2019 American Control Conference (ACC)
Philadelphia, PA, USA, July 10-12, 2019

A Randomized Block Coordinate Iterative Regularized Subgradient Method for High-dimensional Ill-posed Convex Optimization

Harshal Kaushik[1] and Farzad Yousefian[2]

***Abstract*— Motivated by ill-posed optimization problems arising in image processing, we consider a bilevel optimization model, where we seek among the optimal solutions of the inner level problem, a solution that minimizes a secondary metric. Minimal norm gradient, sequential averaging, and iterative regularization appear among the known schemes developed for addressing this class of problems. However, to the best of our knowledge, none of these schemes address nondifferentiability and high-dimensionality of the solution space. Motivated by this gap, we consider the case where the solution space has a block structure and both objective functions are nondifferentiable. We develop a randomized block coordinate iterative regularized subgradient scheme (RB-IRG). Under a uniform distribution for selecting the blocks and a careful choice of the stepsize and regularization sequences, we establish the convergence of the sequence generated by RB-IRG scheme to the unique solution of the bilevel problem of interest in an almost sure sense. Furthermore, we derive a convergence rate of $\mathcal{O}\left(\frac{\sqrt{d}}{k^{0.5-\delta}}\right)$ in terms of the expected objective value of the inner level problem, where d denotes the number of blocks and $\delta > 0$ is an arbitrary small scalar. We demonstrate the performance of RB-IRG algorithm in solving the ill-posed problems arising in image processing.**

I. Introduction

In this work, we are interested in computing the optimal solution of a bilevel problem given as,

$$\begin{aligned} &\text{minimize } g(x) \\ &\text{s.t.} \quad x \in \operatorname{argmin}\{f(x) : x \in X\}, \end{aligned} \tag{P_f^g}$$

where functions f and g are defined as $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$. In particular, we consider the case where the set X is assumed to have a block structure, i.e., $X = \prod_{i=1}^{d} X_i$, where $X_i \subseteq \mathbb{R}^{n_i}$ and $\sum_{i=1}^{d} n_i = n$. Our work in solving (P_f^g) is inspired by the revived interest in the block coordinate first-order schemes in the recent years. These schemes have recently gained popularity due to their potential of solving large-scale optimization problems. In particular, they proved to be effective when the size of the solution space is of the order $10^8 - 10^{12}$ [21], [23]. While the cyclic coordinate schemes are common to make the selection of the blocks, recently the focus has been shifted to randomized strategies due to theoretical [21], [23], [26] and practical advantages they offer in solving the large scale optimizaiton problems [9], [17], [25]. Next, we state the assumptions on (P_f^g) precisely and then present classes of problems that motivate our research in addressing (P_f^g).

Assumption 1: Let the following hold:

(a) Any block i of set X ($X_i \subseteq \mathbb{R}^{n_i}$) is assumed to be nonempty, closed, and convex for all $i = 1, \ldots, d$.
(b) $f : \mathbb{R}^n \to (-\infty, \infty]$ is a nondifferentiable, proper, and convex function.
(c) $g : \mathbb{R}^n \to (-\infty, \infty]$ is a nondifferentiable, proper, and μ-strongly convex function ($\mu > 0$).
(d) $X \subseteq \operatorname{int}(\operatorname{dom}(f) \cap \operatorname{dom}(g))$.

A. Motivating examples

Here we present two applications of the formulation (P_f^g).
(i) **High-dimensional nonlinear constrained optimization**: Consider the following problem with nonlinear constraints,

$$\begin{aligned} &\text{minimize } g(x) \\ &\text{s.t. } h_i(x) \le 0 \quad \text{for } i = 1, \ldots, m \\ &x \in X \triangleq \prod_{i=1}^{d} X_i. \end{aligned} \tag{1}$$

Given that: (i) functions $h_i : \mathbb{R}^n \to \mathbb{R}$ are convex; (ii) function $g : \mathbb{R}^n \to \mathbb{R}$ is convex; (iii) $X \subseteq \mathbb{R}^n$ is convex, satisfying Assumption 1 (a), and that (iv) the feasible set of (1) is nonempty, then problem (1) can be equivalently written in a bilevel structure as,

$$\begin{aligned} &\text{minimize } g(x) \\ &\text{s.t.} \quad x \in \operatorname{argmin}\left\{ f(x) \triangleq \sum_{i=1}^{m} \max\{0, h_i(x)\} : x \in X \right\}. \end{aligned}$$

(ii) **Ill-posed optimization**: Linear inverse problems arising in image deblurring can be cast as [10],

$$\begin{aligned} &\text{minimize } \ \|Ax - b\|_2^2 \\ &\text{s.t.} \quad x \in \mathbb{R}^n, \end{aligned} \tag{2}$$

where A is a blurring operator ($A \in \mathbb{R}^{m \times n}$), b is the given blurred image ($b \in \mathbb{R}^m$), and x is a deblurred image ($x \in \mathbb{R}^n$). This is an ill-posed problem in the sense that there may be multiple solutions or the optimal solution x may be very sensitive to the perturbations in the input b. To address the ill-posedness, and to induce sparsity and stability, problem (2) can be reformulated in a bilevel structure as following (e.g., see [15]),

$$\begin{aligned} &\text{minimize } \|x\|_2^2 + \|x\|_1 \\ &\text{s.t.} \quad x \in \operatorname{argmin}\left\{ \|Ax - b\|_2^2 : x \in \mathbb{R}^n \right\}. \end{aligned} \tag{3}$$

[1] Harshal Kaushik is a graduate student at the Department of Industrial Engineering and Management, Oklahoma State University, Stillwater, OK 74074, USA, harshal.kaushik@okstate.edu

[2] Farzad Yousefian is an assistant professor at the Department of Industrial Engineering and Management, Oklahoma State University, Stillwater, OK 74074, USA, farzad.yousefian@okstate.edu

978-1-5386-7926-5/$31.00 ©2019 AACC

TABLE I: Comparison of schemes for solving the bilevel optimization problem (P_f^g)

Ref.	Problem formulation	Assumption	Scheme	Scale	Metric	Rate
[27]	minimize $g(x)$ s.t. $x \in \text{argmin}\{f(x) : x \in X\}$	f and g both smooth and convex	Iterative regularized	Standard scale	–	–
[4]	minimize $g(x)$ s.t. $x \in \text{argmin}\{f(x) : x \in X\}$	f convex, Lipschitz cont. g strongly conv.	Minimal norm gradient	Standard scale	Inner level	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$
[24]	minimize $g(x)$ $x \in \text{argmin}\{f_1(x) + f_2(x) : x \in \mathbb{R}^n\}$	f_1, f_2 convex. g is strongly convex. f_1, g Lipschitz cont.	Sequential averaging	Standard scale	Inner level	$\mathcal{O}\left(\frac{1}{k}\right)$
[31]	minimize $\|x\|$ s. t. $x \in \text{SOL}(X, F)$, where $F(x) \triangleq f(x, \xi)$	F is monotone and continuous.	aRSSA$_{l,r}$	Standard scale	Outer level	$\mathcal{O}\left(\frac{1}{k^{1/6-\delta}}\right)$
[29]	$\min\limits_{x_i \in X_i, z \in Z} \sum_{i=1}^N f_i(x_i)$ s.t. $Dx + Hz = 0$	f_i is convex and possibly nonsmooth.	Asynchronous ADMM	Standard Scale	Feasibility	$\mathcal{O}\left(\frac{1}{k}\right)$
[2]	for $\mathcal{G}(\mathcal{N}, \mathcal{E})$, $\min \sum_{i \in \mathcal{N}} \xi_i(x) + f_i(x)$ s.t. $x \in \mathbb{R}^n$, $x_i = x_j$ for all $(i, j) \in \mathcal{E}$	ξ_i, f_i are convex, f_i Lipschitz continuous.	Distributed proximal gradient	Standard Scale	Feasibility	$\mathcal{O}(1/k)$
[30]	$\min\limits_{x} g_1(x) + g_2(x)$ s.t. $Ax = b$	g_1 is convex and Lipschitz continuous. g_2 is convex.	Linear ADMM	Standard Scale	Feasibility Optimality	$\mathcal{O}(1/k^2)$
This work	minimize $g(x)$ $x \in \text{argmin}\{f(x) : x \in X\}$; $X = \prod_{i=1}^d X_i$	f is convex and g is strongly convex.	Randomized block iterative reg. subgradient	Large Scale	Feasibility	$\mathcal{O}\left(\frac{\sqrt{d}}{k^{0.5-\delta}}\right)$

B. Existing methods

Minimal norm gradient, sequential averaging, and iterative regularization are some of the known schemes developed for addressing problem (P_f^g) and are described next. Given $\eta > 0$, consider the following regularized problem

$$\begin{aligned} \text{minimize} \quad & f(x) + \eta g(x) \\ \text{s.t.} \quad & x \in X. \end{aligned} \tag{P_η}$$

Tikhonov in [28] showed that under some assumptions, the solution of regularized problem (P_η) converges to the solution of the inner level problem of (P_f^g) as the regularization parameter η goes to zero. Later, the threshold value of η, under which the solution of (P_η) is the same as the solution of the inner level problem of (P_f^g), was studied under the area of *exact regularization* [14], [16], [20]. There have been numerous theoretical studies in the '80s, '90s [5], [6], [8], [13], [16], [20] and more recently [7], [11] on finding the suitable η, but in practice there is not much guidance on tuning this parameter. Finding a suitable η necessitates solving a sequence of problems (P_η) for η_k, where $\eta_k \to 0$. This *two-loop* scheme is significantly inefficient, especially in high dimensional spaces.

In the past decade, interest has been shifted to solving the bilevel problem (P_f^g) using *single-loop schemes*. Solodov in [27] showed that for both functions g and f in (P_f^g) with Lipschitz gradient, and f to be a composite function with the indicator function, solutions to (P_f^g) can be found by an iterative regularized gradient descent with sequence $\eta_k \to 0$ and $\sum_{k=1}^\infty \eta_k = \infty$. In (P_η), when g is ℓ_2 norm in variational inequality regimes, Yousefian et al. [31] showed that solution to (P_f^g) can be found by employing an iterative regularized smoothing gradient scheme.

In 2014, the minimal norm gradient (MNG) scheme was proposed [4]. MNG is a *two-loop* scheme where an optimization problem needs to be solved at each iteration k, making MNG to be computationally expensive for the large scale problems. Later, in [24] a sequential averaging scheme (BiG-SAM) was developed with a rate of convergence $\mathcal{O}(1/k)$. Recently in [15], a general iterative regularized algorithm based on a primal-dual diagonal descent method was proposed to solve (P_f^g).

In all the aforementioned papers, the missing part is addressing the high-dimensional structure, which is common in the high-resolution image processing applications such as hyper-spectral imaging. Our goal is to bridge this gap by developing a *single-loop* randomized block coordinate iterative regularized subgradient scheme.

High-dimensional nonlinear constrained optimization (1) is another motivating example to our work. One of the popular primal-dual methods is Alternating Direction Method of Multipliers (ADMM) [2], [29], [30]. One of the underlying assumptions for ADMM is the linearity of the constraints. In our work, bilevel problem (1) addresses the nonlinearity in the constraints and the high-dimensionality of the space.

C. Main contributions

To highlight the contribution of our work and its distinction from the other methods, we provide TABLE I. Specifically, the contributions are pointed out as following:
(i) To address high-dimensional, ill-posed optimization problem (P_f^g) with nondifferentiable objective function, we develop a *single-loop* first-order subgradient scheme (RB-IRG).
(ii) We establish the convergence of the sequence generated from (RB-IRG) scheme to the unique solution of (P_f^g).
(iii) We show that the generated sequence converges nonasymptotically with a rate $\mathcal{O}\left(\frac{\sqrt{d}}{k^{0.5-\delta}}\right)$, with respect to the inner level function of problem (P_f^g).

The rest of the paper is organized as follows. In Section II, we propose the (RB-IRG) scheme and give the preliminaries required in the convergence analysis. Section III is for showing the convergence of the sequence generated by the (RB-IRG) scheme to the unique solution of problem (P_f^g). In Section IV, we discuss the convergence rate analysis. In Section V, we apply the (RB-IRG) algorithm to an image deblurring example and discuss the computational effectiveness of our scheme. Concluding remarks are provided in Section VI.

Notation: Vector x is assumed to be a column vector ($x \in \mathbb{R}^n$), x^T is the transpose. $x^{(i)}$ denotes the i^{th} block of x. X_i denotes the i^{th} block of dimensions for set X. $\|x\|$ denotes the Euclidean vector norm, i.e., $\|x\| = \sqrt{x^T x}$. $\mathcal{P}_S(s)$ is used for the Euclidean projection of vector s on a set S, i.e., $\| s - \mathcal{P}_S(s)\| = \min_{y \in S} \|s - y\|$. a.s. used for 'almost surely'. $\mathcal{F}_k$ denotes the set of variables $\{i_0, \ldots, i_{k-1}\}$. For a random variable i_k, $\mathsf{Prob}(i_k = i)$ is p_{i_k}. $\tilde{\nabla}$ denotes the subgradient and ∂ denotes the subdifferential set. $\tilde{\nabla}_i f(x)$ is the i^{th} block of $\tilde{\nabla} f(x)$. $\min_{1 \leq i \leq d}\{\mathsf{p}_i\}$ is denoted by p_{min} and $\max_{1 \leq i \leq d}\{\mathsf{p}_i\}$ is denoted by p_{max}. Interior of set A is denoted by $int(A)$.

II. Algorithm outline

Here we explain the (RB-IRG) scheme and the required preliminaries for the convergence and rate analysis.

A. Proposed scheme RB-IRG

Here, a randomized block coordinate iterative regularized subgradient scheme (RB-IRG) is proposed for solving (P_f^g). In RB-IRG scheme, both the sequences of regularization parameter η_k and stepsize parameter γ_k are in terms of iteration k. The update rules of γ_k and γ_k are finalized later (in Theorem 2). To address the high-dimensionality, at each iteration we update a random block of the iterate x_k. Selection of block i_k at iteration k is governed by Assumption 2. Finally, averaging is employed which will be helpful in deriving the rate statement.

Algorithm 1 Randomized block iterative regularized gradient descent (RB-IRG) algorithm

```
Initialization:
Set k = 0, select a point $x_0 \in X$, parameters $\gamma_0 > 0$,
and $\eta_0 > 0, S_0 = \gamma_0^r$, and $\bar{x}_0 = x_0$.
for k = 0, 1, ..., N-1 do
    $i_k$ is generated by Assumption 2.
    Compute $\tilde{\nabla}_i f(x_k) \in \partial f\left(x_k^{(i)}\right)$ and $\tilde{\nabla}_i g(x_k) \in \partial g\left(x_k^{(i)}\right)$ for $x_k^{(i)} \in X_i$.
    Update $x_{k+1}^{i_k}$:=
    $\begin{cases} \mathcal{P}_{X_i}\left(x_k^{(i)} - \gamma_k\left(\tilde{\nabla}_i f\left(x_k\right) + \eta_k \tilde{\nabla}_i g\left(x_k\right)\right)\right) & \text{if } i = i_k. \\ x_k^{(i)} \quad \text{if } i \neq i_k. \end{cases}$    (RB-IRG)
    Update $\overline{x}_k$ as following,
    $S_{k+1} = S_k + \gamma_{k+1}^r, \quad \overline{x}_{k+1} = \frac{S_k \overline{x}_k + \gamma_{k+1}^r x_{k+1}}{S_{k+1}}.$    (4)
end for
```

Assumption 2: **(Random sample i_k)** Random variable i_k is generated at each iteration k from an i.i.d. distribution governed by probability p_{i_k} where prob($i_k = i$) = $\mathsf{p}_{i_k} > 0$, and $\sum_{i=1}^d \mathsf{p}_{i_k} = 1$.

B. Preliminaries

In this subsection, we list all the required preliminaries in Remarks 1-4 and Lemmata 1-3. This will be used in the convergence and rate analysis in the next sections. Throughout the paper, we use x_g^* and $x_{\eta_k}^*$ to denote the unique minimizers of (P_f^g) and (P_η), respectively.

Remark 1: From Assumptions 1 (b, c), the objective function of (P_η), is a strongly convex. The feasible region of (P_η) is closed and convex (from Assumption 1(a)). Therefore (P_η) has a unique minimizer. (cf. Ch. 2 of [12]). Similarly, we can claim that (P_f^g) has a unique minimizer.

Remark 2: In problem (P_f^g), for any $x_1, x_2 \in X$, for a convex function f and μ-strongly convex function g,

$$\left(\tilde{\nabla} f(x_1) - \tilde{\nabla} f(x_2)\right)^T (x_1 - x_2) \geq 0,$$

$$\left(\tilde{\nabla} g(x_1) - \tilde{\nabla} g(x_2)\right)^T (x_1 - x_2) \geq \mu \|x_1 - x_2\|^2.$$

The following lemma is used in proving the convergence.

Lemma 1: **(Lemma 10, pg. 49 of [22])** Let $\{v_k\}$ be a sequence of nonnegative random variables, where $E[v_0] < \infty$, and let $\{\alpha_k\}$ and $\{\beta_k\}$ be deterministic scalar sequences such that: $\mathsf{E}[v_{k+1}|v_0, \ ..., v_k] \leq (1-\alpha_k)v_k + \beta_k$ for all $k \geq 0$, $0 \leq \alpha_k \leq 1$, $\beta_k \geq 0$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \beta_k < \infty$, $\lim_{k \to \infty} \frac{\beta_k}{\alpha_k} = 0$. Then, $v_k \to 0$, a.s., and $\lim\limits_{k \to \infty} \mathsf{E}[v_k] = 0$.

The next result will be used in our analysis.

Lemma 2: **(Theorem 6, pg. 75 of [18])** Let $\{u_t\}$ ($\subset \mathbb{R}^n$) be a convergent sequence such that it has a limit point $\hat{u} \in \mathbb{R}^n$ and consider another sequence $\{\alpha_k\}$ of positive numbers such that $\sum_{k=0}^{\infty} \alpha_k = \infty$. Suppose v_k is given by $v_k = \frac{\sum_{t=0}^{k-1}(\alpha_t u_t)}{\sum_{t=0}^{k-1} \alpha_t}$, for all $k \geq 1$. Then $\lim\limits_{k \to \infty} v_k = \hat{u}$.

Remark 3: From Assumption 1 (b, c, d), for all $x \in X$, the set $\partial f(x)$ is nonempty and bounded (cf. Ch. 3 of [3]). Similarly $\partial g(x)$ is nonempty and bounded for all $x \in X$.

Remark 4: From Remark 3, let us say that for any $x^{(i)} \in X_i$, there exists a scalar $C_{f,i}$ such that $\left\|\tilde{\nabla}_i f(x)\right\| \leq C_{f,i}$. Let $C_f \triangleq \sqrt{\sum_{i=1}^d C_{f,i}^2}$. Now we have, $\left\|\tilde{\nabla} f(x)\right\| \leq C_f$ for all $x \in X$. Similarly, $\left\|\tilde{\nabla} g(x)\right\| \leq C_g$ for all $x \in X$.

In the following lemma, we present the properties of $\{x_{\eta_k}^*\}$, which denotes the of solution of (P_η) for $\eta \in \{\eta_k\}$.

Lemma 3: **(Bound on $\mathbf{x}_{\eta_k}^*$, see Prop. 1 of [1])** Consider problem (P_f^g) and (P_η). Let Assumption 1 hold. Then, for the sequence $\{x_{\eta_k}^*\}$, and x_g^* for any $k \geq 1$, we have

(a) $\left\|x_{\eta_k}^* - x_{\eta_{k-1}}^*\right\| \leq \frac{C_g}{\mu}\left|\frac{\eta_{k-1}}{\eta_k} - 1\right|$.

(b) When $\{\eta_k\}$ goes to zero, $\{x_{\eta_k}^*\}$ converges to x_g^*.

Our objective is to show $\|x_{k+1} - x_g^*\| \to 0$. Now from the triangle inequality, $\|x_{k+1} - x_{\eta_k}^*\| \to 0$ and $\|x_{\eta_k}^* - x_g^*\| \to 0$. We know $\|x_{\eta_k}^* - x_g^*\| \to 0$ as $\eta_k \to 0$. Our main objective is to show $\|x_{k+1} - x_{\eta_k}^*\| \to 0$. Next we define an error function which will be used in the convergence analysis.

Definition 1: Let Assumption 2 hold. Then for any $x, y \in \mathbb{R}^n$, function $\mathcal{L}(x, y) = \sum_{i=1}^d \mathsf{p}_i^{-1}\left\|x^{(i)} - y^{(i)}\right\|^2$.

The following corollary holds from Definition 1.

Corollary 1: Consider Definition 1, p_{max} and p_{min} as defined in the notation, and let Assumption 2 hold. Then for any $x, y \in \mathbb{R}^n$, $\mathsf{p}_{max}\mathcal{L}(x, y) \leq \|x - y\|^2 \leq \mathsf{p}_{min}\mathcal{L}(x, y)$.

III. Convergence analysis of RB-IRG scheme

Here we begin with deriving a recursive error bound, that will be used later to show the almost sure convergence.

Lemma 4: **(Recursive relation for $\mathcal{L}\left(x_{k+1}, x^*_{\eta_k}\right)$)** Consider problem ($P_f^g$) and ($P_\eta$). Let Assumptions 1 and 2 hold. Let $\{x_k\}$ be the sequence generated from Algorithm 1. Let positive sequences $\{\gamma_k\}$, and $\{\eta_k\}$ be non-increasing and $\gamma_0\,\eta_0 < 1/\mu \mathsf{p}_{min}$. Then the following relation holds,

$$\mathsf{E}\big[\mathcal{L}\left(x_{k+1}, x^*_{\eta_k}\right)|\mathcal{F}_k\big] \leq (1-\mu\gamma_k\eta_k\mathsf{p}_{min})\,\mathcal{L}\left(x_k, x^*_{\eta_{k-1}}\right) + \frac{2C_g^2}{\mathsf{p}^2_{min}\mu^3\gamma_k\eta_k}\left(\frac{\eta_{k-1}}{\eta_k}-1\right)^2 + 2\gamma_k^2(C_f^2+\eta_0^2C_g^2).$$

Proof: Due to space limitation, find the proof in the arXiv version [19], Lemma 4. ■

A. Convergence analysis

Remark 5: Throughout the analysis, we assume that blocks are randomly selected using a uniform distribution.

Assumption 3: Let the following hold:

(a) $\{\gamma_k\}$ and $\{\eta_k\}$ are positive sequences for $k \geq 0$ converging to zero such that $\gamma_0\eta_0 < \frac{d}{\mu}$;

(b) $\sum_{k=0}^{\infty}\gamma_k\eta_k = \infty$; (c) $\sum_{k=0}^{\infty}\left(\frac{1}{\gamma_k\eta_k}\right)\left(\frac{\eta_{k-1}}{\eta_k}-1\right)^2 < \infty$;

(d) $\sum_{k=0}^{\infty}\gamma_k^2 < \infty$; (e) $\lim_{k\to\infty}\left(\frac{1}{\gamma_k^2\eta_k^2}\right)\left(\frac{\eta_{k-1}}{\eta_k}-1\right)^2 = 0$;

(f) $\lim_{k\to\infty}\frac{\gamma_k}{\eta_k} = 0$.

Next, we show the a.s. convergence of the sequence $\{x_k\}$.

Theorem 1: **(a.s. convergence of $\{x_k\}$)** Consider (P_f^g) and (P_η). Let Assumption 3 hold. Consider the sequence $\{x_k\}$ is obtained by Algorithm 1, and the sequence $\{x^*_{\eta_k}\}$ suppose obtained by solving (P_η). Then, $\mathcal{L}\left(x_k, x^*_{\eta_{k-1}}\right)$ goes to zero a.s. and $\lim\limits_{k\to\infty}\mathsf{E}\Big[\mathcal{L}\left(x_k, x^*_{\eta_{k-1}}\right)\Big] = 0$.

Proof: We apply Lemma 1 to the result of Lemma 4. $v_k \triangleq \mathcal{L}\left(x_k, x^*_{\eta_{k-1}}\right)$, $\alpha_k \triangleq \frac{\mu\gamma_k\eta_k}{d}$, $\beta_k \triangleq \left(\frac{2d^2}{\mu\gamma_k\eta_k}\right)\frac{C_g^2}{\mu^2}\left(\frac{\eta_{k-1}}{\eta_k}-1\right)^2 + 2\gamma_k^2(C_f^2+\eta_0^2C_g^2)$. Now, in order to claim the convergence of v_k, we show that all conditions of Lemma 1 hold. Note that $\mathsf{p}_i = 1/d$. From Assumption 3 (a), definition of $\{\gamma_k\}$, $\{\eta_k\}$, and from $\gamma_0\eta_0 < \frac{d}{\mu}$, the first condition of Lemma 1 is satisfied. Now consider sequence β_k. From Assumption 3 (a), sequences $\{\gamma_k\}$, $\{\eta_k\}$ and the constant μ are positive, so the second condition of Lemma 1 is satisfied. Now in $\sum_{k=0}^{\infty}\alpha_k$, i.e. $\sum_{k=0}^{\infty}\frac{\mu\gamma_k\eta_k}{d}$. From Assumption 3(b), the third condition of Lemma 1 holds. Now from the definition of β_k and from Assumption 3(c) and (d), the fourth condition of Lemma 1 holds. Finally consider $\lim_{k\to\infty}\left(\frac{\beta_k}{\alpha_k}\right) = 0$. Using the definition of β_k and Assumption 3(e, f), condition 5 of Lemma 1 holds. Thus we get the required result. ■

Next in Lemma 5 we give the choice of sequences γ_k and η_k that satisfy Assumption 3.

Lemma 5: Let Assumption 2 hold. Then sequences $\{\gamma_k\}$ and $\{\eta_k\}$ given by $\gamma_k = \gamma_0(k+1)^{-a}$ and $\eta_k = \eta_0(k+1)^{-b}$ where a, and b satisfy, $a > 0, \quad b > 0, \quad a+b < 1, \quad b < a, \quad a > 0.5$, where $\gamma_0 > 0$ and $\eta_0 > 0$. Then $\{\gamma_k\}$ and $\{\eta_k\}$ satisfy Assumption 3.

Proof: Similar to the proof of Lemma 5 in [31]. Omitted because of the space requirements. ■

Next, we show the a.s. convergence of the sequence $\{\bar{x}_k\}$.

Theorem 2: **(a.s. convergence of $\{\bar{x}_k\}$)** Consider problem (P_f^g). Let γ_k and η_k be the sequences defined by Lemma 5 where $\gamma_0 > 0$, $\eta_0 > 0$, and $ar < 1$. Then $\{\bar{x}_k\}$ converges to the unique solution of (P_f^g), x_g^* a.s.

Proof: From $\lambda_{t,k} = \gamma_t^r/\sum_{j=0}^{k}\gamma_j^r$, $\left\|\bar{x}_k - x_g^*\right\| = \left\|\sum_{t=0}^{k}\lambda_{t,k}x_t - \sum_{t=0}^{k}\lambda_{t,k}x_g^*\right\| = \left\|\sum_{t=0}^{k}\lambda_{t,k}\left(x_t - x_g^*\right)\right\|$. Using the triangle inequality, $\left\|\bar{x}_k - x_g^*\right\| \leq \sum_{t=0}^{k}\lambda_{t,k}\left\|x_t - x_g^*\right\|$. From definition of $\lambda_{t,k}$, $\left\|\bar{x}_k - x_g^*\right\| \leq \frac{\sum_{t=0}^{k}\gamma_t^r\left\|x_t - x_g^*\right\|}{\sum_{j=0}^{k}\gamma_j^r}$. Comparing with Lemma 2, $\alpha_k \triangleq \gamma_k^r$, $u_k \triangleq \left\|x_k - x_g^*\right\|$, $v_{k+1} \triangleq \sum_{t=0}^{k}\gamma_t^r\left\|x_t - x_g^*\right\|$. Consider $\sum_{k=0}^{\infty}\alpha_t$, i.e. $\sum_{k=0}^{\infty}(1+k)^{-at}$. Now, we have $at < 1$, so $\sum_{k=0}^{\infty}(1+k)^{-at} = \infty$. From Theorem 1, $\left\|x_k - x_g^*\right\| \to 0$ a.s. Therefore, Using Lemma 2, we get the required result. ■

IV. Rate of convergence

In this section, first we derive the rate of convergence for the sequence generated from RB-IRG scheme.

Lemma 6: **(Feasibility error bound for Algorithm 1)** Consider problem (P_f^g) and $\{\bar{x}_k\}$, the sequence generated by Algorithm 1. Let Assumption 1 hold, $r(<1)$ be an arbitrary scalar, and γ_k be a non-increasing sequence. Let η_k be a non-increasing sequence and X to be bounded, i.e. $\|x\| \leq M$ for all $x \in X$ for some $M > 0$. Then for any $z \in X$, the following holds,

$$\mathsf{E}[f\left(\bar{x}_N\right)] - f(z) \leq \left(\sum_{i=0}^{N-1}\gamma_i^r\right)^{-1}\left(2M_g\sum_{k=0}^{N-1}\gamma_k^r\eta_k + 2\mathsf{p}_{max}^{-1}M^2\left(\gamma_0^{r-1}+\gamma_{N-1}^{r-1}\right) + \left(\sum_{k=0}^{N-1}\gamma_k^{r+1}\right)\left(C_f^2+C_g^2\eta_0^2\right)\right),$$

where $M_g(>0)$ is a scalar such that $g(x) \leq M_g$ for all $x \in X$.

Proof:

Due to space limitation, find the proof in the arXiv version [19], Lemma 6. ■

Next, we state Lemma 7 (see Lemma 9, pg. 418 of [31]) and use it in Theorem 3 to derive the rate of convergence.

Lemma 7: For a scalar $\alpha \neq -1$ and integers l, N, where $0 \leq l \leq N-1$, we have

$$\frac{N^{\alpha+1}-(l+1)^{\alpha+1}}{\alpha+1} \leq \sum_{k=l}^{N-1}(k+1)^{\alpha} \leq (l+1)^{\alpha} + \frac{(N+1)^{\alpha+1}-(l+1)^{\alpha+1}}{\alpha+1}.$$

In Theorem 3, we show the rate of convergence for the sequence generated from RB-IRG scheme.

Theorem 3: Consider problem (P_f^g) and the sequence generated from Algorithm 1 $\{\bar{x}_N\}$. Let Assumptions 1, and 2 with a uniform distribution. Let the sequence $\{\gamma_k\}$ and $\{\eta_k\}$ are given by the following, $\gamma_k = \gamma_0/(k+1)^{0.5+0.1\delta}$ and

$\eta_k = \eta_0/(k+1)^{0.5-\delta}$, such that $\gamma_0 \triangleq \gamma\sqrt{d}$, for some $\gamma > 0$, $\eta_0 > 0,\ \gamma\eta_0 < \frac{\sqrt{d}}{\mu},\ 0 < \delta < 0.5,$ and $r < 1$. Then,
(i) Sequence $\{\bar{x}_N\}$ converges to x_g^* almost surely.
(ii) $\mathsf{E}[f(\bar{x}_N)] \to f^*$ with the rate $\mathcal{O}\left(\sqrt{d}/N^{0.5-\delta}\right)$.

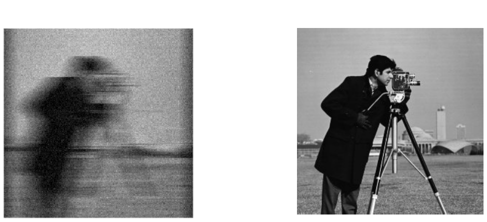
Proof: (i) Consider the sequences given for γ_k and η_k. By denoting $a = 0.5 + 0.1\delta$ and $b = 0.5 - \delta$, we have, $\gamma_k = \gamma_0/(k+1)^a$ and $\eta_k = \eta_0/(k+1)^b$. Also we know that $0 < \delta < 0.5$ and $r < 1$. Therefore, we have: $a, b > 0,\ b < a,\ 0.5 < a < 0.55,\ 0 < b < 0.5,\ a + b < 1$ and $ar < 1$. So, γ_k and η_k satisfy all the conditions of Lemma 2.
(ii) Substituting γ_k, η_k, and $z := x_g^*$ in Lemma 6, we obtain,

$$\mathsf{E}[f(\bar{x}_N)] - f^* \leq \left(\gamma_0^r \sum_{i=0}^{N-1} \frac{1}{(k+1)^{ar}}\right)^{-1} \left(2\mathsf{p}_{max}^{-1} M^2 \gamma_0^{r-1} \left(N^{a(1-r)} + 1\right) + \gamma_0^r \left(2M_g\eta_0 \sum_{k=0}^{N-1} \frac{1}{(k+1)^{ar+b}} + \left(C_f^2 + C_g^2\eta_0^2\right)\gamma_0 \underbrace{\sum_{k=0}^{N-1} \frac{1}{(k+1)^{ar+a}}}_{\text{term-1}}\right)\right).$$

modifying term-1 in equation above, and expanding terms,

$$\mathsf{E}[f(\bar{x}_N)] - f^* \leq 2\mathsf{p}_{max}^{-1} M^2 \gamma_0^{-1} \left(\underbrace{\left(\sum_{i=0}^{N-1} \frac{1}{(k+1)^{ar}}\right)^{-1} N^{a(1-r)}}_{\text{term-3}} + \underbrace{\left(\sum_{i=0}^{N-1} \frac{1}{(k+1)^{ar}}\right)^{-1}}_{\text{term-2}}\right) + \left(2M_g\eta_0 + \gamma_0\left(C_f^2 + C_g^2\eta_0^2\right)\right).$$

The above equation can also be written as
$\mathsf{E}[f(\bar{x}_N)] - f^* \leq 2dM^2\gamma^{-1}d^{-0.5}\,(\text{term-3} + \text{term-2}) + \text{term-4}\left(2M_g\eta_0 + \gamma\sqrt{d}\left(C_f^2 + C_g^2\eta_0^2\right)\right)$.
From Lemma 7, we have, term-2 $\leq \frac{1-ar}{N^{-ar+1}-1} = \mathcal{O}\left(N^{-(1-ar)}\right)$, term-3 $\leq \frac{(1-ar)N^{a(1-r)}}{N^{-ar+1}-1} = \mathcal{O}\left(N^{-(1-a)}\right)$, term-4 $\leq \left(\frac{1-ar}{N^{-ar+1}-1}\right)\left(1 + \frac{(N+1)^{1-(ar+b)}-1}{1-(ar+b)}\right) = \mathcal{O}\left(N^{-(1-ar)}\right) + \mathcal{O}\left(N^{-b}\right)$. Now, substituting bounds of terms-2, 3, and 4, we have,
$\mathsf{E}[f(\bar{x}_N)] - f^* \leq \mathcal{O}\left(\sqrt{d}\mathsf{max}\left\{N^{-(1-ar)}, N^{-(1-a)}, N^{-b}\right\}\right)$
$= \mathcal{O}\left(\sqrt{d}N^{-\mathsf{min}\{1-ar,\ 1-a,\ b\}}\right)$.
From definitions of $a, r,$ and δ, we obtain the result. ■


(a) Blurred image

(b) Original image

Fig. 1: Blurred and original image of cameraman

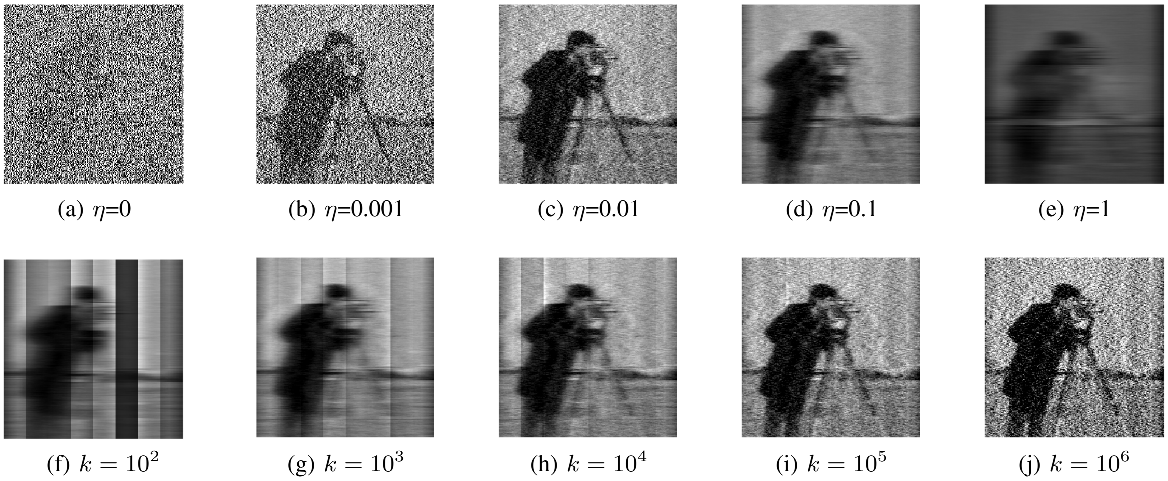
V. Numerical results

In the literature, one of the ways to address the ill-posedness in image deblurring is employing the regularization technique. The ill-posed problem (2) is converted into the regularized problem (P_η) by, for example, substituting functions $f(x) := \|Ax - b\|_2^2$, and $g(x) := \|x\|_2^2 + \|x\|_1$. As the value of regularization parameter η changes, a different optimization problem (P_η) is solved. The basic idea is, $\eta(\in (0, +\infty))$ governs the way by which solutions of linear inverse problem (2) are approximated by (P_η). This is a *two-loop* scheme, as explained earlier in the introduction. It is computationally inefficient to find a suitable regularization parameter η. In this section, we address this challenge in an image deblurring using problem formulation (3), by successfully avoiding the conventional *two-loop* regularization technique. The values of regularization parameter η_k and stepsize parameter γ_k are updated iteratively, explained below in the inference.

We are provided with the blurred noisy image Fig. 1(a), which is further converted into the column vector b. Our objective is to get the original image, Fig. 1 (a) using image deblurring. Here we compare two ways of deblurring: standard regularization, and using RB-IRG (proposed algorithm). **Inference:** Fig. 2(a)–(e) show the deblurred images obtained by conventional regularization at different η for 10^5 iterations. Fig. 2(f)–(j) show the deblurred images using RB-IRG scheme with stopping at different iteration. Our scheme (RB-IRG) is computationally efficient, because unlike the case of *two-loop* regularization, in (RB-IRG) we implement the scheme once. A question then is: what iteration k we should stop the scheme at? Stopping at a suitable iteration k is desired because that governs the deblurred image quality. In particular, this *single-loop* scheme seems to be promising because one can generate images after every fixed number of iterations and stop the implementation once the generated deblurred image is good enough. Note that, image deblurring is used for a toy example here, to demonstrate the performance of (RB-IRG) scheme. This application can be extended for deblurring of images with a higher resolution. In this work, our intention is to demonstrate the performance of RB-IRG scheme on the well-known example of cameraman.

VI. Concluding remarks

We address ill-posed optimization problem with a high dimensional solution space and nondifferentiable objective function. A randomized block coordinate iterative regularized subgradient (RB-IRG) scheme is developed to address problem (P_f^g). We establish the convergence of the sequence generated from RB-IRG scheme to the unique solution of (P_f^g) in an almost sure sense. Furthermore, we derive a rate of convergence $\mathcal{O}\left(\frac{\sqrt{d}}{k^{0.5-\delta}}\right)$, with respect to the inner level objective of the bilevel problem (P_f^g). Our ground assumptions in the convergence proof and rate analysis are mild, such that f and g can be nondifferentiable functions. Demonstration of RB-IRG on an image deblurring example shows that the proposed *single-loop* scheme computationally performs


(a) η=0 (b) η=0.001 (c) η=0.01 (d) η=0.1 (e) η=1

(f) $k = 10^2$ (g) $k = 10^3$ (h) $k = 10^4$ (i) $k = 10^5$ (j) $k = 10^6$

Fig. 2: First row: Image deblurring using the regularization technique with different values of η, running for 10^5 iterations. Second row: Image deblurring using RB-IRG scheme, stopping at different iterations k.

well compared to the conventional *(two-loop)* regularization schemes.

References

[1] M. Amini and F. Yousefian, "An Iterative Regularized Mirror Descent Method for Ill-posed Nondifferentiable Stochastic Optimization," arXiv:1901.09506

[2] N. S. Aybat, Z. Wang, T. Lin, and S. Ma, "Distributed linearized alternating direction method of multipliers for composite convex consensus optimization," *IEEE Transactions On Automatic Control*, vol. 63, no. 1, pp. 5-19, 2018.

[3] A. Beck, *First-order methods in optimization, MOS-SIAM Series on Optimization*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2017.

[4] A. Beck and S. Sabach, "A first-order method for finding minimal norm-like solutions of convex optimization problems," *Mathematical Programming*, vol. 147, no. 2, pp. 25-46, 2014.

[5] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*, Academic Press, New York, 1982.

[6] D. P. Bertsekas, "Necessary and sufficient conditions for a penalty method to be exact," *Mathematical Programming*, vol. 9, no. 1, pp. 87-99, 1975.

[7] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, Belmont, MA, 2003.

[8] J. V. Burke, "An exact penalization viewpoint of constrained optimization," *SIAM Journal of Control And Optimization*, vol. 29, no. 4, pp. 968-998, 1991.

[9] K.-W. Chang, C.-J. Hsieh, and C.-J. Lin, "Coordinate descent method for large-scale ℓ_2-loss linear support vector machines," *Journal of Machine Learning Research*, vol. 9, pp. 1969-1398, 2008.

[10] K. Cohen, A. Nedić, and R. Srikant, "On Projected Stochastic Gradient Descent Algorithm with Weighted Averaging for Least Squares Regression," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5974-5981, 2017.

[11] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust region methods*, MPS-SIAM Series on Optimization, Society of Industrial and Applied Mathematics, Philadelphia, 2000.

[12] F. Facchinei and J. S. Pang, *Finite-dimensional variational inequalities and complementarity problems*, Springer-Verlag New York, New York, 2003.

[13] R. Fletcher, *An ℓ_1 Penalty method for nonlinear constraints, in numerical optimization 1984*, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., Philadelphia, 1985, Society of Industrial and Applied Mathematics, pp. 26-40.

[14] M. P. Friedlander and P. Tseng, "Exact regularization of convex programs," *SIAM Journal of Optimization*, vol. 18, no. 4, pp. 1326-1350, 2007.

[15] G. Garrigos, L. Rosasco, and S. Villa, "Iterative regularization via dual diagonal descent," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 2, pp. 189-215, 2018.

[16] S. -P. Han and O. L. Mangasarian, "Exact penalty function in nonlinear programming," *Mathematical Programming*, vol. 17, no. 1, pp. 251-269, 1979.

[17] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S.-S. Keerthi, and S. Sundararajan, "A dual coordinate descent method for large-scale linear svm," *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 2008.

[18] K. Knopp, *Theory and applications of infinite series, Blackie & Son Ltd.*, Glasgow, Great Britain, 1951.

[19] H. Kaushik and F. Yousefian, "A randomized block coordinate iterative regularized gradient method for high-dimensional ill-posed convex optimization," arXiv:1809.10035

[20] O. L. Mangasarian, "Sufficiency of exact penalty minimization," *SIAM Journal on Control and Optimization*, vol. 23, no. 1, pp. 30-37, 1985.

[21] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341-362, 2012.

[22] B. T. Polyak, *Introduction to optimization*, Optimization Software, Inc., New York, 1987.

[23] P. Ricktárik and M. Takáč, "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function," *Mathematical Programming*, vol. 144, no. 2, pp. 1-38, 2014.

[24] S. Sabach and S. Shtern, "A first order method for solving convex bilevel optimization problems," *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 640-660, 2017.

[25] S. Shalev-Shwartz and A. Tewari, "Stochastic methods for ℓ_1 regularized loss minimization," *Journal of Machine Learning Research*, vol. 12, pp. 1865-1892, 2011.

[26] S. Shalev-Shwartz and T. Zhang, "Stochastic dual coordinate ascent methods for regularized loss minimization," *Journal of Machine Learning Research*, vol. 14, pp. 597-599, 2013.

[27] M. Solodov, "An explicit descent method for bilevel convex optimization," *Journal of Convex Analysis*, vol. 14, no. 2, pp. 227-238, 2007.

[28] A. N. Tikhonov and V. Y. Arsenin, *Solutions of ill-posed problems*, V. H. Winston and Sons, Washington, D. C., 1977. Translated from Russian.

[29] E. Wei, A. Ozdaglar, "On the $\mathcal{O}(1/k)$ convergence of asynchronous distributed alternating direction method of multipliers," *Global Conference On Signal And Information Processing (IEEE)*, 2013.

[30] Y. Xu, "Accelerated first-order primal dual proximal methods for linearly constrained composite convex programming," *SIAM Journal Of Optimization*, vol. 27, no. 3, pp. 1459-1484, 2017.

[31] F. Yousefian, A. Nedić, and U. V. Shanbhag, "On smoothing, regularization and averaging in stochastic approximation methods for stochastic variational inequality problems," *Mathematical Programming*, vol. 165, no. 1, pp. 391-431, 2017.