Convergence of the Randomized Kaczmarz Algorithm in Hilbert Space

Xin Guo

The Hong Kong Polytechnic University

27 May 2017, Fudan University, Shanghai
International Conference on Computational Harmonic Analysis 2017

Joint work with Junhong Lin and Ding-Xuan Zhou

Supported in part by Research Grants Council of Hong Kong

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 $k \geq 1$,

$$\begin{aligned} x_{k+1} &= x_k + (y_i - \langle a_i, x_k \rangle) \frac{a_i}{\|a_i\|^2} \\ &= x_k + c_k^* a_i, \\ c_k^* &= \arg\min_{c \in \mathbb{R}} \left(\langle x_k + c a_i, a_i \rangle - y_i \right)^2 . \end{aligned}$$

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

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

Randomized Kaczmarz algorithm

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

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

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

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

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

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

- $y_k = \langle \varphi_k, x \rangle + \epsilon_k$, ϵ_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 $T \to \infty$.
- (Bach & Moulines '13):

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

- Lin & Zhou '15: $\lim_{k\to\infty} \mathbb{E}[\|x_{k+1} - x\|^2] = 0$ 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 '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

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

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

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

  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, ... Reproducing property: $\langle f, K_w \rangle_K = f(w)$.
- It is understood that the regularization parameter λ 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 (β_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, $\varphi = K_X/\|K_X\|_K$, 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.
- 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,

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

  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 $\eta_k \equiv \eta \in (0, 1]$. 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

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

- 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 ρ is concentrated on only one point φ^*, 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_η, S_L, T, and R_L: $\mathrm{HS}(H) \to \mathrm{HS}(H)$ by

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

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

Now set $x_1 = 0$.

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

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

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x\|^2] =& \mathbb{E}[(x_k - x)^T (I - \eta P_k)^2 (x_k - x)] \\ =& \cdots\cdots\cdots \\ =& \mathbb{E}[(0 - x)^T (I - \eta P_1) \cdots (I - \eta P_k)^2 \cdots (I - \eta P_1)(0 - x)] \\ =& \mathrm{Tr}\mathbb{E}[(I - \eta P_1) \cdots (I - \eta P_k)^2 \cdots (I - \eta P_1)(x \otimes x)] \\ =& \mathrm{Tr}[Q_\eta^k (x \otimes x)] \\ =& \mathrm{Tr}[Q_\eta^k (I - Q_\eta)^{2s} (I - Q_\eta)^{-2s} (x \otimes x)] \end{aligned}$$

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

$$\langle A, T(A)\rangle_{\mathrm{HS}} = \mathbb{E}\mathrm{Tr}(A^T PAP) = \mathbb{E}\mathrm{Tr}(PA^T PPAP),$$

$$\mathrm{Tr}(PA^T PPAP) \le \begin{cases} \mathrm{Tr}(A^T PPA) \\ \mathrm{Tr}(PA^T AP) \end{cases}.$$

Therefore $T \le S_L$, in the sense that $\langle A, T(A)\rangle_{\mathrm{HS}} \le \langle A, S_L(A)\rangle_{\mathrm{HS}}$ 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 $R_L \le S_L$. We have

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

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

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

Consider the covergence in weak sense,

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

Therefore

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

□

The Loewner-Heinz Inequality

Theorem (Loewner-Heinz)

Let A and B be two positive semi-definite (PSD) operators such that $A \leq B$ (i.e. $B - A$ is PSD). Then for any $0 < r \leq 1$, $A^r \leq B^r$.

- see text books: Bhatia 1997, Horn & Johnson 1985, etc.
- proof: $A \leq B \Rightarrow (sI + A)^{-1} \geq (sI + B)^{-1}$ for $s > 0$,
  $\Rightarrow A(sI + A)^{-1} \leq B(B + sI)^{-1}$.

$$t^r = \frac{\sin \pi r}{\pi} \int_0^\infty \frac{s^{r-1} t}{s + t} ds, \quad \forall t > 0, 0 < r < 1.$$

- in linear algebra, it is well known that $A \leq B \Rightarrow \mathrm{Im}(A) \subset \mathrm{Im}(B)$
- $0 \leq A \leq B \Rightarrow \|B^{-r/2} A^{r/2}\|_{\mathsf{op}} \leq 1$
- $\forall r > 1/2$, $\sup\{\|B^{-r} A^r\|_{\mathsf{op}} : A, B \in \mathbb{R}^{2\times 2}, 0 \leq A \leq B\} = \infty$ (constructive proof done with the help of Chen-Di Wang)

Strong convergence for the noiseless case is still an interesting open problem. However, we have the following example.

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 ρ be a discrete probability distribution such that $\rho(e_i) = q_i$. Assume $\mathrm{Var}(\epsilon) = \sigma^2 > 0$ in the model $y_j = \langle \psi_j, x \rangle + \epsilon_j$ for the algorithm

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

Then $\lim_{k\to\infty} \mathbb{E}[\|x_{k+1} - x\|^2] = \infty$, and this limit is independent of x (in particular, it holds true even for $x = 0$). Meanwhile, if $x \in L^{s_1}(H)$ with $s_1 \geq 0$ and $u \in L^{s_2}(H)$ with $s_2 > 0$, one has that $\limsup_{k\to\infty} \mathbb{E}[\langle u, x_{k+1} - x \rangle^2] = O(\eta)$ as $\eta \to 0^+$.

Recall [Needell 2010]: In the finite dimensional case with step sizes $\eta = 1$, $\mathbb{E}[\|x_{k+1} - x\|] \sim \|A^{-1}\|\|A\|_F$ as $k \to \infty$.

Regularization by relaxation

If we allow that η_k be small enough, we can obtain strong convergence in H.

Theorem (G & Lin & Zhou)

Let T be a positive integer or infinity. Suppose the step sizes satisfy $\sum_{j=1}^{T} \eta_j^2 = \eta_{2,T} < 1$ and the slope vector satisfies $x \in L^{s_1}(H)$ for some $s_1 > 0$. Then for $1 \leq k \leq T$, one has

$$\mathbb{E}[\|x_{k+1} - x\|^2] \leq \frac{4(s_1^{s_1}+1)^2 \|L^{-s_1} x\|^2}{1 + \left(\sum_{j=1}^{k} \eta_j\right)^{2s_1}} + \frac{\|x\|^2 + \sigma^2 \mathrm{Tr}(L)}{1 - \eta_{2,T}} \sum_{j=1}^{k} \eta_j^2.$$

Theorem (G & Lin & Zhou)

Suppose $\mathrm{Tr}(L^{s_*}) < \infty$ *for some* $0 < s_* \leq 1$, *and that the slope vector satisfies* $x \in L^{s_1}(H)$ *for some* $s_1 > 0$. *Define the step sizes*

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

in the k*'th iteration for* $1 \leq k \leq T$. *Then*

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

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.

About the finite-trace assumption $\mathrm{Tr}(L^{s_*}) < \infty$, $0 < s_* \leq 1$

- trivial when $s_* = 1$ since $L = \mathbb{E}[\phi \otimes \phi]$ is of trace class
- usually in literature the effective dimension $\mathcal{N}(\lambda) = \mathrm{Tr}(L(L + \lambda I)^{-1})$ is used to characterize the complexity of the hypothesis space $\mathcal{H}_K$ (Caponnetto & De Vito, 2007; Blanchard & Krämer, 2010).

Theorem

Let L be a positive semi-definite operator of trace class, with $\mathcal{N}(\lambda) := \mathrm{Tr}(L(L + \lambda I)^{-1})$. Let $0 < s < 1$. We have

$$\mathrm{Tr}(L^s) = \frac{\sin \pi s}{\pi} \int_0^\infty t^{s-1} \mathcal{N}(t) dt,$$

therefore,

- *If $\mathrm{Tr}(L^s) < \infty$, then $\mathcal{N}(\lambda) = O(\lambda^{-s})$ as $\lambda \to 0^+$;*
- *If $\mathcal{N}(\lambda) = O(\lambda^{-s})$, then for any $\epsilon > 0$, $\mathrm{Tr}(L^{s+\epsilon}) < \infty$.*

Theorem (G & Lin & Zhou)

Suppose $x \in L^{s_1}(H)$ *and* $u \in L^{s_2}(H)$ *with* $s_1 \geq 0$, $s_2 \geq 0$, *and* $s_1 + s_2 > 0$. *Set constant step size* $\eta_k \equiv T^{-\omega}$ *for* $k = 1, \ldots, T$, *where*

$$\omega = \begin{cases} \frac{1+4s_1}{3+4s_1}, & 0 \leq s_2 < 1/4, \\ \frac{2s_1+2s_2}{1+2s_1+2s_2}, & s_2 \geq 1/4. \end{cases}$$

Then

$$\mathbb{E}[\langle u, x_{T+1} - x \rangle^2] \leq C_3 \begin{cases} T^{-(4s_1+4s_2)/(3+4s_1)}, & 0 \leq s_2 < 1/4, \\ T^{-(1+4s_1)/(3+4s_1)} \log(T+1), & s_2 = 1/4, \\ T^{-(2s_1+2s_2)/(1+2s_1+2s_2)}, & s_2 > 1/4, \end{cases}$$

where C_3 *is a constant independent of* T. *If we replace* u *by the random vector* $\psi \sim (H, \rho)$, *assume* $\mathrm{Tr}(L^{2-4s_0}) < \infty$ *for some* $s_0 \in [1/4, 1/2)$, *and set* $\eta_k \equiv T^{-(2s_1+2s_0)/(1+2s_1+2s_0)}$, *then*

$$\mathbb{E}[\langle \psi, x_{T+1} - x \rangle^2] \leq C_3'(s_0) \begin{cases} T^{-(1+4s_1)/(3+4s_1)} \log(T+1), & s_0 = 1/4, \\ T^{-(2s_1+2s_0)/(1+2s_1+2s_0)}, & s_0 \in (1/4, 1/2), \end{cases}$$

where $C_3'(s_0)$ *is a constant independent of* T.

Theorem (G & Lin & Zhou)

Suppose $x \in L^{s_1}(H)$ *and* $u \in L^{s_2}(H)$ *with* $s_1 \geq 0$, $s_2 \geq 1/4$, *and* $s_1 + s_2 > 1/2$. *Let* $\eta_k = \eta_1 k^{-\omega}$ *with* $\omega = (2s_1 + 2s_2)/(1 + 2s_1 + 2s_2)$ *for* $k \geq 2$. *Then for any* $k \geq 1$, *we have*

$$\mathbb{E}[\langle u, x_{k+1} - x\rangle^2] \leq C_5 \begin{cases} (k+1)^{-\omega}\log(k+1), & s_2 = 1/4, \\ (k+1)^{-\omega}, & s_2 > 1/4, \end{cases}$$

where C_5 *is a constant independent of* k. *If we replace* u *by the random vector* $\psi \sim (H, \rho)$, *assume* $\mathrm{Tr}(L^{2-4s_0}) < \infty$ *for some* $s_0 \in [1/4, 1/2)$, *and set* $\eta_k = \eta_1 k^{-\omega}$ *with* $\omega = (2s_1 + 2s_0)/(1 + 2s_1 + 2s_0)$, *then for* $k \geq 1$,

$$\mathbb{E}[\langle \psi, x_{k+1} - x\rangle^2] \leq C_5' \begin{cases} (k+1)^{-\omega}\log(k+1), & s_0 = 1/4, \\ (k+1)^{-\omega}, & 1/4 < s_0 < 1/2, \end{cases}$$

where C_5' *is a constant independent of* k.

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; Polyak-Ruppert averaging; Kaczmarz in Banach space; ...
- Xin GUO, `x.guo@polyu.edu.hk`

Thank you!