

# Convergence of the Randomized Kaczmarz Algorithm in Hilbert Space

Xin Guo

The Hong Kong Polytechnic University

12 December 2016, ICSMAADS2016, Hangzhou

Joint work with Junhong Lin and Ding-Xuan Zhou

# Outline

- The Classical Kaczmarz Algorithms
- Randomized Kaczmarz Algorithm in Hilbert Space
- Motivating Applications
- Convergence Analysis
- Summary and Future Works

# The classical Kaczmarz algorithm

Consider a linear equation system  $Ax = y$ :  $x \in \mathbb{R}^d$ ,  $y \in \mathbb{R}^m$ , and  $A$  is a matrix with dimension  $m \times d$ .

The classical Kaczmarz algorithm:  $x_1 = 0$  and for  $k \geq 1$ ,

$$x_{k+1} = x_k + (y_i - \langle a_i, x_k \rangle) \frac{a_i}{\|a_i\|^2},$$

- $i = 1 + ((k - 1) \bmod m)$
- $a_1, \dots, a_m$ : rows of  $A$ . So  $\langle a_i, x \rangle = y_i$ .
- The algorithm and its convergence: (Kaczmarz 1937).
- step size: 1
- projection of error

$$x_{k+1} - x = \left( I - \frac{a_i}{\|a_i\|} \otimes \frac{a_i}{\|a_i\|} \right) (x_k - x)$$

# Randomized Kaczmarz algorithm

$$x_{k+1} = x_k + (y_k - \langle \varphi_k, x_k \rangle) \frac{\varphi_k}{\|\varphi_k\|^2}$$

- $\{\varphi_k\}_{k=1}^{\infty} \stackrel{iid}{\sim} (\mathbb{R}^d, \rho)$ : random measures
- $y_k = \langle \varphi_k, x \rangle$
- A special example:  $\rho$  supported on the row vectors  $\{a_1, \dots, a_m\}$  of the matrix  $A$ , with  $\rho(a_i) \propto \|a_i\|^2$ .

$$\mathbb{E} [\|x_{k+1} - x\|^2] \leq \left(1 - \frac{1}{\|A\|_F^2 \|A^{-1}\|_2^2}\right)^k \|x\|^2,$$

Strohmer & Vershynin '09; for general  $\rho$ , Zouzias & Freris '13, Gower & Richtárik '15.

- For general  $\rho$ ,  $\lim_{k \rightarrow \infty} C^k \|x_k - x\|^2 = 0$  almost surely for some  $C > 1$  (Chen & Powell '12).

- Recall the definition:

$$x_{k+1} = x_k + (y_k - \langle \varphi_k, x_k \rangle) \frac{\varphi_k}{\|\varphi_k\|^2}.$$

Since  $\frac{y_k}{\|\varphi_k\|} = \left\langle \frac{\varphi_k}{\|\varphi_k\|}, x \right\rangle$ , it is convenient to assume  $\|\varphi_k\| = 1$ .

# Relaxed randomized Kaczmarz algorithm

$$x_{k+1} = x_k + \eta_k (y_k - \langle \varphi_k, x_k \rangle) \varphi_k$$

- $y_k = \langle \varphi_k, x \rangle + \epsilon_k$ ,  $\epsilon_k$ : centered independent noise,  $\|\varphi_k\| = 1$
- $0 < \eta_k < 1$ : necessary to guarantee a convergence
- Needell '10: With step size  $\eta_k \equiv 1$ ,  $\mathbb{E}[\|x_{T+1} - x\|] = O(\|A^{-1}\| \|A\|_F)$  as  $T \rightarrow \infty$ .
- (Bach & Moulines '13):

$$\mathbb{E} \left[ \left\langle a, \frac{1}{k} \sum_{j=2}^{k+1} x_j - x \right\rangle^2 \right] = O\left(\frac{\dim H}{n}\right)$$

- Lin & Zhou '15:  $\lim_{k \rightarrow \infty} \mathbb{E}[\|x_{k+1} - x\|^2] = 0$  if and only if  $\lim_{k \rightarrow \infty} \eta_k = 0$  and  $\sum_{k=1}^{\infty} \eta_k = \infty$ . In this case, one further has

$$\sum_{k=1}^{\infty} \sqrt{\mathbb{E}[\|x_{k+1} - x\|^2]} = \infty.$$

- Lin & Zhou '15:  $\mathbb{E}[\|x_{k+1} - x\|^2] = O(k^{-\theta})$  for  $\eta_k \sim k^{-\theta}$  and  $\theta \in (0, 1)$ .

# Randomized Kaczmarz algorithm in Hilbert space

$$x_{k+1} = x_k + \eta_k (y_k - \langle \varphi_k, x_k \rangle) \varphi_k.$$

- $\{\varphi_k\}_{k=1}^{\infty} \stackrel{iid}{\sim} \rho$  on  $H$ , a general Hilbert space;  $\|\varphi_k\| = 1$ .
- $y_k = \langle \varphi_k, x \rangle$
- It is closely related to the online learning algorithms (online gradient descent algorithm): let  $\{(w_k, y_k)\}_k$  be a sample of input-output pairs. Let  $f_1 = 0$  and

$$f_{k+1} = f_k + \eta_k (y_k - \langle f_k, K_{w_k} \rangle_K) K_{w_k} - \eta_k \lambda f_k.$$

$K$ : a Mercer kernel;  $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K, \|\cdot\|_K)$  is the corresponding reproducing kernel Hilbert space;  $K_w : u \mapsto K(w, u)$ , a function in  $\mathcal{H}_K$ .

**Some Reference:** Cesa-Bianchi & Long & Warmuth '96, Vapnik '98, Kivinen & Smola & Williamson '04, Pontil & Ying & Zhou '05, Smale & Yao '06, Ying & Zhou '06, ...

- It is understood that the regularization parameter  $\lambda$  is not necessary (Ying & Pontil 07).
- Usually the regularity assumption  $f_\rho = L_K^s(\mathcal{H}_K)$ ,  $s > 0$ , is made.

# Application to functional data analysis

- Functional linear models ( $\beta_0$ : unknown slope function)

$$Y^* = \alpha_0 + \int_{\mathcal{T}} X^*(t)\beta_0(t)dt + \epsilon.$$

- $\mathcal{T}$ : a compact domain, e.g., an interval or a square.  $Y^* \in \mathbb{R}$ .
- $X^*(t)$ : a square integrable stochastic process over  $\mathcal{T}$  with covariance

$$C(s, t) = \mathbb{E}[(X(s) - \mathbb{E}[X(s)])(X(t) - \mathbb{E}[X(t)])].$$

- Data:  $\{(X_k, Y_k)\}_k$ , iid copies of  $(X^*, Y^*)$ .
- a large literature [James 2002, Cardot&Ferraty&Sarda 2003, Ramsay&Silverman 2005, Yao&Müller&Wang 2005, Ferraty&Vieu 2006, Cai&Hall 2006, Li&Hsing 2007, Hall&Horowitz 2007, ...]
- Reproducing kernel approaches [Yuan&Cai 2010, Cai&Yuan 2012]

$$(\hat{\alpha}_0, \hat{\beta}_0) = \arg \min_{\alpha \in \mathbb{R}, \beta \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (Y_i - F(X_i))^2 + \lambda J(F) \right\},$$

where  $F(X) = \alpha + \int_{\mathcal{T}} X(t)\beta(t)dt$ , and  $J(F) = \|\beta\|_K^2$ .



- We assume  $\alpha_0 = 0$  for simplicity
- Consider the noise-free case  $\epsilon = 0$
- Apply randomized Kaczmarz algorithm with RKHS

$$x_{k+1} = x_k + \eta_k \left( Y_k^* - \int_{\mathcal{T}} x_k(t) X_k^*(t) dt \right) \frac{a_k^*}{\|a_k^*\|_K^2}.$$

- $a_k^* = \int_{\mathcal{T}} X_k^*(s) K_s ds$
- Write  $\varphi_k = a_k^* / \|a_k^*\|_K$ ,  $y_k = Y_k^* / \|a_k^*\|_K$ . One has

$$x_{k+1} = x_k + \eta_k (y_k - \langle x_k, \varphi_k \rangle) \varphi_k.$$

## The assumption $x \in L^s(H)$ , $s > 0$

- $L := \mathbb{E}[\varphi_k \otimes \varphi_k]$  where for any  $u \in H$ ,  $(\varphi_k \otimes \varphi_k)u := \langle \varphi_k, u \rangle \varphi_k$ .
- $L$  is symmetric, positive semi-definite, Hilbert-Schmidt, and of trace class.
- For online learning algorithms,  $\varphi = K_X / \|K_X\|_K$  with  $X$  having some unknown distribution, and

$$L = \mathbb{E} \frac{K_X \otimes K_X}{\|K_X\|^2} = \mathbb{E} \frac{K_X \otimes K_X}{K(X, X)}.$$

When  $K_X / \|K_X\|_K$  and  $\|K_X\|_K$  are independent, for example when  $K(x, x) \equiv 1$ ,  $L$  is the integral operator  $L_K$  that is widely used in literature, and  $x \in L^s_K(\mathcal{H}_K)$  is a widely used assumption.

- It is an interesting open problem that under which (all?) circumstance would one get a better kernel simply by normalization.
- In functional linear regression problems,

$$L = \mathbb{E} \frac{a^* \otimes a^*}{\|a^*\|_K^2} = \mathbb{E} \frac{\int_{\mathcal{J}} \int_{\mathcal{J}} X^*(t) X^*(r) (K_t \otimes K_r) dt dr}{\int_{\mathcal{J}} \int_{\mathcal{J}} X^*(t) X^*(r) K(t, r) dt dr}.$$

when  $\|a^*\|_K$  and  $a^* / \|a^*\|_K$  are independent,  $L$  is the operator  $L_{K^{1/2} C K^{1/2}}$  studied in literature (Yuan-Cai 2010, Cai-Yuan 2012), where  $C(t, r) = \mathbb{E}[X^*(t) X^*(r)]$ .

Consider the noise-free model  $y_k = \langle \varphi_k, x \rangle$  with  $\{\varphi_k\}_{k=1}^{\infty} \stackrel{iid}{\sim} \rho$  and  $\|\varphi_k\| = 1$ . Let  $L = \mathbb{E}[\varphi_k \otimes \varphi_k]$ .

## Theorem (rate of weak convergence, G-Lin-Zhou)

Let  $\eta_k = \eta \in (0, 1]$ . For any  $x \in L^{s_1}(H)$  and  $u \in L^{s_2}(H)$  with some  $s_1, s_2 \in [0, 1/4]$ , we have

$$\mathbb{E}[\langle u, x_{k+1} - x \rangle^2] \leq \|L^{-s_2}u\|^2 \|L^{-s_1}x\|^2 (\eta k)^{-2s_1-2s_2};$$

If instead of a fixed  $u$ ,  $x_{k+1} - x$  is measured by an independent random vector  $\alpha \sim (H, \rho)$ , one has

$$\mathbb{E}[\langle \alpha, x_{k+1} - x \rangle^2] \leq \text{Tr}(L) \|L^{-s_1}x\|^2 (\eta k)^{-2s_1-\frac{1}{2}}.$$

- Remark: when  $H = \mathcal{H}_K$  and  $\alpha = K_X$  with  $X$  being random,

$$\mathbb{E}[\langle \alpha, x_{k+1} - x \rangle^2] = \mathbb{E}[\|x_{k+1} - x\|_{L_2}^2].$$

- Example: suppose  $\rho$  is concentrated on only one point  $\varphi^*$ , then  $L = \varphi^* \otimes \varphi^*$ , and  $\varphi_k = \pm \varphi^*$  for all  $k$ . So the randomized Kaczmarz algorithm converges if and only if  $x = C\varphi^*$  for some constant  $C$ . Even if  $x \perp \varphi^*$ ,  $\langle u, x_{k+1} - x \rangle$  still converges if  $u = C'\varphi^*$ .

*Outline of the proof.*

$$x_{k+1} = x_k + (y_k - \langle \varphi_k, x_k \rangle) \varphi_k,$$

$$\begin{aligned} x_{k+1} - x &= x_k - x + \langle \varphi_k, x - x_k \rangle \varphi_k \\ &= (I - \varphi_k \otimes \varphi_k)(x_k - x). \end{aligned}$$

Let  $P_k = I - \varphi_k \otimes \varphi_k$ , then  $P_1, \dots, P_k, \dots$  is a sequence of i.i.d. random orthogonal projections, and

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x\|^2] &= \mathbb{E}[(x_k - x)^T P_k^2 (x_k - x)] \\ &= \mathbb{E}[(x_{k-1} - x)^T P_{k-1} P_k^2 P_{k-1} (x_{k-1} - x)] \\ &= \dots \dots \dots \\ &= \mathbb{E}[(0 - x)^T P_1 \dots P_k^2 \dots P_1 (0 - x)]. \end{aligned}$$

Strong convergence is not expected.

## Example

Let  $\{e_i\}_{i=1}^{\infty}$  be an orthonormal basis of  $H$ . Let  $q_1 \geq q_2 \geq \dots > 0$  and  $\sum_{i=1}^{\infty} q_i = 1$ . Let  $\rho$  be a discrete probability distribution such that  $\rho(e_i) = q_i$ . Assume  $\text{Var}(\epsilon) = \sigma^2 > 0$  the model  $y = \langle \varphi, x \rangle + \epsilon$  and set  $\eta_k \equiv \eta \in (0, 1]$  for Algorithm

$$x_{k+1} = x_k + \eta_k (y_k - \langle \varphi_k, x_k \rangle) \varphi_k.$$

Then  $\lim_{k \rightarrow \infty} \mathbb{E}[\|x_{k+1} - x\|^2] = \infty$ . Note that if  $x \in L^{s_1}(H)$  and  $u \in L^{s_2}(H)$  with some  $s_1, s_2 \in [0, \frac{1}{4}]$ , we still have

$$\mathbb{E}[\langle u, x_{k+1} - x \rangle^2] = \mathcal{O}((\eta k)^{-2s_1 - 2s_2} + \sigma^2 \eta^{1 - 2s_2}).$$

Recall [Needell 2010]: In the finite dimensional case with step sizes  $\eta_k \equiv 1$ ,

$$\mathbb{E}[\|x_{k+1} - x\|] \sim \|A^{-1}\| \|A\|_F \text{ as } k \rightarrow \infty.$$

# Regularization by relaxation

If we allow that  $\eta_k$  be small enough, we can obtain strong convergence in  $H$ .

## Theorem (G-Lin-Zhou)

Let  $M$  be a positive integer or infinity. Suppose the step sizes satisfy  $\sum_{j=1}^M \eta_j^2 = \eta_{2,M} < 1$  and the slope vector satisfies  $x \in L^{s_1}(H)$ . Then for  $1 \leq k \leq M$ , one has

$$\mathbb{E}[\|x_{k+1} - x\|^2] \leq \frac{2\|x\|^2 \|L^{-s_1}x\|^2 s_1^{2s_1}}{\|L^{-s_1}x\|^2 s_1^{2s_1} + \|x\|^2 e^{2s_1} \left(\sum_{j=1}^k \eta_j\right)^{2s_1}} + \frac{\|x\|^2 + \sigma^2 \text{Tr}(L)}{1 - \eta_{2,M}} \sum_{j=1}^k \eta_j^2.$$

## Theorem (G & Lin & Zhou)

Suppose  $x_\rho \in L^{s_1}(H)$  for some  $s_1 > 0$ , and  $\text{Tr}(L^{s_*}) < \infty$  for some  $0 < s_* < 1$ . Let  $2 \leq T < \infty$  be the number of iterations with constant step size

$$\eta_k \equiv \eta = \begin{cases} T^{-(2s_1+s_*)/(1+2s_1+s_*)}, & s_1 + s_* \geq 1, \\ T^{-(1+s_1)/(2+s_1)}, & s_1 + s_* < 1. \end{cases}$$

Then

$$\mathbb{E}[\|x_{T+1} - x_\rho\|^2] \leq C_1 \begin{cases} T^{-2s_1/(1+2s_1+s_*)}, & s_1 + s_* \geq 1, \\ T^{-2s_1/(2+s_1)}, & s_1 + s_* < 1, \end{cases}$$

where  $C_1$  is independent of  $T$ .

Note: the error estimate in this theorem is much better than the bound given in Ying & Pontil '07, and is arbitrarily close to the minimax optimal convergence rate given by G & Fan & Zhou '16.

# Summary, and Future Work

- The classical randomized Kaczmarz algorithm is generalized to Hilbert space inspired by the “regularity” assumption from learning theory.
- Polynomial convergence is obtained regardless of the minimum eigenvalue of the coefficient matrix, for classical Kaczmarz algorithm;
- Future works: minimax rates; a better connection and understanding of the applications in functional data; mini-batch; averaging; Kaczmarz in Banach space; ...
- Xin GUO, x.guo@polyu.edu.hk

THANK YOU!