On the convergence of a stochastic approximation method for structured bi-level optimization

Nicolas Couellan, Wenjuan Wang

On the convergence of a stochastic approximation method for structured bi-level optimization

Nicolas Couellan*[1] and Wenjuan Wang[2]

[1] *ENAC, Université de Toulouse, 7 Avenue Edouard Belin, 31400 Toulouse, France*
[2] *Department of Computing, Bournemouth University, Fern Barrow, Poole, Dorset, BH12 5BB, United Kingdom*

November 16, 2018

Abstract

We analyze the convergence of stochastic gradient methods for well structured 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 if gradient approximations are chosen so as to satisfy a sufficient decrease condition. 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

*nicolas.couellan@enac.fr

gradients (see Assumptions A1-A4).

Observe that the form of Problem (1) is very specific and well structured when compared to general bi-level optimization problems. In (1), the outer objective does not depend explicitely on the outer variable x. The outer level is not either subject to any constraint. The instance (1) is one of the simplest form of bilevel programming. The difficulty that arises is therefore not in the structure of the levels but in the dimensions of the finite sums contained in both objectives. We consider that these sums are so large that evaluating them would be very expensive or even impossible. This is usually the setting of risk minimisation problems in learning problems in the statistics or image processing communities.

There are many applications of bi-level optimization [1]. Bi-level programming problems are generally difficult to solve when little is known on the objective functions [3]. Extensive research has been done in the field of bi-level programming and researchers have proposed several numerical strategies to find approximate solutions of these problems [3, 7]. One common method is to replace the inner problem by its KKT optimality conditions. Descent techniques based on gradient, sub-gradient or trust region steps have also been proposed [12, 3]. 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, 11, 8]. 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.

Integrating randomness in the bi-level problem to account for uncertainty in the problem parameters is also an active research topic and is known as stochastic bilevel programming [6, 10]. This work is different from the stochastic setting we are proposing. In this research randomness is incorporated in the problem but the solution process is usually deterministic where as in our technique as we will see below, we calculate random approximations of large gradient information to carry out cheap optimization moves. This concept is known as stochastic approximation rather than stochastic optimization.

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 and almost sure convergence of the minimization process can be proven. These techniques date back from the 1950's[17] and had been well studied since [21, 15, 13]. Recently, with the increasing interest in data science, they have been revisited through the eye of the machine learning research community. However, so far, very little investigation has been made on the use of

these stochastic approximation methods in the context of bi-level optimization. In [4, 5], we have proposed to design and experiment 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. Extensive experimentation has been made in this context with significant training time reduction when compared to other techniques. The technique is successful in practice but no convergence results have been established so far for these 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. The function $G(x, .) : y \to G(x, y)$ is strictly convex and Lipschitz continuous with constant L_G. Its gradient ∇G is Lipschitz continuous with constant $L_{\nabla G}$ and $\nabla^2_y G(x, y)$ is well conditioned.
- A3 The function $F : x \to F(y(x))$ is bounded below.
- A4 The fonction $y : x \to y(x)$ 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 with respect to the variable y 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. We also require that the Hessian $\nabla^2_y G(x, y)$ is well conditioned to ensure, as we will see later, that the implicit function is also Lipschitz continuous. Assumption A3 requires that the funtion $F : x \to F(y(x))$ is bounded from below 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. Out of context, these assumptions may seem strong but, in practice, in the type of problems we address, usually there are satisfied. In the last part of the article, we check and discuss 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 ($i \in \{1, \ldots, N\}$) 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 approximation 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 an approximation of the gradient of the objective (or both objectives if both are decomposable) and making use of bi-level differentiation. If the gradient approximations are chosen so as to satisfy a sufficient decrease condition, we show that under the assumptions on F and G above, both methods converge in expectation towards a stationary point of Problem (1). Note that if the choice of gradient approximations ensures that the approximation of the outer objective gradient (with respect to the outer variable) is an unbiased estimate of the true gradient, our technique reduces to the standard stochastic gradient technique [17]. However, except for a special class of inner objective functions, the choice of an unbiased estimate is not trivial and this is the reason why we allow some bias in the choice of the approximation as long as the sufficient decrease condition is satisfied. We would also like to emphasize that one may find that the convergence in expectation is a rather weak convergence result as opposed to an almost sure convergence. However, to the best of our knowledge, this work is a first attempt to address stochastic approximation to bilevel optimization and we argue that convergence in expectation is a first step towards a more complete convergence analysis of these stochastic bilevel methods and we leave the analysis of stronger convergence result as an open perspective.

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 Preliminaries

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) [19] 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 $\nabla^2_y G(x^*, y^*)$ 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 $\nabla^2_v G(u, v)$ is invertible and furthermore,

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

Therefore, we can write

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

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

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

Since ∇G is Lipschitz continuous, we have that $\left\|\nabla^2 G(x, y)\right\|$ is bounded, meaning that $\left\|\nabla^2_y G(x, y)\right\|$ and $\left\|\nabla^2_{xy} G(x, y)\right\|$ are also bounded. From Assumption A2, we know that $\nabla^2_y G(x, y)$ is well conditioned, therefore there exists $c > 0$ such that $\| \left[\nabla^2_y G(x, y)\right]^{-1} \| \|\nabla^2_y G(x, y)\| \leq c$. Since $\nabla^2_y G(x, y)$ is bounded and non singular, we can further say that $\| \left[\nabla^2_y G(x, y)\right]^{-1} \|$ is 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 $L^2_y L_{\nabla_y F} + L_F L_{\nabla_x y}$.*

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 $\tilde{g}_i$ as an estimate of $\nabla_x [F(y(x))]$ to compute a stochastic move. To ensure convergence, the expected angle between $\tilde{g}_i$ and $\nabla_x [F(y(x))]$ must be sufficiently positive. The BSG_o procedure is summarized in Algorithm (1).

Algorithm 1 BSG_o Algorithm

Choose x_0 and $\alpha_0 > 0$
$k \leftarrow 0$
while $\|\nabla_x [F_i(y(x_k))]\| \geq 0$ **do**
  Pick i randomly and uniformly in $\{1, \ldots, N\}$
  Compute $y(x_k) = \underset{\bar{y} \in \mathbb{R}^m}{\operatorname{argmin}} G(x_k, \bar{y})$
  Choose $\tilde{g}_i$ such that $E[\tilde{g}_i | x_k]^\top \nabla_x [F(y(x_k))] \geq \mu \|[\nabla_x F(y(x_k))]\|^2$
  $x_{k+1} \leftarrow x_k - \alpha_k \tilde{g}_i$
  Update α_k
  $k \leftarrow k + 1$
end while

Note that the special case

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

is an unbiased estimate of $\nabla_x [F(y(x_k))]$ under the differentiability and Lipschitz assumptions made above [18] and the algorithm reduces to the standard Robbins and Monro algorithm for finding the roots of the function $x \to \nabla_x [F(y(x))]$ [17]. However, we address a more general case where $\tilde{g}_i$ is not required to be an unbiased estimator but should only require sufficient expected decrease of the function. This will also be critical later when we consider inner stochastic approximation as it will be difficult to ensure 'unbiasness' of the gradient approximation.

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

$$\varepsilon_k = \tilde{g}_i - \nabla_x [F(y(x_k))]$$

We state and prove the following convergence theorem:

Theorem 3.1 *Suppose that:*

1. *Assumptions A1-A4 are satisfied,*
2. $\exists \nu, \mu > 0$ *such that* $\forall i \in \{1, \ldots, N\}$, $\forall k > 0$,

$$\|E[\tilde{g}_i | x_k]\| \leq \nu \|\nabla_x [F(y(x_k))]\|$$

and

$$E[\tilde{g}_i | x_k]^\top [F(y(x_k))] \geq \mu \|[F(y(x_k))]\|^2$$

,

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

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

4. $\forall k > 0$, 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_o *algorithm converges in expectation to a stationary point of the function* $x \to F(y(x))$*, i.e.*

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

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

$$x_{k+1} = x_k - \alpha_k \tilde{g}_i$$

where i is randomly chosen in $\{1, \ldots, N\}$.
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, we can write the following inequality

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

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

$$\begin{aligned}
E[F(y(x_{k+1}))] &\leq F(y(x_k)) + \alpha_k \left(\frac{L_{\nabla_x F}}{2} \alpha_k - \mu\right) \|\nabla_x \left[F(y(x_k))\right] \|^2 \\
&\quad + \frac{L_{\nabla_x F}}{2} \alpha_k^2 E\left[\|\varepsilon_k\|^2\right] \qquad (6) \\
&. \qquad (7)
\end{aligned}$$

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

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

Observe that if $\forall k > 0$, α_k is chosen so as to ensure that $0 < \alpha_k < \frac{2\mu}{L_{\nabla_x F} + D}$, 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 (8) to pairs of iterates starting from (x_1, x_2) to some iterates (x_{K-1}, x_K) for any $K > 2$, we get:

$$E\left[F(y(x_0))\right] - E\left[F(y(x_1))\right] \geq \alpha_0 \left(\mu - \alpha_0 \frac{L_{\nabla_x F} + D}{2}\right) E\left[\|\nabla_x \left[F(y(x_0))\right]\|^2\right]$$

$$E\left[F(y(x_1))\right] - E\left[F(y(x_2))\right] \geq \alpha_1 \left(\mu - \alpha_1 \frac{L_{\nabla_x F} + D}{2}\right) E\left[\|\nabla_x \left[F(y(x_1))\right]\|^2\right]$$

$$\cdots$$

$$E\left[F(y(x_{K-1}))\right] - E\left[F(y(x_K))\right] \geq \alpha_{K-1} \left(\mu - \alpha_{K-1} \frac{L_{\nabla_x F} + D}{2}\right) \times E\left[\|\nabla_x \left[F(y(x_{K-1}))\right]\|^2\right]$$

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

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

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

$$\sum_{k=1}^{K-1} \alpha_k \left(\mu - \alpha_k \frac{L_{\nabla_x F} + D}{2}\right) E\left[\|\nabla_x \left[F(y(x_{k-1}))\right]\|^2\right] \leq M \quad . \tag{9}$$

Let now $s_k = \alpha_k(\mu - \frac{L_{\nabla_x F}}{2}\alpha_k)$. 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 (9), we can write

$$\sum_{k=0}^{\infty} s_k E\left[\|\nabla_x \left[F(y(x_{k-1}))\right]\|^2\right] \leq M < \infty, \tag{10}$$

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

$$E\left[\|\nabla_x \left[F(y(x_{k-1}))\right]\|^2\right] \geq \hat{\epsilon}, \tag{11}$$

implying

$$\sum_{k=0}^{\infty} s_k E\left[\|\nabla_x \left[F(y(x_{k-1}))\right]\|^2\right] \geq \hat{\epsilon} \sum_{k=0}^{\infty} s_k = \infty. \tag{12}$$

The inequality (12) contradicts inequality (9) meaning that the assumption (11) 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 [2], 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

$$\begin{cases} E\left[\|\nabla_x \left[F(y(x_k))\right]\|\right] < \check{\epsilon}/2 \\ \check{\epsilon}/2 \leq E\left[\|\nabla_x \left[F(y(x_l))\right]\|\right] \leq \check{\epsilon} \qquad \forall l \in \mathbb{N} \text{ such that } k < l < i^{(k)} \\ \check{\epsilon} < E\left[\|\nabla_x \left[F(y(x_{i^{(k)}}))\right]\|\right] \end{cases} \tag{13}$$

On one hand, from (13) and Lemma 2.2, observe that

$$\begin{aligned} \frac{\check{\epsilon}}{2} &\leq E\left[\|\nabla_x \left[F(y(x_{i^{(k)}})\right]\|\right] - E\left[\|\nabla_x \left[F(y(x_k)\right]\|\right] \\ &= E\left[\|\nabla_x \left[F(y(x_{i^{(k)}})\right]\| - \|\nabla_x \left[F(y(x_k))\right]\|\right] \\ &\leq E\left[\|\nabla_x \left[F(y(x_{i^{(k)}})\right] - \nabla_x \left[F(y(x_k))\right]\|\right] \\ &\leq L_{\nabla_x F} E\left[\|x_{i^{(k)}} - x_k\|\right] \\ &\leq L_{\nabla_x F} \sum_{l=k}^{i^{(k)}-1} \alpha_l E\left[\|\tilde{g}_l | x_l\|\right] \qquad (14) \\ &\leq \nu L_{\nabla_x F} \sum_{l=k}^{i^{(k)}-1} \alpha_l \|\nabla_x \left[F(y(x_l))\right]\| \qquad (15) \end{aligned}$$

Taking the expectation in the right hand side of (15) over all possible realizations of the random variable x_l (for $l = k, \ldots, i^{(k)} - 1$), we obtain

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

Hence,

$$\liminf_{k\to\infty} \sum_{l=k}^{i^{(k)}-1} \alpha_l \geq \frac{1}{2\nu L_{\nabla_x F}}. \tag{16}$$

On the other hand, from (8) and (13), we can write

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

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 (16). As a consequence, the statement that *there exists* $\check{\epsilon} > 0$ *such that (13) 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 (3) 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 [2, 16] and more recently in [20]. Intuitively, it is reasonable to assume that if $\|\nabla_x \left[F(y(x_k))\right]\|$ is small, there is little noise 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:

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

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 (2.1) to compute an approximation $\tilde{g}_i^j$ of $\nabla_x\left[F(y(x))\right]$ as:

$$\tilde{g}_i^j = -A_j\tilde{h}_i$$

where, similarly as in section 3, to ensure convergence, we require that the matrix A_j and the vector $\tilde{h}_i$ are chosen so as to ensure the following sufficient decrease condition :

$$E[\tilde{g}_i^j|x_k]^\top\nabla_x\left[F(y(x_k))\right] \geq \mu\|\nabla_x\left[F(y(x_k))\right]\|^2$$

for some positive constant μ. This condition can also be written as:

$$-E[\tilde{h}_i|x_k]^\top E[A_j|x_k]^\top H\nabla_y F(y_j^k) \geq \mu\|H\nabla_y F(y_j^k)\|^2 \tag{17}$$

where $y_j^k = \underset{\bar{y}\in\mathbb{R}^m}{\operatorname{argmin}} G_j(x_k,\bar{y})$ and $H = [\nabla_y^2 G(x_k,y_j^k)]^{-1}\nabla_{xy}^2 G(x_k,y_j^k)$.
The condition (17) is a coupling condition between the upper level($\nabla_y F$) and the lower level (H). In practice, it is not easy to check whether such condition is satisfied or not. For this reason, in the next proposition, we give further conditions on each of the levels separately that also ensure (17).

Proposition 4.1 *Under Assumption A2, if there exists ρ and γ two positive constants such that $\gamma\rho \geq \mu\|H\|^2$ and such that for all $k>0$, $\tilde{h}_i$ and A_j the following conditions are satisfied*

$$-E[\tilde{h}_i|x_k]^\top\nabla_y F(y_j^k) \geq \rho\|\nabla_y F(y_j^k)\|^2 \tag{18}$$

and

$$E[A_j|x_k]^\top H \succeq \gamma I_m \tag{19}$$

then (17) is satisfied.

Proof Since $\gamma\rho \geq \mu\|H\|^2$, we can write

$$\gamma\rho\|\nabla_y F(y_j^k)\|^2 \geq \mu\|H\|^2\|\nabla_y F(y_j^k)\|^2 \geq \mu\|H\nabla_y F(y_j^k)\|^2.$$

From (18), we can further write

$$-\gamma E[\tilde{h}_i|xk]^\top\nabla_y F_i(y_j^k) \geq \mu\|H\nabla_y F(y_j^k)\|^2,$$

which, using (19) leads to

$$-E[\tilde{h}_i|xk]^\top E[A_j|x_k]^\top H\nabla_y F_i(y_j^k) \geq \mu\|H\nabla_y F(y_j^k)\|^2,$$

being condition (17). ∎

Furthermore, we also show that under Assumption A2, the norm of the gradient approximation $\tilde{g}_i^j$ remains bounded by the norm of $\nabla_x\left[F(y(x_k))\right]$ if the norm of $\tilde{h}_i$ is also bounded. This result is summarized in the following proposition:

Proposition 4.2 *Under Assumption A2, if there exists two positive constants M and ν such that for all $j \in \{1, \ldots, J\}$, $\|A_j\| \leq M\|H\|$ and for all $k > 0$,*

$$\|E[\tilde{h}_i|x_k]\| \leq \nu\|\nabla_y \left[F(y(x_k))\right]\|$$

then, there exists a positive constant C_ν such that

$$\|E[\tilde{g}_i^j|x_k]\| \leq C_\nu\|\nabla_x \left[F(y(x_k))\right]\|$$

Proof For all $k > 0$, for all $(i, j) \in \{1, \ldots, N\} \times \{1, \ldots, J\}$, we have

$$\begin{aligned}\|E[\tilde{g}_i^j|x_k]\| &\leq \|E[A_j|x_k]\|\|E[\tilde{h}_i|x_k]\| \\ &\leq \nu\|E[A_j|x_k]\|\|\nabla_y \left[F(y_j^k)\right]\|\end{aligned}$$

Since $\|A_j\| \leq M\|H\|$, we can write

$$\|E[\tilde{g}_i^j|x_k]\| \leq \nu M\|H\|\|\nabla_y \left[F(y_j^k)\right]\|$$

We also know that there exists $\lambda > 1$ a constant (with respect to i and k) that satisfies the following:

$$\|\nabla_y \left[F(y_j^k)\right]\|\|H\| = \lambda\|H\nabla_y \left[F(y_j^k)\right]\|,$$

Therefore,

$$\begin{aligned}\|E[\tilde{g}_i^j|x_k]\| &\leq \lambda\nu M\|H\nabla_x \left[F(x_k, y_j^k)\right]\| \\ &\leq C_\nu\|\nabla_x \left[F(y(x_k))\right]\|\end{aligned}$$

where $C_\nu = \lambda\nu M$. ∎

The *BSG* algorithm is summarized in Algorithm (2).

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

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

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

Theorem 4.3 *Suppose that:*

1. *Assumptions A1-A4 are satisfied,*
2. *$\exists\nu > 0$ such that $\forall i \in \{1, \ldots, N\}$, $\forall k > 0$,*

$$\|E[\tilde{h}_i|x_k]\| \leq \nu\|\nabla_x \left[F(y(x_k))\right]\|$$

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 $\{1, \ldots, N\}$
    Pick j randomly and uniformly in $\{1, \ldots, J\}$
    Compute $y^{(j)}(x_k) = \underset{\bar{y} \in \mathbb{R}^m}{\operatorname{argmin}} G_j(x_k, \bar{y})$
    Choose $\tilde{g}_i^j$ such that $E[\tilde{g}_i^j | x_k]^\top \nabla_x \left[F(y(x))\right] \geq \mu \|\nabla_x \left[F(y(x))\right] \|^2$.
    $x_{k+1} \leftarrow x_k - \alpha_k \tilde{g}_i^j$
    Update α_k
    $k \leftarrow k + 1$
end while

3. $\exists \mu > 0$ *such that* $\forall (i,j) \in \{1, \ldots, N\} \times \{1, \ldots, J\}$, $\forall k > 0$,

$$E[\tilde{h}_i | x_k]^\top E[A_j | x_k]^\top H \nabla_y F(y_j^k) \geq \mu \|H \nabla_y F(y_j^k)\|$$

4. $\forall j \in \{1, \ldots, J\}$, $\exists M > 0, \|A_j\| \leq M \|H\|$

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

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

6. $\forall k > 0$, 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

$$x_{k+1} \leftarrow x_k - \alpha_k \tilde{g}_i^j$$

where i and j are randomly chosen at each iteration k in $\{1, ..., N\}$ and $\{1, ..., J\}$ respectively and $\tilde{g}_i^j = -A_j \tilde{h}_i$ with $\tilde{h}_i$ and A_j satisfying the assumptions of Theorem 4.3.
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

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

Using the sufficient decrease condition, we can further bound the previous expression as follows

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

Taking the expectation again over all 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 (14) can be re-written as follows

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

where $\tilde{g}_{i_l}^{j_l}$ is chosen as $\tilde{g}_{i_l}^{j_l} = -A_{j_l}\tilde{h}_{i_l}$ with i_l and j_l are picked randomly and uniformly in the sets $\{1,\ldots,N\} \times \{1,\ldots,J\}$.
Using Proposition 4.2, 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

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

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:

$$\min_{\zeta} \left[r(\zeta) + \delta \sum_{j=1}^{J} \phi_j(\zeta^\top x_j) \right]$$

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 $\delta > 0$ 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 δ. Tuning δ 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 [9]). For this reason, stochastic bi-level optimization may be preferred [4]. The bi-level problem resulting from learning the hyperparameter δ can be written as follows:

$$\begin{aligned} \min_{\delta} \quad & \sum_{i=1}^{N} \phi_i(\bar{\zeta}(\delta)^\top x_i^v) \\ s.t. \quad & \bar{\zeta}(\delta) = \operatorname*{argmin}_{\zeta} \left[r(\zeta) + \delta \sum_{j=1}^{J} \phi_j(\zeta^\top x_j) \right] \end{aligned} \tag{20}$$

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 (20) where ϕ_j are standard loss functions as often used in machine learning applications (see Table 1) and r is the commonly used squared L_2-norm

(i.e. $r(\zeta) = \frac{1}{2}\|\zeta\|_2^2$). If the BSG algorithm, $\tilde{h}_i$ and A_j are taken for example as

$$\tilde{h}_i = \phi_i(\bar{\zeta}(\delta)^\top x_i^v)$$

and

$$A_j = [\nabla_y^2 G_j(x_k, y_j^k)]^{-1} \nabla_{xy}^2 G_j(x_k, y_j^k),$$

it is easy to check that $\tilde{g}_i^j = -A_j \tilde{h}_i$ is an unbiased estimate of $\nabla_x [F(y(x))]$ and the sufficient decrease condition (17) is satisfied.
Whenever ϕ_j is non differentiable as in the case of the hinge loss, observe that one can replace in the inner and outer objectives $\sum_{j=1}^{J} \phi_j(\zeta^\top x_j)$ and $\sum_{i=1}^{N} \phi_i(\bar{\zeta}(\delta)^\top x_i^v)$ by the following sums $\sum_{j=1}^{J_e}(1 - y_j \zeta^\top x_j)$ and $\sum_{i=1}^{N_e}(1 - y_j \bar{\zeta}(\delta)^\top x_i^v)$ where J_e and N_e are the number of training and validation error vectors, vectors with non zero losses, as explained in [4]. 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 $\zeta^\top x_j$ 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 $\zeta \to \left[r(\zeta) + \delta \sum_{j=1}^{J} \phi_j(\zeta^\top x_j) \right]$ is strictly convex, except for the Support Vector Machine (SVM) case where strict convexity can be ensured by adding an extra attribute to the data as explained in [14] 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 $\bar{\zeta} : \delta \to \bar{\zeta}(\delta)$. Remember that

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

It is easy to see that $\nabla_\zeta^2 G(\delta, \zeta(\delta)) = I$ where I is the identity matrix and that $\nabla_{\delta\zeta}^2 G(\delta, \zeta(\delta))$ is a constant vector independent of δ and ζ for all loss functions in Table 1 (ex: $\nabla_{\delta\zeta}^2 G(\delta, \zeta(\delta)) = \sum_{i=1}^{J_e} y_j x_j$ 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 same stochastic bi-level technique was also used to adjust

the level of input data uncertainty to be integrated in a robust SVM classification model [5].

In these applications, the stochastic bi-level technique has shown significant savings in computing times when compared to other alternative strategies that sometime are not even applicable when the size of the dataset is extremely large. For these types of large problems that require an outer optimization level to tune hyperparameters, the proposed technique is often the only optimization-based method available.

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.

The bi-level formulation we have considered is a very simple instance of bi-level programming but the difficulty resides in the size of the problems we have been considering. The stochastic approximation methods we have proposed allows sampling or approximation of both inner and outer objective functions. This setting is specifically interesting when simple but large optimization problems require hyperparameter optimization as this is tha case in large machine learning or image processing problems.

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

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

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

[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] S. Dempe, S. Ivanov, and A. Naumov. Reduction of the bilevel stochastic optimization problem with quantile objective function to a mixed-integer problem. *Appl. Stoch. Model. Bus. Ind.*, 33(5):544–554, 2017.

[7] S. Dempe, V. Kalashnikov, G. Perez-Valdes, and N. Kalashnykova. *Bilevel programming problems, theory, algorithms and applications to energy networks*. Springer, 2015.

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

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

[10] R. Kovacevic and G. Pflug. Electricity swing option procing by stochastic bilevel optimization: a survey and new approaches. *European Journal of Operational Research*, 237:389–774, 2014.

[11] 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.

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

[13] L. Ljung, G. Pflug, and H. Walk. *Stochastic approximation and optimization of random systems*. Springer Basel AG, 2015.

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

[15] G. Pflug. *Optimization of Stochastic Models*. Kluwer Academic Publisher, Norwell, Mass., 1996.

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

[17] H. Robbins and S. Monro. On a stochastic approximation method. *Ann. Math. Stat.*, 22, 1951.

[18] R. Rubinstein and A. Shapiro. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method.* John Wiley & Sons, Chichester, England, 1993.

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

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

[21] M. Wasan. *Stochastic Approximation.* Cambridge University Press, 1969.