# On the Convergence of Randomized Kaczmarz Algorithm in Hilbert Space

Xin Guo

The Hong Kong Polytechnic University

July 07, 2016, Oberwolfach

Joint work with Junhong Lin and Ding-Xuan Zhou

# Outline

- The Classical Kaczmarz Algorithms
- Randomized Kaczmarz Algorithm in Hilbert Space
- Motivating Applications
- Strong Convergence and Divergence with Noisy Data
- 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 approximates the solution iteratively by

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

- $i = 1 + ((k-1) \mod m)$
- $a_1, \ldots, a_m$: rows of $A$.
- The algorithm and its convergence: [Kaczmarz 1937].
- step size: 1

# Randomized Kaczmarz algorithm

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

- $\{\varphi_k\}_{k=1}^{\infty} \overset{iid}{\sim} (\mathbb{R}^d, \rho)$: 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 $\rho(a_i) \propto \|a_i\|^2$.

$$\mathbb{E}\left[\|x_{k+1} - x\|^2\right] \leq (1 - \frac{1}{\|A\|_F^2 \|A^{-1}\|_2^2})^k \|x_1 - x\|^2.$$

[Strohmer&Vershynin 2009]

- When the system $Ax = y$ is overdetermined, [Zouzias-Freris 2013]:

$$\mathbb{E}\left[\|x_{k+1} - x_{\mathsf{LS}}\|^2\right] \leq \left(1 - \frac{1}{\|A\|_F^2 \|A^\dagger\|^2}\right)^k \|x_{\mathsf{LS}}\|^2.$$

- For general $\rho$, $\lim_{k\to\infty} C^k \|x_k - x\|^2 = 0$ almost surely for some $C > 1$ (Chen-Powell 2012).
- 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

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

- $\{\varphi_k\}_{k=1}^{\infty} \sim (S^{d-1}, \rho)$ (Recall: sample: $\{\varphi_k, y_k\}$, so we assume $\varphi_k \in S^{d-1}$ simply for the sake of simplicity in analysis, and one may normalize the data before carrying out the algorithm!)
- $y_k = \langle \varphi_k, x \rangle + \epsilon_k$, $\epsilon_k$: centered independent noise
- $0 < \eta_k < 1$: necessary to guarantee a convergence
- [Needell 2010]: With step size $\eta_k \equiv 1$, $\mathbb{E}[\|x_{T+1} - x\|] \sim \|A^{-1}\|\|A\|_F$ as $T \to \infty$.

- [Lin-Zhou 2015]: $\lim_{k\to\infty} \mathbb{E}[\|x_{k+1} - x\|^2] = 0$ if and only if $\lim_{k\to\infty} \eta_k = 0$ 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 2015]: $\mathbb{E}[\|x_{k+1} - x\|^2] = O(k^{-\theta})$ for $\eta_k \sim k^{-\theta}$ and $\theta \in (0, 1)$.

# Randomized Kaczmarz algorithm in Hilbert space

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

- $\{\varphi_k\}_{k=1}^{\infty} \overset{iid}{\sim} \rho$ on $H$, a general Hilbert space; $\|\varphi_k\| = 1$.
- $y_k = \langle \varphi_k, x \rangle$

- It is closely related to online learning algorithm (online gradient descent algorithm)

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

$K$: a Mercer kernel; $(\mathcal{H}_K, \langle \cdot, \cdot \rangle)$ is the corresponding reproducing kernel Hilbert space; $K_x : u \mapsto K(x, u)$, a function in $\mathcal{H}_K$.
[Cesa-Bianchi&Long&Warmuth 1996, Vapnik 2998, Kivinen&Smola&Williamson 2004, Pontil&Ying&Zhou 2005, Smale&Yao 2006, Ying&Zhou 2006, ...]

- It is understood that the regularization parameter $\lambda$ is not necessary (Ying&Pontil 2007).
- Usually the regularity assumption $f_\rho = L_K^s(\mathcal{H}_K)$, $s > 0$, is made.

# Application to functional data analysis

- Functional linear model ($\beta_0$: 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

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

- 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 $x \in L_K^s(\mathcal{H}_K)$ is a widely used assumption.

In functional linear regression problems,

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

when $\|a^*\|_K$ and $a^*/\|a^*\|_K$ are independent, $L$ is the operator $L_{K^{1/2}CK^{1/2}}$ 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 $\{\varphi_k\}_{k=1}^{\infty} \overset{iid}{\sim} \rho$ 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*

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

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.*

Let $\varphi \sim (H, \rho)$, so $\|\varphi\| = 1$. Let $P = \varphi \otimes \varphi$ be an orthogonal projection. we have $\mathbb{E}[P] = L$. Define $Q_\eta, S_L$, and $T$: $\mathrm{HS}(H) \to \mathrm{HS}(H)$ by

$$
\begin{aligned}
Q_\eta(A) &= \mathbb{E}[(1 - \eta\varphi \otimes \varphi)A(1 - \eta\varphi \otimes \varphi)], \\
S_L(A) &= \frac{1}{2}LA + \frac{1}{2}AL, \\
T(A) &= \mathbb{E}[PAP].
\end{aligned}
$$

They are all symmetric. $T$ is of trace class. We have $Q_\eta = I - 2\eta S_L + \eta^2 T$.

For any $A \in \mathrm{HS}(H)$,

$$\langle A, T(A)\rangle_{\mathrm{HS}} = \mathbb{E}\mathrm{Tr}(A^T PAP) = \mathbb{E}\mathrm{Tr}(PA^T PPAP),$$
$$\mathrm{Tr}(PA^T PPAP) \leq \begin{cases} \mathrm{Tr}(A^T PPA) \\ \mathrm{Tr}(PA^T AP) \end{cases}.$$

Therefore $T \leq S_L$, in the sense that $\langle A, T(A)\rangle_{\mathrm{HS}} \leq \langle A, S_L(A)\rangle_{\mathrm{HS}}$ for any $A \in \mathrm{HS}(H)$.
On the other hand, we write $\{(\lambda_i, \phi_i)\}_i$ as the normalized eigensystem of $L$, then

the eigensystem of $R_L$ is $\{(\sqrt{\lambda_i \lambda_j}, \phi_i \otimes \phi_j)\}_{i,j}$, and
the eigensystem of $S_L$ is $\{((\lambda_i + \lambda_j)/2, \phi_i \otimes \phi_j)\}_{i,j}$.

Therefore $R_L \leq S_L$. We have

$$\eta R_L \leq (2\eta - \eta^2) R_L \leq (2\eta - \eta^2) S_L + \eta^2 (S_L - T) = I - Q_\eta.$$

Now set $x_1 = 0$.

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

we have

$$\begin{aligned} \mathbb{E}[(x_{k+1} - x) \otimes (x_{k+1} - x)] &= \mathbb{E}[Q_\eta((x_k - x) \otimes (x_k - x))] \\ &= \cdots = Q_\eta^k (x \otimes x), \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}[\langle u, x_{k+1} - x \rangle^2] &= \mathbb{E}\, \langle u, (x_{k+1} - x) \otimes (x_{k+1} - x) u \rangle \\ &= \left\langle u, Q_\eta^k (x \otimes x) u \right\rangle = \mathrm{Tr}((u \otimes u) Q_\eta^k (x \otimes x)). \end{aligned}$$

By the Löwner-Heinz inequality, for any $s \in [0, 1/4]$, $(\eta R_L)^{4s} \leq (I - Q_\eta)^{4s}$. So $\|(I - Q_\eta)^{-2s} R_L^{2s}(A)\|_{\mathrm{HS}} \leq \eta^{-2s} \|A\|_{\mathrm{HS}}$ for any $A \in \mathrm{HS}(H)$.
On the other hand, $x \in L^s(H)$ implies that

$$\|R_L^{-2s}(x \otimes x)\|_{\mathrm{HS}} = \|L^{-s} x\|^2.$$

Therefore

$$\begin{aligned}
& \mathrm{Tr}((u \otimes u) Q_\eta^K (x \otimes x)) \\
\leq\ & \|(I - Q_\eta)^{-2s_2} R_L^{2s_2} R_L^{-2s_2}(u \otimes u)\|_{\mathrm{HS}} \\
& \times \|(I - Q_\eta)^{-2s_1} R_L^{2s_1} R_L^{-2s_1}(x \otimes x)\|_{\mathrm{HS}} \\
& \times \|Q_\eta^k (1 - Q_\eta)^{2s_1 + 2s_2}\|_{\mathrm{op}} \\
\leq\ & \eta^{-2s_2} \|L^{-s_2} u\|^2 \eta^{-2s_1} \|L^{-s_1} x\|^2 \sup_{0 \leq \lambda \leq 1} \lambda^k (1 - \lambda)^{2s_1 + 2s_2} \\
\leq\ & \|L^{-s_2} u\|^2 \|L^{-s_1} x\|^2 (\eta k)^{-2s_1 - 2s_2}.
\end{aligned}$$

No strong convergence, ... But this seems OK.

## Example

Let $\{e_i\}_{i=1}^{\infty}$ be an orthonormal basis of $H$. Let $q_1 \geq q_2 \geq \ldots > 0$ and $\sum_{i=1}^{\infty} q_i = 1$. Let $\rho$ be a discrete probability distribution such that $\rho(e_i) = q_i$. Assume $\mathrm{Var}(\epsilon) = \sigma^2 > 0$ 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 $\lim_{k\to\infty} \mathbb{E}[\|x_{k+1} - x\|^2] = \infty$. 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 finite dimension case with step size $\eta_k \equiv 1$, $\mathbb{E}[\|x_{k+1} - x\|] \sim \|A^{-1}\| \|A\|_F$ as $k \to \infty$.

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*

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

## Corollary

*Suppose $x \in L^{s_1}(H)$ for some $s_1 > 0$. Let $M < \infty$ be the number of iterations with constant step size $\eta_k \equiv M^{-(2s_1+1)/(2s_1+2)}$. Then*

$$\mathbb{E}[\|x_{k+1} - x\|^2] \leq C_1 M^{-\frac{s_1}{s_1+1}},$$

*where $C_1$ is independent of $s_1$ or $M$.*

Remark: this rate is consistent with the result in [Ying&Pontil 2007].

# Some related results

- [Bach&Moulines 2013]:

$$\mathbb{E}\left[\langle a, x_{k+1} - x\rangle^2\right] = \mathcal{O}\left(\frac{d}{n}\right),$$

  where $d = \dim H$.

- [Gower&Richtárik 2015]:

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

- Some detailed and more careful comparison coming soon.

# 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;
- Soon coming: Simulations;
- 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!