Improved Incremental Constraint Projection Method by Sampling Multiple Constraints

Jialin Liu, Yuantao Gu*, and Mengdi Wang†

November 20, 2014

Abstract

We focus on stochastic optimization problem and variational inequalities problem in which the objective function is expectation of a randomized convex real-value function, and the feasible set is the intersection of large number convex sets. We propose ICPM-SMC as a general framework which involves *Incremental Constraints Projection Method* (ICPM) as one of its special schemes. Then we analyze its convergence rate and convergence stability. For rate, we uses different *expected errors* as criteria for different cases; for stability, we propose three criteria. Use these tools, we find that we can control the parameters and sampling scheme in ICPM-SMC to get particular convergence rate or convergence stability.

1 Introduction

Consider the following *Stochastic Optimization* (SO) Problem,

$$\begin{aligned} \min_{x} \ & \Big\{ F(x) = \mathbf{E}[f(x;v)] \Big\} \\ \text{s.t. } & x \in X = \cap_{i=1}^{m} X_i \end{aligned} \tag{1}$$

where $F : \Re^n \mapsto \Re, f(\cdot; v) : \Re^n \mapsto \Re$ are real-valued convex functions, the constraints X_i are convex closed sets, and v is a random variable.

Problems (1) have a long history and many methods are proposed. For an unconstrained problem (let $X = \Re^n$), *Stochastic Gradient Descent* (SGD) is an important method, (see [1–3]) which picks an arbitrary initial point x_0, and iterate like follows:

$$x_{k+1} = x_k - \alpha_k g(x_k, v_k) \tag{2}$$

*Jialin Liu and Yuantao Gu are with Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, CHINA. This work was supported by National 973 Program of China (Grant No. 2013CB329201) and National Natural Science Foundation of China (NSFC 61371137).

†Mengdi Wang is with Department of Operations Research and Financial Engineering, Princeton University, Princeton 08544, USA.

where $g(x_k, v_k) \in \partial f(x_k, v_k)$ denotes one if its sub-gradients, scalar α_k is the step size at the k-th iterate, and v_k is a random variable picked from the domain of v at the k-th iterate. SGD method is so important that many randomized optimization algorithm are all based on this method, such as SAG, SVGR, RCDC (see [4–6]). However, they can only solve unconstrained problems.

To solve constrained optimization problems, to the best of our knowledge, there exist three methods. The first one is to add penalty for constraints violation, which can formulate an constrained problem into an unconstrained one, such as Penalty Function Method, Interior Point Method (see [7–9]). The second one is primal-dual methods, which iterates on both primal and dual problems and converges to the saddle point of the Lagrange function. Such as DA, ADMM (see [10–12]). The third one, which we consider mainly, is projection methods (see [13–15]). For a simple constraint set X, we can take the projection on X at every iterate, which modifies (2) as the following:

$$x_{k+1} = \Pi_X(x_k - \alpha_k g(x_k, v_k)) \tag{3}$$

where Π_X means projection on X.

However, if X is a complex set like $X = \cap_{i=1}^m X_i$, to calculate projection on X is expensive. We should consider a more practical method for this case. *Incremental Constraint Projection Method* (ICPM) is proposed [16,17], which randomly picks one X_i (let $\omega_k = X_i$) from the constraint sets and take projection on the simple X_i at each step:

$$\begin{aligned} y_{k+1} =& x_k - \alpha_k g(x_k, v_k) \\ x_{k+1} =& \Pi_{\omega_k} y_{k+1}. \end{aligned} \tag{4}$$

If the feasible set X is the intersection of many simple sets (i.e. subspaces), ICPM is very cheap for one step. However, it still can be improved:

- The convergence rate is slow, it takes large number (10^5-10^6) of steps to achieve an small error (10^{-3}-10^{-4}).
- The convergence process is not stable because of sampling randomly.

Our work is to improve ICPM from two aspects: *convergence rate* and *convergence stability*. We propose *Incremental Constraint Projection Method by Sampling Multiple Constraints* (ICPM-SMC), which samples several constraints $\omega_{k,i}, i = 1, 2, \ldots, M_k$ instead of only one ω_k. We estimate the convergence rate and stability for ICPM-SMC, and proof that new algorithm does better than ICPM. MDPM, a special scheme of ICPM-SMC, has a faster convergence rate than ICPM; ICAPM, another special scheme of ICPM-SMC, has a better convergence stability.

The remainder of this paper is organized as the following. Section 2 introduces ICPM-SMC. In section 3, we give some most basic assumptions and lemmas which will be used in convergence analysis. In section 4, we analyze the convergence rate of the new algorithms. In section 5, we analyze the convergence stability of the algorithms. In section 6, we give the final conclusion. All the proofs are given in appendix.

Algorithm 1: Incremental Constraint Projection Method by Sampling Multiple Constraints (ICPM-SMC)

Choose an arbitrary $x_0 \in \Re^n$ and positive integers $\{M_k\}$;

for $k = 0, 1, 2, \ldots$ **do**

(1) Sample a random cost function $g(x_k; v_k)$;

(2) Update using a gradient descent:

$$y_{k+1} = x_k - \alpha_k g(x_k; v_k) \ ; \tag{5}$$

(3) Sample M_k constraints $\{\omega_{k,i}\}_{i=1}^{M_k}$ independently from $\{X_i\}_{i=1}^{m}$ according to a uniform distribution.

(4) Calculate x_{k+1} as the averages of random projections:

$$x_{k+1} = \sum_{i=1}^{M_k} w_k(i) \Pi_{\omega_{k,i}} y_{k+1} \ ; \tag{6}$$

2 Algorithms

Consider ICPM (4), we only sample one constraint from $\{X_i\}_{i=1}^{m}$, which causes large variance and low convergence rate if m is quite large. In order to improve the convergence rate and stability of ICPM, we revise ICPM and propose ICPM-SMC as given by Algorithm 1.

At each iteration of ICPM-SMC, we first take a stochastic gradient descent step starting from x_k, then we sample a number of constraints $\omega_{k,i}, i = 1, 2, \ldots, M_k$, and take a weighted average of the projections as the next iterate x_{k+1}. It is easy to see that the proposed ICPM-SMC contains the ICPM as a special case when $M_k = 1$ for all k.

For ICPM-SMC, we can design different $w_k(i)$ to get different properties we want. Consider two special weights distribution. One is unbiased distributed as:

$$w_k(i) = \frac{1}{M_k}, \quad \forall i = 1, 2, \ldots, M_k, \tag{7}$$

which can specialize (6) as the following one:

$$x_{k+1} = \frac{1}{M_k} \sum_{i=1}^{M_k} \Pi_{\omega_{k,i}} y_{k+1}, \tag{8}$$

which is called *Incremental Constraint Averaging Projection Method* (ICAPM), see Figure 1 for graphical visualization of the ICAPM procedure.

Intuitively, by taking average of random projections, we may reduce the variance at every iteration and keep the next iterate concentrated around its expectation (see Theorem 5 for a formal statement). As illustrated by Figure 1, this prevents the next iterate x_{k+1} from randomly jumping into a distant constraint set. While improving the stability of iterates,

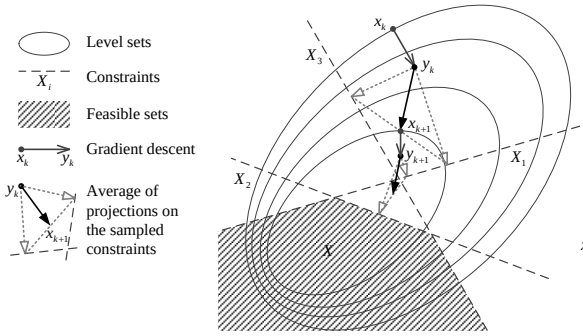

Figure 1: ICAPM algorithm

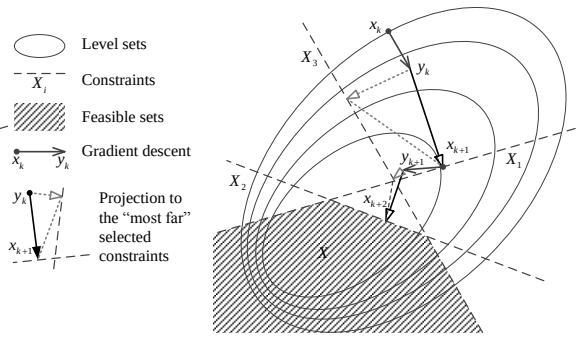

Figure 2: MDPM algorithm

the averaging scheme is computationally efficient, as calculating each random projection still involves only one simple set X_i.

Another special case for ICPM-SMC is *Max Distance Projection Method* (MDPM), in which is weight distribution $w_k(i)$ is defined as:

$$w_k(i) = \begin{cases} 1, i = \arg\max\limits_j \{\|x_k - \Pi_{\omega_{k,j}} x_k\|^2\} \\ 0, \quad \text{otherwise} \end{cases} \tag{9}$$

which can specialize (6) as the following one:

$$x_{k+1} = \Pi_{\omega_k} y_{k+1}, \quad \text{where } \omega_k = \arg\max_i \{\|x_k - \Pi_{\omega_{k,i}} x_k\|^2\}. \tag{10}$$

In MDPM, we take the projection which is the most distant from x_k as the next iterate, see Figure 2 for graphical visualization of the MDPM procedure. By taking the most distant projection, we guarantee the expected convergence rate can be more fast (see Theorem 2-4 for a formal statement). This scheme prevents the expected next iterate from moving slowly at the current iterate. As the averaging scheme, the max distance scheme is also efficient because we only choose one simple set from $\{X_i\}_{i=1}^m$.

3 Assumptions and Preliminaries

Before the analysis of Algorithm 1, we give some basic notation. Suppose there exists at least one optimal solution x^* to problem (1), i.e.,

$$\mathbf{E}[f(x; v)] \geq \mathbf{E}[f(x^*; v)], \quad \forall\ x \in \cap_{i=1}^m X_i.$$

To estimate the convergence rate, the distance to a given optimal solution x^* is defined:

$$e^2(x_k) := \|x_k - x^*\|^2. \tag{11}$$

To estimate the convergence rate to feasible set, the distance to the feasible set X is defined:

$$d^2(x_k) := d^2(x_k, X) := \|x_k - \Pi_X x_k\|^2. \tag{12}$$

To simplify the notation of the distance to the super-set of feasible set X, we define the projection and distance:

$$\begin{aligned} \Pi_j x :=& \Pi_{X_j} x, \\ d(x_k, X_j) :=& \|x_k - \Pi_j x_k\|. \end{aligned} \tag{13}$$

Define $\mathscr{F}_k$:

$$\mathscr{F}_k := \{v_t, \omega_{t,i}, x_t, y_t \mid |t = 1, 2, \ldots, k, i = 1, 2, \ldots, M_k\}, \tag{14}$$

as the collection of random variables that are revealed up to the kth iterations.
For simplicity, we define the sub-gradient $G(x)$ of $F(x)$:

$$G(x) = \widetilde{\nabla} F(x) \in \partial F(x). \tag{15}$$

Before analysis, we consider some basic inequalities which are the basis of the whole paper.

Lemma 1 (Basic Iterative Bounds for Error) *Suppose sequence $\{x_k\}$ is generated by ICPM-SMC (Algorithm 1), we can get the following inequalities for all $k \geq 0$:*

(a) For arbitrary optimal point x^*, squared distance to the optimal point $e^2(x_k) = \|x_k - x^*\|^2$ fits the following:

$$e^2(x_{k+1}) \leq e^2(x_k) - 2\alpha_k g'(x_k, v_k)(x_k - x^*) - \frac{1}{2}\sum_{i=1}^{M_k} w_k(i) d^2(x_k, \omega_{k,i}) + 5\alpha_k^2 \|g(x_k, v_k)\|^2 \tag{16}$$

(b) Squared distance to the feasible set $d^2(x_k) = \|x_k - \Pi_X x_k\|^2$ fits the following:

$$d^2(x_{k+1}) \leq (1+\epsilon) d^2(x_k) + (5 + 1/\epsilon)\alpha_k^2 \|g(x_k, v_k)\|^2 - \frac{1}{2}\sum_{i=1}^{M_k} w_k(i) d^2(x_k, \omega_{k,i}) \tag{17}$$

where ϵ is an arbitrary positive scalar.

Proof Proof (a) and (b) separately.

Proof of (a) Consider the squared error:

$$\begin{aligned} & \|x_{k+1} - x^*\|^2 \\ =& \|\sum_{i=1}^{M_k} w_k(i) \Pi_{\omega_{k,i}} y_{k+1} - x^*\|^2 \\ =& \|y_{k+1} - x^*\|^2 + \|\sum_{i=1}^{M_k} w_k(i) \Pi_{\omega_{k,i}} y_{k+1} - y_{k+1}\|^2 + 2(y_{k+1} - x^*)^T (\sum_{i=1}^{M_k} w_k(i) \Pi_{\omega_{k,i}} y_{k+1} - y_{k+1}) \\ =& \|y_{k+1} - x^*\|^2 + \|\sum_{i=1}^{M_k} w_k(i) \Pi_{\omega_{k,i}} y_{k+1} - y_{k+1}\|^2 - 2\sum_{i=1}^{M_k} w_k(i)(y_{k+1} - x^*)^T (y_{k+1} - \Pi_{\omega_{k,i}} y_{k+1}) \end{aligned}$$

Consider the third term on the right side, based on $x^* \in X \subset X_{\omega_{k,i}}$, we have:

$$(y_{k+1} - x^*)^T(y_{k+1} - \Pi_{\omega_{k,i}} y_{k+1}) \geq (y_{k+1} - \Pi_{\omega_{k,i}} y_{k+1})^T(y_{k+1} - \Pi_{\omega_{k,i}} y_{k+1}) = \|y_{k+1} - \Pi_{\omega_{k,i}} y_{k+1}\|^2$$

Combined with former inequality,

$$\|x_{k+1} - x^*\|^2 \leq \|y_{k+1} - x^*\|^2 + \|\sum_{i=1}^{M_k} w_k(i)\Pi_{\omega_{k,i}} y_{k+1} - y_{k+1}\|^2 - 2\sum_{i=1}^{M_k} w_k(i)\|y_{k+1} - \Pi_{\omega_{k,i}} y_{k+1}\|^2$$

Consider the second term, use *Jensen Inequality*, we can get:

$$\|\sum_{i=1}^{M_k} w_k(i)\Pi_{\omega_{k,i}} y_{k+1} - y_{k+1}\|^2 \leq \sum_{i=1}^{M_k} w_k(i)\|(\Pi_{\omega_{k,i}} y_{k+1} - y_{k+1})\|^2$$

Thus, we have,

$$\|x_{k+1} - x^*\|^2 \leq \|y_{k+1} - x^*\|^2 - \sum_{i=1}^{M_k} w_k(i)\|y_{k+1} - \Pi_{\omega_{k,i}} y_{k+1}\|^2 \tag{18}$$

We use $\|y - \Pi_S y\|^2 \leq 2\|x - \Pi_S x\|^2 + 8\|x - y\|^2$[1], get the following holds for all $i = 1, 2, \ldots, M_k$:

$$\|y_{k+1} - \Pi_{\omega_{k,i}} y_{k+1}\|^2 \geq \frac{1}{2}\|x_k - \Pi_{\omega_{k,i}} x_k\|^2 - 4\|y_{k+1} - x_k\|^2$$

Combined with (18), we can get the following:

$$\|x_{k+1} - x^*\|^2 \leq \|y_{k+1} - x^*\|^2 - \frac{1}{2}\sum_{i=1}^{M_k} w_k(i)\|x_k - \Pi_{\omega_{k,i}} x_k\|^2 + 4\|y_{k+1} - x_k\|^2$$

Consider the first term, we can get:

$$\|y_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha_k g'(x_k, v_k)(x_k - x^*) + \alpha_k^2\|g(x_k, v_k)\|^2.$$

Then (a) is proved.

Proof of (b) Consider the squared distance to the feasible set X:

$$d^2(x_{k+1}) \leq \|x_{k+1} - \Pi_X y_{k+1}\|^2 = \|\sum_{i=1}^{M_k} w_k(i)(\Pi_{\omega_{k,i}} y_{k+1} - \Pi_X y_{k+1})\|^2$$

Based on *Jensen Inequality*, we can get:

$$\|\sum_{i=1}^{M_k} w_k(i)(\Pi_{\omega_{k,i}} y_{k+1} - \Pi_X y_{k+1})\|^2 \leq \sum_{i=1}^{M_k} w_k(i)\|\Pi_{\omega_{k,i}} y_{k+1} - \Pi_X y_{k+1}\|^2$$

Consider each term:

$$\begin{aligned}\|\Pi_{\omega_{k,i}} y_{k+1} - \Pi_X y_{k+1}\|^2 =& \|\Pi_{\omega_{k,i}} y_{k+1} - y_{k+1} + y_{k+1} - \Pi_X y_{k+1}\|^2 \\ \leq& \|y_{k+1} - \Pi_X y_{k+1}\|^2 - \|\Pi_{\omega_{k,i}} y_{k+1} - y_{k+1}\|^2\end{aligned}$$

[1]Proved in [16], Lemma 1 (b)

Consider the first term, use *Cauchy-Schwarz inequality*, we can get:

$$\begin{aligned}\|y_{k+1} - \Pi_X y_{k+1}\|^2 \leq & \|y_{k+1} - \Pi_X x_k\|^2 \\ = & \|y_{k+1} - x_k + x_k - \Pi_X x_k\|^2 \\ \leq & (1 + 1/\epsilon)\|y_{k+1} - x_k\|^2 + (1 + \epsilon)\|x_k - \Pi_X x_k\|^2\end{aligned}$$

where ϵ is an arbitrary scalar.
Combine the above inequalities, we can get:

$$\|\Pi_{\omega_{k,i}} y_{k+1} - \Pi_X y_{k+1}\|^2 \leq (1 + \epsilon)d^2(x_k) + (1 + 1/\epsilon)\alpha_k^2\|g(x_k, v_k)\|^2 - \|\Pi_{\omega_{k,i}} y_{k+1} - y_{k+1}\|^2$$

We use $\|y - \Pi_S y\|^2 \leq 2\|x - \Pi_S x\|^2 + 8\|x - y\|$, we can get:

$$\|\Pi_{\omega_{k,i}} y_{k+1} - \Pi_X y_{k+1}\|^2 \leq (1 + \epsilon)d^2(x_k) + (5 + 1/\epsilon)\alpha_k^2\|g(x_k, v_k)\|^2 - \frac{1}{2}\|\Pi_{\omega_{k,i}} x_k - x_k\|^2$$

Thus,

$$d^2(x_{k+1}) \leq (1 + \epsilon)d^2(x_k) + (5 + 1/\epsilon)\alpha_k^2\|g(x_k, v_k)\|^2 - \frac{1}{2}\sum_{i=1}^{M_k} w_k(i)\|\Pi_{\omega_{k,i}} x_k - x_k\|^2$$

Part (b) is proved. ∎

To get ICPM-SMC convergent, we need some basic assumptions given bellow.

Assumption 1 (Basic Assumptions) *(a) The expected objective function $F(\cdot)$ is convex.*

(b) The feasible set $X = \cap_{i=1}^m X_i$ is nonempty and there exists a constant scalar η such that for any $x \in \Re^n$:

$$\|x - \Pi_X x\|^2 \leq \eta \max_{i=1,\ldots,m} \|x - \Pi_i x\|^2. \tag{19}$$

(c) The sample scheme for objective functions is conditionally unbiased: for any $x \in \Re^n$,

$$\mathbf{E}[g(x, v_k)|\mathscr{F}_k] = G(x) \in \partial F(x). \tag{20}$$

(d) Function $g(x, v_k)$ fits the follows for $k \geq 0$ with probability 1:

$$\mathbf{E}[\|g(x, v_k) - g(y, v_k)\|^2|\mathscr{F}_k] \leq L^2(\|x - y\|^2 + 1), \quad \forall x, y \in R^n \tag{21}$$

and

$$\mathbf{E}[\|g(x^*, v_k)\||\mathscr{F}_k] \leq B. \tag{22}$$

(e) Sampling scheme on constraints is "nearly unbiased":

$$\inf_{k\geq 0} P(\omega_{k,i} = X_j|F_k) \geq \frac{\rho}{m}, \quad \forall j = 1,\ldots,m,\ \forall i = 1,2,\ldots,M_k. \tag{23}$$

In Assumption 1, (a) and (c) are trivial and very basic assumptions, so we consider (b) and (d). Assumption (b) is called *Linear Regularity*. If the constraints X_is are all linear sets (subspaces of $\Re^n$, half-spaces .etc), Linear Regularity obviously holds. Assumption (d) is a limitation for the cost function $g(x,v)$. It allows g to be Lipchitz continuous, then objective function $f(\cdot,v)$ in Problem 1 is smooth and has a Lipchitz continuous gradient. It also allows nonsmooth objective function with a bounded subgradient. The whole Assumption 1 is also used for analysis of ICPM [16].

Use these assumptions, we can estimate the expected error $\mathbf{E}[e^2(x_{k+1})|\mathscr{F}_k]$ or $\mathbf{E}[d^2(x_{k+1})|\mathscr{F}_k]$ based on the inequalities in Lemma 1 as the following.

Corollary 1 (Iterative Bounds for Expected Error) *Suppose sequence $\{x_k\}$ is generated by ICPM-SMC (Algorithm 1), if Assumption 1 holds, we can get the following inequalities for all $k \geq 0$:*

(a) For any optimal point x^*, squared distance to the optimal point $e^2(x_k) = \|x_k - x^*\|^2$ fits the following:

$$\begin{aligned}\mathbf{E}[e^2(x_{k+1})|\mathscr{F}_k] \leq & (1+10L^2\alpha_k^2)e^2(x_k) + 10\alpha_k^2(L^2+B^2) \\ & - 2\alpha_k G'(x_k)(x_k - x^*) - \frac{1}{2}\sum_{i=1}^{M_k} w_k(i)\mathbf{E}[d^2(x_k,\omega_{k,i})|\mathscr{F}_k]\end{aligned} \tag{24}$$

(b) Squared distance to the feasible set $d^2(x_k) = \|x_k - \Pi_X x_k\|^2$ fits the following:

$$\mathbf{E}[d^2(x_{k+1})|\mathscr{F}_k] \leq (1+\epsilon)d^2(x_k) + 2(5+1/\epsilon)\alpha_k^2(L^2e^2(x_k)+L^2+B^2) - \frac{1}{2}\sum_{i=1}^{M_k} w_k(i)\mathbf{E}[d^2(x_k,\omega_{k,i})|\mathscr{F}_k] \tag{25}$$

where ϵ is an arbitrary positive scalar.

This corollary is derived from Lemma 1.

Proof Take conditional expectation on $\mathscr{F}_k$ of $e^2(x_{k+1})$, use (16), we can get:

$$\begin{aligned}\mathbf{E}[e^2(x_{k+1})|\mathscr{F}_k] \leq & e^2(x_k) - 2\alpha_k\mathbf{E}[g'(x_k,v_k)|\mathscr{F}_k](x_k - x^*) \\ & - \frac{1}{2}\sum_{i=1}^{M_k} w_k(i)\mathbf{E}[d^2(x_k,\omega_{k,i})|\mathscr{F}_k] + 5\alpha_k^2\mathbf{E}[\|g(x_k,v_k)\|^2|\mathscr{F}_k]\end{aligned}.$$

Based on Assumption 1(c), the second term can be simplified as:

$$2\alpha_k\mathbf{E}[g'(x_k,v_k)|\mathscr{F}_k](x_k - x^*) = 2\alpha_k G'(x_k)(x_k - x^*)$$

Based on Assumption 1(d), the third term can be bounded as:

$$\mathbf{E}[\|g(x_k,v_k)\|^2|\mathscr{F}_k] \le 2\mathbf{E}[\|g(x_k,v_k)-g(x^*,v_k)\|^2|\mathscr{F}_k]+2\mathbf{E}[\|g(x^*,v_k)\|^2|\mathscr{F}_k] \le 2L^2(\|x_k-x^*\|^2+1)+2B^2$$

Combine the above all, we can get (24). Take the conditional expectation of $d^2(x_{k+1})$, we can easily get (25) by following the same line. ∎

Assumption 2 (Strongly Convexity) *The function $F(x)$ is* strongly convex *with a constant $\sigma > 0$, if its sub-gradient $G(x)$ satisfies:*

$$(g(x) - g(y))'(x - y) \ge \sigma\|x - y\|^2 \tag{26}$$

holds for all $x, y \in \Re^n$.

Corollary 2 (Iterative Bounds for Expected Error for Strongly Convex Functions) *Suppose sequence $\{x_k\}$ is generated by ICPM-SMC (Algorithm 1), if Assumption 1 and 2 holds, let x^* denotes an arbitrary optimal point and $e^2(x_k) := \|x_k - x^*\|^2$, we can get the following inequalities for all $k \ge 0$:*

$$e^2(x_{k+1}) \le (1-2\alpha_k\sigma)e^2(x_k)-2\alpha_k h_k'(x_k-x^*)+2B\alpha_k d(x_k)-\frac{1}{2}\sum_{i=1}^{M_k} w_k(i)d^2(x_k,\omega_{k,i})+5\alpha_k^2\|g(x_k,v_k)\|^2, \tag{27}$$

where $h_k := g(x_k, v_k) - G(x_k)$. And we can bound its expectation on $\mathscr{F}_k$,

$$\mathbf{E}[e^2(x_{k+1})|\mathscr{F}_k] \le (1-2\alpha_k\sigma+10L^2\alpha_k^2)e^2(x_k)+2B\alpha_k d(x_k)-\frac{1}{2}\sum_{i=1}^{M_k} w_k(i)\mathbf{E}[d^2(x_k,\omega_{k,i})|\mathscr{F}_k]+10(L^2+B^2)\alpha_k^2. \tag{28}$$

Proof Consider (16), the second term is a random variable, it can be estimated by x_k:

$$\begin{aligned} g'(x_k,v_k)(x_k-x^*) =& G'(x_k)(x_k-x^*) + (g(x_k,v_k)-G(x_k))'(x_k-x^*) \\ =& G'(x_k)(x_k-x^*) + h_k'(x_k-x^*) \\ =& (G(x_k)-G(x^*))'(x_k-x^*) + G'(x^*)(x_k-x^*) + h_k'(x_k-x^*) \end{aligned}$$

where the first term can be bounded by the following using strongly convex assumption,

$$(G(x_k) - G(x^*))'(x_k - x^*) \ge \sigma\|x_k - x^*\|^2,$$

and the second term can be bounded by the following using Convexity of X and Assumption 1(d),

$$G'(x^*)(x_k-x^*) \ge G'(x^*)(x_k-\Pi_X x_k) \ge -\|G(x^*)\|d(x_k) \ge -\mathbf{E}[\|g(x^*,v_k)\||\mathscr{F}_k]d(x_k) \ge -B(x_k)$$

Combine all the above and (16), we can get (27).
Take expectation on (27), where $\mathbf{E}[h_k|\mathscr{F}_k] = 0$, we can get (28). ∎

4 Convergence and Convergence Rate

In this section, we proof the convergence (Section 4.1) of ICPM-SMC under the basic assumptions, then we analyze the convergence rate in non-strongly convex case and strongly convex case separately (Section 4.2-4.4), and give a summary (Section 4.5) at last.

Before the analysis, we consider an important term included in (24), (25) and (28): $\mathbf{E}[d^2(x_k, \omega_{k,i})|\mathscr{F}_k]$ and get the following lemma.

Lemma 2 *Suppose that $\{x_k\}$ is generated by ICPM-SMC, if Assumption 1 holds, there exists a constant scalar $C \in (0,1)$, so that the following holds for any $k \geq 0$,*

$$\mathbf{E}[\sum_{i=1}^{M_k} w_k(i)\|x_k - \Pi_{\omega_{k,i}} x_k\|^2 | \mathscr{F}_k] \geq C d^2(x_k), \quad w.p.1, \tag{29}$$

where C fits:

(a) In general case, $C \geq \frac{\rho}{m\eta}$.

(b) If the weight distribution $w_k(i)$ fits MDPM (9), and the sampling scheme is "sampling with replacement", then we can give a larger lower bound for C as $C \geq \frac{1}{\eta}[1-(1-\frac{\rho}{m})^{M_k}]$.

(c) If the weight distribution $w_k(i)$ fits MDPM (9), and the sampling scheme is "sampling without replacement", then we can give a larger lower bound for C as $C \geq \frac{1}{\eta}[1-\prod_{i=1}^{M_k}(1-\frac{\rho}{m-i+1})]$.

Proof Consider the three cases separately.

Proof of (a) The general case is proved in [16] (see (26) in [16]), because expectation $\mathbf{E}[d^2(x_k, \omega_{k,i})|\mathscr{F}_k]$ is the same as $\mathbf{E}[d^2(x_k, \omega_k)|\mathscr{F}_k]$ in [16].

$$\mathbf{E}[\|x_k - \Pi_{\omega_{k,i}} x_k\|^2 | \mathscr{F}_k] \geq \frac{\rho}{m\eta} d^2(x_k), \quad w.p.1.$$

Combined with $\mathbf{E}[\sum_{i=0}^{M_k} w_k(i)\|x_k - \Pi_{\omega_{k,i}} x_k\|^2 | \mathscr{F}_k] = \sum_{i=0}^{M_k} w_k(i)\mathbf{E}[\|x_k - \Pi_{\omega_{k,i}} x_k\|^2 | \mathscr{F}_k]$, we can get inequality (29) where $C \geq \frac{\rho}{m\eta}$. Part (a) is proved.

Proof of (b) Consider MDPM (10), we can get:

$$\mathbf{E}[\sum_{i=0}^{M_k} w_k(i)\|x_k - \Pi_{\omega_{k,i}} x_k\|^2 | \mathscr{F}_k] = \mathbf{E}[\|x_k - \Pi_{\omega_k} x_k\|^2 | \mathscr{F}_k], \tag{30}$$

where $\omega_k = \arg\max_i \{\|x_k - \Pi_{\omega_{k,i}} x_k\|^2\}$.
Let $\omega_{\max} = \arg\max_i \{\|x_k - \Pi_{X_i} x_k\|^2\}, i = 1, 2, \ldots, m$ (All the constraints, not only the chosen ones), then

$$d^2(x_k, \omega_{\max}) = \max_{j=1,2,\ldots,m} \{d^2(x_k, X_j)\} \geq \frac{1}{\eta} d^2(x_k, X)$$

Based on the sample scheme, we can get:

$$\begin{aligned} P(\omega_k = \omega_{\max}) =& P(\omega_{\max} \in \{\omega_{k,i} | i = 1, 2, \ldots, M_k\}) \\ =& 1 - \prod_{i=1}^{M_k} (1 - P(\omega_{k,i} = \omega_{\max})) \\ \geq& 1 - (1 - \frac{\rho}{m})^{M_k} \end{aligned}$$

Thus, the expectation can be bounded as:

$$\mathbf{E}[\|\Pi_{\omega_k} x_k - x_k\|^2 | \mathscr{F}_k] \geq P(\omega_k = \omega_{\max}) d^2(x_k, \omega_{\max}) \geq \frac{1}{\eta}[1 - (1 - \frac{\rho}{m})^{M_k}].$$

Combined with (30), we can get Part (b) of Lemma 2.

Proof of (c) Consider "sampling without replacement", the probability should be revised as:

$$\begin{aligned} P(\omega_k = \omega_{\max}) =& 1 - \prod_{i=1}^{M_k} (1 - P(\omega_{k,i} = \omega_{\max})) \\ \geq& 1 - (1 - \frac{\rho}{m})(1 - \frac{\rho}{m-1}) \ldots (1 - \frac{\rho}{m - M_k + 1}) = 1 - \prod_{i=1}^{M_k} (1 - \frac{\rho}{m - i + 1}), \end{aligned}$$

which can derive Part (c) of this lemma. ∎

In fact, constant C is important in the convergence rate analysis. In general case, ICPM-SMC has the same lower bound of C with ICPM, while if the weight distribution $w_k(i)$ fits MDPM, we can get a larger lower bound. If we sample the constraints with replacement, the lower bound can be larger than general case; if we sample without replacement, the lower bound can be larger than than with replacement. And we can conclude intuitively from the inequalities that larger C leads to a faster convergence. This is the motivation of the following subsections. We will analyze strictly how C influence the convergence rate.

4.1 Convergence Proof

Use Corollary 1, we can get that $\{x_k\}$, generated by ICPM, is almost surely convergent under Assumption 1.

Theorem 1 (Almost Sure Convergence) *Suppose that the sequence $\{x_k\}$ is generated by ICPM-SMC (Algorithm 1). Let Assumption 1 hold, and let the stepsize α_k satisfy:*

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{31}$$

Then $\{x_k\}$ converges almost surely to a random point in the set of optimal solutions for Problem (1).

PROOF Combine (24), (25) and (29), we can get:

$$\mathbf{E}[e^2(x_{k+1})|\mathscr{F}_k] \leq (1+10L^2\alpha_k^2)e^2(x_k)+10\alpha_k^2(L^2+B^2)-2\alpha_k G'(x_k)(x_k-x^*)-\frac{C}{2}d^2(x_k), \quad (32)$$

and

$$\mathbf{E}[d^2(x_{k+1})|\mathscr{F}_k] \leq (1+\epsilon)d^2(x_k) + 2(5+1/\epsilon)\alpha_k^2(L^2e^2(x_k) + L^2 + B^2) - \frac{C}{2}d^2(x_k), \quad (33)$$

where ϵ is an arbitrary scalar.

Combined with $G'(x_k)(x_k - x^*) \geq F(x_k) - F(x^*)$, we can use (32) directly in *Coupled Supermartingale Convergence Theorem* (see Section 3 of [16]), and get:

$$x_k \xrightarrow{a.s.} x^*,$$

where x^* is a random point in the optimal set for Problem 1. ∎

Using constant step size If we let the step size constant, we cannot give a convergence guarantee, but we can give a error bound (see section 4.3, which the iterate $\{x_k\}$ will eventually get into.

An extension to Variational Inequalities(VI) Problem In fact, ICPM-SMC can be used in VI problems[2] as well. We can get a similar convergence theorem as Theorem 1. Article [17] talks about ICPM in VI problem, while it assumes $G(x)$ is *strongly monotone*. Our result can be regarded as an extension from *strongly monotone* objective function to *strictly monotone* objective function. See Appendix A for details.

4.2 EPO Error (without strongly convex assumption)

First of all, we consider the convergence rate without strongly convex assumption. In this case, we usually analyze the objective error: $f(x_k) - f(x^*)$ for Problem 1, not the solution error $\|x_k - x^*\|$, because x^* is not unique and $\|x_k - x^*\|^2$ is not surely convergent to zero. In our paper, the algorithms are randomized, so we should analyze its expected objective error: $\mathbf{E}[f(x_k) - f(x^*)]$. But $f(x_k) - f(x^*)$ is not always non-negative because x_k is not surely feasible. Then we define *Expected Projected Objective Error* (EPO Error): $\mathbf{E}[f(\Pi_X x_k) - f(x^*)]$, which is a non-negative error and a good criterion for convergence rate in general situation. And we also define its ergodic formulation (Ergodic EPO Error): $\mathbf{E}[f(\widetilde{x}_k) - f(x^*)]$, where $\widetilde{x}$ is the ergodic iterate defined as

$$\widetilde{x}_n := \frac{1}{\sum_{i=0}^k \alpha_i} \sum_{i=0}^k \alpha_i \Pi_X x_i. \quad (34)$$

[2]VI problem is a kind of problem more general than convex optimization problem, see [21] as a reference for VI problems, and [17] talks about stochastic VI problems.

Consider Ergodic EPO Error, use the convexity of f, we have:

$$\mathbf{E}[f(\widetilde{x}_k)-f(x^*)] = \mathbf{E}[f(\frac{1}{\sum_{i=0}^k \alpha_i}\sum_{i=0}^k \alpha_i \Pi_X x_i)-f(x^*)] \leq \frac{1}{\sum_{i=0}^k \alpha_i}\sum_{i=0}^k \alpha_i \mathbf{E}[f(\Pi_X x_i)-f(x^*)]. \tag{35}$$

Theorem 2 (EPO Error Estimation) *Suppose $\{x_k\}$ is generated by ICPM-SMC (Algorithm 1), if Assumption 1 holds, EPO Error and Ergodic EPO Error for all $k \geq 0$:*

$$\min_{0\leq i\leq k}\Big\{\mathbf{E}[\big(F(\Pi_X x_i)-F(x^*)\big)]\Big\} \leq \lambda_k^{-1}\exp(AL^2\Lambda_k)\big(e^2(x_0)+A(L^2+B^2)\Lambda_k\big), \tag{36}$$

and

$$\mathbf{E}[\big(F(\widetilde{x}_k)-F(x^*)\big)] \leq \lambda_k^{-1}\exp(AL^2\Lambda_k)\big(e^2(x_0)+A(L^2+B^2)\Lambda_k\big), \tag{37}$$

where $\lambda_k = \sum_{i=0}^k \alpha_k$, $\Lambda_k = \sum_{i=0}^k \alpha_k^2$ and A is an important constant that $A = 10 + 16/C^2$, C is a constant related to sampling scheme and sampling number (discussed in Lemma 2).

Proof In this proof, we can use (32) directly. Consider the third term of (32), use convexity of $F(x)$ in Problem 1, we can get:

$$\begin{aligned}
G'(x_k)(x_k-x^*) \geq& F(x_k)-F(x^*)\\
=& F(\Pi_X x_k)-F(x^*)+F(x_k)-F(\Pi_X x_k)\\
\geq& F(\Pi_X x_k)-F(x^*)-G'(\Pi_X x_k)(x_k-\Pi_X x_k)\\
=& F(\Pi_X x_k)-F(x^*)-(G(\Pi_X x_k)-G(x^*)+G(x^*))'(x_k-\Pi_X x_k)\\
\geq& F(\Pi_X x_k)-F(x^*)-L(\|\Pi_X x_k-x^*\|^2+1)\|x_k-\Pi_X x_k\|-B\|x_k-\Pi_X x_k\|\\
\geq& F(\Pi_X x_k)-F(x^*)-Le(x_k)d(x_k)-(L+B)d(x_k),
\end{aligned}$$

where the fifth step is based on

$$\begin{aligned}
\|G(x)-G(y)\| =& \sqrt{\|\mathbf{E}[g(x,v_k)|\mathscr{F}_K]-\mathbf{E}[g(y,v_k)|\mathscr{F}_K]\|^2} \leq \sqrt{\mathbf{E}[\|g(x,v_k)-g(y,v_k)\|^2}\\
=& \sqrt{L^2\|x-y\|^2+1} \leq L\|x-y\|+1
\end{aligned}$$

and

$$G(x) = \sqrt{\|\mathbf{E}[g(x,v_k)|\mathscr{F}_k]\|^2} \leq \sqrt{\mathbf{E}[\|g(x,v_k)\|^2|\mathscr{F}_k]} \leq B.$$

Combine it with (32), we can get the following inequality:

$$\begin{aligned}
\mathbf{E}[e^2(x_{k+1})|\mathscr{F}_k] \leq& (1+10L^2\alpha_k^2)e^2(x_k)+10\alpha_k^2(L^2+B^2)-2\alpha_k\big(F(\Pi_X x_k)-F(x^*)\big)\\
&+2\alpha_k\big(Le(x_k)d(x_k)+(L+B)d(x_k)\big)-\frac{C}{2}d^2(x_k),
\end{aligned}$$

which combined with

$$2\alpha_k Le(x_k)d(x_k)-\frac{C}{4}d^2(x_k) \leq \Big(16\frac{L^2}{C^2}\Big)\alpha_k^2 e^2(x_k)$$

and

$$2\alpha_k(L+B)d(x_k)-\frac{C}{4}d^2(x_k) \leq \frac{16}{C^2}(L^2+B^2),$$

we can get :

$$\mathbf{E}[e^2(x_{k+1})|\mathscr{F}_k] \leq (1 + AL^2\alpha_k^2)e^2(x_k) + A(L^2 + B^2)\alpha_k^2 - 2\alpha_k\big(F(\Pi_X x_k) - F(x^*)\big), \quad (38)$$

where $A = 10 + 16/C^2$.

Consider (38), take total expectation, we have:

$$\mathbf{E}[e^2(x_{k+1})] \leq (1 + AL^2\alpha_k^2)\mathbf{E}[e^2(x_k)] + A(L^2 + B^2)\alpha_k^2 - 2\alpha_k\mathbf{E}[\big(F(\Pi_X x_k) - F(x^*)\big)],$$

which inducted into:

$$\mathbf{E}[e^2(x_{k+1})] \leq \Big(\prod_{i=0}^{k}\big(1+AL^2\alpha_i^2\big)\Big)\Big(e^2(x_0)+A(L^2+B^2)\sum_{i=0}^{k}\alpha_i^2\Big)-\sum_{i=0}^{k}2\alpha_i\mathbf{E}[\big(F(\Pi_X x_i)-F(x^*)\big)].$$

Let $\lambda_k = \sum_{i=0}^{k}\alpha_k$ and $\Lambda_k = \sum_{i=0}^{k}\alpha_k^2$, we can get:

$$\begin{aligned}\sum_{i=0}^{k}2\alpha_i\mathbf{E}[\big(F(\Pi_X x_i) - F(x^*)\big)] \leq& \Big(\prod_{i=0}^{k}\big(1 + AL^2\alpha_i^2\big)\Big)\Big(e^2(x_0) + A(L^2 + B^2)\sum_{i=0}^{k}\alpha_i^2\Big),\\ \leq& \exp(AL^2\Lambda_k)\big(e^2(x_0) + A(L^2 + B^2)\Lambda_k\big)\end{aligned}$$

and obviously,

$$\lambda_k \min_{0\leq i\leq k}\Big\{\mathbf{E}[\big(F(\Pi_X x_i) - F(x^*)\big)]\Big\} \leq \sum_{i=0}^{k}2\alpha_i\mathbf{E}[\big(F(\Pi_X x_i) - F(x^*)\big)]$$

and based on (35), we can get:

$$\mathbf{E}[\big(F(\widetilde{x}_k) - F(x^*)\big)] \leq \lambda_k^{-1}\sum_{i=0}^{k}2\alpha_i\mathbf{E}[\big(F(\Pi_X x_i) - F(x^*)\big)],$$

thus, we can get (36) and (37) easily. ∎

About Parameters From (36) and (37), we can get that the error is $O(\frac{\exp(\Lambda_k)}{\lambda_k})$. This rate is in the same order for ICPM and ICPM-SMC. While the parameter in the error bound is quite different if the sampling scheme and sampling number is different. See the parameter in this bound, C could be controlled by sampling scheme and sampling number (see Lemma 2). Larger C leads to smaller A based on $A = 10 + 16/C^2$, and then leads to a smaller error bound for every $k \geq 0$, which means faster convergence rate. Thus, from Lemma 2, we can conclude that "Max Distance" weight distribution (9) behaves better in expected convergence rate, and sampling without replacement is better than that with replacement.

Special Case when $\alpha_k = k^{-\alpha}$ Let $\alpha_k = k^{-\alpha}$, we have $\lambda_k = \sum_{i=0}^{k} \alpha_i = O(k^{1-\alpha})$ and $\Lambda_k = O(k^{1-2\alpha})$, use them in (36) and (37) directly, then the error bound can be estimated as:

$$(k^{1-\alpha})^{-1} \exp(AL^2 k^{1-2\alpha}) \big(e^2(x_0) + A(L^2 + B^2)k^{1-2\alpha}\big) = O(k^{\alpha-1}) + O(k^{-\alpha}). \tag{39}$$

In this case, if $\alpha = 1/2$, then $\alpha_k = \frac{1}{\sqrt{k}}$, the convergence will be fast: $O(\frac{1}{\sqrt{k}})$. But such a step size is not squared summable, thus we cannot let $\alpha = 1/2$, we should let $\alpha \in (1/2, 1]$ and $\alpha \to 1/2$.

4.3 Expected Solution Error (For Strongly Convex Case)

In strongly convex situation, besides objective error $f(x_k) - f(x^*)$, we often analyze the solution error $\|x_k - x^*\|$, where x^* is the unique optimal solution, and the error is convergent to zero if the algorithm converges to x^*. For a stochastic algorithm, we analyze the convergence rate of expected squared solution error $\mathbf{E}[\|x_k - x^*\|^2]$.

Before giving the convergence rate, we consider an important lemma which will also be used in Section 5.

Lemma 3 *Suppose $\{\delta_k\}$ and $\{\alpha_k\}$ are non-negative sequences, and $\{\alpha_k\}$ is non-increasing, if the following holds for all $k \geq 0$,*

$$\delta_{k+1} \leq (1 - \mu\alpha_k + M\alpha_k^2)\delta_k + N\alpha_k^2, \tag{40}$$

where μ, M and N are non-negative constants, we can get a upper bound for δ_k:

$$\delta_k \leq \delta_0 \exp(-\mu\lambda_k) \exp(M\Lambda_k) + NI(M) \exp(2M\Lambda_{k_0}) \exp\Big(-\frac{\mu}{2}\lambda_k\Big) + N \sum_{j=1}^{k} \Big(\prod_{i=j+1}^{k} \big(1 - \frac{\mu}{2}\alpha_i\big)\Big)\alpha_j^2, \tag{41}$$

where $\lambda_k = \sum_{i=0}^{k} \alpha_k$, $\Lambda_k = \sum_{i=0}^{k} \alpha_k^2$, $k_0 = \inf_{k \in N}\{\alpha_k \leq \frac{\mu}{2M}\}$ and

$$I(M) \leq \begin{cases} 0, & \text{if } M = 0, \\ \dfrac{1}{M}, & \text{if } M \neq 0. \end{cases} \tag{42}$$

Corollary 3 *Define a function $\phi_\beta(t)$:*

$$\phi_\beta(t) \leq \begin{cases} \dfrac{k^\beta + 1}{\beta} & , \quad \text{if } \beta < 0, \\ \log(t) & , \quad \text{if } \beta = 0. \\ \dfrac{k^\beta - 1}{\beta} & , \quad \text{if } \beta > 0, \end{cases} \tag{43}$$

Based on this function, let $\alpha_k = Rk^{-\alpha}$ in Lemma 3, we can give a more specific bound for δ_k:

$$\delta_k \leq \begin{cases} 2\exp\big(2MR^2\phi_{1-2\alpha}(k)\big)\exp\Big(-\frac{\mu R}{8}k^{1-\alpha}\Big)\Big(\delta_0 + NI(M)\Big) + \frac{4RN}{\mu}k^{-\alpha} & , \quad \textit{if } 0 \leq \alpha < 1, \\ \exp(MR^2)\Big(\delta_0 + NI(M)\Big)k^{-\frac{\mu R}{2}} + NR^2\phi_{\frac{\mu R}{4}-1}(k)k^{-\frac{\mu R}{4}} & , \quad \textit{if } \alpha = 1, \end{cases} \tag{44}$$

where $I(M)$ is defined in (42).

Using constant step size Lemma 3 and Corollary 3 have been proved in [20] (see the proof of Theorem 1 in it). Consider two important special cases: $\alpha = 0$ and $\alpha = 1$. If $\alpha = 0$, the step size $\alpha_k = R$ is a constant. Based on (44), we can get that in this case,

$$\delta_k \leq O\bigg(\exp\Big(2MR^2\big(1 - \frac{\mu}{16MR}\big)k\Big)\bigg) + \frac{4RN}{\mu},$$

which reveals that: if $\frac{\mu}{R} \geq 16M$, δ_k can be bounded by a constant $\frac{4RN}{\mu}$.

Using step size of $\alpha_k = R/(k+1)$ Consider another case $\alpha = 1$, the step size $\alpha_k = \frac{R}{k+1}$. In this case, the convergence rate is sensitive to the parameter μ and R. Based on (44), we can get that in this case,

$$\delta_k \leq \begin{cases} O(k^{-\frac{\mu R}{4}}) & , \quad \text{if } 0 < \mu R < 4, \\ O(\frac{\log(k)}{k}) & , \quad \text{if } \mu R = 4, \\ O(\frac{1}{k}) & , \quad \text{if } \mu R > 4. \end{cases}$$

This property is also studied in [15], which talks a sequence satisfying the following recursion:

$$\delta_{k+1} \leq (1 - \frac{p}{k+1})\delta_k + \frac{d}{(k+1)^2},$$

and gives a similar result as (44) when $\alpha = 1$.

Use Lemma 3 and Corollary 3, we can get upper bound estimation of $\mathbf{E}[\|x_k - x^*\|^2]$.

Theorem 3 (Expected Solution Error) *Suppose $\{x_k\}$ is generated by ICPM-SMC (Algorithm 1), if Assumption 1 and 2 hold and step size α_k is non-increasing, $\mathbf{E}[e^2(x_k)] = \mathbf{E}[\|x_k - x^*\|^2]$ can be bounded as:*

$$\mathbf{E}[e^2(x_k)] \leq e^2(x_0)A_{1,k} + NA_{2,k}, \tag{45}$$

where

$$N = 10(L^2 + B^2) + \frac{4B^2}{C^2}, \tag{46}$$

$$A_{1,k} = \exp(-2\sigma\lambda_k)\exp(10L^2\Lambda_k), \tag{47}$$

and

$$A_{2,k} = \frac{1}{10L^2}\exp(20L^2\Lambda_{k_0})\exp(-\sigma\lambda_k) + \sum_{j=1}^{k}\Big(\prod_{i=j+1}^{k}(1-\sigma\alpha_i)\Big)\alpha_j^2. \tag{48}$$

PROOF Combine (28) and (29), take total expectation on it, we can get

$$\mathbf{E}[e^2(x_{k+1})|\mathscr{F}_k] \le (1-2\alpha_k\sigma+10L^2\alpha_k^2)e^2(x_k) + 2B\alpha_k d(x_k) - \frac{C}{2}d^2(x_k) + 10(L^2+B^2)\alpha_k^2.$$

Consider

$$2B\alpha_k d(x_k) - \frac{C}{2}d^2(x_k) \le \frac{4B^2}{C^2}\alpha_k^2,$$

take total expectation, we can get:

$$\mathbf{E}[e^2(x_{k+1})] \le (1-2\alpha_k\sigma+10L^2\alpha_k^2)\mathbf{E}[e^2(x_k)] + \Big(10(L^2+B^2)+\frac{4B^2}{C^2}\Big)\alpha_k^2. \tag{49}$$

We can use this inequality directly into (40), let $\delta_k = \mathbf{E}[d^2(x_k)]$, $\mu = 2\sigma$, $M = 10L^2$ and $N = 10(L^2+B^2)+\frac{4B^2}{C^2}$, then we can get (45). ∎

About Parameters From (45-48), we can get some parameters that influence the convergence rate: σ, L, B and C. The former three are decided by the properties of the problem itself, while C can be controlled by the algorithm. Based on $N = 10(L^2+B^2)+\frac{4B^2}{C^2}$, larger C gives a smaller N and leads to a smaller expected error, which means a faster convergence rate. From Lemma 2, we can get the principle of C, then we can get the same conclusion as non-strongly convex case (Section 4.2) on how sampling scheme and weight distribution $w_k(i)$ influence the convergence rate.

Another interesting case is using constant step size $\alpha_k = R$. According to Corollary 3, we cannot guarantee convergence, but we can give an error upper bound $\frac{2R}{\sigma}N$. Thus, larger C leads to a smaller error bound when using constant step size.

4.4 Expected Distance to Feasible Set

Except for $\|x_k - x^*\|$ and $f(\Pi_X x_k) - f(x^*)$, another rate we concern is $\|x_k - \Pi_X x_k\|$, which means in what rate x_k goes to the feasible set X. For stochastic algorithm, we analyze $\mathbf{E}[d^2(x_k)] = \mathbf{E}[\|x_k - \Pi_X x_k\|^2]$ to give this rate. Before analysis, we give a useful lemma.

Lemma 4 *Suppose $\{\delta_k\}$ and $\{\alpha_k\}$ are non-negative sequences, if the following holds for all $k \ge 0$,*

$$\delta_{k+1} \le (1-\delta)\delta_k + N\alpha_k^2, \tag{50}$$

where σ and N are non-negative constants.

(a) If α_k is non-increasing: $\alpha_{k+1} \le \alpha_k$, we can get a upper bound for δ_k:

$$\delta_k \le \delta_0(1-\delta)^k + N(1-\delta)^{k/2}\Lambda_{k/2} + \frac{N}{\delta}\alpha_{k/2}^2, \tag{51}$$

where $\Lambda_k = \sum_{i=0}^{k}\alpha_k^2$.

(b) If α_k is "slowly decreasing": there exists a K_0, so that for any $k \geq K_0$,

$$\alpha_{k+1}^2 \geq \big(1 - \frac{\delta}{2}\big)\alpha_k^2 \tag{52}$$

holds, we can get another upper bound for δ_k:

$$\delta_k \leq \frac{2N}{\delta}\alpha_k^2 + \Big(\delta_{K_0} - \frac{2N}{\delta}\alpha_{K_0}^2\Big)(1-\delta)^{k-K_0}. \tag{53}$$

PROOF Consider non-increasing case, apply (50) by k times, we can get:

$$\delta_k \leq \delta_0(1-\delta)^k + N\sum_{i=0}^{k-1}(1-\delta)^{k-i}\alpha_i^2.$$

Consider the last term, for any $m \in \{1, 2, \ldots, k-1\}$, we can split it into two terms:

$$\begin{aligned}\sum_{i=0}^{k-1}(1-\delta)^{k-i}\alpha_i^2 =& \sum_{i=0}^{m}(1-\delta)^{k-i}\alpha_i^2 + \sum_{i=m+1}^{k-1}(1-\delta)^{k-i}\alpha_i^2 \\ \leq& (1-\delta)^{k-m}\sum_{i=0}^{m}\alpha_i^2 + \alpha_m^2\sum_{i=m+1}^{k-1}(1-\sigma)^{k-i} \\ \leq& (1-\delta)^{k-m}\sum_{i=0}^{m}\alpha_i^2 + \frac{\alpha_m^2}{\delta}.\end{aligned}$$

Let $m = [k/2]$, we can get (51). (a) is proved.

Consider (b), we should rewrite (50) as (for $k \geq K_0$):

$$\begin{aligned}\delta_{k+1} \leq& (1-\delta)\delta_k + N\alpha_k^2 \\ =& (1-\delta)\delta_k + \frac{2}{\delta}\Big(1-\frac{\delta}{2}\Big)N\alpha_k^2 - \frac{2}{\delta}(1-\delta)N\alpha_k^2 \\ \leq& (1-\delta)\delta_k + \frac{2}{\delta}N\alpha_{k+1}^2 + \frac{2}{\delta}(1-\delta)N\alpha_k^2,\end{aligned}$$

and then

$$\delta_{k+1} - \frac{2N}{\delta}\alpha_{k+1}^2 \leq (1-\delta)\Big(\delta_k - \frac{2N}{\delta}\alpha_k^2\Big).$$

Apply it by $k - K_0$ times, we can get (53). (b) is proved. ∎

Remark Whether in case (a) or (b), we can estimate the bound as following:

$$\delta_k \leq O\big((1-\delta)^k\big) + N \cdot O(\alpha_k^2), \tag{54}$$

where N is an important parameter in the convergence rate. (Because we usually use a step size like $\alpha_k = k^{-\alpha}$, it obviously satisfies $\alpha_{k/2} = O(\alpha_k)$.)

Use this lemma, we can give an upper bound for $\mathbf{E}[d^2(x_k)] = \mathbf{E}[\|x_k - \Pi_X x_k\|^2]$.

Theorem 4 (Expected Distance to the Feasible Set) *Suppose $\{x_k\}$ is generated by ICPM-SMC (Algorithm 1), if Assumption 1 holds and step size α_k satisfies "non-increasing" or "slowly decreasing" (defined in Lemma 4), we can get that $\mathbf{E}[e^2(x_k)]$ is bounded by a constant, and $\mathbf{E}[d^2(x_k)]$ is bounded by the following sequence:*

$$\delta_k \leq O\Big((1-\frac{C}{4})^k\Big) + \Big(5+\frac{4}{C}\Big)\cdot O(\alpha_k^2), \tag{55}$$

where C is a constant related to the sampling scheme (discussed in Lemma 2).

Proof In this section, we don't assume $F(x)$ is strongly convex, thus we can use (32) and (33), take total expectation, we can get:

$$\mathbf{E}[e^2(x_{k+1})] \leq (1+10L^2\alpha_k^2)\mathbf{E}[e^2(x_k)] + 10\alpha_k^2(L^2+B^2), \tag{56}$$

and

$$\mathbf{E}[d^2(x_{k+1})] \leq (1+\epsilon)\mathbf{E}[d^2(x_k)] + 2\Big(5+\frac{1}{\epsilon}\Big)\big(L^2\mathbf{E}[e^2(x_k)] + L^2 + B^2\big)\alpha_k^2 - \frac{C}{2}\mathbf{E}[d^2(x_k)]. \tag{57}$$

Apply (56) by k times, we can get $\mathbf{E}[e^2(x_k)]$ is bounded:

$$\begin{aligned}\mathbf{E}[e^2(x_k)] \leq& \prod_{i=0}^{k}\big(1+10L^2\alpha_i^2\big)\Big(e^2(x_0) + 10(L^2+B^2)\sum_{i=0}^{k}\alpha_i^2\Big)\\ \leq& \exp(10L^2\Lambda_\infty)\Big(e^2(x_0) + 10(L^2+B^2)\Lambda_\infty\Big),\end{aligned}$$

where $\Lambda_\infty = \sum_{i=0}^{\infty}\alpha_i^2 < \infty$. To simplify, we use A denote $A := \max_{k\geq 0}\mathbf{E}[e^2(x_k)]$, then (57) can be rewritten as:

$$\mathbf{E}[d^2(x_{k+1})] \leq (1+\epsilon)\mathbf{E}[d^2(x_k)] + 2\Big(5+\frac{1}{\epsilon}\Big)\Big((A+1)L^2+B^2\Big)\alpha_k^2 - \frac{C}{2}\mathbf{E}[d^2(x_k)].$$

Let $\epsilon = C/4$, we can get:

$$\mathbf{E}[d^2(x_{k+1})] \leq \Big(1-\frac{C}{4}\Big)\mathbf{E}[d^2(x_k)] + 2\Big(5+\frac{4}{C}\Big)\big((A+1)L^2+B^2\big)\alpha_k^2.$$

Use Lemma 4, we can get (55). ∎

From (55), we can get that larger C can also guarantee a smaller distance to the feasible set. Since we often use a "not summable but squared summable" sequence (see (31)) as the step size, then $(1-\delta)^k = o(\alpha_k^2)$ holds, thus we can estimate $\mathbf{E}[d^2(x_k)]$ by $O(\alpha_k^2)$.

4.5 A Brief Summary

We analyzed *EPO error* $\mathbf{E}[f(\Pi_X x_k) - f(x^*)]$, *expected solution error* $\mathbf{E}[\|x_k - x^*\|^2]$ and *expected distance to feasible set* $\mathbf{E}[\|x_k - \Pi_X x_k\|^2]$ separately and gave the rates in Section 4.2-4.4. In all the rates, we notice there exists an important parameter: C. We organized all the results about convergence rate of expectation in Table 4.5.

Type of Error	Conditions	Orders of Rate ($\alpha_k = Rk^{-\alpha}$)	Factors Involving C
$\mathbf{E}[f(\Pi_X x_k) - f(x^*)]$	General	$O\big(k^{\alpha-1}\big) + O\big(k^{-\alpha}\big)$	$10 + \frac{16}{C^2}$
$\mathbf{E}[\|x_k - x^*\|^2]$	Strongly Convex $F(x)$	$O\Big(\frac{1}{k}\Big)$, if $\alpha = 1$ and R sufficient large	$10(L^2 + B^2) + \frac{4B^2}{C^2}$
$\mathbf{E}[\|x_k - \Pi_X x_k\|^2]$	General	$O\big(k^{-2\alpha}\big)$	$5 + \frac{4}{C}$

Table 1: Summary of Convergence Rate

See the right column of Table 4.5, increasing C can only reduce a part of the error because in $10 + \frac{16}{C^2}$, $10(L^2 + B^2) + \frac{4B^2}{C^2}$ and $5 + \frac{4}{C}$, $1/C$ is just one of the two terms in the parameter. See Lemma 2, in fact, if we just use ICPM (let $M_k = 1$ for all $k \geq 0$ in ICPM-SMC), $1/C = m\eta/\rho$ is quite large because m is often large. Thus, $1/C$ is the main item. That is to say, ICPM-SMC is quite effective.

Consider the case of Lemma 2 (b), we use "Max Distance" weight distribution and Sampling with Replacement. If $M_k = 1$, $C = \rho/(m\eta)$ is the same as ICPM; if $M_k \to \infty$, $C \to 1$, which is the same order as the other term in $5 + 4/C$, $10 + 16/C^2$, etc; if $M_k \approx m$, $C \to 1 - e^{-\rho}$, which is also $O(1)$. Assume $1 \ll M_k \ll m$, we have

$$1-(1-\frac{\rho}{m})^{M_k} \approx 1-(1-\rho\frac{M_k}{m}+\rho^2\frac{M_k(M_k-1)}{2}\frac{1}{m^2}) = M_k\frac{\rho}{m}-\rho^2\frac{M_k(M_k-1)}{2}\frac{1}{m^2} \approx M_k\frac{\rho}{m},$$

which is almost M_k times as ICPM. Then consider calculating one step of MDPM and calculating M_k steps of ICPM. They have almost the same effect when $1 \ll M_k \ll m$ according to the above conclusion, however, one step of MDPM needs calculate only one step of sub-gradient and M steps of projection, while M steps of ICPM needs to calculate M steps of sub-gradient and M steps of projection. That is to say, MDPM use less expense to get almost the same effect as ICPM.

5 Convergence Stability

For randomized algorithms, uncertainty is an obstacle of practical use. In this section, we say a stochastic algorithm is more *stable* if its uncertainty is reduced.

In this section, we propose 3 criteria to judge the convergence stability of a stochastic algorithm: *Conditional Variance*, *Boundedness Probability* and *Exceeding Length*. The first one is analyzed without "strongly convex" assumption, while the other two are both based on "strongly convex" assumption because they analyze solution error $\|x_k - x^*\|^2$, not the objective error. In these sections, we contrast ICPM and ICPM-SMC in these criteria.

5.1 Conditional Variance

In this section, we use *Conditional Variance* to analyze ICPM-SMC:

$$Var[x_{k+1}|\mathscr{F}_k] = \mathbf{E}[\|x_{k+1} - \mathbf{E}[x_{k+1}|\mathscr{F}_k]\|^2|\mathscr{F}_k]$$

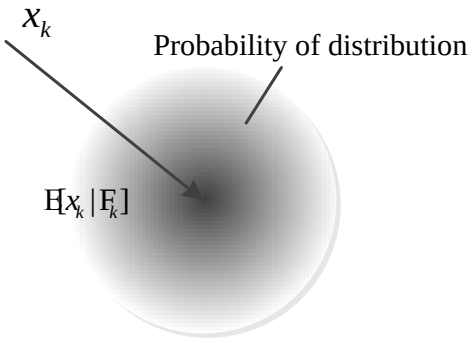

Figure 3: Conditional Variance

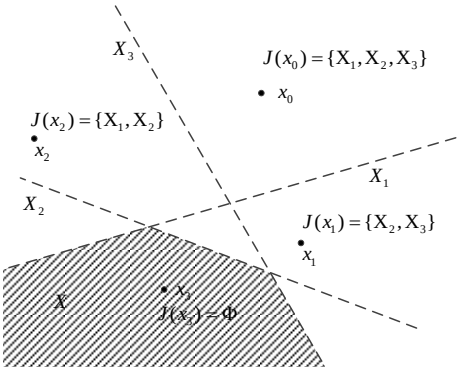

Figure 4: Examples for $J(x)$

If the variance is enough small, we can say x_{k+1} will not go too far away form its expectation on the condition of $\mathscr{F}_k$, that is what we call stability. We give this an illustration in Figure 3.

Now we give a definition (also used in following sections) $J(x)$:

$$J(x) = \{j | x \in X_j, j = 1, 2, \ldots, m\}, \tag{58}$$

which is a set involving the index of constraints that x satisfies. We give an intuitive image for this definition in Figure 4.

Obviously, if we consider the variance on the condition of $\mathscr{F}_k$, $J(x_k)$ is a deterministic set and $\#J(x_k)$, which means the number of $J(x_k)$, is a deterministic scalar. While $J(y_{k+1})$ is a random set and $\#J(y_{k+1})$ is a random variable. So we usually use its expectation $\mathbf{E}[\#J(y_{k+1})|\mathscr{F}_k]$, which is a part of J_k in the following theorem.

Theorem 5 (Conditional Variance) *Let Assumption 1 hold. Suppose that $\{x_k\}$ is generated by ICPM-SMC, we can give an upper bound on the conditional variance per iteration:*

$$Var[x_{k+1}|\mathscr{F}_k] \leq \frac{16}{\overline{M}_k} J_k \|x_k - \Pi_X x_k\|^2 + \left(\frac{64}{\overline{M}_k} J_k + 2\right) A^2 \alpha_k^2, \tag{59}$$

where A is a scalar such that $\mathbf{E}[\|g(x_k, v_k)\||\mathscr{F}_k] \leq A$ with probability 1, $\|x_k - \Pi_X x_k\|^2$ is the distance from x_k to the feasible set X. And then we see two important factors J_k and $\overline{M}_k$.

Here J_k is defined as:

$$J_k = P_{max} \mathbf{E}\big[\#J(y_{k+1})\big|\mathscr{F}_k\big] \tag{60}$$

where $\#J(y_{k+1})$ means the number of set $J(y_{k+1})$, and P_{max} is a constant defined as

$$P_{max} = \max_{j=1,\ldots,m} \{P(\omega_{k,i} = X_j | \mathscr{F}_k)\} = 1 - \rho\Big(1 - \frac{1}{m}\Big), \tag{61}$$

which means the upper bound of probability for sampling one particular constraint.

And M_k is decided by the weight distribution and sampling scheme. We consider it in three cases:

(a) For general ICPM-SMC, $\overline{M}_k \geq 1$.

(b) If we use average weight distribution (ICAPM, see (8)), and sample $\omega_{k,i}$ *with replacement,* $\overline{M}_k = M_k$.

(c) If we use average weight distribution (ICAPM, see (8)), and sample $\omega_{k,i}$ *without replacement,* $\overline{M}_k = \frac{M_k(m-1)}{m-M_k}$.

PROOF Firstly, we consider the variance of $x_{k+1,i} = \Pi_{\omega_{k,i}} y_{k+1}$.

Variance of every item $x_{k+1,i}$ Split the variance of $x_{k+1,i}$ into two parts, one caused by randomness of v_k, another caused by randomness of $\omega_{k,i}$s.

$$\begin{aligned}
&\|x_{k+1,i} - \mathbf{E}[x_{k+1,i}|\mathscr{F}_k]\| \\
=&\|x_{k+1,i} - \mathbf{E}[\mathbf{E}[x_{k+1,i}|y_{k+1}]|\mathscr{F}_k]\| \\
\leq&\|x_{k+1,i} - \mathbf{E}[x_{k+1,i}|y_{k+1}]\| + \|\mathbf{E}[x_{k+1,i}|y_{k+1}] - \mathbf{E}[\mathbf{E}[x_{k+1,i}|y_{k+1}]|\mathscr{F}_k]\|
\end{aligned} \tag{62}$$

Define $J(x) = \{j | x \in X_j, j = 1, 2, \ldots, m\}$ which involves the index of constraints that x satisfies, the first term of (62) can be bounded as:

$$\begin{aligned}
&\|x_{k+1,i} - \mathbf{E}[x_{k+1,i}|y_{k+1}]\| \\
\leq&\|\Pi_{\omega_k,i} y_{k+1} - \sum_{j=1}^{m} \Pi_{X_j} y_{k+1} P\big(\omega_{k,i} = X_j|\mathscr{F}_k\big)\| \\
\leq& \sum_{j\in J(y_{k+1})} P\big(\omega_{k,i} = X_j|\mathscr{F}_k\big)\big(\|\Pi_{\omega_k,i} y_{k+1} - \Pi_{X_j} y_{k+1}\|\big) \\
&+ \sum_{j\notin J(y_{k+1})} P\big(\omega_{k,i} = X_j|\mathscr{F}_k\big)\big(\|\Pi_{\omega_k,i} y_{k+1} - \Pi_{X_j} y_{k+1}\|\big)
\end{aligned} .$$

Based on Assumption 1 (e), we can get the probability:

$$\frac{\rho}{m} \leq P\big(\omega_{k,i} = X_j|\mathscr{F}_k\big) \leq 1 - \rho\Big(1 - \frac{1}{m}\Big).$$

To simplify, let $P_{\max} = 1 - \rho\Big(1 - \frac{1}{m}\Big)$.

If $\omega_{k,i} \in J(y_{k+1})$, we have:

$$\begin{aligned}
&\sum_{j\in J(y_{k+1})} P\big(\omega_{k,i} = X_j|\mathscr{F}_k\big)\big(\|\Pi_{\omega_k,i} y_{k+1} - \Pi_{X_j} y_{k+1}\|\big) \\
=& \sum_{j\in J(y_{k+1})} P\big(\omega_{k,i} = X_j|\mathscr{F}_k\big)\big(\|\Pi_{\omega_k,i} y_{k+1} - y_{k+1} + y_{k+1} - \Pi_{X_j} y_{k+1}\|\big) \\
\leq& \sum_{j\in J(y_{k+1})} P_{\max}\Big(\|\Pi_{\omega_k,i} y_{k+1} - y_{k+1}\| + \|y_{k+1} - \Pi_{X_j} y_{k+1}\|\Big) \\
\leq&2P_{\max}\big(\#J(y_{k+1})\big)\|y_{k+1} - \Pi_X y_{k+1}\| = 2P_{\max}\big(\#J(y_{k+1})\big)d(y_{k+1})
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{j\notin J(y_{k+1})} P\big(\omega_{k,i}=X_j|\mathscr{F}_k\big)\big(\|\Pi_{\omega_k,i}y_{k+1}-\Pi_{X_j}y_{k+1}\|\big)\\
=&\sum_{j\notin J(y_{k+1})} P\big(\omega_{k,i}=X_j|\mathscr{F}_k\big)\big(\|\Pi_{\omega_k,i}y_{k+1}-y_{k+1}\|\big)\\
\leq& P_{\max}\big(m-\#J(y_{k+1})\big)d(y_{k+1})
\end{aligned}$$

Thus, if $\omega_{k,i}\in J(y_{k+1})$, we can get:

$$\begin{aligned}
&\|x_{k+1,i}-\mathbf{E}[x_{k+1,i}|y_{k+1}]\|\\
\leq& 2P_{\max}\big(\#J(y_{k+1})\big)d(y_{k+1})+P_{\max}\big(m-\#J(y_{k+1})\big)d(y_{k+1})\\
=& P_{\max}\big(m+\#J(y_{k+1})\big)d(y_{k+1})
\end{aligned}$$

In the same way, if $\omega_{k,i}\notin J(y_{k+1})$, we can get:

$$\|x_{k+1,i}-\mathbf{E}[x_{k+1,i}|y_{k+1}]\|\leq P_{\max}\big(\#J(y_{k+1})\big)d(y_{k+1})$$

Thus,

$$\begin{aligned}
&\mathbf{E}[\|x_{k+1,i}-\mathbf{E}[x_{k+1,i}|y_{k+1}]\|^2|y_{k+1}]\\
\leq& P\big(\omega_{k,i}\in J(y_{k+1})\big)\Big(P_{\max}\big(m+\#J(y_{k+1})\big)\Big)^2 d^2(y_{k+1})\\
&+P\big(\omega_{k,i}\notin J(y_{k+1})\big)\Big(P_{\max}\big(\#J(y_{k+1})\big)\Big)^2 d^2(y_{k+1})\\
=& P_{\max}\big(\#J(y_{k+1})\big)\Big(P_{\max}\big(m+\#J(y_{k+1})\big)\Big)^2 d^2(y_{k+1})\\
&+P_{\max}\big(m-\#J(y_{k+1})\big)\Big(P_{\max}\big(\#J(y_{k+1})\big)\Big)^2 d^2(y_{k+1})\\
\leq& 4P_{\max}\big(\#J(y_{k+1})\big)d^2(y_{k+1})\\
\leq& 4P_{\max}\big(\#J(y_{k+1})\big)\big(2d^2(x_k)+8\alpha_k^2\|g(x_k,v_k)\|^2\big),
\end{aligned}$$

where the last step uses $d^2(x)\leq 2d^2(y)+\|x-y\|^2$. Take expectation on the condition of $\mathscr{F}_k$, let $J_k=\mathbf{E}\big[P_{\max}\big(\#J(y_{k+1})\big)\big|\mathscr{F}_k\big]$, we get:

$$\begin{aligned}
\mathbf{E}\big[\big\|x_{k+1,i}-\mathbf{E}[x_{k+1,i}|y_{k+1}]\big\|^2\big|\mathscr{F}_k\big]=&\mathbf{E}\Big[\mathbf{E}\big[\|x_{k+1,i}-\mathbf{E}[x_{k+1,i}|y_{k+1}]\big\|^2\big|y_{k+1}\big]\Big|\mathscr{F}_k\Big]\\
\leq&\mathbf{E}\Big[8P_{\max}\big(\#J(y_{k+1})\big)\big(d^2(x_k)+4\alpha_k^2\|g(x_k,v_k)\|^2\big)\Big|\mathscr{F}_k\Big]\\
\leq&8J_k\big(d^2(x_k)+4\alpha_k^2A^2\big)
\end{aligned}$$

Let $Y=y_{k+1}$ regarded as a random variable here. The second term of (62) can be bounded

as:

$$\begin{aligned}
&\mathbf{E}\Big[\Big\|\mathbf{E}[x_{k+1,i}|Y]-\mathbf{E}\big[\mathbf{E}[x_{k+1,i}|Y]\big|\mathscr{F}_k\big]\Big\|^2\Big|\mathscr{F}_k\Big]\\
\leq&\mathbf{E}\Big[\Big\|\mathbf{E}[x_{k+1,i}|Y]-\mathbf{E}\big[x_{k+1,i}\big|\mathbf{E}[Y|\mathscr{F}_k]\big]\Big\|^2\Big|\mathscr{F}_k\Big]\\
=&\mathbf{E}\Big[\Big\|\sum_{j=1}^m P\big(\omega_{k,i}=X_j|\mathscr{F}_k\big)\Pi_{X_j}Y-\sum_{j=1}^m P\big(\omega_{k,i}=X_j|\mathscr{F}_k\big)\Big(\Pi_{X_j}\mathbf{E}[Y|\mathscr{F}_k]\Big)\Big\|^2\Big|\mathscr{F}_k\Big]\\
=&\mathbf{E}\Big[\Big\|\sum_{i=1}^m P\big(\omega_{k,i}=X_j|\mathscr{F}_k\big)\Big(\Pi_iY-\Pi_i\mathbf{E}[Y|\mathscr{F}_k]\Big)\Big\|^2\Big|\mathscr{F}_k\Big]\\
\leq&\mathbf{E}\Big[\Big\|\sum_{i=1}^m P\big(\omega_{k,i}=X_j|\mathscr{F}_k\big)\big(Y-\mathbf{E}[Y|\mathscr{F}_k]\big)\Big\|^2\Big|\mathscr{F}_k\Big]\\
=&\mathbf{E}\Big[\big\|\alpha_k\big(g(x_k,v_k)-G(x_k)\big)\big\|^2\Big|\mathscr{F}_k\Big]\leq 2A^2\alpha_k^2,
\end{aligned}$$

where the second step we use $\mathbf{E}[|X-\mathbf{E}[X]|]\leq\mathbf{E}[|X-a|]$ for arbitrary a.

The variance can be bounded as

$$\begin{aligned}
Var[x_{k+1,i}|\mathscr{F}_k]\leq&2\mathbf{E}\big[\big\|x_{k+1,i}-\mathbf{E}[x_{k+1,i}|y_{k+1}]\big\|^2\big|\mathscr{F}_k\big]+2\mathbf{E}\Big[\big\|E[x_{k+1,i}|Y]-\mathbf{E}\big[\mathbf{E}[x_{k+1,i}|Y]\big|\mathscr{F}_k\big]\big\|^2\Big|\mathscr{F}_k\Big]\\
\leq&16J_kd^2(x_k)+64J_kA^2\alpha_k^2+2A^2\alpha_k^2
\end{aligned}$$

Proof of (a) Consider the following term:

$$\begin{aligned}
&\|x_{k+1}-\mathbf{E}[x_{k+1}|\mathscr{F}_k]\|\\
=&\|\sum_{i=1}^{M_k}w_k(i)x_{k+1,i}-\mathbf{E}[\mathbf{E}[x_{k+1}|y_{k+1}]|\mathscr{F}_k]\|\\
\leq&\|\sum_{i=1}^{M_k}w_k(i)(x_{k+1,i}-\mathbf{E}[x_{k+1,i}|y_{k+1}])\|+\|\mathbf{E}[x_{k+1}|y_{k+1}]-\mathbf{E}[\mathbf{E}[x_{k+1}|y_{k+1}]|\mathscr{F}_k]\|
\end{aligned}$$

Consider the first term, use *Jensen Inequality*,

$$\|\sum_{i=1}^{M_k}w_k(i)(x_{k+1,i}-\mathbf{E}[x_{k+1,i}|y_{k+1}])\|\leq\sum_{i=1}^{M_k}w_k(i)\|x_{k+1,i}-\mathbf{E}[x_{k+1,i}|y_{k+1}]\|=\|x_{k+1,i}-\mathbf{E}[x_{k+1,i}|y_{k+1}]\|$$

Then we follow the same line as the proof of $x_{k+1,i}$ and get the variance of x_{k+1} can be bounded by the same bound as $x_{k+1,i}$, which equals to (59) in general cases ($\overline{M}_k=1$). Part(a) is proved.

Proof of (b) Consider the special situation of "average weight distribution" (ICAPM, see (7) and (8)). If we use sampling with replacement, every $\omega_{k,i}$ is iid. Based on iid of $\omega_{k,i}$, we can get:

$$\mathbf{E}[\|\frac{1}{M_k}\sum_{i=1}^{M_k}(x_{k+1,i}-\mathbf{E}[x_{k+1,i}|y_{k+1}])\|^2|y_{k+1}]=\frac{1}{M_k}\|x_{k+1,i}-\mathbf{E}[x_{k+1,i}|y_{k+1}]\|^2.$$

Thus, the whole variance is

$$Var[x_{k+1,i}|\mathscr{F}_k] \leq \frac{16}{M_k} J_k d^2(x_k) + \frac{64}{M_k} J_k A^2 \alpha_k^2 + 2A^2\alpha_k^2,$$

which equals to (59) when $\overline{M}_k = M_k$. Part (b) is proved.

Proof of (c) If we use sampling without replacement, the variance can be reduced furthermore. As we know, sampling without replacement can give the following result:

$$\mathbf{E}[\|\frac{1}{M_k}\sum_{i=1}^{M_k}(x_{k+1,i} - \mathbf{E}[x_{k+1,i}|y_{k+1}])\|^2|y_{k+1}] = \frac{1}{M_k}\frac{m-M_k}{m-1}\|x_{k+1,i} - \mathbf{E}[x_{k+1,i}|y_{k+1}]\|^2,$$

which equals to 0 when $M_k \geq m$, and give a smaller upper bound for variance. Follow the same line above, we can get the whole variance and part (c) is proved. ∎

About J_k If we use unbiased sampling scheme (i.e. $\rho = 1$ in (23)), let $\rho = 1$ in (61), we get $P_{\max} = 1/m$, then we have:

$$J_k = \mathbf{E}\big[\frac{\#J(y_{k+1})}{m}\big|\mathscr{F}_k\big] \leq 1.$$

Thus, for unbiased sampling scheme, we can give a simplified bound:

$$Var[x_{k+1}|\mathscr{F}_k] \leq \frac{16}{\overline{M}_k}\|x_k - \Pi_X x_k\|^2 + \left(\frac{64}{\overline{M}_k} + 2\right) A^2\alpha_k^2. \tag{63}$$

About $\overline{M}_k$ If we use ICAPM (average weight distribution), we can guarantee a larger lower bound for $\overline{M}_k$. Furthermore, if we use "sampling without replacement", we can improve the lower bound further. If we sample with replacement, the variance factor $\frac{1}{M_k} \to 0$ when $M_k \to 0$; while if without replacement, the factor $\frac{1}{M_k}\frac{m-M_k}{m-1} = 0$ when $M_k \geq m$. The second one gives a smaller variance.

5.2 Boundedness Probability

In Section 4, we discussed expected error $\mathbf{E}[\|x_k - x^*\|^2]$, $\mathbf{E}[f(\Pi_X x_k) - f(x^*)]$ and $\mathbf{E}[\|x_k - \Pi_X x_k\|^2]$, not the error $\|x_k - x^*\|^2$,etc. itself. In article [17], the error is estimated by

$$\min_{0\leq i\leq k}\|x_k - x^*\|^2.$$

These two methods can estimate convergence rate, however, it's not effective because the uncertainty caused by large variance makes the algorithm not "stable". Take Theorem 3 as an example. Suppose $\alpha_k = \frac{R}{k+1}$ and R is sufficient large, then we can get the following based on Theorem 3:

$$\mathbf{E}[\|x_k - x^*\|^2] \leq O(\frac{1}{k}).$$

Or we can get the another bound based on Proposition 2 in [17]:

$$\min_{0\le i\le k}\|x_i-x^*\|^2\le O(\frac{1}{k}).$$

Hence, if the variance is quite large, the error $\|x_k-x^*\|^2$ maybe quite large. While we don't know x^* before the result output, so we cannot determine which one of $\{x_i|i=0,\dots,m\}$ is the most optimal solution. If we output the last one x_k or a random one, it may be quite far to the optimal point x^*. That's the reason why we have to analyze

$$\|x_k-x^*\|^2,\quad \forall 0\le k\le T$$

as a criterion. In fact, for a stochastic algorithm, we cannot get almost surely bound for it, we can only estimate the probability of being bounded, that's what we called "Boundedness Probability", which could be an important criterion for convergence stability. In this case, for example, we should analyze $P\big(\|x_k-x^*\|^2\le O(\frac{1}{k}),\forall 0\le k\le T\big)$.

Theorem 6 (Boundedness Probability) *Suppose $\{x_k\}$ is generated by ICPM-SMC, if Assumption 1 and 2 hold, we also assume $g(x,v)=G(x)$ for clear result, we can get for any $T\ge 0$:*

$$e^2(x_k)\le e^2(x_0)\exp(-4\sigma\lambda_k)+\Big(5A^2+\frac{16B^2}{C^2}\Big)\sum_{j=1}^{k}\Big(\prod_{i=j+1}^{k}(1-\sigma\alpha_i)\Big)\alpha_j^2,\quad \forall 0\le k\le T,$$
$$w.p.\ \prod_{k=1}^{T}\Big(1-\frac{4}{C^2}\frac{1}{\overline{M}_k}\Big), \tag{64}$$

where "w.p." means "with probability at least", $A=\max_{k\ge 0}\|g(x_k,v_k)\|$ since x_k is convergent, $\overline{M}_k$ is a scalar depended on the weight distribution and sampling scheme (discussed in Theorem 5).

Proof Our proof line has two steps: firstly we get a recursion like (40) with a probability, and then apply it iteratively to get a bound with probability.

Consider (27), where has two random terms: $\sum_{i=1}^{M_k}w_k(i)\|\Pi_{\omega_{k,i}}x_k-x_k\|^2$ and $\|g(x_k,v_k)\|^2$. For x_k is convergent, we assume the second one is bounded and let $A=\max_{k\ge 0}\|g(x_k,v_k)\|$.

Now consider the first one, let $d_{k,i}=\|\Pi_{\omega_{k,i}}x_k-x_k\|^2$ and $d_k=\sum_{i=1}^{M_k}w_k(i)d_i$. By (2), $\mathbf{E}[d_k|\mathscr{F}_k]\ge Cd^2(x_k)$. Since the variance $Var[d_{k,i}|\mathscr{F}_k]$ can be bounded by $d^4(x_k)$:

$$Var[d_{k,i}|\mathscr{F}_k]\le\mathbf{E}[(d_{k,i})^2|\mathscr{F}_k]=\mathbf{E}[\|\Pi_{\omega_{k,i}}x_k-x_k\|^4|\mathscr{F}_k]\le d^4(x_k),$$

we can get variance of d_k by following almost the same method as the proof of Theorem 5:

$$Var[d_k|\mathscr{F}_k]\le\frac{1}{\overline{M}_k}d^4(x_k),$$

where $\overline{M}_k$ is the same with that in Theorem 5.

Consider *Chebyshev Inequality*:

$$P\big(|X - EX| \leq \epsilon\big) \geq 1 - \frac{Var(X)}{\epsilon^2}.$$

Let $X = d_k$ and $\epsilon = \frac{C}{2}d^2(x_k)$, we can get:

$$P\Big(X \geq \frac{C}{2}d^2(x_k)\Big) \geq 1 - \frac{4}{C^2}\frac{1}{\overline{M}_k}.$$

In another way, we can rewrite it as:

$$\sum_{i=1}^{M_k} w_k(i)\|\Pi_{\omega_{k,i}}x_k - x_k\|^2 \geq \frac{C}{2}d^2(x_k), \quad \text{w.p. } \Big(1 - \frac{4}{C^2}\frac{1}{\overline{M}_k}\Big), \tag{65}$$

where "w.p." means "with probability at least". Use (65) and (27), we can get:

$$e^2(x_{k+1}) \leq (1-2\alpha_k\sigma)e^2(x_k) - 2\alpha_k h_k'(x_k - x^*) + 2B\alpha_k d(x_k) - \frac{C}{4}d^2(x_k) + 5A^2\alpha_k^2, \quad \text{w.p. } \Big(1 - \frac{4}{C^2}\frac{1}{\overline{M}_k}\Big).$$

Assume $g(x, v) = G(x)$, that is to say $h_k = 0$ for all $k \geq 0$, and it's obviously that

$$2B\alpha_k d(x_k) - \frac{C}{4}d^2(x_k) \leq \Big(\frac{4B}{C}\Big)^2\alpha_k^2,$$

which gives that:

$$e^2(x_{k+1}) \leq (1 - 2\alpha_k\sigma)e^2(x_k) + \Big(5A^2 + \frac{16B^2}{C^2}\Big)\alpha_k^2, \quad \text{w.p. } \Big(1 - \frac{4}{C^2}\frac{1}{\overline{M}_k}\Big). \tag{66}$$

Use (66) directly in Lemma 3, consider the recursion holds for all $0 \leq i \leq k$, we can get (64). ∎

Special case when $\alpha_k = R/(k+1)$ Based on (3) and (64), assume $R > \frac{2}{\sigma}$, we can get following for all $T \geq 0$:

$$e^2(x_k) \leq e^2(x_0)k^{-\sigma R} + \Big(5A^2 + \frac{16B^2}{C^2}\Big)R^2k^{-1}, \forall 0 \leq k \leq T, \quad \text{w.p. } \prod_{k=1}^{T}\Big(1 - \frac{4}{C^2}\frac{1}{\overline{M}_k}\Big). \tag{67}$$

From (67), we can find that the error can be bounded by $O(1/k)$ with a probability. For a fixed T, the probability $\prod_{k=1}^{T}\Big(1 - \frac{4}{C^2}\frac{1}{\overline{M}_k}\Big)$ is approximate to 1 if $\overline{M}_k$ is large. If we just use ICPM, $\overline{M}_k = 1$, the probability $\big(1 - \frac{4}{C^2}\big)^T \to 0$ if $T \to \infty$. If we use ICAPM ($\overline{M}_k = M_k$) and sample constraints with replacement, we can let $M_k = O(T^2)$ to make the probability limits to 1 when $T \to \infty$. If we sample constraints without replacement, we can use finite samples to make the probability limits to 1.

Special case when $\alpha_k = R$ is a constant Based on (3) and (64), we can get:

$$e^2(x_k) \leq e^2(x_0)\exp(-2\sigma Rk) + \Big(5A^2 + \frac{16B^2}{C^2}\Big)R^2, \forall 0 \leq k \leq T, \quad \text{w.p. } \prod_{k=1}^{T}\Big(1 - \frac{4}{C^2}\frac{1}{\overline{M}_k}\Big). \tag{68}$$

For an arbitrary $\epsilon > 0$, let $T_0 = \frac{1}{2\sigma R}\log\big(\frac{e^2(x_0)}{\epsilon}\big)$ in (68), we can get:

$$P\Big(e^2(x_{T_0}) \leq \big(5A^2 + \frac{16B^2}{C^2}\big)R^2 + \epsilon\Big) \geq \prod_{k=1}^{T_0}\Big(1 - \frac{4}{C^2}\frac{1}{\overline{M}_k}\Big),$$

which means that probability of entering an arbitrary bound is largely improved if $\overline{M}_k$ increases.

By the two examples above, we express what is "stable" and why ICPM-SMC (especially ICAPM) can improve the stability of the algorithm.

5.3 Exceeding Length

We bound the error with a probability less than 1, that means it's possible to exceed the bound. Take the case that $\alpha_k = R/(k+1)$ and R is sufficient large as an example (according to (67)). There exists constants C_1 and C_2 so that:

$$e^2(x_k) \leq \frac{C_1}{K+1}, \quad \text{w.p. } \prod_{i=0}^{k}\Big(1 - \frac{C_2}{\overline{M}_k}\Big).$$

What we concern is that at which k the error exceed the bound $e^2(x_k) > \frac{C_1}{k+1}$. We call it "exceeding length". Obviously, longer the length is, more stable the algorithm is.

To simplify our discussion, we give some notations. Define random sequence E_k:

$$E_k = \begin{cases} 1, & \text{if } e^2(x_k) \leq \dfrac{C_1}{k+1}, \\ 0, & \text{if } e^2(x_k) > \dfrac{C_1}{k+1}. \end{cases} \tag{69}$$

Consider the recursion like (66), we write it in the following formulation:

$$e^2(x_{k+1}) \leq h\big(e^2(x_k)\big), \text{w.p. } P_k$$

, where $h(\cdot)$ is a $\Re \to \Re$ map. Define another sequence:

$$\overline{E}_k = \begin{cases} 1, & \text{if } e^2(x_{k+1}) \leq h\big(e^2(x_k)\big), \\ 0, & \text{if } e^2(x_{k+1}) > h\big(e^2(x_k)\big). \end{cases} \tag{70}$$

Obviously, they have the following relationship:

$$P\big(\overline{E}_k = 1\big|\mathscr{F}_k\big) \geq 1 - \frac{4}{C^2}\frac{1}{\overline{M}_k}, \tag{71}$$

and

$$P\big(E_0 = E_1 = \cdots = E_k = 1\big) \geq P\big(\overline{E}_0 = \overline{E}_1 = \ldots \overline{E}_k = 1\big) \tag{72}$$

based on Lemma 40.

Then we define "exceeding length" T_{ex}:

$$T_{\text{ex}} = \min_{k\geq 0}\{k|E_k = 0\}, \tag{73}$$

which is a random variable. The following theorem give an estimation of its expectation $\mathbf{E}[T_{\text{ex}}]$.

Theorem 7 (Expected Exceeding Length) *Suppose $\{x_k\}$ is generated by ICAPM, the assumptions used in Theorem 6 hold, assume we use sampling with replacement and sampling number is constant $M_k = M$, we can get a lower bound estimation of $\mathbf{E}[T_{ex}]$:*

$$\mathbf{E}[T_{ex}] \geq \frac{C^2}{4}M, \tag{74}$$

which means a linear relationship between "exceeding length" and sampling number M.

Proof Consider the distribution of T_{ex},

$$\begin{aligned}
&P\big(T_{\text{ex}} > t\big)\\
=&P\big(E_0 = E_1 = \cdots = E_t = 1\big)\\
\geq&P\big(\overline{E}_0 = \overline{E}_1 = \cdots = \overline{E}_t = 1\big), \quad \text{based on (72)}\\
=&P\big(\overline{E}_0 = 1\big)P\big(\overline{E}_1 = 1\big|\overline{E}_0 = 1\big)\ldots P\big(\overline{E}_t = 1\big|\overline{E}_0 = \cdots = \overline{E}_{t-1} = 1\big),
\end{aligned}$$

where every item can be bounded as:

$$\begin{aligned}
&P\big(\overline{E}_k = 1\big|\overline{E}_0 = \cdots = \overline{E}_{k-1} = 1\big)\\
=&\int_{\mathscr{F}_k} P\big(\overline{E}_k\big|\mathscr{F}_k\big)P\big(\mathscr{F}_k\big|\overline{\mathscr{F}}_k\big)d\Omega, \quad \text{define } \overline{\mathscr{F}}_k = \big\{\Omega\big|\overline{E}_0 = \cdots = \overline{E}_{k-1} = 1\big\}\\
\geq&\Big(1 - \frac{4}{C^2}\frac{1}{\overline{M}_k}\Big)\int_{\mathscr{F}_k} P\big(\mathscr{F}_k\big|\overline{\mathscr{F}}_k\big)d\Omega, \quad \text{based on (71)}\\
=&1 - \frac{4}{C^2}\frac{1}{\overline{M}_k}.
\end{aligned}$$

Thus, use it for $0 \leq k \leq t$, we can get:

$$P\big(T_{\text{ex}} > t\big) \geq \prod_{k=0}^{t}\Big(1 - \frac{4}{C^2}\frac{1}{\overline{M}_k}\Big) \tag{75}$$

Then we can get the lower bound of the expectation:

$$\mathbf{E}[T_{\text{ex}}] = \sum_{t=0}^{\infty} P\big(T_{\text{ex}} > t\big) \geq \sum_{t=0}^{\infty}\Big(\prod_{k=0}^{t}\Big(1 - \frac{4}{C^2}\frac{1}{\overline{M}_k}\Big)\Big),$$

where we let $\overline{M}_k = M_k = M$ is constant, then we can get:

$$\mathbf{E}[T_{\text{ex}}] \geq \sum_{t=0}^{\infty}\Big(1 - \frac{4}{C^2}\frac{1}{M}\Big)^t = \frac{C^2}{4}M.$$

Theorem 7 is proved. ∎

About Conditions in Theorem 7 In this theorem, we assume unbiased weight distribution $w_k(i)$ (7) i.e. ICAPM (8), sampling with replacement and constant sampling scheme. The result shows that ICAPM guarantee the lower bound of "exceeding length" longer, which means more stable property. To give a clear result, we only analyze sampling with replacement and get a $O(M)$ result, that means sampling without replacement behaves even better than this based on the results in Section 5.1 and 5.2, (i.e. gives a more longer "exceeding length"). What's more, if we use an increasing sampling number M_k, we can make the length longer, even infinity in terms of expectation.

6 Conclusion

ICPM-SMC, derived from ICPM algorithm, is a general framework to solve stochastic optimization problem. We can change the parameters in it to get different special algorithms. ICAPM and MDPM are two such algorithms. The conditions to make the algorithm convergent for ICPM-SMC and ICPM are the same. For convergence rate in terms of expectation, they share the same order of rate, while coefficients of the rate are different. MDPM can largely improve the expected convergence rate by using only the "most far" projection. For convergence stability, ICAPM behaves better especially when using "sampling without replacement" and large sample number. Thus, we can control the parameters and sampling scheme in ICPM-SMC to get particular properties we want.

A Using ICPM-SMC in Variational Inequalities Problem

Let's expend SO problem to VI problem. Use the convexity of F, Problem (1) equals to the following problem: find a solution $x^* \in X$ such that,

$$\tilde{\nabla} F(x^*)'(x - x^*) \geq 0$$

holds for all $x \in X$, where $\tilde{\nabla} F(x) \in \partial F(x)$ is a subgradient of F at the point x. To be more general, we consider the *Variational Inequalities* (VI) problem: find a solution $x^* \in X$ such that,

$$G(x^*)'(x - x^*) \geq 0, \quad \text{where } G(x) := \mathbf{E}[g(x, v)] \tag{76}$$

holds for all $x \in X = \cap_{i=1}^m X_i$, where X_i are closed convex sets, and $g(x, v)$ are $\Re^n \to \Re^n$ mappings with random variable v. Obviously, VI problem is more general than SO problem, because we cannot find a $F(x)$ for each $G(x)$ that $G(x) \in \partial F(x), \forall x$, while we can find a $G(x)$ for each $F(x)$ based on the convexity of F.

Similar to Theorem 1, we can get the almost sure convergence of ICPM-SMC for Problem (76).

Corollary 4 (Almost Sure Convergence for VI Problem) *Consider VI problem (76). Suppose that the sequence $\{x_k\}$ is generated by ICPM-SMC (Algorithm 1). Assume $G(x)$ is*

continuous and strictly monotone, other conditions remain the same with SO problem, then we can get the same conclusion: x_k converges almost surely to the unique optimal point for Problem (1).

PROOF Based on strictly monotone assumption, there exists only one optimal point x^*. The basic inequalities still hold. Let's consider (32), the third term can be estimated as:

$$\begin{aligned} G'(x_k)(x_k - x^*) =& G'(\Pi_X x_k)(\Pi_X x_k - x^*) + (G(x_k) - G(\Pi_X x_k))'(\Pi_X x_k - x^*) + G'(x_k)(x_k - \Pi_X x_k) \\ \geq& G'(\Pi_X x_k)(\Pi_X x_k - x^*) - L(d(x_k))e(x_k) - (Le(x_k) + B)d(x_k) \\ \geq& G'(\Pi_X x_k)(\Pi_X x_k - x^*) - 2Ld(x_k)e(x_k) - Bd(x_k), \end{aligned}$$

where the second step is based on continuous assumption on $G(x)$ for Problem 76, which limit $\|G(x) - G(y)\| \leq L(\|x - y\| + 1)$ to $\|G(x) - G(y)\| \leq L\|x - y\|$.

Thus, we can get:

$$\begin{aligned} \mathbf{E}[e^2(x_{k+1})|\mathscr{F}_k] \leq& (1 + 10L^2\alpha_k^2)e^2(x_k) + 10\alpha_k^2(L^2 + B^2) - 2\alpha_k G'(\Pi_X x_k)(\Pi_X x_k - x^*) \\ &+ 4L\alpha_k d(x_k)e(x_k) + 2B\alpha_k d(x_k) - \frac{C}{2}d^2(x_k), \end{aligned}$$

where the last there terms can be estimated as:

$$\begin{aligned} 4L\alpha_k d(x_k)e(x_k) + 2B\alpha_k d(x_k) - \frac{C}{2}d^2(x_k) &= \left(4L\alpha_k d(x_k)e(x_k) - \frac{C}{4}d^2(x_k)\right) + \left(2B\alpha_k d(x_k) - \frac{C}{4}d^2(x_k)\right) \\ &\leq \left(\frac{8L}{C}\right)^2 \alpha_k^2 e^2(x_k) + \left(\frac{4B}{C}\right)^2 \alpha_k^2. \end{aligned}$$

Thus, we can simplify (32) as the following:

$$\mathbf{E}[e^2(x_{k+1})|\mathscr{F}_k] \leq \big(1 + O(\alpha_k^2)\big)e^2(x_k) + O(\alpha_k^2) - 2\alpha_k G'(\Pi_X x_k)(\Pi_X x_k - x^*). \tag{77}$$

Consider (33), let $\epsilon = \frac{C}{4}$, add (77) and (33) together, we can get:

$$\begin{aligned} \mathbf{E}[e^2(x_{k+1}) + d^2(x_{k+1})|\mathscr{F}_k] \leq& \big(1 + O(\alpha_k^2)\big)\big(e^2(x_k) + d^2(x_k)\big) + O(\alpha_k^2) \\ &- 2\alpha_k G'(\Pi_X x_k)(\Pi_X x_k - x^*) - \frac{C}{4}d^2(x_k), \end{aligned} \tag{78}$$

where $\{e^2(x_k)\}$, $\{d^2(x_k)\}$ and $\{G'(\Pi_X x_k)(\Pi_X x_k - x^*)\}$ are all nonnegative sequences, we can use *Supermartingale Convergence Theorem*[3] directly, which is also a lemma to proof "Coupled Supermartingale Convergence Theorem", we can get that: $e^2(x_k) + d^2(x_k)$ is almost surely convergent and the following hold with probability 1:

$$\sum_{k=0}^{\infty} d^2(x_k) < \infty \quad \text{and} \quad \sum_{k=0}^{\infty} 2\alpha_k G'(\Pi_X x_k)(\Pi_X x_k - x^*) < \infty.$$

[3]This theorem is first studied by Robbins and Siegmund in [19] and Dimitri Bertsekas gave this name "Supermartingale Convergence Theorem" in [18].)

Thus, we can get:

$$d^2(x_k) \xrightarrow{a.s.} 0,$$

and based on $\sum_{i=0}^{\infty} \alpha_k < \infty$, we can find a subsequence $\{x_{k_l}\}$ from $\{x_k\}$ that satisfies:

$$G'(\Pi_X x_{k_l})(\Pi_X x_{k_l} - x^*) \xrightarrow{a.s.} 0.$$

Since we have proved that $e^2(x_k) + d^2(x_k)$ is convergent, it's obvious that $\{x_{k_l}\}$ is bounded with probability 1. It has a cluster point $\overline{x}$, (i.e. $\lim_{l\to\infty} x_{k_l} = \overline{x}$). Using the continuity of $d(x)$, $\Pi_X x$ and $G(x)$ then we can get:

$$d^2(\overline{x}) = 0 \quad \text{and} \quad G'(\Pi_X \overline{x})(\Pi_X \overline{x} - x^*) = 0,$$

which is equal to

$$G'(\overline{x})(\overline{x} - x^*) = 0.$$

As we know, $G'(x)(x-x^*) = (G(x)-G(x^*))'(x-x^*)+G'(x^*)(x-x^*)$ for all $x \in \Re^n$, where $(G(x)-G(x^*))'(x-x^*) \geq 0$ and $G'(x^*)(x-x^*) \geq 0$, thus, we have $(G(\overline{x})-G(x^*))'(\overline{x}-x^*) = G'(x^*)(\overline{x} - x^*) = 0$, use the strict monotone assumption of $G(x)$ for Problem 76, we can get $\overline{x} = x^*$. Use (77) lonely, we can get $\|x_k - x^*\|^2$ is almost surely convergent, $\|x_k - \overline{x}\|^2$ is also almost surely convergent and convergent to 0. That is to say,

$$x_k \xrightarrow{a.s.} x^*.$$

Theorem 1 is proved. ∎

Remark Compared to Proposition 1 of [17], we make two extensions: one is expending ICPM to ICPM-SMC, another is weaken the assumption of *strongly monotone* to *strictly monotone*.

References

[1] Le Cun, Leon Bottou Yann, and L. Bottou. "Large scale online learning." *Advances in neural information processing systems* 16 (2004): 217.

[2] O. Bousquet and L. Bottou. "The tradeoffs of large scale learning." *Advances in neural information processing systems.* 2008.

[3] L. Bottou. "Large-scale machine learning with stochastic gradient descent." *Proceedings of COMPSTAT'2010.* Physica-Verlag HD, 2010. 177-186.

[4] N. L. Roux, M. Schmidt, F. Bach. "A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets." *Advances in Neural Information Processing Systems.* 2012: 2663-2671.

[5] R. Johnson, T. Zhang. "Accelerating stochastic gradient descent using predictive variance reduction." *Advances in Neural Information Processing Systems.* 2013: 315-323.

[6] P. Richtrik, M. Tak. "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function." *Mathematical Programming.* 2014, 144(1-2): 1-38.

[7] F. A. Potra, S. J. Wright. "Interior-point methods." *Journal of Computational and Applied Mathematics.* 2000, 124(1): 281-302.

[8] Y. Nesterov, A. Nemirovskii, Y. Ye. "Interior-point polynomial algorithms in convex programming." *Philadelphia: Society for Industrial and Applied Mathematics.* 1994.

[9] D. P. Bertsekas. *Nonlinear programming.* 1999.

[10] C. J. Hsieh, K. W. Chang, C. J. Lin. "A dual coordinate descent method for large-scale linear SVM." *Proceedings of the 25th international conference on Machine learning.* ACM, 2008: 408-415.

[11] Z. Luo, P. Tseng. "On the convergence rate of dual ascent methods for linearly constrained convex minimization." *Mathematics of Operations Research.* 1993, 18(4): 846-867.

[12] S. Boyd. "Alternating direction method of multipliers." *Talk at NIPS Workshop on Optimization and Machine Learning.* 2011.

[13] S. S. Ram, A. Nedic, V. V. Veeravalli. "Incremental stochastic subgradient algorithms for convex optimization." *SIAM Journal on Optimization.* 2009, 20(2): 691-717.

[14] A. Nedic, D. P. Bertsekas. "Incremental subgradient methods for nondifferentiable optimization." *SIAM Journal on Optimization.* 2001, 12(1): 109-138.

[15] A. Nedic, D. P. Bertsekas. "Convergence rate of incremental subgradient algorithms." *Stochastic optimization: algorithms and applications.* Springer US, 2001: 223-264.

[16] M. Wang, D. P. Bertsekas. "Incremental constraint projection-proximal methods for nonsmooth convex optimization." *Technical report,* MIT, 2013.

[17] M. Wang, D. P. Bertsekas. "Incremental constraint projection methods for variational inequalities." *Mathematical Programming,* 2014: 1-43.

[18] D. P. Bertsekas. "Incremental proximal methods for large scale convex optimization." *Mathematical programming,* 2011, 129(2): 163-195.

[19] H. Robbins, D. Siegmund. "A convergence theorem for non negative almost supermartingales and some applications." *Herbert Robbins Selected Papers.* Springer New York, 1985: 111-135.

[20] E. Moulines, F. Bach. "Non-asymptotic analysis of stochastic approximation algorithms for machine learning." *Advances in Neural Information Processing Systems.* 2011: 451-459..

[21] D. Kinderlehrer, G. Stampacchia. *An introduction to variational inequalities and their applications.* Siam, 2000.