convex (Namkoong & Duchi, 2016),

$$\min_{\theta \in \Theta} \max_{P \in \Delta_n} \left\{ \sum_{i=1}^{n} P_i \ell(\theta, X_i) : D_\phi(P||\hat{P}_n) \leq \rho \right\},$$  

(9)

where $\rho > 0$ is a hyper-parameter, $\Delta_n = \{ P \in \mathbb{R}^n; P \geq 0, \sum_{i=1}^{n} P_i = 1 \}$, $\hat{P}_n = (1/n, \ldots, 1/n)$, and $D_\phi(P||Q) = \int \phi(dP/dQ) dQ$ is the Legendre-Fenchel dual of the $\phi$-divergence based on $\phi(t) = \frac{1}{2} t^2 (t - 1)^2$. The min-max formulation is convex and concave when the loss function is convex. Nevertheless, the stochastic optimization algorithms proposed for solving the min-max formulation are not scalable. The reason is that it introduces an $n$-dimensional dual variable $P$ that is restricted on a probability simplex. As a result, the per-iteration cost could be dominated by updating the dual variable that scales as $O(n)$, which is prohibitive when the training set is large. Although one can use a special structure and a stochastic coordinate update on $P$ to reduce the per-iteration cost to $O(\log(n))$ (Namkoong & Duchi, 2016), the iteration complexity could be still blown up by a factor up to $n$ due to the variance in the stochastic gradient on $P$.

As a potential solution to addressing the scalability issue, we consider the following reformulation:

$$F(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(\theta, z_i) + \lambda \frac{\text{Var}_n(\theta, S_n)}{n}$$

$$= \min_{\alpha > 0} \alpha \sum_{i=1}^{n} l(\theta, z_i) + \lambda \left( \frac{\text{Var}_n(\theta, S_n)}{2\alpha} + \frac{\alpha}{2n} \right)$$

$$= \min_{\alpha > 0} \alpha \sum_{i=1}^{n} l(\theta, z_i) + \lambda \left( \frac{\alpha}{2n} \right)$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \left( (\ell(\theta, z_i)) - (E_i[\ell(\theta, z_i)]) \right)^2 \frac{1}{2\alpha}.$$  

(10)

In practice, one usually needs to tune the regularization parameter $\lambda$ in order to achieve the best performance. As a result, we can further simplify the problem by absorbing $\alpha$ into the regularization parameter $\lambda$ and end up with the following formulation by noting $-\frac{1}{2} s^2 = \max_{y \geq 0} \frac{1}{2} y^2 - y s$ for $s \geq 0$:

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} l(\theta, z_i) + \lambda \left( \frac{1}{2n} \sum_{i=1}^{n} (\ell(\theta, z_i))^2 \right)$$

$$+ \lambda \left( \min_{y \geq 0} \frac{1}{2} y^2 - y \sum_{i=1}^{n} l(\theta, z_i) \right).$$  

(11)

It is notable that the above formulation only introduces one additional scalable variable $y \in \mathbb{R}^+$, though the problem might become a non-convex problem of $\theta$. However, when the loss function $l(\theta, z)$ itself is a non-convex function, the min-max formulation (9) also loses its convexity, which makes our inf-projection formulation more favorable.

We conduct experiments to verify the efficacy of the inf-projection formulation and proposed stochastic algorithms in comparison to the stochastic algorithms for solving min-max formulation (9). We perform two experiments on four datasets, i.e., a9a, RCV1, covtype and URL from the libsvm website, whose number of examples are $n = 32561, 581012, 697641$ and $2396130$, respectively (Table 2). For each dataset, we randomly sample 80% as training data and the rest as testing data. We evaluate training error and testing error of our algorithms and baselines versus cpu time.

In the first experiment, we use (convex) logistic loss for $l(\theta, z_i)$ in our inf-projection formulation (11) and min-max formulation (9). We compare our St-SPG with the stochastic algorithm Bandit Mirror Descent (BMD) proposed in (Namkoong & Duchi, 2016). We implement two versions of BMD, one using the standard mirror descent method to update the dual variable $P$ and the other (denoted by BMD-eff) exploiting binary search tree (BST) to update the $P$. To this end, it needs to use a modified constraint on $P$, i.e., $P \in \{ p \in \mathbb{R}_+^n; p_i \geq \delta/n, n^2/2\|p-1/n\|^2 \leq \rho \}$ (see Sec. 4 in (Namkoong & Duchi, 2016)). We tune hyper-parameters from a reasonable range, i.e., for St-SPG, $\lambda \in \{10^{-5.2}; \gamma, \mu \in \{10^{-3.3}\}$. For BMD and BMD-eff, we use step sizes $\eta_P \in \{10^{-8}; \eta_P \in \{10^{-5}\}$ for updating $P$, step size $\eta_\theta \in \{10^{-5}\}$ for updating $\theta$, $\rho \in \{n \times 10^{-3.3}\}$ and fix $\delta = 10^{-5}$. Training and testing errors against cpu time (s) of the three algorithms on four datasets are reported in Figure 1.

In the second experiment, we use (non-convex) truncated logistic loss in (11) and (9). In particular, the truncated loss function is given by $\phi(l((\theta, z_i))) = \alpha \log(1 + l((\theta, z_i))/\alpha)$, where $l$ is logistic loss and we set $\alpha = \sqrt{10n}$ as suggested in (Xu et al., 2018b). Since the loss is non-convex, we compare MSPG with proximally guided stochastic mirror descent (PGSMD) (Rafique et al., 2018) and its efficient variant (denoted by PGSMD-eff) for solving the min-max formulation that is non-convex and concave, where the efficient variant is implemented with the same modified constraint on $P$ and BST as BMD-eff. For MSPG, we tune $\lambda \in \{10^{-5.2}\}$, the step size parameter $c$ in Proposition 1 from $\{10^{-5.2}\}$. Hyper-parameters of PGSMD and PGSMD-eff including $\eta_P, \eta_\theta, \rho$ and $\delta$ are selected in the same range as in the first experiment. The weak convexity parameter $\rho_{wc}$ are chosen from $\{10^{-5.5}\}$. Training and testing errors against cpu time (s) of the three algorithms on four datasets are reported in Figure 2.

We can observe two conclusions from the results of both experiments. First, the training and testing errors from solving the inf-projection formulation (11) converge to a
close or even a lower level compared to that from solving the min-max formulation (9), which verifies the efficacy of the inf-projection formulation. Second, the proposed stochastic algorithms have significant improvement in the convergence time of training/testing errors, especially on large datasets, covtype, RCV1 and URL, which can be verified by comparing convergence of training/testing errors against cpu time.

6. Conclusion

In this paper, we design and analyze stochastic optimization algorithms for a family of inf-projection minimization problems. We show that the concerned inf-projection structure covers a variety of special cases, including DC functions and bi-convex functions as special cases (non-smooth functions in Section 4) and another family of inf-projection formulations (smooth functions in Section 3). We develop stochastic optimization algorithms for those problems with theoretical guarantees of their first-order convergence for finding a (nearly) \( \epsilon \)-stationary solution at \( O(1/\epsilon^{4/\nu}) \). To the best of our knowledge, this is the first work to provide comprehensive convergence analysis for stochastic optimization of non-convex inf-projection minimization problems. Additionally, to verify the significance of our inf-projection formulation, we investigate an important machine learning problem, variance-based regularization, and compare our algorithms with baselines for min-max formulation (distributionally robust optimization). Empirical results demonstrate the significance and effectiveness of our proposed algorithms.
Stochastic Optimization for Non-convex Inf-Projection Problems

References

Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1-2):459–494, August 2014. ISSN 0025-5610.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Chen, Z., Yang, T., Yi, J., Zhou, B., and Chen, E. Universal stagewise learning for non-convex problems with convergence on averaged solutions. *CoRR*, abs/1808.06296, 2018.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *CoRR*, abs/1803.06523, 2018a.

Davis, D. and Drusvyatskiy, D. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *CoRR*, abs/1802.02988, 2018b.

Davis, D., Grimmer, B. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *arXiv preprint arXiv:1707.03505*, 2017.

Davis, D., Edmunds, B., and Udell, M. The sound of apalm clapping: Faster nonsmooth nonconvex optimization with stochastic asynchronous palm. In *Advances in Neural Information Processing Systems*, pp. 226–234, 2016.

Driggs, D., Tang, J., Davies, M., and Schönlieb, C.-B. Spring: A fast stochastic proximal alternating method for non-smooth non-convex optimization. *arXiv preprint arXiv:2002.12266*, 2020.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program.*, 155(1-2):267–305, 2016.

Gorski, J., Pfeuffer, F., and Klamroth, K. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, 66(3):373–407, Dec 2007.

Hong, M., Razaviyayn, M., Luo, Z.-Q., and Pang, J.-S. A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing. *IEEE Signal Processing Magazine*, 33(1):57–77, 2015.

Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 1675–1685. Curran Associates, Inc., 2017.

Kumar, M. P., Packer, B., and Koller, D. Self-paced learning for latent variable models. In *Neural Information Processing Systems* 23, pp. 1189–1197, 2010.

Maurer, A. and Pontil, M. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2208–2216, 2016.

Namkoong, H. and Duchi, J. C. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2975–2984, 2017.

Nesterov, Y. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, Aug 2015. ISSN 1436-4646. doi: 10.1007/s10107-014-0790-0. URL https://doi.org/10.1007/s10107-014-0790-0.

Nitanda, A. and Suzuki, T. Stochastic difference of convex algorithm and its application to training deep boltzmann machines. In *Artificial Intelligence and Statistics*, pp. 470–478, 2017.

Rafique, H., Liu, M., Lin, Q., and Yang, T. Non-convex minimax optimization: Provable algorithms and applications in machine learning. *CoRR*, abs/1810.02060, 2018.

Rockafellar, R. and Wets, R. J.-B. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Shah, S., Yadav, A. K., Castillo, C. D., Jacobs, D. W., Studer, C., and Goldstein, T. Biconvex relaxation for semidefinite programming in computer vision. In *ECCV (6)*, volume 9910 of *Lecture Notes in Computer Science*, pp. 717–735. Springer, 2016.

Shalev-Shwartz, S. and Singer, Y. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. *Machine learning*, 80(2-3):141–163, 2010.

Thi, H. A. L., Le, H. M., Phan, D. N., and Tran, B. Stochastic DCA for the large-sum of non-convex functions problem and its application to group variable selection in classification. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3394–3403, International Convention Centre, Sydney, Australia, 06–11 Aug
Xu, Y. and Yin, W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

Xu, Y., Qi, Q., Lin, Q., Jin, R., and Yang, T. Stochastic optimization for dc functions and non-smooth non-convex regularizers with non-asymptotic convergence. *arXiv preprint arXiv:1811.11829*, 2018a.

Xu, Y., Zhu, S., Yang, S., Zhang, C., Jin, R., and Yang, T. Learning with non-convex truncated losses by SGD. *CoRR*, abs/1805.07880, 2018b. URL http://arxiv.org/abs/1805.07880.

Xu, Y., Jin, R., and Yang, T. Stochastic proximal gradient methods for non-smooth non-convex regularized problems. *CoRR*, abs/1902.07672, 2019.

Zhao, P. and Zhang, T. Stochastic optimization with importance sampling for regularized loss minimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 1–9, 2015.