RandomShuffle Beats SGD after Finite Epochs

Jeff HaoChen
Tsinghua University

Suvrit Sra
Massachusetts Institute of Technology
Introduction

• Goal: to minimize the function

\[ F(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \]
Introduction

- SGD with replacement: (often appears in algorithm analysis)
  
  - $x_k = x_{k-1} - \gamma \nabla f_{s(k)}(x_{k-1})$
  
  - $s(k)$ uniformly random from $[n]$, $1 \leq k \leq T$

- SGD without replacement: (often appears in reality)
  
  - $x^t_k = x^t_{k-1} - \gamma \nabla f_{\sigma_t(k)}(x^t_{k-1})$
  
  - $\sigma_t$ uniformly from random permutation of $[n]$, $1 \leq k \leq n$
Introduction

• SGD with replacement: (often appears in algorithm analysis)
  
  \[ x_k = x_{k-1} - \gamma \nabla f_{s(k)}(x_{k-1}) \]
  
  \( s(k) \) uniformly random from \([n], 1 \leq k \leq T \)

• SGD without replacement: (often appears in reality)
  
  \[ x^t_k = x^t_{k-1} - \gamma \nabla f_{\sigma_t(k)}(x^t_{k-1}) \]
  
  \( \sigma_t \) uniformly from random permutation of \([n], 1 \leq k \leq n \)
Introduction

• SGD with replacement: (often appears in algorithm analysis)
  
  • $x_k = x_{k-1} - \gamma \nabla f_{s(k)}(x_{k-1})$
  
  • $s(k)$ uniformly random from $[n]$, $1 \leq k \leq T$

• SGD without replacement: (often appears in reality)
  
  • $x^t_k = x^t_{k-1} - \gamma \nabla f_{\sigma_t(k)}(x^t_{k-1})$
  
  • $\sigma_t$ uniformly from random permutation of $[n]$, $1 \leq k \leq n$
Introduction

• So a natural question: *which one is better?*

• A Numerical Comparison: *(Bottou, 2009)*
Introduction

• So a natural question: *which one is better?*

• A Numerical Comparison: (*Bottou, 2009*)
Introduction

• Why?

• Intuitively, we should prefer RandomShuffle for the following two reasons:
  • It uses more “information” in one epoch (by visiting each component)
  • It has smaller variance for one epoch

• However, what is a rigorous proof?
Introduction

• Why?

• Intuitively, we should prefer RandomShuffle for the following two reasons:
  • It uses more “information” in one epoch (by visiting each component)
  • It has smaller variance for one epoch

• However, what is a rigorous proof?
Introduction

• Why?

• Intuitively, we should prefer RandomShuffle for the following two reasons:
  • It uses more “information” in one epoch (by visiting each component)
  • It has smaller variance for one epoch

• However, what is a rigorous proof?
A Brief History

• Under **strong structure**, we can convert this problem into matrix inequality: (Recht and Ré, 2012)

\[
\begin{align*}
\text{Assume the problem is quadratic: } f_i(x) &= (a_i^T x - y_i)^2 \\
\text{Then “RandomShuffle is better than SGD after one epoch” is true under conjecture:}
\end{align*}
\]

\[
\begin{bmatrix}
\|
\begin{bmatrix}
E_{wo} \\
\prod_{j=1}^{k} A_{i_{k-j+1}} \\
\prod_{j=1}^{k} A_{i_j}
\end{bmatrix}
\| \\
\end{bmatrix}
\leq
\begin{bmatrix}
\|
\begin{bmatrix}
\prod_{j=1}^{k} A_{i_{k-j+1}} \\
\prod_{j=1}^{k} A_{i_j}
\end{bmatrix}
\|
\end{bmatrix}
\]

• Which we still don’t know how to prove yet 😞
A Brief History

• Under **strong structure**, we can convert this problem into matrix inequality: (Recht and Ré, 2012)

\[ \text{Under strong structure, we can convert this problem into matrix inequality:} \]

• Assume the problem is quadratic: \( f_i(x) = (a_i^T x - y_i)^2 \)

• Then “RandomShuffle is better than SGD after one epoch” is true under conjecture:

\[ \begin{array}{c}
\mathbb{E}_{w_0} \left[ \prod_{j=1}^{k} A_{i_{k-j+1}} \prod_{j=1}^{k} A_{i_{j}} \right] \leq \mathbb{E}_{w_r} \left[ \prod_{j=1}^{k} A_{i_{k-j+1}} \prod_{j=1}^{k} A_{i_{j}} \right]
\end{array} \]

• Which we still don’t know how to prove yet 😞
A Brief History

- Under **strong structure**, we can convert this problem into matrix inequality: (Recht and Ré, 2012)

- Assume the problem is quadratic: $f_i(x) = (a_i^T x - y_i)^2$

- Then “RandomShuffle is better than SGD after one epoch” is true under conjecture:

\[
\| E_{\text{wo}} \left[ \prod_{j=1}^{k} A_{i_{k-j+1}} \prod_{j=1}^{k} A_{i_{j}} \right] \| \leq E_{\text{wr}} \left[ \prod_{j=1}^{k} A_{i_{k-j+1}} \prod_{j=1}^{k} A_{i_{j}} \right] \]

- Which we still don’t know how to prove yet 😞
A Brief History

• Under **strong structure**, we can convert this problem into matrix inequality: (Recht and Ré, 2012)

\[
\left(\begin{array}{c}
\text{Recht} \\
\text{and Ré, 2012}
\end{array}\right)
\]

• Assume the problem is quadratic: 

\[
f_i(x) = (a_i^T x - y_i)^2
\]

• Then “RandomShuffle is better than SGD after one epoch” is true under conjecture:

\[
\begin{aligned}
\mathbb{E}_{\text{wo}} \left[ \prod_{j=1}^{k} A_{i_{k-j+1}} \prod_{j=1}^{k} A_{ij} \right] &\leq \mathbb{E}_{\text{wr}} \left[ \prod_{j=1}^{k} A_{i_{k-j+1}} \prod_{j=1}^{k} A_{ij} \right]
\end{aligned}
\]

• Which we still don’t know how to prove yet 😞
A Brief History

• What about the more general situation?

• We can try to show with a better convergence bound!
  • The hope is: prove a faster worst-case convergence rate of RandomShuffle

• A well-known fact: SGD converges with rate $O \left( \frac{1}{T} \right)$:
  
  • $\mathbb{E}[\| x_T - x^* \|^2] \leq O \left( \frac{1}{T} \right)$
A Brief History

• What about the more general situation?

• We can try to show with a better convergence bound!
  
  • The hope is: prove a faster worst-case convergence rate of RandomShuffle

• A well-known fact: SGD converges with rate $O \left( \frac{1}{T} \right)$:

  • $\mathbb{E}[\| x_T - x^* \|^2] \leq O \left( \frac{1}{T} \right)$
A Brief History

• What about the more general situation?

• We can try to show with a better convergence bound!
  • The hope is: prove a faster worst-case convergence rate of RandomShuffle

• A well-known fact: SGD converges with rate $O \left( \frac{1}{T} \right)$:
  • $\mathbb{E}[\| x_T - x^* \|^2] \leq O \left( \frac{1}{T} \right)$
A Brief History

• One of the recent breakthrough: (Gürbüzbalaban, 2015)
  
  • Asymptotically RandomShuffle has convergence rate $O\left(\frac{1}{T^2}\right)$
  
  • But not sure what happen after finite epochs

• In contrast, there is a non-asymptotic result: (Shamir, 2016)
  
  • RandomShuffle is no worse than SGD, with provably $O\left(\frac{1}{T}\right)$ convergence rate
  
  • But cannot show that RandomShuffle is really faster
A Brief History

• One of the recent breakthrough: (Gürbüzbalaban, 2015)
  - Asymptotically RandomShuffle has convergence rate $O \left( \frac{1}{T^2} \right)$
  - But not sure what happen after finite epochs

• In contrast, there is a non-asymptotic result: (Shamir, 2016)
  - RandomShuffle is no worse than SGD, with provably $O \left( \frac{1}{T} \right)$ convergence rate
  - But cannot show that RandomShuffle is really faster
A Brief History

• One of the recent breakthrough: (Gürbüzbalaban, 2015)
  
  • **Asymptotically** RandomShuffle has convergence rate $O\left(\frac{1}{T^2}\right)$
  
  • But not sure what happen after finite epochs

• In contrast, there is a non-asymptotic result: (Shamir, 2016)
  
  • RandomShuffle is **no worse** than SGD, with provably $O\left(\frac{1}{T}\right)$ convergence rate
  
  • But cannot show that RandomShuffle is really faster

What happens in between?
Summary of results

We analyze RandomShuffle in the following settings:

- Strongly convex, Lipschitz Hessian
- Sparse data
- Vanishing variance
- Nonconvex, under PL condition
- Smooth convex
Summary of results

We analyze RandomShuffle in the following settings:

- Strongly convex, **Lipschitz Hessian**
- Sparse data
- Vanishing variance
- Nonconvex, under PL condition
- Smooth convex

Dheeraj Nagaraj et al. get rid of this constraint
Summary of results

We analyze RandomShuffle in the following settings:

- Strongly convex, Lipschitz Hessian
- Sparse data
- Vanishing variance
- Nonconvex, under PL condition
- Smooth convex

this talk
Summary of results

We analyze RandomShuffle in the following settings:

• Strongly convex, Lipschitz Hessian

• Sparse data

• Vanishing variance

• Nonconvex, under PL condition

• Smooth convex
First attempt: try to prove a tighter bound!

• Can we show a non-asymptotic bound better than \( O \left( \frac{1}{T} \right) \)? E.g., \( O \left( \frac{1}{T^{1+\delta}} \right) \)?

• If we can, then everything is solved 😊

• ......unless we cannot 😞

**Theorem 3.** Given the information of \( \mu, L, G \). Under the assumption of constant step sizes, no step size choice for RANDOMSHUFFLE leads to a convergence rate \( o \left( \frac{1}{T} \right) \) for any \( T \geq n \), if we do not allow \( n \) to appear in the bound.
First attempt: try to prove a tighter bound!

• Can we show a non-asymptotic bound better than $O \left( \frac{1}{T} \right)$? E.g., $O \left( \frac{1}{T^{1+\delta}} \right)$?

• If we can, then everything is solved 😊

• ......unless we cannot 😞

**Theorem 3.** Given the information of $\mu, L, G$. Under the assumption of constant step sizes, no step size choice for RANDOMSHUFFLE leads to a convergence rate $o \left( \frac{1}{T} \right)$ for any $T \geq n$, if we do not allow $n$ to appear in the bound.
First attempt: try to prove a tighter bound!

- Can we show a non-asymptotic bound better than $O\left(\frac{1}{T}\right)$? E.g., $O\left(\frac{1}{T^{1+\delta}}\right)$?

- If we can, then everything is solved 😊

- ......unless we cannot 😞

**Theorem 3.** Given the information of $\mu$, $L$, $G$. Under the assumption of constant step sizes, no step size choice for RANDOMSHUFFLE leads to a convergence rate $o\left(\frac{1}{T}\right)$ for any $T \geq n$, if we do not allow $n$ to appear in the bound.
Proof of the theorem

**Theorem 3.** Given the information of $\mu, L, G$. Under the assumption of constant step sizes, no step size choice for RANOMSHUFFLE leads to a convergence rate $o\left(\frac{1}{T}\right)$ for any $T \geq n$, if we do not allow $n$ to appear in the bound.

- We only consider the case when $T = n$, i.e., we run one epoch of the algorithm.

- We prove the theorem with a counter-example:
  - Recall function $F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$
  - We set $f_i(x) = \begin{cases} 
  \frac{1}{2} (x - b)' A(x - b), & i \text{ odd,} \\
  \frac{1}{2} (x + b)' A(x + b), & i \text{ even.}
  \end{cases}$
  - $A$ and $b$ to be determined later...
Proof of the theorem

Theorem 3. Given the information of $\mu$, $L$, $G$. Under the assumption of constant step sizes, no step size choice for RANDOMSHUFFLE leads to a convergence rate $o\left(\frac{1}{T}\right)$ for any $T \geq n$, if we do not allow $n$ to appear in the bound.

• We only consider the case when $T = n$, i.e., we run one epoch of the algorithm

• We prove the theorem with a counter-example:

  • Recall function $F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$

  • We set $f_i(x) = \begin{cases} \frac{1}{2} (x - b)'A(x - b), & i \text{ odd}, \\ \frac{1}{2} (x + b)'A(x + b), & i \text{ even}. \end{cases}$

  • $A$ and $b$ to be determined later...
Proof of the theorem

• Step 1: Calculate the error

\[
\mathbb{E}\left[||x_T - x^*||^2\right] = \left|(I - \gamma A)^T(x_0 - x^*)\right|^2 + \mathbb{E}\left[||\sum_{t=1}^{T}(-1)^{\sigma(t)}\gamma(I - \gamma A)^{T-t}Ab||^2\right]
\]

\(P\)

\(Q\)

• Step 2: Simplify via eigenvector basis decomposition

\[P = \sum_{i=1}^{d} (1 - \gamma \lambda_i)^2 p_i^2, \quad Q = \gamma^2 \sum_{i=1}^{d} q_i^2 \lambda_i^2 \mathbb{E}\left[\sum_{t=1}^{T}(-1)^{\sigma(t)}(1 - \gamma \lambda_i)^{T-t}\right]^2\]

• Step 3: Construct a contradiction

• For contradiction, assume there is \(\gamma\) dependent on \(T\) achieving convergence \(o\left(\frac{1}{T}\right)\)

\[
\frac{\gamma T}{2 - \gamma \lambda_i} = \frac{1}{\lambda_i} + o(1)
\]
Proof of the theorem

• Step 1: Calculate the error

$$\mathbb{E} \left[ \| x_T - x^* \|^2 \right] = \left\| (I - \gamma A)^T (x_0 - x^*) \right\|^2 + \mathbb{E} \left[ \left\| \sum_{t=1}^T (-1)^{\sigma(t)} \gamma (I - \gamma A)^{T-t} Ab \right\|^2 \right]$$

• Step 2: Simplify via eigenvector basis decomposition

$$P = \sum_{i=1}^d (1 - \gamma \lambda_i)^2 p_i^2, \quad Q = \gamma^2 \sum_{i=1}^d q_i^2 \lambda_i^2 \mathbb{E} \left[ \left( \sum_{t=1}^T (-1)^{\sigma(t)} (1 - \gamma \lambda_i)^{T-t} \right)^2 \right]$$

• Step 3: Construct a contradiction

For contradiction, assume there is $\gamma$ dependent on $T$ achieving convergence $o \left( \frac{1}{T} \right)$

$$\frac{\gamma T}{2 - \gamma \lambda_i} = \frac{1}{\lambda_i} + o(1)$$
Proof of the theorem

• Step 1: Calculate the error

\[
\begin{align*}
\mathbb{E} \left[ \left| \left| x_T - x^* \right| \right|^2 \right] &= \left| \left| (I - \gamma A)^T (x_0 - x^*) \right| \right|^2 + \mathbb{E} \left[ \left| \sum_{t=1}^{T} (-1)^{\sigma(t)} \gamma (I - \gamma A)^{T-t} A b \right| \right]^2
\end{align*}
\]

\[
\begin{align*}
\text{P} \quad \text{Q}
\end{align*}
\]

• Step 2: Simplify via eigenvector basis decomposition

\[
\begin{align*}
P &= \sum_{i=1}^{d} (1 - \gamma \lambda_i)^{2T} p_i^2, \quad Q = \gamma^2 \sum_{i=1}^{d} q_i^2 \lambda_i^2 \mathbb{E} \left[ \sum_{t=1}^{T} (-1)^{\sigma(t)} (1 - \gamma \lambda_i)^{T-t} \right]^2
\end{align*}
\]

• Step 3: Construct a contradiction

• For contradiction, assume there is \( \gamma \) dependent on \( T \) achieving convergence \( o \left( \frac{1}{T} \right) \)

\[
\frac{\gamma T}{2 - \gamma \lambda_i} = \frac{1}{\lambda_i} + o(1)
\]
Proof of the theorem

• Step 1: Calculate the error

\[ \mathbb{E} \left[ \| x_T - x^* \|^2 \right] = \left\| (I - \gamma A)^T (x_0 - x^*) \right\|^2 + \mathbb{E} \left[ \left\| \sum_{t=1}^{T} (-1)^{\sigma(t)} \gamma (I - \gamma A)^{T-t} A b \right\|^2 \right] \]

• Step 2: Simplify via eigenvector basis decomposition

\[ P = \sum_{i=1}^{d} (1 - \gamma \lambda_i)^2 p_i^2, \quad Q = \gamma^2 \sum_{i=1}^{d} q_i^2 \lambda_i^2 \mathbb{E} \left[ \sum_{t=1}^{T} (-1)^{\sigma(t)} (1 - \gamma \lambda_i)^{T-t} \right]^2 \]

• Step 3: Construct a contradiction

• For contradiction, assume there is \( \gamma \) dependent on \( T \) achieving convergence \( o \left( \frac{1}{T} \right) \)

\[ \frac{\gamma T}{2 - \gamma \lambda_i} = \frac{1}{\lambda_i} + o(1) \]
Proof of the theorem

• Step 1: Calculate the error

\[
\begin{align*}
\mathbb{E} \left[ \left| x_T - x^* \right|^2 \right] &= \left\| (I - \gamma A)^T (x_0 - x^*) \right\|^2 + \mathbb{E} \left[ \left\| \sum_{t=1}^{T} (-1)^{\sigma(t)} \gamma (I - \gamma A)^{T-t} A b \right\|^2 \right] \\
&= P + Q
\end{align*}
\]

• Step 2: Simplify via eigenvector basis decomposition

\[
\begin{align*}
P &= \sum_{i=1}^{d} (1 - \gamma \lambda_i)^2 p_i^2, \\
Q &= \gamma^2 \sum_{i=1}^{d} q_i^2 \lambda_i^2 \mathbb{E} \left[ \sum_{t=1}^{T} (-1)^{\sigma(t)} (1 - \gamma \lambda_i)^{T-t} \right]^2
\end{align*}
\]

• Step 3: Construct a contradiction

• For contradiction, assume there is \( \gamma \) dependent on \( T \) achieving convergence \( o \left( \frac{1}{T} \right) \)

\[
\frac{\gamma T}{2 - \gamma \lambda_i} = \frac{1}{\lambda_i} + o(1)
\]

\( \implies \) Cannot be true for different \( \lambda_i \)!
What to do next?

**Theorem 3.** Given the information of $\mu, L, G$. Under the assumption of constant step sizes, no step size choice for `RandomShuffle` leads to a convergence rate $o \left( \frac{1}{T} \right)$ for any $T \geq n$, if we do not allow $n$ to appear in the bound.

• This means the best non-asymptotic rate we can hope is $O \left( \frac{1}{T} \right)$

  Short Time: $O \left( \frac{1}{T} \right)$  
  Long Time: $O \left( \frac{1}{T^2} \right)$

  What happens in between?

• Key step: introduce $n$ into the bound

  • The hope is if we can show bound like $O \left( \frac{n}{T^2} \right)$, RandomShuffle behaves better 😊
What to do next?

**Theorem 3.** Given the information of $\mu$, $L$, $G$. Under the assumption of constant step sizes, no step size choice for RANDOMSHUFFLE leads to a convergence rate $o\left(\frac{1}{T}\right)$ for any $T \geq n$, if we do not allow $n$ to appear in the bound.

- This means the best non-asymptotic rate we can hope is $O\left(\frac{1}{T}\right)$

  Short Time: $O\left(\frac{1}{T}\right)$  
  Long Time: $O\left(\frac{1}{T^2}\right)$

  What happens in between?

- Key step: introduce $n$ into the bound

  - The hope is if we can show bound like $O\left(\frac{n}{T^2}\right)$, RandomShuffle behaves better 😊
Bounds dependent on $n$

For general second order differentiable functions with Lipschitz Hessian:

**Theorem 2.** Define constant $C = \max \left\{ \frac{32}{\mu^2} (L_H L_D + 3L_H G), 12(1 + \frac{L}{\mu}) \right\}$. So long as $\frac{T}{\log T} > Cn$, with step size $\eta = \frac{8\log T}{T^\mu}$, RANDOMSHUFFLE achieves convergence rate:

$$\mathbb{E}[\|x_T - x^*\|^2] \leq O\left(\frac{1}{T^2} + \frac{n^3}{T^3}\right).$$
Bounds dependent on $n$

• On one hand, RandomShuffle converges with

$$O\left(\frac{1}{T^2} + \frac{n^3}{T^3}\right)$$

• On the other hand, SGD converges with

$$O\left(\frac{1}{T}\right)$$

• So the take away is:

RandomShuffle is provably better than SGD after $O(\sqrt{n})$ epochs!
Bounds dependent on $n$

• On one hand, RandomShuffle converges with

$$O \left( \frac{1}{T^2} + \frac{n^3}{T^3} \right)$$

• On the other hand, SGD converges with

$$O \left( \frac{1}{T} \right)$$

• So the take away is:

RandomShuffle is provably better than SGD after $O(\sqrt{n})$ epochs!
Summary of results

We analyze RandomShuffle in the following settings:

• Strongly convex, Lipschitz Hessian

• Sparse data

• Vanishing variance

• Nonconvex, under PL condition

• Smooth convex
Sparse setting

• A sparse problem can be written as:

\[ F(x) = \sum_{i=1}^{n} f_i(x_{e_i}) \]

• Where each \( e_i \) is a subset of all the dimensions \([d]\)

• Consider a graph with \( n \) nodes, with edge \((i, j)\) if \( e_i \cap e_j \neq \emptyset \)

• Define the sparsity level of the problem:

\[ \rho = \frac{\max_{1 \leq i \leq n} |\{e_j : e_i \cap e_j \neq \emptyset\}|}{n} \]
Sparse setting

• A sparse problem can be written as:

\[ F(x) = \sum_{i=1}^{n} f_i(x_{e_i}) \]

• Where each \( e_i \) is a subset of all the dimensions \([d]\)

• Consider a graph with \( n \) nodes, with edge \((i, j)\) if \( e_i \cap e_j \neq \emptyset \)

• Define the sparsity level of the problem:

\[ \rho = \frac{\max_{1 \leq i \leq n} |\{e_j : e_i \cap e_j \neq \emptyset\}|}{n} \]
Sparse setting

• A fact about sparsity:

\[
\frac{1}{n} \leq \rho \leq 1
\]

• We have the following improved bound for sparse problem:

**Theorem 4.** Define constant \( C = \max \left\{ \frac{32}{\mu^2} (L_H LD + 3L_H G), 12(1 + \frac{L}{\mu}) \right\} \). So long as \( \frac{T}{\log T} > C' n \), with step size \( \eta = \frac{8 \log T}{T \mu} \), RANDOMSHUFFLE achieves convergence rate:

\[
\mathbb{E}[\|x_T - x^*\|^2] \leq O\left(\frac{1}{T^2} + \frac{\rho^2 n^3}{T^3}\right).
\]

• As a corollary, when \( \rho = O\left(\frac{1}{n}\right) \), there is a \( O\left(\frac{1}{T^2}\right) \) convergence rate!
Sparse setting

• A fact about sparsity:

\[
\frac{1}{n} \leq \rho \leq 1
\]

• We have the following improved bound for sparse problem:

**Theorem 4.** Define constant \( C = \max \left\{ \frac{32}{\mu^2} (L_H LD + 3L_H G), 12(1 + \frac{L}{\mu}) \right\} \). So long as \( \frac{T}{\log T} > C'n \), with step size \( \eta = \frac{8\log T}{T\mu} \), RANDOMSHUFFLE achieves convergence rate:

\[
\mathbb{E}[\|x_T - x^*\|^2] \leq O\left(\frac{1}{T^2} + \frac{\rho^2 n^3}{T^3}\right).
\]

• As a corollary, when \( \rho = O\left(\frac{1}{n}\right) \), there is a \( O\left(\frac{1}{T^2}\right) \) convergence rate!
Sparse setting

• A fact about sparsity:

\[ \frac{1}{n} \leq \rho \leq 1 \]

• We have the following improved bound for sparse problem:

**Theorem 4.** Define constant \( C = \max \left\{ \frac{32}{\mu^2}(L_H LD + 3L_H G), 12(1 + \frac{L}{\mu}) \right\} \). So long as \( \frac{T}{\log T} > Cn \), with step size \( \eta = \frac{8\log T}{T\mu} \), RANDOMSHUFFLE achieves convergence rate:

\[ \mathbb{E}[\|x_T - x^*\|^2] \leq O\left(\frac{1}{T^2} + \frac{\rho^2 n^3}{T^3}\right). \]

• As a corollary, when \( \rho = O\left(\frac{1}{n}\right) \), there is a \( O\left(\frac{1}{T^2}\right) \) convergence rate!
Summary of results

We analyze RandomShuffle in the following settings:

• Strongly convex, Lipschitz Hessian
• Sparse data
• Vanishing variance
• Nonconvex, under PL condition
• Smooth convex
When Variance Vanishes

• When the variance vanishes at the optimality

\[ f_i(x^*) = 0, \quad \forall i \]

• Given \( n \) pairs of numbers \( 0 \leq \mu_i \leq L_i \), a optimal solution \( x^* \in \mathbb{R}^d \) and an initial upper bound on distance \( R \)

• A valid problem is defined as \( n \) functions and an initial point \( x_0 \) such that:
  
  • \( f_i \) is \( \mu_i \)-strongly convex, \( L_i \)-Lipschitz continuous
  
  • \( f_i'(x^*) = 0 \)
  
  • \( \| x_0 - x^* \|_2 \leq R \)
When Variance Vanishes

• When the variance vanishes at the optimality
  \[ f_i(x^*) = 0, \quad \forall \ i \]

• Given \( n \) pairs of numbers \( 0 \leq \mu_i \leq L_i \), a optimal solution \( x^* \in \mathbb{R}^d \) and an initial upper bound on distance \( R \)

• A valid problem is defined as \( n \) functions and an initial point \( x_0 \) such that:
  • \( f_i \) is \( \mu_i \)-strongly convex, \( L_i \)-Lipschitz continuous
  • \( f_i'(x^*) = 0 \)
  • \( \| x_0 - x^* \|_2 \leq R \)
When Variance Vanishes

• When the variance vanishes at the optimality
  \[ f_i(x^*) = 0, \quad \forall i \]

• Given \( n \) pairs of numbers \( 0 \leq \mu_i \leq L_i \), a optimal solution \( x^* \in \mathbb{R}^d \) and an initial upper bound on distance \( R \)

• A valid problem is defined as \( n \) functions and an initial point \( x_0 \) such that:
  
  • \( f_i \) is \( \mu_i \)-strongly convex, \( L_i \)-Lipschitz continuous
  
  • \( f_i'(x^*) = 0 \)
  
  • \( \| x_0 - x^* \|_2 \leq R \)
Theorem 5. Given constants \((\mu_1, L_1), \ldots, (\mu_n, L_n)\) such that \(0 \leq \mu_i \leq L_i\), a dimension \(d\), a point \(x^* \in \mathbb{R}^d\) and an upper bound of initial distance \(\|x_0 - x^*\|_2 \leq R\). Let \(\mathcal{P}\) be the set of valid problems. For step size \(\gamma \leq \min_i \left\{ \frac{2}{L_i + \mu_i} \right\}\) and any \(T \geq 1\), there is

\[
\max_{P \in \mathcal{P}} \mathbb{E} \left[ \|X_{RS} - x^*\|^2 \right] \leq \max_{P \in \mathcal{P}} \mathbb{E} \left[ \|X_{SGD} - x^*\|^2 \right].
\]
When Variance Vanishes

**Theorem 5.** Given constants $(\mu_1, L_1), \ldots, (\mu_n, L_n)$ such that $0 \leq \mu_i \leq L_i$, a dimension $d$, a point $x^* \in \mathbb{R}^d$ and an upper bound of initial distance $\|x_0 - x^*\|_2 \leq R$. Let $\mathcal{P}$ be the set of valid problems. For step size $\gamma \leq \min_i \left\{ \frac{2}{L_i + \mu_i} \right\}$ and any $T \geq 1$, there is

$$\max_{P \in \mathcal{P}} \mathbb{E} \left[ \|X_{RS} - x^*\|^2 \right] \leq \max_{P \in \mathcal{P}} \mathbb{E} \left[ \|X_{SGD} - x^*\|^2 \right].$$

RandomShuffle is provably better than SGD after **ANY** number of iterations!
Thanks!