Working Paper

Proximal point mappings and constraint aggregation principle

Mikhail Davidson

WP-96-102
August 1996

IIASA
International Institute for Applied Systems Analysis ◻ A-2361 Laxenburg ◻ Austria
Telephone: +43 2236 807 ◻ Fax: +43 2236 71313 ◻ E-Mail: info@iiasa.ac.at

Proximal point mappings and constraint aggregation principle

Mikhail Davidson

WP-96-102
August 1996

IIASA International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 807 □ Fax: +43 2236 71313 □ E-Mail: info@iiasa.ac.at

Abstract

In the present paper we study a regularization techniques for the constraint aggregation method for solving large scale convex optimization problems. Idea of the constraint aggregation is in replacing the set of original constraints by a single one which is a certain linear combination of them. This makes the resulting relaxed problem much easier to solve. However, previous algorithms that used this scheme exhibited quite a slow convergence. The motivation for the present work was to make an attempt to improve the convergence by using the idea of constraint aggregation in the framework of proximal point method. In the paper we propose the regularized constraint aggregation method and conduct its convergence analysis. Estimates for the rate of convergence of the trajectory to the feasible set and to the optimal solution set are derived under certain regularity assumptions. These estimates appear to be better than those for the method without regularization. Comparative numerical tests of both algorithms are reported.

Key words: Regularization, proximal-point method, constraint aggregation, nonsmooth optimization, error bounds.

Proximal point mappings and constraint aggregation principle

Mikhail Davidson

1 Introduction

We are interested in solving the following optimization problem

$$\min_{x \in X} f(x) \tag{1}$$

subject to

$$Ax \leq b, \tag{2}$$

where $x \in \mathbf{R}^n$, $f : \mathbf{R}^n \to \mathbf{R}$ is convex and possibly nonsmooth, A is a $m \times n$ matrix, $b \in R^m$, and the set X is convex and compact. We also assume that the dimension of the problem (the numbers n and m) is very large and therefore direct use of classical iterative nonsmooth optimization methods for this problem, such as *bundle methods* (see, e.g. [1, 2]) or *projected subgradient methods* (see [3, 4, 5]), in practice encounter substantial difficulties because of necessity to solve a very complicated subproblem at each iteration.

The issue of increasing dimension is characteristic for optimization nowdays. Modern applications in modelling (particularly, modelling in presence of uncertainty), stochastic programming provide examples of such problems. Therefore, there is a need for special techniques addressing these issues.

Recently, in [6] there has been suggested an approach reffered to as *constraint aggregation principle* aimed at overcoming these difficulties. We shall assume that the structure of X is simple and the main difficulty comes from the large number of constraints (2). The idea of the constraint aggregation is in replacing (2) with a certain linear combination of the constraints of (2) which makes the resulting problem much easier to solve. More specifically, let y^k be the current solution approximation, then the subproblem to be solved (as suggested in [6]) is

$$\min_{x \in X} f(x)$$

subject to

$$\langle [Ay^k - b]_+, \ Ax - b \rangle \leq 0,$$

where $[\cdot]_+$ denotes $\max[\cdot, \ 0]$ (componentwice). The subsequent solution approximation y^{k+1} is obtained by moving from y^k towards the solution of the relaxed subproblem z^k:

$$y^{k+1} = y^k + \tau_k (z^k - y^k),$$

where $\tau_k \geq 0$ are such that $\tau_k \to 0$ and $\sum_k \tau_k = +\infty$. In sequel the method described will be reffered to as *constraint aggregation method (CAM)*. Constraint aggregation (however in a somewhat different form) has been used in earlier work [7] for numerically solving certain control problems.

One drawback of CAM is slow convergence. The reason for this is that the solution z^k to the relaxed subproblem can diviate very far from the current solution approximation and the "average course" of the trajectory, which dictates quite strong conditions on the stepsize and causes slow decrease in the norm of the residual $|[Ay^k - b]_+|$.

A way to improve the situation may be to use certain regularization techniques. This is the subject of the present paper. We introduce the regularized version of CAM and study its behavior. Namely, we make an attempt to use the constraint aggregation principle in the general framework of the proximal point method.

The paper is organized as follows. In section 1 the regularized constraint aggregation method is introduced and it is proved that the trajectory of the method is convergent to the set of optimal solutions of problem (1)-(2) without any regularity assumptions about the problem's data. In the rest of the paper we study the convergence properties of the method under the regularity assumptions regarding (a) the feasible set and (b) the set of optimal solutions of problem (1)-(2). In section 3, using only the first type assumptions we prove that the trajectory is converging to an optimal point. In section 4 we consider a special case where $f(x) \equiv 0$, i.e. the problem reduces to solving a system of inequalities. Here, linear convergence rate estimates are obtained and a simple criterion of consistency of the system is derived. Note that this criterion does not use the regularity assumptions. Section 5 is devoted to deriving convergence rate estimate of the trajectory of the method to the set of optimal solutions in the general case (where f is not necessarily zero). For this we need at first to obtain the convergence rate of the trajectory to the feasible set. These rates appear to be assymptotically better than the corresponding rates for CAM. They are also essentially dependent on the regularity constants of the set of optima and the feasible set of the problem in question. The latter provides an explanation for certain qualitative fenomena in the behavior of the regularized method, see section 7. Section 7 also contains numerical results of comparative tests of both algorithms.

2 Regularized constraint aggregation method

2.1 Definition of the algorithm RCAM

Choose the sequence of numbers $\tau_k \geq 0$ such that $\tau_k \to 0$ and $\sum_k \tau_k = +\infty$. Let the starting point x^0 be an arbitrary point from X and $x^k \in X$ be a current solution approximation. Define x^{k+1} as

$$x^{k+1} := \arg\min_{x \in X}[\tau_k f(x) + |x - x^k|^2/2] \tag{3}$$

subject to

$$\langle [Ax^k - b]_+,\ Ax - b \rangle \leq 0.$$

Throughout the paper the algorithm described will be referred to as *regularized constraint aggregation method (RCAM)*. Note that in RCAM the parameters τ_k are no longer the stepsizes but rather they play a role in a way similar to inverse penalty multipliers.

2.2 Convergence analysis

In this section we prove that the algorithm defined above is convergent to the set of optimal solutions of (1)-(2). Below we will use the following notation

$$\mathcal{X} := \{x \in X | Ax \leq b\}$$

is the constraint set of (1)-(2);

$$\mathcal{X}^* := \{x \in \mathcal{X} | f(x) \leq f_*\}$$

is the optimal solution set of (1)-(2) (with f_* being the optimum of (1)-(2)).

In the proof we will use the Euclidean distance function $\rho(\cdot, \mathcal{X}^*)$ to the optimal solution set $\mathcal{X}^*$ as a merit function. Let x^* be an arbitrary point from $\mathcal{X}^*$, then

$$\rho(x^{k+1}, \mathcal{X}^*) \leq |x^{k+1} - x^*|,$$

and the following arguments provide an estimate for the right-hand side of the latter inequality. We have

$$|x^{k+1} - x^*|^2 = |x^k - x^*|^2 + 2\langle x^{k+1} - x^k,\ x^k - x^*\rangle + |x^{k+1} - x^k|^2. \tag{4}$$

KKT optimality conditons in the subproblem (3) yield

$$\langle \tau_k \partial f(x^{k+1}) + x^{k+1} - x^k + \lambda A^T[Ax^k - b]_+,\ x - x^{k+1}\rangle \geq 0 \tag{5}$$

for all $x \in X$, where λ is an optimal dual multiplier to the aggregate constraint. Set $x = x^*$, then using nonnegativity of λ, complementarity slackness, and the fact that x^* is admissible for the relaxed set in (3), one obtains

$$\begin{aligned}\lambda\langle A^T[Ax^k - b]_+,\ x^* - x^{k+1}\rangle = \lambda\langle [Ax^k - b]_+,\ Ax^* - b\rangle + \\ \lambda\langle [Ax^k - b]_+,\ b - Ax^{k+1}\rangle \leq 0.\end{aligned} \tag{6}$$

Using this estimate in (5) we obtain

$$\langle \tau_k \partial f(x^{k+1}),\ x^* - x^{k+1}\rangle + \langle x^{k+1} - x^k,\ x^* - x^{k+1}\rangle \geq 0.$$

The second term in (4) can be rewritten as follows

$$\langle x^{k+1} - x^k,\ x^k - x^*\rangle = -|x^{k+1} - x^k|^2 + \langle x^{k+1} - x^k,\ x^{k+1} - x^*\rangle$$

and we can use the previous estimate to obtain

$$\begin{aligned}\rho^2(x^{k+1}, \mathcal{X}^*) \leq |x^{k+1} - x^*|^2 = |x^k - x^*|^2 + 2\langle x^{k+1} - x^k,\ x^{k+1} - x^*\rangle - \\ |x^{k+1} - x^k|^2 \leq |x^k - x^*|^2 + 2\tau_k(f_* - f(x^{k+1})) - |x^{k+1} - x^k|^2\end{aligned}$$

for each $x^* \in X^*$. Thus,

$$\rho^2(x^{k+1}, \mathcal{X}^*) \leq \rho^2(x^k, \mathcal{X}^*) + 2\tau_k(f_* - f(x^{k+1})) - |x^{k+1} - x^k|^2. \tag{7}$$

Let us fix a sequence of numbers $\delta_k \geq 0$ such that

$$\delta_k \to 0, \quad \sum_k \delta_k = +\infty, \quad \lim_{k\to\infty} \delta_k/\tau_k = 0,$$

and define two sets of indexes

$$K_- := \{k = 1, 2, \ldots \mid 2\tau_k(f(x^{k+1}) - f_*) + |x^{k+1} - x^k|^2 \geq \delta_k\};$$
$$K_+ := \{k = 1, 2, \ldots\} \setminus K_-.$$

Evidently, K_+ is infinite. Indeed, suppose K_+ is finite and $|K_+| \leq N$. Then for all $k \geq N$ one has

$$\rho^2(x^{k+1}, \mathcal{X}^*) \leq \rho^2(x^k, \mathcal{X}^*) - \delta_k.$$

Summing up these inequalities for each k and recalling that the series of δ_k is devergent we arrive at a contradiction with the compactness of X. Thus, K_+ contains an infinite number of elements.

Next we obtain a bound for the norm $|[Ax^{k+1} - b]_+|^2$ via $|x^{k+1} - x^k|^2$:

$$|[Ax^{k+1} - b]_+|^2 = |[Ax^{k+1} - Ax^k + Ax^k - b]_+|^2 \leq |A(x^{k+1} - x^k) + [Ax^k - b]_+|^2 =$$
$$|[Ax^k - b]_+|^2 + 2\langle A(x^{k+1} - x^k),\ [Ax^k - b]_+\rangle + |A(x^{k+1} - x^k)|^2.$$

The scalar product can be bounded as follows

$$\langle Ax^{k+1} - b,\ [Ax^k - b]_+\rangle + \langle b - Ax^k,\ [Ax^k - b]_+\rangle \leq -|[Ax^k - b]_+|^2$$

which upon the substitution and rearrangement yields

$$|[Ax^{k+1} - b]_+|^2 + |[Ax^k - b]_+|^2 \leq |A(x^{k+1} - x^k)|^2 \leq |A|^2\ |x^{k+1} - x^k|^2. \tag{8}$$

For each $k \in K_+$, therefore, one has

$$2\tau_k(f(x^{k+1}) - f_*) + |[Ax^{k+1} - b]_+|^2/|A|^2 \leq \delta_k.$$

Let $\{k_t | t = 1, 2, \ldots\}$ be an arbitrary infinite subsequence of indexes from the set K_+. By definition of the sequence of δ_k one has

$$f(x^{k_t+1}) + \frac{|[Ax^{k_t+1} - b]_+|^2}{2|A|^2\tau_{k_t}} - f_* \to 0, \quad t \to \infty. \tag{9}$$

Since the difference $f(x^{k_t+1}) - f_*$ is bounded then the norm $|Ax^{k_t+1} - b|^2$ tends to zero as t tends to infinity which means that the limitting set of the subsequence of x^{k_t+1} belongs to $\mathcal{X}$. In fact, a stronger assertion holds, namely

$$\frac{|[Ax^{k_t+1} - b]_+|^2}{\tau_{k_t}} \to 0, \quad t \to \infty. \tag{10}$$

Indeed, assume the contrary, i.e.,

$$\limsup \frac{|[Ax^{k_t+1} - b]_+|^2}{\tau_{k_t}} = \alpha > 0.$$

Therefore, from (9),

$$\liminf f(x^{k_t+1}) - f_* \leq -\alpha.$$

At the same time, since the limitting set of $\{x^{k_t+1}\}$ belongs to $\mathcal{X}$ we arrive at a contradiction.

Now, formulas (9) and (10) immediately imply that $f(x^{k_t+1})-f_* \to 0$, i.e. the limitting set of $\{x^{k_t+1}\}$ belongs, in fact, to $\mathcal{X}^*$:

$$\rho(x^{k_t+1}, \mathcal{X}^*) \to 0, \quad t \to \infty.$$

Next we prove that the whole sequence of $x^k, \ k = 1, 2 \ldots$ is converging to the optimal solution set $\mathcal{X}^*$. By definition of K_- and K_+, the latter formulae and (7) provide that

$$\rho^2(x^{k+1}, \mathcal{X}^*) \le \max[\kappa(\delta_k), \ \rho^2(x^k, \mathcal{X}^*) - \delta_k]$$

where $\kappa(\delta_k)$ tends to zero together with δ_k. This is wellknown Wasan's inequality [8] that implies $\rho^2(x^k, \mathcal{X}^*) \to 0$.

Thus, we conclude that the whole sequence $\{x^k\}$ is converging to the optimal solution set $\mathcal{X}^*$.

3 Regular case

This section and the following sections are devoted to the convergence analysis of the algorithm in the case where the problem (1)-(2) satisfies certain regularity assumptions. These assumptions postulate the error bounds relating the distance to solution set of a system of inequalities or optimal solution set of an optimization problem to the residual of this system or the optimality gap calculated at some point. We will be interested in regularity of the feasible set $\mathcal{X}$ of problem (1)-(2) and the optimal solution set $\mathcal{X}^*$ of this problem.

Definition 1. *The set $\mathcal{X}$ is said to be α-regular, $\alpha \ge 1$, if there exists a constant l_1 such that for all $x \in X$*

$$|[Ax - b]_+| \ge l_1 \rho^\alpha(x, \mathcal{X}).$$

Definition 2. *The problem (1)-(2) is said to be α-regular, $\alpha \ge 1$, if there exists a constant l_2 such that for all $x \in \mathcal{X}$*

$$f(x) - f_* \ge l_2 \rho^\alpha(x, \mathcal{X}^*).$$

Thus, the assumption of α-regularity requires that the norm of the residual of the system of inequalities or the optimality gap be bounded from below with a polynomial of degree α of the distance function.

In the case $\mathcal{X} = \mathbf{R}^n$ the notion of α-regularity (as in definition 2) was introduced in [9]. In [10], the case of arbitrary convex constraint set was considered. In the case $\alpha = 1$, definition 2 coincides with the concept of the *weak sharp minimum*, see, for example, [11] and [12].

In many cases of practical importance the assumption of regularity for the constraint set $\mathcal{X}$ is less restrictive than that for the set $\mathcal{X}^*$. Furthemore, it is often the case that for the constant l_1, in practice, reasonable lower bounds can be constructed while the constant l_2 is often much smaller than l_1 and it is much more difficult to bound it from below.

As will be shown later convergence rate estimates for RCAM depend essentially on l_1 and l_2 and difference between these constants can provide some insight on the behavior of the trajectory generated by the algorithm.

In this section we prove that only under the assumption of α-regularity for the feasible set $\mathcal{X}$, the trajectory is convergent to an optimal point. In other words, this property is invariant with regard to the objective function (within the class of Lipschitzian functions).

Theorem 1. *Let $f(x)$ be Lipschitzian with the constant L and the set $\mathcal{X}$ be α-regular with $\alpha \geq 1$. Define the sequence of numbers $\tau_k \geq 0$ as follows*

$$\tau_k \to 0, \ \sum \tau_k = +\infty, \ \ \sum \tau_k^{2\alpha/(2\alpha-1)} < \infty.$$

Then the sequence of $x^k \in X$ generated by the algorithm is convergent to some point $y^* \in \mathcal{X}^*$.

Proof. In the previous section it was proved that the sequence $\{x^k\}$ is converging to the optimal solution set $\mathcal{X}^*$. Let y^* be an arbitrary limit point of $\{x^k\}$. Similar to the previous section we obtain

$$|x^{k+1} - y^*|^2 \leq |x^k - y^*|^2 + 2\tau_k(f_* - f(x^{k+1})) - |x^{k+1} - x^k|^2,$$

and using (8) we can write this as follows

$$|x^{k+1} - y^*|^2 \leq |x^k - y^*|^2 + 2\tau_k(f_* - f(x^{k+1})) - |[Ax^{k+1} - b]_+|^2/|A|^2.$$

For $f(x^{k+1})$ one can write the following estimates

$$f(x^{k+1}) \geq f(y^{k+1}) - L\rho(x^{k+1}, \mathcal{X}) \geq f_* - L\rho(x^{k+1}, \mathcal{X}),$$

where $\bar{y}^{k+1} \in \mathcal{X}$ minimizes the distance from x^{k+1} to $\mathcal{X}$.

Then using regularity of the set $\mathcal{X}$ we arrive at the following inequality

$$|x^{k+1} - y^*|^2 \leq |x^k - y^*|^2 + 2\tau_k L\rho_{k+1} - l_1^2\rho_{k+1}^{2\alpha}/|A|^2,$$

where $\rho_{k+1} \equiv \rho(x^{k+1}, \mathcal{X})$. Maximizing the right-hand side with respect to ρ_{k+1} one obtains

$$|x^{k+1} - y^*|^2 \leq |x^k - y^*|^2 + \text{const}\,\tau_k^{2\alpha/(2\alpha-1)}, \tag{11}$$

where const is a positive constant depending only on L, l_1, α, and $|A|$. Since by the assumption, the series of $\tau_k^{2\alpha/(2\alpha-1)}$ is convergent then for arbitrary small positive ϵ there exists a sufficiently large N such that for all $p \geq 1$ one has

$$\text{const} \sum_{k=N}^{N+p} \tau^{2\alpha/(2\alpha-1)} \leq \epsilon$$

and hence

$$|x^{N+p} - y^*|^2 \leq |x^N - y^*|^2 + \epsilon,$$

which is obtained by summing up the inequalities (11) starting from $k = N$ to $k = N + p-1$. Without loss of generality we may assume that N is chosen such that $|x^N - y^*|^2 \leq \epsilon$. This is because there exists a subsequence of x^k converging to y^*. But then

$$|x^{N+p} - y^*|^2 \leq 2\epsilon,$$

for all $p \geq 1$ which means that the limitting set of the entire sequence $\{x^k\}$ consists of a single point y^*. The proof is complete.

4 System of inequalities

In this section we consider the special case of problem (1)-(2) where $f(x) \equiv 0$, i.e. the problem is to find a point $z^* \in \mathcal{X}$. The iterative solution procedure to be suggested is also a special case of the algorithm RCAM. Namely, let x^k be the current solution approximation, then define x^{k+1} as follows

$$x^{k+1} := \arg\min_{x \in X} |x - x^k|^2/2 \qquad (12)$$

subject to

$$\langle [Ax^k - b]_+, \; Ax - b \rangle \leq 0.$$

It turns out that in the case of the system of inequalities much stronger convergence properties of the algorithm can be obtained.

Let us first assume that $\mathcal{X}$ is nonempty and z^* be an arbitrary point in $\mathcal{X}$. Using the arguments similar to those from section 2 we obtain

$$|x^{k+1} - z^*|^2 \leq |x^k - z^*|^2 - |x^{k+1} - x^k|^2. \qquad (13)$$

From here it follows that the sequence of $|x^k - z^*|$ is nonincreasing and thus, it is convergent. In other words, if z^* is a limit point of the sequence of x^k, then the limiting set of $\{x^k\}$ consists of the unique point z^* which means that in the case $f(x) \equiv 0$, the algorithm is converging to a point without additional regularity assumptions.

Now let us show that under the regularity assumptions, the rate of convergence of the algorithm can be established. Namely, suppose that the set $\mathcal{X}$ is α-regular with $\alpha = 1$. For example, when X is a linearly constrained set, from Hoffman's lemma, [13], it follows that the assumption is true. For some recent generalizations of the Hoffman's result see, for example, [14].

Using estimate (8) we obtain

$$|x^{k+1} - z^*|^2 \leq |x^k - z^*|^2 - |[Ax^k - b]_+|^2/|A|^2.$$

From the regularity it follows

$$|x^{k+1} - z^*|^2 \leq (1 - l_1^2/|A|^2)|x^k - z^*|^2$$

which means that the algorithm is linearly convergent.

Now let us consider the case where the set $\mathcal{X}$ may be empty. The following property holds.

Assertion. *If for each $k = 1, 2, \ldots$ the subproblem (3) (with $f(x) \equiv 0$) is solvable (i.e., the relaxed set is nonempty) then*

$$\mathcal{X} \neq \emptyset \Leftrightarrow \sum_{k=1}^{T} |x^{k+1} - x^k|^2 \leq d^2 \qquad (14)$$

for every $T = 1, 2, \ldots$, where d is the diameter of the set X.

Proof. If $\mathcal{X}$ is not an emptyset then the boundedness of the sum in (14) follows directly from (13). To prove the inverse implication, assume that $\mathcal{X}$ is empty. Hence, for some $\delta > 0$ and all $x \in X$ we have $|Ax - b| \geq \delta$. Taking into account (8) and summig up the corresponding inequalities up to a sufficiently large T we arrive at a contradiction with the boundedness of the sum.

Remark. In the assertion we did not use the regularity assumptions regarding the set $\mathcal{X}$. Under 1-regularity one can show the length of the path of the method $\sum_k |x^{k+1} - x^k|$ will be bounded with qd, where q is a constant dependent on the regularity of the set $\mathcal{X}$. This allows one to strengthen the assertion replacing the bound in (14) with

$$\sum_{k=1}^{T} |x^{k+1} - x^k| \leq qd, \quad T = 1, 2, \ldots$$

Similar estimates of the length of the path can be found in [15].

The assertion gives us a simple criterion for identifying whether the system of inequalities defining the set $\mathcal{X}$ is consistent. Given a positive tolerance δ one has two alternatives. The first one is where $|[Ax - b]_+| \geq \delta$ for each $x \in X$, i.e., there is no a δ-feasible solution to the system. In this case, after at most d^2/δ^2 steps either the sum in (14) will exceed d^2 or an infeasible subproblem will be encountered and the fact that $\mathcal{X}$ is empty will be identified. The second alternative is where the set $\mathcal{X}$ is nonempty and here, according to the assertion, after at most d^2/δ^2 steps the δ-feasible point will be found.

5 Convergence rate estimates: general case

In this section we again consider the general case where the function $f(x)$ is not necessarily zero. The purpose is to give an estimate for the rate of convergence of the trajectory of the method to the optimal solution set $\mathcal{X}^*$ under the regularity assumptions introduced in section 3. Here we restrict ourselves with the case of 1-regularity.

The section is divided into two subsections. In the first one the rate of convergence to the feasible set is estimated. The second one uses this result in order to provide a bound for the rate of convergence to the optimal solution set.

Throughout the rest of the paper we impose an additional requirement for the choise of the sequence of τ_k:

$$\lim_{k \to \infty} \tau_{k+1}/\tau_k \geq 1. \tag{15}$$

Note, for example, that for every $\beta \in (0, 1]$ and positive N and s, the sequence of $\tau_k = s/(k + N)^\beta$ satisfies this requirement.

5.1 Convergence to the feasible set

We start by providing a bound for the distance between the solutions to subproblems (3) and (12), respectively. Since τ_k tends to zero as k tends to infinity, the distance between the solutions is decreasing. It turns out that the following estimate holds

$$|x^{k+1} - y^k| \leq \tau_k L, \tag{16}$$

where y^k denotes the orthogonal projection of x^k onto $\mathcal{X}^k$, the relaxed set in subproblem (3), and L is the Lipscitzian constant of the function $f(x)$. To show this, let us substitute y^k for x in KKT optimality conditions (5):

$$\langle \tau_k \partial f(x^{k+1}) + x^{k+1} - x^k + \lambda A^T [Ax^k - b]_+,\ y^k - x^{k+1} \rangle \geq 0. \tag{17}$$

Let us estimate the terms in the latter inequality separately. Substituting y^k for x^* in (6) one analogously obtains

$$\lambda \langle A^T [Ax^k - b]_+,\ y^k - x^{k+1} \rangle \leq 0.$$

Furthermore,

$$\langle x^{k+1} - x^k,\ y^k - x^{k+1}\rangle = -|x^{k+1} - y^k|^2 + \langle y^k - x^k,\ y^k - x^{k+1}\rangle \leq$$
$$-|x^{k+1} - y^k|^2.$$

In the latter inequality we used the condition for y^k to be the projection of x^k onto $\mathcal{X}^k$:

$$\langle x^k - y^k,\ x - y^k\rangle \leq 0, \quad x \in \mathcal{X}^k$$

Substituting these estimates in (17) and using convexity of f we obtain

$$|x^{k+1} - y^k|^2 \leq \tau_k(f(y^k) - f(x^{k+1})) \leq \tau_k L|x^{k+1} - y^k|$$

which implies the desired estimate.

Now let us turn to estimating the distance of trajectory to the feasible set. Denote by $\bar{y}^k \equiv \mathrm{Proj}_{\mathcal{X}}(x^k)$ the orthogonal projection of the current iterate x^k onto the feasible set $\mathcal{X}$. Then

$$\rho^2(x^{k+1}, \mathcal{X}) \leq |x^{k+1} - \bar{y}^k|^2 = |x^k - \bar{y}^k|^2 +$$
$$2\langle x^{k+1} - x^k,\ x^k - \bar{y}^k\rangle + |x^{k+1} - x^k|^2 \tag{18}$$

Let us rewrite the second term:

$$\langle x^{k+1} - x^k,\ x^k - \bar{y}^k\rangle = -|x^{k+1} - x^k|^2 + \langle x^{k+1} - x^k,\ x^{k+1} - \bar{y}^k\rangle.$$

To estimate the latter scalar product we will again employ optimality conditions (5) with $x = \bar{y}^k$:

$$\langle \tau_k \partial f(x^{k+1}) + x^{k+1} - x^k + \lambda A^T[Ax^k - b]_+,\ \bar{y}^k - x^{k+1}\rangle \geq 0.$$

Substituting $\bar{y}^k$ for y^k in (6) one has

$$\lambda\langle A^T[Ax^k - b]_+,\ \bar{y}^k - x^{k+1}\rangle \leq 0.$$

Hence, by convexity of $f(x)$

$$\langle x^{k+1} - x^k,\ x^{k+1} - \bar{y}^k\rangle \leq \tau_k(f(\bar{y}^k) - f(x^{k+1})).$$

Thus, substituting these estimates into (18), using (16), and the fact that $f(x)$ is Lipschitzian we obtain

$$|x^{k+1} - \bar{y}^k|^2 \leq |x^k - \bar{y}^k|^2 + \tau_k(f(\bar{y}^k) - f(x^{k+1})) - |x^{k+1} - x^k|^2 =$$
$$|x^k - \bar{y}^k|^2 + \tau_k(f(y^k) - f(x^{k+1})) + \tau_k(f(\bar{y}^k) - f(y^k)) - |x^{k+1} - x^k|^2 \leq$$
$$|x^k - \bar{y}^k|^2 + \tau_k^2 L^2 + \tau_k L|\bar{y}^k - y^k| - |x^{k+1} - x^k|^2.$$

Using (8) and the regularity assumption regarding the feasible set $\mathcal{X}$, one gets the estimate

$$l_1|x^k - \bar{y}^k| \leq |A|\ |x^{k+1} - x^k|.$$

Besides, by definition of y^k and using (13) one has

$$|y^k - \bar{y}^k| \leq |x^k - \bar{y}^k|.$$

(Note that here $\bar{y}^k$ and y^k stand for z^* and x^{k+1} from (13), respectively.) Finally, the estimate for the $\rho(x^{k+1}, \mathcal{X})$ appears to be as follows

$$\rho^2(x^{k+1}, \mathcal{X}) \le \kappa \rho^2(x^k, \mathcal{X}) + \tau_k^2 L^2 + \tau_k L \rho(x^k, \mathcal{X}), \tag{19}$$

where $\kappa = (1 - l_1^2/|A|^2) < 1$.

Consider two cases: the first one is where

$$\tau_k L \le \frac{1-\kappa}{4} \rho(x^k, \mathcal{X}). \tag{20}$$

Then, the following estimate holds

$$\rho^2(x^{k+1}, \mathcal{X}) \le \frac{\kappa+1}{2} \rho^2(x^k, \mathcal{X}),$$

(we have used that $(1-\kappa)/4 < 1$). Secondly, assume that in (20) the inverse inequality holds, which means that $\rho(x^k, \mathcal{X})$ is bounded from above with $C\tau_k$, where $C = 4L/(1-\kappa)$ is a certain constant independent of the iteration number. Using formula (19) one can estimate $\rho(x^{k+1}, \mathcal{X})$:

$$\rho^2(x^{k+1}, \mathcal{X}) \le \tau_k^2 (\kappa C^2 + L^2 + LC).$$

Let us show that the expression in braces at the right-hand side does not exceed qC^2 with some $q < 1$. Indeed, using the definition of C

$$C^2 - \kappa C^2 - L^2 - LC = \frac{16L^2}{(1-\kappa)^2} - \frac{16\kappa L^2}{(1-\kappa)^2} - L^2 - \frac{4L^2}{1-\kappa} =$$
$$\frac{12L^2}{1-\kappa} - L^2 \ge \frac{11L^2}{1-\kappa}.$$

From here we derive that

$$\frac{\kappa C^2 + L^2 + LC}{C^2} \le \frac{C^2 - 11L^2/(1-\kappa)}{C^2} =$$
$$1 - \frac{11(1-\kappa)}{16} \equiv q < 1.$$

Therefore,

$$\rho^2(x^{k+1}, \mathcal{X}) \le \tau_k^2 q C^2.$$

Using the requirement (15) for the sequence of τ_k one concludes that starting with some K and for all $k \ge K$ the following is true: $q^{1/2} \le \tau_{k+1}/\tau_k$; and hence

$$\rho^2(x^{k+1}, \mathcal{X}) \le \tau_{k+1}^2 C^2.$$

This means that if for some $k \ge K$ the second opportunity realizes then the estimate

$$\rho(x^t, \mathcal{X}) \le C\tau_t$$

will hold for all $t \ge k$. If for all $k \ge K$ the first alternative takes place, the sequence of $\rho(x^k, \mathcal{X})$ is decreasing at least at a linear rate with the coefficient

$$\left(\frac{1+\kappa}{2}\right)^{1/2} = \left(1 - \frac{l_1^2}{2|A|^2}\right)^{1/2}.$$

Summarizing the arguments above we arrive at the following assertion

Theorem 2. *Let $\mathcal{X}$ be 1-regular and the sequence of τ_k in the definition of RCAM satisfy additional requirement (15). Then for all $t \geq K$ one has*

$$\rho(x^t, \mathcal{X}) \leq \max\left[C\tau_t,\ \left(1 - \frac{l_1^2}{2|A|^2}\right)^{\frac{t-K}{2}} \rho(x^K, \mathcal{X})\right], \tag{21}$$

where K is such that

$$\left(1 - \frac{l_1^2}{2|A|^2}\right)^{1/2} \leq \frac{\tau_{k+1}}{\tau_k}$$

for all $k \geq K$ and

$$C = \frac{4L}{1-\kappa} = \frac{4L|A|^2}{l_1^2}.$$

Hence, it turns out that starting from a sufficiently large k, the rate of convergence of the trajectory to the feasible set is greater than in the constraint aggregation method without regularization.

5.2 Convergence to the optimal solution set

In this subsection, based on the results of the previous one, we establish the assymptotic rate of convergence of the trajectory of the method to the optimal solution set of the problem (1)-(2).

We will need the following auxiliary property.

Lemma. *Let the parameters $\gamma \geq 0$, $\beta \geq 0$, $p \geq 0$, $s \geq 0$, and θ be fixed and satisfy the relationships below. Let the sequence $\{\rho_k\}$, $\rho_k \geq 0$, be such that starting with some T*

$$\rho_{k+1}^2 \leq \rho_k^2 - \gamma\tau_k\rho_k + \beta\tau_k^2, \quad k = T,\ T+1, \dots, \tag{22}$$

with $\rho_T \leq 2\beta s/(\gamma(T+p))$, where

$$\tau_k = \frac{s}{k+p}, \quad k = T,\ T+1, \dots,$$

$$\frac{p^2}{(1+p)^2} \geq \max[\theta,\ 1-\theta], \quad \theta = 1 - \frac{\gamma^2}{4\beta}, \quad \gamma^2 < 4\beta.$$

Then, for all $k = 0, 1, \dots$

$$\rho_k \leq \frac{2\beta s}{\gamma(k+p)}. \tag{23}$$

Proof. Denote $M = 2\beta s/\gamma$. By the assumption, for $k = T$ the assertion is true. Suppose that it is true for some $k > T$ and we prove it for $k+1$. Maximum of the right-hand side of (22) with respect to ρ_k is attained either at $\rho_k = 0$ or $\rho_k = M/(k+p)$ (the latter is by the assumption of induction). If $\rho_k = 0$ maximizes the right-hand side of (22) then

$$\rho_{k+1}^2 \leq \frac{\beta s^2}{(k+p)^2}.$$

Let us check that the latter ratio is less than $M^2/(k+p+1)^2$. Indeed,

$$\frac{M^2}{(k+p+1)^2} - \frac{\beta s^2}{(k+p)^2} = \frac{\beta s^2}{(k+p+1)^2}\left(\frac{4\beta}{\gamma^2} - \frac{(k+p+1)^2}{(k+p)^2}\right) \geq$$
$$\frac{\beta s^2}{(k+p+1)^2}\left(\frac{1}{1-\theta} - \frac{(p+1)^2}{p^2}\right) \geq 0.$$

The latter is by the definition of p.

Consider the case where the maximum is attained at $\rho_k = M/(k+p)$. We have

$$\rho_{k+1}^2 \leq \frac{M^2 - \gamma s M + \beta s^2}{(k+p)^2}.$$

It is sufficient to prove that

$$M^2 - \gamma s M + \beta s^2 \leq M^2\theta \tag{24}$$

since by definition of p

$$M^2\theta \leq M^2\frac{p^2}{(p+1)^2} \leq M^2\frac{(k+p)^2}{(k+p+1)^2}, \quad k = 0, 1, \ldots,$$

which would imply the desired estimate. Inequality (24) is checked straightforwardly using the definitions of M and θ. The proof is complete.

Now we can prove the rate of convergence of the trajectory of the method.

Theorem 3. *Let the set $\mathcal{X}$ and problem (1)-(2) be 1-regular and the sequence of τ_k be chosen as in the lemma. Then there exists index K_1 (depending only on l_1) and s such that for all $k \geq K_1$ the estimate (23) holds with*

$$\rho_k = \rho(x^k, \mathcal{X}^*), \ \gamma = 2l_2, \ \beta = \max[\gamma^2/4, \ LC + l_2C + L^2],$$

and p as specified in the lemma.

Proof. Let us employ the estimate (7) obtained earlier. The difference $f(x^{k+1}) - f_*$ can be rewritten as follows

$$f(x^{k+1}) - f_* = f(\bar{y}^k) - f_* + f(x^{k+1}) - f(\bar{y}^k),$$

For the first difference, using the 1-regularity assumption for the problem (1)-(2), we have

$$f(\bar{y}^k) - f_* \geq l_2\rho(\bar{y}^k, \mathcal{X}^*) \geq l_2\rho(x^k, \mathcal{X}^*) - l_2\rho(x^k, \mathcal{X}).$$

The second difference can be bounded as follows

$$f(\bar{y}^k) - f(x^{k+1}) = f(\bar{y}^k) - f(y^k) + f(y^k) - f(x^{k+1}) \leq$$
$$L(|x^{k+1} - y^k| + |y^k - \bar{y}^k|) \leq L(|x^{k+1} - y^k| + \rho(x^k, \mathcal{X})).$$

Using these bounds we arrive at the following estimate for $\rho(x^{k+1}, \mathcal{X}^*)$:

$$\rho^2(x^{k+1}, \mathcal{X}^*) \leq \rho^2(x^k, \mathcal{X}^*) - 2\tau_k l_2\rho(x^k, \mathcal{X}^*) + 2\tau_k(L + l_2)\rho(x^k, \mathcal{X}) +$$
$$2\tau_k L|x^{k+1} - y^k|.$$

Set the sequence of τ_k as specified in the lemma. From the estimate (21) of the rate of convergence of trajectory to the feasible set it follows that for all $k \geq K$ one has

$$\rho(x^k, \mathcal{X}) \le \max[C\tau_k,\ d[(1+\kappa)/2]^{(k-K)/2}],$$

where d is the diameter of the set X. It is clear that for some $K_1 \ge K$ and all $k \ge K_1$ this maximum is attained at $C\tau_k$. Then, using (16) one can further bound the distance to the optimal solution set as follows

$$\rho^2(x^{k+1}, \mathcal{X}^*) \le \rho^2(x^k, \mathcal{X}^*) - 2\tau_k l_2 \rho(x^k, \mathcal{X}^*) + 2\tau_k^2(LC + l_2C + L^2)$$

for all $k \ge K_1$. Finally, to estimate $\rho(x^k, \mathcal{X}^*)$ one can apply lemma with the parameters as specified in the theorem and setting $T = K_1$ To ensure that the initial condition for the ρ_T is satisfied, one has to choose s sufficiently large. The proof is complete.

Thus the rate of convergence of the algorithm to the set of optimal solutions of problem (1)-(2) is established.

6 Numerical results

This section presents the results of numerical tests of the regularized constraint aggregation method for problem (1)-(2). In reviewing these results two important issues should be addressed. The first issue is the comparative performance of RCAM and the method without aggregation (CAM). The second issue is the behavior of the trajectory of the method under different strategies for choosing the sequence of τ_k.

In all the tests we have used the dual transportaion problem with the same set of data as in [16]:

$$\max \left[\sum_{i=1}^{N} s_i w_i - \sum_{j=1}^{N} d_j v_j \right] \tag{25}$$

$$w_i - v_j \le a_{ij},\ i = 1, 2 \dots, N,\ j = 1, 2, \dots, N, \tag{26}$$

where $N = 48$. The aggregate constraint was constructed by convolving (26) for all $i = 1, 2 \dots, N,\ j = 1, 2, \dots, N$. The set X was formed by the box constraints $w_i,\ v_j \in [0, M]$ (for the upper bound M the value 3000 was choosen as one of the possible variants suggested in [16]). The initial approximation for the method was obtained by minimizing the objective function (25) over X (as in [16]).

Table 1 summarizes the results of comparative tests. In all cases the starting values for the residual and the function were $|[Ax^0 - b]_+| = 9.78e+4$ and $f(x^0) = 7.27e+6$. The upper part of table 1 presents the results for RCAM and CAM with $\tau_k = 1/k,\ k = 1, 2, \dots$, and the lower part with $\tau_k = 5/k,\ k = 1, 2, \dots$. The optimal value of the objectve function $f_* = 638565$.

Table 1, ($\tau_k = 1/k$)

iter. no.	RCAM		CAM	
	$\lvert[Ax^k - b]_+\rvert$	$f(x^k)$	$\lvert[Ax^k - b]_+\rvert$	$f(x^k)$
50	8.02	5.47e+5	1.01e+4	2.24e+6
100	5.2	5.54e+5	6.45e+3	1.81e+6
500	0.79	5.67e+5	2.13e+3	1.14e+6
1000	0.53	5.73e+5	1.34e+3	9.82e+5

Table 1, ($\tau_k = 5/k$)

iter. no.	RCAM		CAM	
	$\|[Ax^k - b]_+\|$	$f(x^k)$	$\|[Ax^k - b]_+\|$	$f(x^k)$
50	51.2	1.10e+6	1.07e+4	2.87e+6
100	24.6	6.11e+5	1.07e+4	2.29e+6
500	5.14	6.13e+5	3.18e+3	1.32e+6
1000	2.49	6.15e+5	1.96e+3	1.10e+6

From table 1 it is seen that RCAM clearly outperforms CAM. Another important remark is that RCAM tends to be very sensitive to the choice of the sequence of τ_k which suggests a possibility of better tuning the method. The tests have shown that the behavior of the trajectory had a certain speciality: decrease in the norm of the residual was comparatively fast as opposed to the decrease in the optimality gap of the objective function. In other words, there is an effect of "glueing up" of the trajectory to the feasible set.

This speciality can be easily explained from the point of view of the convergence rate results from the previous section. The estimates obtained are essentially dependent on the regularity constants l_1 and l_2 of the fesible set $\mathcal{X}$ and the optimal solutions set $\mathcal{X}^*$, respectively. Constant l_1 is determined by a nondegeneracy measure of the constraint matrix of (26), and, moreover, it can be easily bounded from below using the Sleiter condition. At the same time, constant l_2 is determined by a nondegeneracy measure of the matrix defining the optimal solutions set and the Sleiter condition can not be used. Hence, the feasible set is "more regular" than the set of optima, and thus, the rate of convergence to the feasible set is greater than that to the set of optima, which explains the effect of "glueing up".

In view of these observations the following strategies for choosing the sequence of τ_k were suggested

$$\tau_k = \begin{cases} 1/\log(k+1), & k = 1, 2, \ldots, T, \\ 1/(k - T + 1), & k > T, \end{cases}$$

where T =100, 200, 300, and 500. The idea is to make the sequence $\{\tau_k\}$ tend to zero more smoothly in order to increase the role of the objective function term in the subproblem (3).

The results are summarized in the following table (the columns correspond to the specified values of T).

Table 2, (T =100, 200)

iter. no.	$\|[Ax^k - b]_+\|$	$f(x^k)$	$\|[Ax^k - b]_+\|$	$f(x^k)$
50	126	6.24e+5	126	6.24e+5
500	1.16	6.16e+5	1.59	6.27e+5
1000	0.51	6.17e+5	0.69	6.28e+5

Table 2, (T =300, 500)

iter. no.	$\|[Ax^k - b]_+\|$	$f(x^k)$	$\|[Ax^k - b]_+\|$	$f(x^k)$
50	126	6.24e+5	126	6.24e+5
500	3.01	6.31e+5	833	7.80e+5
1000	0.75	6.31e+5	0.96	6.35e+5

Table 2 shows that the more the T, the smaller the gap $|f(x^{1000}) - f_*|$.

7 Conclusions

Convergence analysis conducted in the paper shows that the regularized constraint aggregation method outperforms (at least assymptotically) the method without regularization. The numerical results (even under limited amount of iterations performed as compared to the theoretical bounds) also provide the evidence that RCAM behaves better than CAM.

Convergence rate estimates obtained also allow one to relate certain qualitative properties in the behavior of the method, namely, the effect of "glueing up" the trajectory to the feasible set, to the fact that the feasible set is more regular than the set of optimal solutions. The presence of such an effect suggests that the choice of the sequence of the parameters τ_k should be more specific in order to acheive a better convergence to the set of optima. In connection with this, further analytical work and more extensive computational tests are required, which together make a subject for future research.

Acknowledgments

The author wishes to thank Prof. Andrzej Ruszczynski for suggesting the topick of the paper and constant help and advice during the work.

References

[1] J.-B.Hiriart-Urruty and C.Lemarechal, Convex Analysis and Minimization Algorithms, Springer-Verlag, Berlin, 1993.

[2] K.C.Kiwiel, Methods of Descent for Nondifferentiable Optimization, Springer-Verlag, Berlin, 1985.

[3] Yu.M.Ermoliev, Methods of Stochastic Programming, Nauka, Moscow, 1976 (in Russian).

[4] B.T.Polyak, A general method for solving extremal problems, Doklady Acad. Nauk SSSR, 174(1967), pp.33-36 (in Russian).

[5] N.Z.Shor, Minimization Method for Nondifferentiable Functions, Springer-Verlag, Berlin, 1985.

[6] Yu.M.Ermoliev, A.Kryazhimskii, A.Ruszczynski, Constraint Aggregation Principle in Convex Optimization, working paper WP-95-015, IIASA, Laxenburg, 1995.

[7] A.V.Kryazhimskii and Yu.S.Osipov, To a regularization of a convex extremal problem with inaccurately given constraints. An application to an optimal control problem with state constraints, in Methods of Positional and Program Control, Urals Sci. Center, Sverdlovsk, 1987, pp.34-54 (in Russian).

[8] M.T.Wasan, Stochastic Approximation, Cambridge Univ. Press, Cambridge, 1969.

[9] B.W.Kort and D.P.Bertsekas Combined proximal dual and penalty methods for convex programming, SIAM J. Control Optim., 14(1976), pp. 268-294.

[10] N.M.Novikova, Iterative Stochastic Methods for Solving Variational Problems of Mathematical Physics and Operations Research, Journal of Math. Sciences (Contemprorary Mathematics and Its Applications, Vol.3), New York - London, Plenum Publ. Corp., 1994, N1, p.1-125.

[11] J.V.Burke and M.C.Ferris Weak sharp minima in mathematical programming, SIAM J. Control and Optim., 31(1993), pp. 1340-1359.

[12] M.C.Ferris and O.L.Mangasarian, Minimum principle sufficiency, Mathematical programming, 57(1992), pp. 1-14.

[13] A.J.Hoffman On approximate solutions of systems of linear inequalities Journal of Research of National Bureau of Standards, 49(1952), pp. 263-265.

[14] J.V.Burke and P.Tseng A unified analysis of Hoffman's bound via Fenchel duality, SIAM J. Optim., 34(1996), pp.

[15] U.G.Brannlund On Relaxation Methods for Nonsmooth Convex Optimization, PhD thesis, Dept. of Mathematics, Royal Institute of Technology, Stockholm, 1993.

[16] R.Rozycki Constraint aggregation principle: Application to a dual transportation problem, working paper WP-95-103, IIASA, Laxenburg, 1995.