On the convergence of stochastic bi-level gradient methods

Nicolas Couellan * Wenjuan Wang

Institut de Mathématiques de Toulouse; UMR5219,
Université de Toulouse ; CNRS, UPS IMT,
F-31062 Toulouse Cedex 9, France

February 15, 2016

Abstract

We analyze the convergence of stochastic gradient methods for bi-level optimization problems. We address two specific cases: first when the outer objective function can be expressed as a finite sum of independent terms, and next when both the outer and inner objective functions can be expressed as finite sums of independent terms. We assume Lipschitz continuity and differentiability of both objectives as well as convexity of the inner objective and consider diminishing steps sizes. We show that, under these conditions and some other assumptions on the implicit function and the variance of the gradient errors, both methods converge in expectation to a stationary point of the problem. We also discuss the satisfaction of our assumptions in machine learning problems where these methods can be nicely applied to automatically tune hyperparameters when the loss functions are very large sums of error terms.

1 Introduction

We consider bi-level optimization problems of the following form:

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & F(y) \\ s.t. & y(x) = \operatorname{argmin}\limits_{\bar{y}\in\mathbb{R}^m} \quad G(x,\bar{y}) \end{array} \tag{1}$$

in which $F(y) : \mathbb{R}^m \to \mathbb{R}$, $G(x,y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, and $x \in \mathbb{R}^n, y \in \mathbb{R}^m$. We assume that n and m are large, F or both F and G are finite sums of independent terms and we have prior knowledge on the regularity of F, G and their

*Email: nicolas.couellan@math.univ-toulouse.fr; Corresponding author

gradients (see Assumptions A1-A4).

There are many applications of bi-level optimization [2]. Bi-level programming problems are generally difficult to solve when little is known on the objectives functions [1]. One common method is to replace the inner problem by its KKT optimality conditions. Descent techniques based on steepest or trust region steps have also been proposed [9, 1]. In recent years, bi-level optimization problems in the form of (1) have been proposed as a framework to model parameter selection in machine learning [4, 5, 8, 6]. The inner problem consists in minimizing a regularized empirical risk for given values of model hyperparameters while the outer problem minimizes a validation error on unseen data over the complete set of hyperparameter values. The volumes of datasets that one has to deal with are often large, leading to large scale bi-level optimization problems.

In machine learning problems, stochastic gradient methods have been the main battle horse to address large scale data. As the objective function can be separated into one regularization term and a large sum of loss terms, the idea is to perform successive optimization moves with respect to one or several randomly chosen data points at a time. Under right assumptions, the convergence in expectation of the minimization process can be proven. In [4, 5], similar ideas have been proposed to design a stochastic gradient algorithm for the specific case of bi-level optimization where both inner and outer objectives can be seen as large finite sums. Results show significant training time reduction when compared to other state-of-the-art techniques. In this article, we propose to analyze the convergence properties of these algorithms. Our initial motivation resides in machine learning applications, however the results are also valid for any problem of the form of (1) satisfying the following assumptions on functions F, G and $y : x \to y(x)$:

Assumptions:

- A1 The function $F(y) : y \to F(y)$ is Lipschitz continuous with constant L_F and has Lipschitz gradient with constant $L_{\nabla_y F}$.
- A2 The function $G(x, y) : (x, y) \to G(x, y)$ is twice differentiable, strictly convex and it is Lipschitz continuous with constant L_G. Its gradient ∇G is Lipschitz continuous with constant $L_{\nabla G}$.
- A3 The function $x \to F(y(x))$ is bounded below.
- A4 The fonction $y :\to y(x)$ has Lipschitz gradient with constant $L_{\nabla_x y}$

Assumption A1 requires Lipschitz regularity on F and its gradient with respect to the variable y. Assumption A2 requires similar regularity on G and and also strict convexity to ensure that the inner problem has a unique solution. Relaxing this assumption would make the bi-level problem (1) a much more complex problem as the solution set of the inner problem would not be a single point but a continuous or discrete set of points. The results that we will develop here

would therefore not be valid anymore. Assumption A3 requires that there exists a solution to the problem whereas assumption A4 necessitates also regularity of the gradient of the implicit function defined by y, the solution of the inner problem, as a function of x. In the last part of the article, we check the satisfaction of these assumptions in specific machine learning applications.

Two algorithms are considered: the bi-level stochastic gradient algorithms with outer approximation of function F when F can be decomposed into a sum of independent F_i ($i \in \{1, \dots, N\}$) and the bi-level stochastic algorithm with inner and outer approximations where both outer and inner objectives functions can be decomposed into a sum of independent terms (i.e. $F = \frac{1}{N}\sum_{i=1}^{N} F_i$ and $G = \frac{1}{J}\sum_{j=1}^{J} G_j$). For these two cases, we consider bi-level techniques based on stochastic gradient methods. The methods perform optimization moves along a stochastic estimate of the gradient of the outer objective function with respect to the outer variable x. The estimate is computed by taking only one random term of the objective (or both objectives if both are decomposable) and making use of bi-level differentiation. We show, that under the assumptions on F and G above, that both methods converge in expectation towards a stationary point of Problem (1).

The article is organized as follows: In Section 2, we first state a general result that we will use throughout the sequel of the article. In Section 3, we prove the convergence of the bi-level stochastic gradient technique with outer approximation. Next, in Section 4, we prove the convergence of the bi-level stochastic gradient technique with inner and outer approximations. Section 5 discusses the application of these convergence results in the machine learning context. Section 6 gives some concluding remarks.

2 Preliminary results

Under the assumptions (A1)-(A4) on functions F, G and y, we state and prove two intermediate results.

2.1 Bi-level differentiation

We first calculate the gradient of the outer objective function in Problem (1) with respect to the variable x using the chain rule for derivatives:

$$\nabla_x \left[F(y(x))\right] = \nabla_y F(y)^\top \nabla y(x). \tag{2}$$

Recall that the implicit function theorem (IFT) [12] states that, if:

- (x^*, y^*) is an optimal solution of the inner problem in (1), meaning that $\nabla_y G(x^*, y^*) = 0$,
- G is C^2 and $\nabla_y^2 G(x^*, y^*)$ is invertible,

there exists an open set $U \subset \mathbb{R}^n$, an open set $V \subset \mathbb{R}^m$ such that $(x^*, y^*) \in U \times V$ and a C^1-function y such that:

- $\forall (u, v) \in U \times V$, $\nabla_v G(u, v) = 0 \Rightarrow v = y(u)$.
- $\forall u \in U$, we have $\nabla_v G(u, y(u)) = 0$.
- $\forall (u, v) \in U \times V$, the matrix $\nabla_v^2 G(u, v)$ is invertible and furthermore,

$$\nabla y(u) = -\left[\nabla_v^2 G(u, y(u))\right]^{-1} \nabla_{vu}^2 G(u, y(u)) \tag{3}$$

Therefore, we can write

$$\nabla_x \left[F(y(x))\right] = -\nabla_y F(y)^\top \left[\nabla_y^2 G(x, y)\right]^{-1} \nabla_{xy}^2 G(x, y). \tag{4}$$

The strict convexity of G ensures a unique solution of $\nabla_y G(x, y) = 0$ and therefore the possibility to express $\nabla_y(x)$ uniquely everywhere, meaning that we can replace the constrained bi-level problem by an unconstrained optimization problem by expressing y as a function of x.

2.2 Lipschitz differentiability of $x \to F(y(x))$

Here, we use the previous result to prove that the implicit function $y : x \to y(x)$ is Lipschitz continuous and that the function $x \to F(y(x))$ is Lipschitz differentiable. This last result will be important in the analysis of convergence of the bi-level stochastic gradient methods as we will see in Section 3 and 4.

Lemma 2.1 *Under assumption A2 above, the implicit function y defined by $y : x \to y(x)$ is Lipschitz continuous.*

Proof We have

$$\nabla_x y(x) = -\left[\nabla_y^2 G(x, y)\right]^{-1} \nabla_{xy}^2 G(x, y) \tag{5}$$

Since ∇G is Lipschitz continuous, we have that $\left\|\nabla^2 G(x, y)\right\|$ is bounded, meaning that $\left\|\nabla_y^2 G(x, y)\right\|$ and $\left\|\nabla_{xy}^2 G(x, y)\right\|$ are also bounded. In (5), G being strictly convex, $\nabla_y^2 G(x, y)$ is non singular and since $\left\|\nabla_{xy}^2 G(x, y)\right|$ is bounded, we have that $\|\nabla_x y(x)\|$ is also bounded, proving that y is Lipschitz continuous. ∎

Lemma 2.2 *Assuming A1, A2, A4 above, the function defined by $F : x \to F(y(x))$ is differentiable with Lipschitz continuous gradient and Lipschitz constant $L_y^2 L_{\nabla_y F} + L_F L_{\nabla_x y}$.*

Proof Clearly, from the definition of $F : x \to F(y(x))$ as a composition of the differentiable function $y \to F(y)$ and $y : x \to y(x)$ (where the existence of $\nabla_x y(x)$ is ensured by IFT), $F : x \to F(y(x))$ is differentiable.

Additionally, $\forall (x, x') \in \mathbb{R}^{n \times n}$, we have

$$\begin{aligned}
\|\nabla_x [F(y(x))] - \nabla_x [F(y(x'))]\| &= \left\|\nabla_y F(y(x))^\top \nabla_x y(x) - \nabla_y F(y(x'))^\top \nabla_x y(x')\right\| \\
&= \left\|\nabla_y F(y(x))^\top \nabla_x y(x) - \nabla_y F(y(x'))^\top \nabla_x y(x)\right. \\
&\quad \left. -\nabla_y F(y(x'))^\top \nabla_x y(x') + \nabla_y F(y(x'))^\top \nabla_x y(x)\right\| \\
&\leq \|\nabla_y F(y(x)) - \nabla_y F(y(x'))\| \, \|\nabla_x y(x)\| \\
&\quad + \|\nabla_y F(y(x'))\| \, \|\nabla_x y(x) - \nabla_x y(x')\| .
\end{aligned}$$

Since ∇F is Lipschitz continuous with respect to y, we have

$$\|\nabla_y F(y(x)) - \nabla_y F(y(x'))\| \leq L_{\nabla_y F} \|y(x) - y(x')\|$$

and $\|\nabla_y F(y(x))\|$ is bounded by L_F.

Lemma 2.1 states also that y is Lipschitz continuous, therefore $\exists L_y > 0$ such that $\|y(x) - y(x')\| \leq L_y \|x - x'\|$ and $\|\nabla_x y(x)\| \leq L_y$. Moreover, assumption A4 ensures that $\|\nabla_x y(x) - \nabla_x y(x')\| \leq L_{\nabla_x y} \|x - x'\|$. Using these bounds in the above inequality, we have:

$$\begin{aligned}
\|\nabla_x [F(y(x))] - \nabla_x [F(y(x'))]\| &\leq L_{\nabla_y F} \|y(x) - y(x')\| \, \|\nabla_x y(x)\| \\
&\quad + L_F L_{\nabla_x y} \|x - x'\| \\
&\leq L_y L_{\nabla_y F} L_y \|x - x'\| + L_F L_{\nabla_x y} \|x - x'\| \\
&\leq (L_y^2 L_{\nabla_y F} + L_F L_{\nabla_x y}) \|x - x'\|,
\end{aligned}$$

proving Lemma 2.2. ∎

3 Convergence of the bi-level stochastic gradient method with outer approximation

In this section, we consider outer level objective function of the form :

$$F(y(x)) = \frac{1}{N} \sum_{i=1}^{N} F_i(y(x))$$

If all F_i ($\forall i = 1, \ldots, N$) are Lipschitz continuous and Lipschitz differentiable, assumption A1 is satisfied and Lemma (2.2) applies, meaning that the function $x \to F(y(x))$ is Lipschitz differentiable.

The principle of the bi-level stochastic gradient method with outer approximation (BSG_o) is to randomly choose one $i \in \{1, \ldots, N\}$ at each iteration and use $\nabla_x [F_i(y(x))]$ as an unbiased estimate of $\nabla_x [F(y(x))]$ to compute a stochastic move. The BSG_o is summarized in Algorithm (1).

Algorithm 1 BSG_o Algorithm

1: Choose x_0 and $\alpha_0 > 0$
2: $k \leftarrow 0$
3: **while** $\|\nabla_x [F_i(y(x_k))]\| \geq 0$ **do**
4: Pick i randomly and uniformly in $\{1, \ldots, N\}$
5: Compute

$$\nabla_x [F_i(y(x_k))] = -\nabla_y F_i(y(x_k))^\top \left[\nabla_y^2 G(x_k, y(x_k))\right]^{-1} \nabla_{xy}^2 G(x_k, y(xk))$$

6: $x_{k+1} \leftarrow x_k - \alpha_k \nabla_x [F_i(y(x_k))]$
7: Update α_k
8: $k \leftarrow k + 1$
9: **end while**

At each iteration k, let ε_k be the error between the estimate $\nabla_x [F_i(y(x_k))]$ and the true gradient $\nabla_x [F(y(x_k))]$,

$$\varepsilon_k = \nabla_x [F_i(y(x_k))] - \nabla_x [F(y(x_k))]$$

We state and prove the following convergence theorem:

Theorem 3.1 *Suppose that:*

1. *Assumptions A1-A4 are satisfied,*
2. *$\exists D > 0$ such that $\forall k > 0$, ε_k satisfies the following inequality*

$$E\left[\|\varepsilon_k\|^2\right] \leq D \|\nabla_x [F(y(x_k))]\|^2,$$

3. *$\forall k > 0$, α_k is chosen such that*

$$\sum_{k=0}^{\infty} \alpha_k = \infty \qquad \text{and} \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty,$$

Then the sequence $\{x_k\}$ generated by the BSG_o algorithm converges in expectation to a stationary point of the function $x \to F(y(x))$, i.e.

$$\lim_{k \to \infty} E[\|\nabla_x [F(y(x_k)] \|] = 0$$

Proof let x_k be a sequence of iterates generated by BSG_o, we have

$$x_{k+1} = x_k - \alpha_k \nabla_x [F_i(y(x_k))]$$

where i is randomly chosen in $\{1, \ldots, N\}$ and

$$\nabla_x \left[F_i(y(x))\right] = -\nabla_y F_i(y)^\top \left[\nabla_y^2 G(x,y)\right]^{-1} \nabla_{xy}^2 G(x,y).$$

Given the current iterate x_k, picking randomly one i and a direction $-\alpha_k \nabla_x \left[F_i(y(x_k))\right]$ leads to many choices of possible moves. On the average, we have:

$$\begin{aligned} E\left[\nabla_x \left[F_i(y(x_k))\right] | x_k\right] &= \frac{1}{N} \sum_{i=1}^{N} \nabla_x \left[F_i(y(x_k))\right] \\ &= \nabla_x \left[\frac{1}{N} \sum_{i=1}^{N} F_i(y(x_k))\right] \\ &= \nabla_x \left[F(y(x_k))\right]. \end{aligned}$$

We therefore see that, at x_k, $\nabla_x \left[F_i(y(x_k))\right]$ is an unbiased estimate of $\nabla_x \left[F(y(x_k))\right]$. This implies that:

$$E[\varepsilon_k | x_k] = E\left[\nabla_x \left[F_i(y(x_k))\right] | x_k\right] - \nabla_x \left[F(y(x_k))\right] = 0$$

From Lemma 2.2, we know that $\nabla_x F$ is Lipschitz continuous with Lipschitz constant $L_{\nabla_x F} = L_y^2 L_{\nabla_y F} + L_F L_{\nabla_x y}$. Therefore, using from now on the notation $< u, v >$ for the inner product of two vectors u and v, we can write the following inequality

$$\begin{aligned} E[F(y(x_{k+1}))|x_k] &\leq E\left[F(y(x_k))|x_k\right] + E\left[< \nabla_x \left[F(y(x_k))\right], x_{k+1} - x_k > |x_k\right] \\ &\quad + \frac{L_{\nabla_x F}}{2} E\left[\|x_{k+1} - x_k\|^2 | x_k\right] \\ &= F(y(x_k)) + < E\left[\nabla_x \left[F(y(x_k))\right] | x_k\right], -\alpha_k E\left[\nabla_x \left[F_i(y(x_k))\right] | x_k\right] > \\ &\quad + \frac{L_{\nabla_x F}}{2} \alpha_k^2 E\left[\|\nabla_x \left[F_i(y(x_k))\right]\|^2 | x_k\right] \\ &= F(y(x_k)) + < \nabla_x \left[F(y(x_k))\right], -\alpha_k E\left[\nabla_x \left[F(y(x_k))\right] + \varepsilon_k | x_k\right] > \\ &\quad + \frac{L_{\nabla_x F}}{2} \alpha_k^2 E\left[\|\nabla_x \left[F(y(x_k))\right] + \varepsilon_k\|^2 | x_k\right]. \end{aligned}$$

Since $E[\varepsilon_k | x_k] = 0$, we can write

$$\begin{aligned} E[F(y(x_{k+1}))|x_k] &\leq F(y(x_k)) - \alpha_k < \nabla_x \left[F(y(x_k))\right], \nabla_x \left[F(y(x_k))\right] > \\ &\quad + \frac{L_{\nabla_x F}}{2} \alpha_k^2 E\left[\|\nabla_x \left[F(y(x_k))\right] + \varepsilon_k\|^2 | x_k\right] \\ &= F(y(x_k)) - \alpha_k \|\nabla_x \left[F(y(x_k))\right]\|^2 \\ &\quad + \frac{L_{\nabla_x F}}{2} \alpha_k^2 E\left[\|\nabla_x \left[F(y(x_k))\right]\|^2 | x_k\right] + \quad \frac{L_{\nabla_x F}}{2} \alpha_k^2 E\left[\|\varepsilon_k\|^2 | x_k\right]. \end{aligned}$$

Taking the expectation again over all realizations of the random variable x_k, we get

$$\begin{aligned} E[F(y(x_{k+1}))] &\leq E\left[F(y(x_k))\right] - \alpha_k E\left[\|\nabla_x \left[F(y(x_k))\right]\|^2\right] \\ &\quad + \frac{L_{\nabla_x F}}{2} \alpha_k^2 E\left[\|\nabla_x \left[F(y(x_k))\right]\|^2\right] + \frac{L_{\nabla_x F}}{2} \alpha_k^2 E\left[\|\varepsilon_k\|^2\right]. \quad (6) \end{aligned}$$

From the fact that $E\left[\|\varepsilon_k\|^2\right] \leq D\|\nabla_x \left[F(y(x_k))\right]\|^2$, we have,

$$E\left[F(y(x_{k+1}))\right] \leq E\left[F(y(x_k))\right] - \alpha_k \left(1 - \alpha_k \frac{L_{\nabla_x F} + D}{2}\right) E\left[\|\nabla_x \left[F(y(x_k))\right]\|^2\right]. \tag{7}$$

Observe that if $\forall k > 0$, α_k is chosen so as to ensure that $0 < \alpha_k < \frac{2}{L_{\nabla_x F} + D}$, then the sequence $\{E[F(y(x_{k+1}))]\}$ is decreasing. As α_k is decreasing, it also implies that for sufficiently large k, $\{E[F(y(x_k))]\}$ will decrease and converge to its infimum as F is bounded below (monotone convergence theorem).

In the remaining part of the proof, we will show that the expected limit point of the sequence $\{x_k\}$ is a stationary point of the function $x \to F(y(x))$.

Applying the above inequality (7) to pairs of iterates starting from (x_1, x_2) to some iterates (x_{K-1}, x_K) for any $K > 2$, we get:

$$\begin{aligned}
E\left[F(y(x_0))\right] - E\left[F(y(x_1))\right] &\geq \alpha_0 \left(1 - \alpha_0 \frac{L_{\nabla_x F} + D}{2}\right) E\left[\|\nabla_x \left[F(y(x_0))\right]\|^2\right] \\
E\left[F(y(x_1))\right] - E\left[F(y(x_2))\right] &\geq \alpha_1 \left(1 - \alpha_1 \frac{L_{\nabla_x F} + D}{2}\right) E\left[\|\nabla_x \left[F(y(x_1))\right]\|^2\right] \\
&\cdots \\
E\left[F(y(x_{K-1}))\right] - E\left[F(y(x_K))\right] &\geq \alpha_{K-1} \left(1 - \alpha_{K-1} \frac{L_{\nabla_x F} + D}{2}\right) \\
&\qquad \times E\left[\|\nabla_x \left[F(y(x_{K-1}))\right]\|^2\right]
\end{aligned}$$

Summing up all the above inequalities, we obtain the following,

$$E\left[F(y(x_0))\right] - E\left[F(y(x_K))\right] \geq \sum_{k=1}^{K-1} \alpha_k \left(1 - \alpha_k \frac{L_{\nabla_x F} + D}{2}\right) E\left[\|\nabla_x \left[F(y(x_k))\right]\|^2\right]$$

From assumption A3, $x \to F(y(x))$ is bounded below. This implies that $E\left[F(y(x_0))\right] - E\left[F(y(x_K))\right]$ is bounded above and $\exists M > 0$ such that $E\left[F(y(x_0))\right] - E\left[F(y(x_K))\right] \leq M$. Hence we can bound the sum in the above inequality as follows

$$\sum_{k=1}^{K-1} \alpha_k \left(1 - \alpha_k \frac{L_{\nabla_x F} + D}{2}\right) E\left[\|\nabla_x \left[F(y(x_{k-1}))\right]\|^2\right] \leq M \quad . \tag{8}$$

Let now $s_k = \alpha_k(1 - \frac{L_{\nabla_x F}}{2}\alpha_k)$. Since $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, s_k satisfies $\sum_{k=0}^{\infty} s_k = \infty$. Taking K to ∞ in (8), we can write

$$\sum_{k=0}^{\infty} s_k E\left[\|\nabla_x \left[F(y(x_{k-1}))\right]\|^2\right] \leq M < \infty, \tag{9}$$

Assume now that $\exists \hat{\epsilon} > 0$ and $\bar{k} \in \mathbb{N}$ such that $\forall k \geq \bar{k}$,

$$E\left[\|\nabla_x \left[F(y(x_{k-1}))\right]\|^2\right] \geq \hat{\epsilon}, \tag{10}$$

implying

$$\sum_{k=0}^{\infty} s_k E\left[\|\nabla_x \left[F(y(x_{k-1}))\right]\|^2\right] \geq \hat{\epsilon} \sum_{k=0}^{\infty} s_k = \infty. \tag{11}$$

The inequality (11) contradicts inequality (8) meaning that the assumption (10) is false. Therefore,

$$\liminf_{k\to\infty} E\left[\|\nabla_x \left[F(y(x_k))\right]\|\right] = 0,$$

Following a similar line of reasoning as in [3], we will now prove that

$$\limsup_{k\to\infty} E\left[\|\nabla_x \left[F(y(x_k))\right]\|\right] = 0.$$

Assume the contrary is true. This means that $\exists \check{\epsilon} > 0$ and $\tilde{k} \in \mathbb{N}$ such that $\forall k \geq \tilde{k}$, $\exists i^{(k)}$ satisfying

$$\begin{cases} E\left[\|\nabla_x \left[F(y(x_k))\right]\|\right] < \check{\epsilon}/2 \\ \check{\epsilon}/2 \leq E\left[\|\nabla_x \left[F(y(x_l))\right]\|\right] \leq \check{\epsilon} \qquad \forall l \in \mathbb{N} \text{ such that } k < l < i^{(k)} \\ \check{\epsilon} < E\left[\|\nabla_x \left[F(y(x_{i^{(k)}}))\right]\|\right] \end{cases} \tag{12}$$

On one hand, from (12) and Lemma 2.2, observe that

$$\begin{aligned} \frac{\check{\epsilon}}{2} &\leq E\left[\|\nabla_x \left[F(y(x_{i^{(k)}})\right]\|\right] - E\left[\|\nabla_x \left[F(y(x_k)\right]\|\right] \\ &= E\left[\|\nabla_x \left[F(y(x_{i^{(k)}})\right]\| - \|\nabla_x \left[F(y(x_k))\right]\|\right] \\ &\leq E\left[\|\nabla_x \left[F(y(x_{i^{(k)}})\right] - \nabla_x \left[F(y(x_k))\right]\|\right] \\ &\leq L_{\nabla_x F} E\left[\|x_{i^{(k)}} - x_k\|\right] \\ &\leq L_{\nabla_x F} \sum_{l=k}^{i^{(k)}-1} \alpha_l E\left[\|\nabla_x \left[F_l(y(x_l))|x_l\right]\|\right] \qquad (13) \\ &= L_{\nabla_x F} \sum_{l=k}^{i^{(k)}-1} \alpha_l \|\nabla_x \left[F(y(x_l))\right]\| \qquad (14) \end{aligned}$$

Taking the expectation in the right hand side of (14) over all possible realizations of the random variable x_l (for $l = k, \ldots, i^{(k)} - 1$), we obtain

$$\begin{aligned} \frac{\check{\epsilon}}{2} &\leq L_{\nabla_x F} \sum_{l=k}^{i^{(k)}-1} \alpha_l E\left[\|\nabla_x \left[F(y(x_l))\right]\|\right] \\ &\leq \check{\epsilon} L_{\nabla_x F} \sum_{l=k}^{i^{(k)}-1} \alpha_l. \end{aligned}$$

Hence,

$$\liminf_{k\to\infty} \sum_{l=k}^{i^{(k)}-1} \alpha_l \geq \frac{1}{2L_{\nabla_x F}}. \tag{15}$$

On the other hand, from (7) and (12), we can write

$$\begin{aligned} E\left[F(y(x_{i^{(k)}}))\right] &\leq E\left[F(y(x_k))\right] - \sum_{l=k}^{i^{(k)}-1} \alpha_l \left(1 - \alpha_l \frac{L_{\nabla_x F} + D}{2}\right) E\left[\|\nabla_x \left[F(y(x_l))\right]\|^2\right] \\ &\leq E\left[F(y(x_k))\right] - \frac{\check{\epsilon}^2}{4} \sum_{l=k}^{i^{(k)}-1} \alpha_l + \frac{\left(L_{\nabla_x F} + D\right)\check{\epsilon}^2}{2} \sum_{l=k}^{i^{(k)}-1} \alpha_l^2. \end{aligned}$$

Since the sequence $\{E\left[F(y(x_k))\right]\}$ converges, and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, we necessarily have that

$$\lim_{k\to\infty} \sum_{l=k}^{i^{(k)}-1} \alpha_l = 0,$$

which contradicts the statement (15). As a consequence, the statement that *there exists* $\check{\epsilon} > 0$ *such that (12) is satisfied* is false and

$$\limsup_{k\to\infty} E\left[\|\nabla_x \left[F(y(x_k))\right]\|\right] = 0,$$

which completes the proof of the convergence of Theorem 3.1. ∎

Note on assumption (2) in Theorem 3.1:

The assumption that the variance of the noise ε is bounded by $E\left[\|\varepsilon_k\|^2\right] \leq D\|\nabla_x \left[F(y(x_k))\right]\|^2$ has also been considered in [3, 11] and more recently in [13]. Intuitively, it is reasonable to assume that if $\|\nabla_x \left[F(y(x_k))\right]\|$ is small, there is little noise (remember also that $E\left[\varepsilon_k\right] = 0$) and that if $\|\nabla_x \left[F(y(x_k))\right]\|$ is growing, the variance of the noise is growing as well (in proportion to its square).

4 Convergence of the bi-level stochastic gradient method with inner and outer approximation

We now consider the case where both outer and inner objective functions can be expressed as finite sums as follows:

$$F(y(x)) = \frac{1}{N}\sum_{i=1}^{N} F_i(y(x)) \qquad G(x,y) = \frac{1}{J}\sum_{j=1}^{J} G_j(x,y)$$

with G_j having the following special structure:

$$G_j(x,y) = h(x,y) + h_j(x,y) \quad \forall j \in 1,\dots,J \tag{16}$$

and where h_j is a linear function with respect to x and y for all $j \in \{1,\dots,J\}$. It is also assumed that the function h shares the same properties of G as defined in assumptions A2. Section 5 discusses applications where G_j possesses this specific structure.

The principle of the bi-level stochastic gradient method with inner and outer approximations (BSG) is to randomly choose one i in $\{1,\dots,N\}$ and one j in $\{1,\dots,J\}$ at each iteration and use $\nabla_x \left[F_i(y^{(j)}(x))\right]$ (where $y^{(j)}(x) = \operatorname{argmin}_{\bar{y}\in\mathbb{R}^m} G_j(x,\bar{y})$) as an approximation of $\nabla_x [F(y(x))]$. Note, as we will see later, that the special structure (16) ensures that $\nabla_x \left[F_i(y^{(j)}(x))\right]$ is an unbiased estimate of $\nabla_x [F(y(x))]$. The BSG algorithm is summarized in Algorithm (2).

Algorithm 2 BSG Algorithm

Choose x_0 and $\alpha_0 > 0$
$k \leftarrow 0$
while $\|\nabla_x \left[F_i(y^{(j)}(x_k))\right]\| \geq 0$ **do**
  Pick i randomly and uniformly in $\{1,\dots,N\}$
  Pick j randomly and uniformly in $\{1,\dots,J\}$
  Compute $y^{(j)}(x_k) = \underset{\bar{y}\in\mathbb{R}^m}{\operatorname{argmin}} G_j(x_k,\bar{y})$
  Compute

$$\nabla_x \left[F_i(y^{(j)}(x_k))\right] = -\nabla_y F_i(y^{(j)}(x_k)(\nabla_y^2 G_j(x_k, y^{(j)}(x_k)))^{-1}\nabla_{xy} G_j(x_k, y^{(j)}(x_k))$$

  $x_{k+1} \leftarrow x_k - \alpha_k \nabla_x \left[F_i(y^{(j)}(x_k))\right]$
  Update α_k
  $k \leftarrow k+1$
end while

At iteration k, let again ε_k be the error between the gradient estimate and the true gradient $\nabla_x [F(y(x))]$:

$$\varepsilon_k = \nabla_x \left[F_i(y^{(j)}(x))\right] - \nabla_x [F(y(x))] .$$

The convergence result for the BSG algorithm is summarized in the following theorem.

Theorem 4.1 *Suppose that:*

1. *Assumptions A1-A4 are satisfied,*

2. $\exists D > 0$ *such that* $\forall k > 0$, ε_k *satisfies the following inequality*

$$E\left[\|\varepsilon_k\|^2\right] \leq D\|\nabla_x \left[F(y(x_k))\right]\|^2,$$

3. $\forall k > 0$, α_k *is chosen such that*

$$\sum_{k=0}^{\infty} \alpha_k = \infty \qquad \textit{and} \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty,$$

Then the sequence $\{x_k\}$ *generated by the BSG algorithm converges in expectation to a stationary point of the function* $x \to F(y(x))$*, i.e.*

$$\lim_{k\to\infty} E[\|\nabla_x \left[F(y(x_k)\right]\|] = 0$$

Proof The sequence of iterates $\{x_k\}$ generated by BSG can be written as

$$x_{k+1} \leftarrow x_k - \alpha_k \nabla_x \left[F_i(y^{(j)}(x_k))\right]$$

where i and j are randomly chosen in $\{1, ..., N\}$ and $\{1, ..., J\}$ respectively. From the chain rule for differentiation and the application of the IFT as explained in Section 2, we have

$$\nabla_x \left[F_i(y^{(j)}(x_k))\right] = -\nabla_y F_i(y^{(j)}(x_k)(\nabla_y^2 G_j(x_k, y^{(j)}(x_k)))^{-1} \nabla_{xy} G_j(x_k, y^{(j)}(x_k))$$

Again, as for the BSG_o algorithm, at x_k, picking randomly one direction opposite to $\nabla_x \left[F_i(y^{(j)}(x_k))\right]$, leads to many choices of moves. On the average, we would get:

$$E\left[\nabla_x \left[F_i(y^{(j)}(x_k))\right] |x_k\right] = E\left[\begin{array}{r} -\nabla_y F_i(y(x_k) \left(\nabla_y^2 G_j(x_k, y(x_k))\right)^{-1} \\ \times \nabla_{xy}^2 G_j(x_k, y(x_k)) \end{array} |x_k\right].$$

From the special structure of $G_j(x, y) = h(x, y) + h_j(x, y)$, we see that $\forall (x, y) \in \mathbb{R}^{n\times m}$, $\nabla_y^2 G(x, y) = \nabla_y^2 G_j(x, y)$, hence

$$\begin{aligned} E\left[\nabla_x \left[F_i(y^{(j)}(x_k))\right] |x_k\right] &= -\frac{1}{N}\sum_{i=1}^{N} \nabla_y F_i(y(x_k)) \left(\nabla_y^2 G(x_k, y(x_k))\right)^{-1} \\ &\qquad \times \frac{1}{J}\sum_{j=1}^{J} (\nabla_{xy}^2 G_j(x_k, y(x_k))) \\ &= -\nabla_y F(y(x_k)) \left(\nabla_y^2 G(x_k, y(x_k))\right)^{-1} \nabla_{xy}^2 G(x_k, y(x_k)) \\ &= \nabla_x \left[F(y(x_k))\right], \end{aligned} \tag{17}$$

which shows that, at the point x_k, $\nabla_x \left[F_i(y^{(j)}(x_k))\right]$ is an unbiased estimate of $\nabla_x \left[F(y(x_k))\right]$, meaning also that

$$E[\varepsilon_k | x_k] = E\left[\nabla_x \left[F_i(y^{(j)}(x_k))\right] |x_k\right] - \nabla_x \left[F(y(x_k))\right] = 0. \tag{18}$$

Lemma 2.2 states that $\nabla_x F$ is Lipschitz continuous (with Lipschitz constant $L_{\nabla_x F} = L_y^2 L_{\nabla_y F} + L_F L_{\nabla_x y}$. Therefore, given x_k, we can bound the value of $F(y(x_{k+1}))$ by a quadratic function above. In expectation, this gives

$$\begin{aligned}
E\left[F(y(x_{k+1}))|x_k\right] &\leq E\left[F(y(x_k))|x_k\right] + E\left[< \nabla_x \left[F(y(x_k))\right], x_{k+1} - x_k > |x_k\right] \\
&+ \frac{L_{\nabla_x F}}{2} E\left[\|x_{k+1} - x_k\|^2 |x_k\right] \\
&= F(y(x_k)) \\
&+ < E\left[\nabla_x \left[F(y(x_k))\right]|x_k\right], -\alpha_k E\left[\nabla_x \left[F_i(y^{(j)}(x_k))\right]|x_k\right] > \\
&+ \frac{L_{\nabla_x F}}{2} \alpha_k^2 E\left[\|\nabla_x \left[F_i(y^{(j)}(x_k))\right]\|^2 |x_k\right] \\
&= F(y(x_k)) \\
&+ < \nabla_x \left[F(y(x_k))\right], -\alpha_k E\left[\nabla_x \left[F(y(x_k))\right] + \varepsilon_k |x_k\right] > \\
&+ \frac{L_{\nabla_x F}}{2} \alpha_k^2 E\left[\|\nabla_x \left[F(y(x_k))\right] + \varepsilon_k\|^2 |x_k\right]
\end{aligned}$$

Using (18) and taking the expectation again over all possible realizations of x_k, the remaining part of the proof is identical to the proof of Theorem 3.1. By exploiting the fact that $E\left[\|\varepsilon_k\|^2\right] \leq D\|\nabla_x \left[F(y(x_k))\right]\|^2$, we can show exactly as before that the sequence $\{E\left[F(y(x_{k+1}))\right]\}$ is decreasing. Observing that, when using inner gradient approximation, the inequality (13) can be re-written as follows

$$\frac{\check{\epsilon}}{2} \leq L_{\nabla_x F} \sum_{l=k}^{i^{(k)}-1} \alpha_l E\left[\|\nabla_x \left[F_l(y^{(l)}(x_l))|x_l\right]\|\right]$$

where F_l and $y^{(l)}$ are two functions picked randomly and uniformly in the sets $\{F_1, \ldots, F_N\}$ and $\{y^{1)}, \ldots, y^{(J)}\}$. Recalling from (17) that

$$E\left[\|\nabla_x \left[F_l(y^{(l)}(x_l))|x_l\right]\|\right] = \|\nabla_x \left[F(y(x_l))\right]\|,$$

and taking the expectation of $\|\nabla_x \left[F(y(x_l))\right]\|$ over all realizations of the random variable x_l, we can write

$$\frac{\check{\epsilon}}{2} \leq L_{\nabla_x F} \sum_{l=k}^{i^{(k)}-1} \alpha_l E\left[\|\nabla_x \left[F(y(x_l))\right]\|\right] \leq \check{\epsilon} L_{\nabla_x F} \sum_{l=k}^{i^{(k)}-1} \alpha_l,$$

and, as before, see that the use of a step length α_k satisfying $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ will also ensure, with the exact same arguments, that

$$\liminf_{k\to\infty} E\left[\|\nabla_x \left[F(y(x_k))\right]\|\right] = \limsup_{k\to\infty} E\left[\|\nabla_x \left[F(y(x_k))\right]\|\right] = 0. \quad \blacksquare$$

Table 1: Piecewise linear loss functions

Loss	ϕ_j
Hinge loss	$\phi_j(z) = max\{0, 1 - y_j z\}$
Absolute deviation loss	$\phi_j(z) = \|y_j - z\|$
ϵ-*insensitive loss*	$\phi_j(z) = max\{0, \|z\| - \epsilon\}$

5 Regularized empirical risk minimization

In this section, we briefly discuss the use of these algorithms and their convergence results in the context of regularized empirical risk minimization (ERM).

Many machine learning problems can be cast as ERM. Basically, one tries to build a model on past observations by minimizing some classification or fitting error. The regularized variant of the problem builds solutions that exhibit nice structure (ex:sparsity) to ensure generalization to unseen data. These problems take the following general form:

$$\min_{\zeta} \left[r(\zeta) + \nu \sum_{j=1}^{J} \phi_j(< \zeta, x_j >) \right]$$

where $x_j \in \mathbb{R}^n$ are the feature vectors of J data points, ϕ_j is a loss function, r a regularization function and $\nu > 0$ an hyperparameter. Table 1 gives examples of ϕ_j that are used for various machine learning problems.
In Problem (5), the trade-off between regularization and classification/fitting is controlled by the hyperparameter ν. Tuning ν when datasets are large (i.e J is large) is a difficult and expensive task if one wants to compute probabilistics bounds or carry out cross-validation procedures (see [7]). For this reason, stochastic bi-level optimization may be preferred [4]. The bi-level problem resulting from learning the hyperparameter ν can be written as follows:

$$\begin{aligned} \min_{\nu} \quad & \sum_{i=1}^{N} \phi_i(< \bar{\zeta}(\nu), x_i^v >) \\ s.t. \quad & \bar{\zeta}(\nu) = \operatorname*{argmin}_{\zeta} \left[r(\zeta) + \nu \sum_{j=1}^{J} \phi_j(< \zeta, x_j >) \right] \end{aligned} \tag{19}$$

where x_i^v for $i \in \{1, \ldots, N\}$ are the validation data (unseen data) on which we are tuning the hyperparameter.

Let us now discuss the applicability of the convergence results of algorithm BSG to Problem (19) where ϕ_j are the ones listed in Table 1 and r is the commonly used squared L_2-norm (i.e. $r(\zeta) = \frac{1}{2}\|\zeta\|_2^2$).

Observe that all proposed loss functions are piecewise linear, therefore G_j in Problem (19) has the form of (16). This ensures that (17) is satisfied and that the gradient estimate computed by BSG is unbiased.

Observe also that for the Support Vector Machine (SVM) hinge loss case, one can replace in the inner and outer objectives $\sum_{j=1}^{J} \phi_j(< \zeta, x_j >)$ and $\sum_{i=1}^{N} \phi_i(< \bar{\zeta}(\nu), x_i^v >)$ by the following sums $\sum_{j=1}^{J_e}(1 - y_j < \zeta, x_j >)$ and $\sum_{i=1}^{N_e}(1 - y_j < \bar{\zeta}(\nu), x_i^v >)$ where J_e and N_e are the number of training and validation error vectors, vectors with non zero losses, as explained in [4]. This way, we get rid of the non differentiability of ϕ_j. In the stochastic approximation practical setting, this only requires checking that the current random pick of data point is an error vector or not, which is computationally inexpensive. For the ϵ-insensitive loss, a simple test on the positivity of $< \zeta, x_j >$ helps also in practice to smoothen the problem.

Clearly, considering differentiable variants of ϕ_j, the functions r and ϕ_j are Lipschitz continuous and Lipschitz differentiable. We can also see that the function $\zeta \rightarrow \left[r(\zeta) + \nu \sum_{j=1}^{J} \phi_j(< \zeta, x_j >) \right]$ is strictly convex, except for the SVM case where strict convexity can be ensured by adding an extra attribute to the data as explained in [10] and solving the SVM in the $(n+1)$-dimensional space. With this setting, assumptions $A1 - A3$ are satisfied.

To check if assumption A4 is satisfied, we need to calculate the derivative of the implicit function $\bar{\zeta} : \nu \rightarrow \bar{\zeta}(\nu)$. Remember that

$$\nabla\zeta(\nu) = -\left[\nabla^2_{\zeta} G(\nu, \zeta(\nu))\right]^{-1} \nabla^2_{\nu\zeta} G(\nu, \zeta(\nu)).$$

It is easy to see that $\nabla^2_{\zeta} G(\nu, \zeta(\nu)) = I$ where I is the identity matrix and that $\nabla^2_{\nu\zeta} G(\nu, \zeta(\nu))$ is a constant vector independent of ν and ζ for all loss functions in Table 1 (ex: $\nabla^2_{\nu\zeta} G(\nu, \zeta(\nu)) = \sum_{i=1}^{J_e} y_j x_j$ for the hinge loss case). Hence, assumption A4 is also satisfied.

The BSG algorithm is therefore applicable to these types of problems. Numerical experiments with BSG for the large scale SVM case with hinge loss can be found in [4]. The bi-level stochastic gradient technique when compared to classical cross validation procedures shows significant computing time savings with similar prediction performance.

6 Conclusions

We have analyzed the convergence of stochastic optimization methods for bi-level optimization problems (of the form of Problem (1)) where either the outer objective function or both outer and inner objective functions can be expressed as finite sums of independent terms. Under assumptions (A1)-(A4), we have shown that convergence to a stationary point of Problem (1) is guaranteed in expectation.

In the machine learning context, optimization is most of the time performed on loss or regularized loss functions and these losses can be expressed as very large sums of terms. Moreover, in this context, tuning model hyperparameters often requires the use of computationally expensive cross-validation procedures combined with a grid search approach. Alternatively, as explained in [4], the overall issue of tuning model parameters on validation data while training, could be expressed as a bi-level optimization problem of the form of Problem (1). The results presented here are therefore giving some expected stationarity guarantees for the bi-level stochastic gradient approach as an efficient alternative to the well established cross-validation procedure among machine learning practitioners.

References

[1] G. S. B. Colson, P. Marcotte. An overview of bilevel optimization. *Annals of Operations Research*, 153:235–256, 2007.

[2] J. Bard. *Practical bilevel optimization: applications and algorithms*. Kluwer Academic Press, 1998.

[3] P. Bersekas and J. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM journal on Optimization*, 10:627–642, 2000.

[4] N. Couellan and W. Wang. Bi-level stochastic gradient for large scale support vector machine. *Neurocomputing*, 2014.

[5] N. Couellan and W. Wang. Uncertainty-safe large scale support vector machines. *Submitted to Machine Learning*, 2015.

[6] P. Du, J. Peng, and T. Terlaky. Self-adaptive support vector machines: modelling and experiments. *Computational Management Science*, 6:41–51, 2009.

[7] I. Guyon. *A practical guide to model selection, Proceedings of the machine learning summer school*. Springer, 2009.

[8] G. Kunapuli, K. Bennett, J. Hu, and J. Pang. Bilevel model selection for support vector machines. *Centre de Recherches Mathématiques, CRM Proceedings and Lectures Notes*, 45, 2008.

[9] J. J. L. Vicente, G. Savard. Descent approaches for quadratic bilevel programming. *Journal of Optimization Theory and Applications*, 81, 1994.

[10] O. Mangasarian and D. Musicant. Successive overrelaxation for support vector machines. *IEEE Transactions on Neural Networks*, 10:1032–1037, 1999.

[11] B. Polyak and Y. Tsypkin. Pseudogradient adaptation and training algorithms. *Automation and Remote Control*, 12:83–94, 1973.

[12] W. Rudin. *Principles of Mathematical Analysis, third edition.* McGraw-Hill,Inc., 1976.

[13] M. Schmidt and N. L. Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *eprint arXiv:1308.6370*, 2013.