Efficient Ordering of Stochastic Gradient Descent

Mid-Term Presentation

March 3, 2023

Team Members

1. EE19B002: Aditya Sharma
2. EE19B022: Gautham Govind A
3. EE19B038: Manikandan Sritharan
4. EE19B142: Akshat Bhandari

Table of Contents

Table of Contents

Up Next...

Gradient Descent: Background

- Gradient Descent is a first-order iterative optimization algorithm for computing local minima of differentiable functions.
- The algorithm and its many variants are used widely in deep learning architectures, thus making them relevant in today's era of very deep networks.

Gradient Descent: Update equation

- According to the algorithm, for a multivariate differentiable function $F(\mathbf{x})$, a local minima can be computed by initializing a vector $\mathbf{a}$ (of appropriate dimension) randomly and iteratively performing the update:

$$\mathbf{a} \longleftarrow \mathbf{a} - \alpha \nabla F(\mathbf{a})$$

provided α, termed the learning rate, is sufficiently small for each update.

Gradient Descent: Illustration

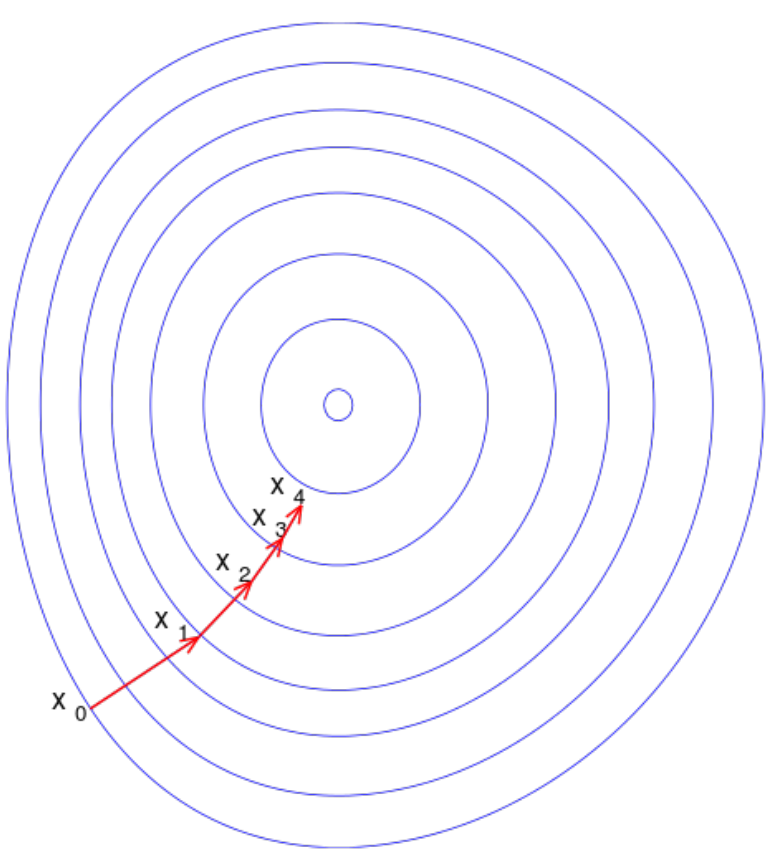


Figure: Illustration of Gradient Descent

Stochastic Gradient Descent: Background

- In many practical applications, the optimization problem has the form:

$$\theta^* = \arg\min_{\theta \in \mathbf{\Theta}} \{ \sum_{i=1}^{n} F(\theta, \mathbf{x^{(i)}}) \}$$

  where $\mathbf{x^{(i)}}$ are samples drawn from the training distribution.
- Though you would ideally like to use gradient descent to optimize, due to several practical constraints, you may not have access to all the samples simultaneously.

Stochastic Gradient Descent: Update equation

- A widely used approach in such a scenario is Stochastic Gradient Descent (SGD), where you solve the problem iteratively by making use of the update equation:

$$\theta_{t+1} = \text{Proj}_{\Theta}(\theta t - \gamma_{t+1} \nabla_{\theta} F(\theta_t, X_{t+1}))$$

  where γ_t is the step size and $\{X_t\}$ is some sequence taking values in $[n]$.
- An interesting avenue for exploration is finding the input sequence $\{X_t\}$ which can provide the best "efficiency".

Markov Chains: Key concepts

- A Markov chain is a stochastic model characterised by the fact that probability of an event depends only on the state attained in the previous event.
- The state vector describes the probability of being in each state and the transition matrix gives the transition probabilities.
- For a certain class of Markov chains, termed Ergodic chains, it turns out that the state vector eventually converges to a unique value termed the steady-state distribution, usually denoted by π.

Markov Chains: Illustration

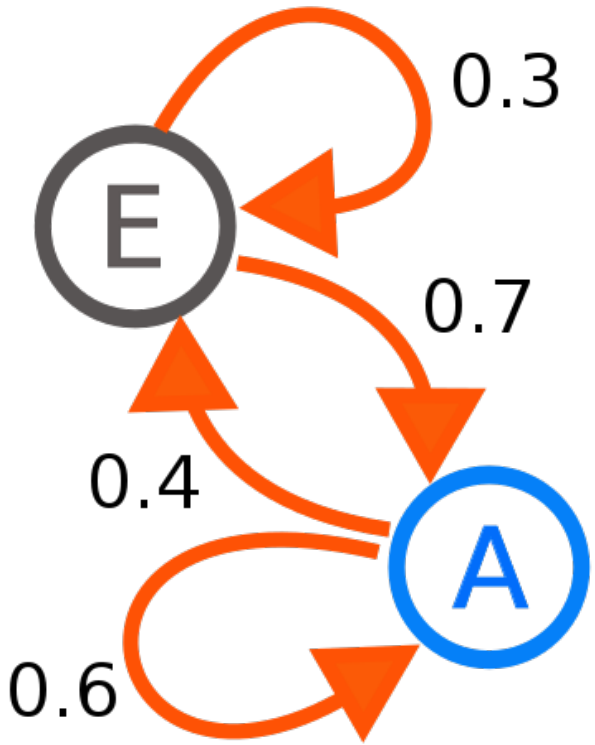


Figure: Example of a Markov Chain

RWSGD: Background

- In one particular version of SGD, termed Random Walk SGD (RWSGD), we decide the input sequence $\{X_t\}_{t\geq 0}$ through a random walk on a Markov chain.
- The key benefit of this setting is that each sample/node can update the parameters using local gradient and then pass the updated parameters instead of divulging the gradient itself.
- This is of particular relevance in privacy critical applications such as medical data analysis.

RWSGD: Analysis

- Existing analysis of RWSGD in literature focuses primarily on **Finite-time bounds** which provide upper bounds for either $\mathbb{E}[f(\tilde{\theta}_t) - f(\theta^*)]$ or $\mathbb{E}[||\nabla f(\theta_t)||_2^2]$.
- The bounds are of the form:

$$\mathbb{E}[||\nabla f(\theta_t)||_2^2] \leq \mathcal{O}\left(\frac{\max\{M, 1/\log(1/\beta)\}}{t^{1-\alpha}}\right)$$

where M depends on local gradients at the minimizer and β is the Second Largest Eigenvalue Modulus (SLEM) of the transition matrix.

Up Next...

Finite-time bounds: Too loose?

- As we have seen in the previous slide, the finite time bounds depend on the value of β or SLEM of the transition matrix.
- The general intuition over the years has been that Markov chains with smaller SLEM lead to faster convergence of the SGD iterations.

Finite-time bounds: Too loose?

- **Claim**: Markov chains with smaller SLEM lead to faster convergence of the SGD iteration.
- **Experiment**: Simulating the RWSGD algorithm with the MHRW, a modification of MHRW (more efficient) and the so called 'fastest mixing Markov chain' (FMMC) as the Markov chain obtained by minimizing the SLEM over the entire class of reversible chains.

Finite-time bounds: Too loose?

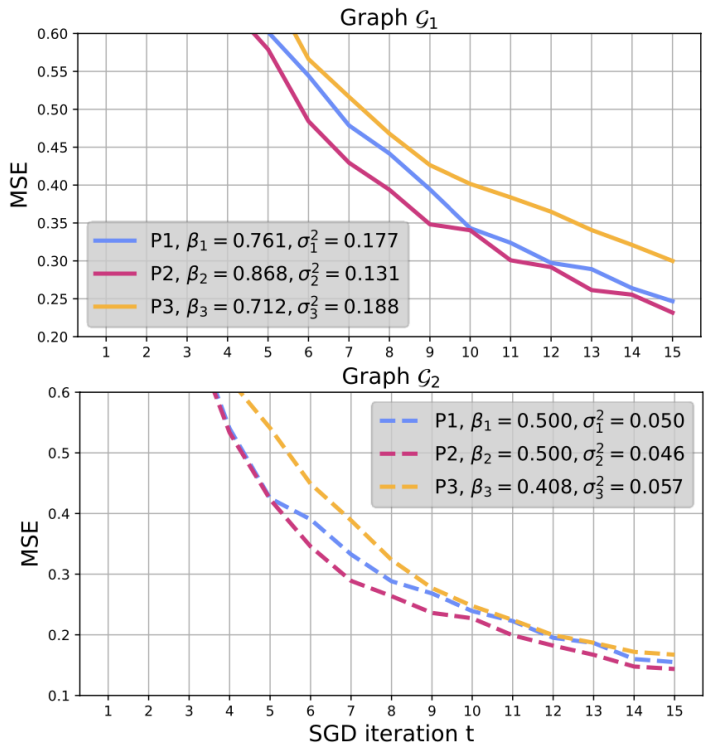


Figure: Comparison of MHRW (P1), Modified MHRW (P2) and FMMC (P3) as inputs for RWSGD on two different graphs

Finite-time bounds: Too loose?

- **Results**:
  1) Even though FMMC is guaranteed to have the least SLEM, it is the worst performing one with largest MSE (mean squared error).

  2) Although MHRW and Modified-MHRW share the same SLEM in the lower plot, they still have performance differences.

  3) Contradictory to the claim, and could be attributed to the finite time results providing upper bounds for all times $t \geq 0$, which may therefore not necessarily be tight.

Asymptotic time bounds

- In addition to mixing time, another widely used metric for characterizing the second order properties of Markov chains is the asymptotic variance (AV).
- For any scalar valued function $g : [n] \to \mathbb{R}$, the estimator $\hat{\mu}_t(g) \triangleq \frac{1}{t}\sum_{i=1}^{t} g(X_i)$, associated with an irreducible Markov chain $\{X_t\}_{t\geq 0}$ with stationary distribution π, is the average of the samples of $g(\cdot)$ obtained along the chain's sample path up to time $t > 0$.
- The **AV of the Markov Chain**, is denoted by $\sigma^2_X(g)$, is then defined as the the limiting variance of the estimator; that is,

$$\sigma^2_X(g) \triangleq \lim_{t\to\infty} t.Var(\hat{\mu}_t(g))$$

Asymptotic time bounds

- For all functions $g(\cdot)$ satisfying $\mathbb{E}_\pi(g^2) < \infty$, the AV is associated with the Central Limit Theorem (CLT) for any Markovian kernel on a finite state space, as the variance of the normally distributed estimates in the limit. More formally, we have

$$\sqrt{t}[\hat{\mu}_t(g) - \mathbb{E}_\pi(g)] \xrightarrow[t\to\infty]{dist} \mathcal{N}(0, \sigma^2_X(g))$$

AV or SLEM?

- In the context of AV, we reconsider the convergence rates of different RWSGD variants.
- It can be seen that while SLEM cannot explain the ordering, a **lower AV implies better convergence rate**.
- It is therefore worthwhile to see if any formal relationship can be established between AV and efficient ordering of input sequences for RWSGD.

AV or SLEM?

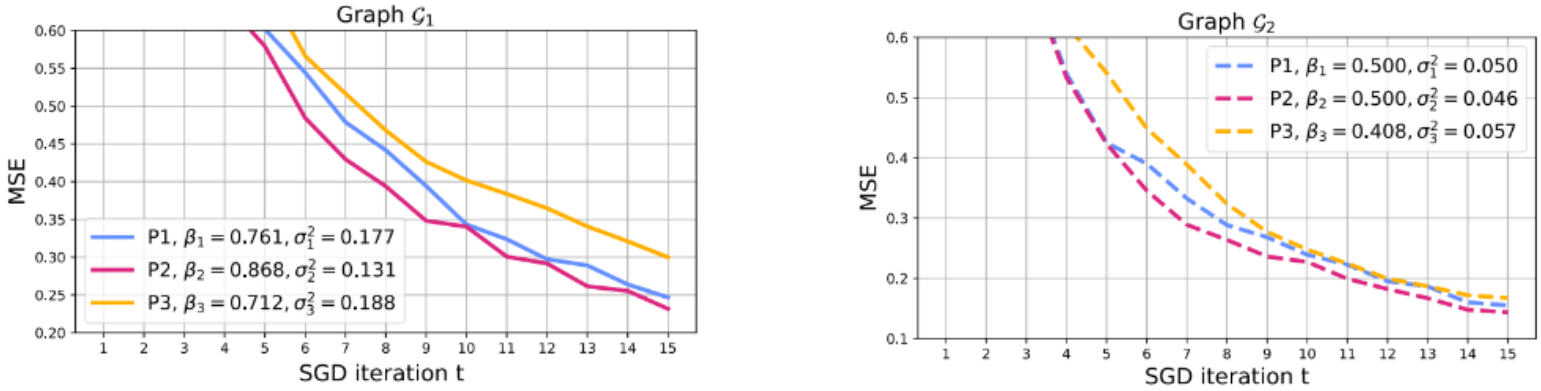


Figure: Comparison of MHRW (P1), Modified-MHRW (P2) and FMMC (P3) as stochastic inputs for RWSGD on two different graphs G1 and G2.

Up Next...

Generalized SGD update rule

- A more general definition of the stationary distribution π would be limit $\pi_i \triangleq \lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t} \mathbb{1}_{\{X_k=i\}}$ (which exists almost surely and is positive for all $i \in [n]$ when $\{X_t\}_{t\geq 0}$ is an irreducible, aperiodic and positive recurrent (ergodic) Markov chain).
- Note such a definition of π allows $\{X_t\}_{t\geq 0}$ to be more general, possibly being non-Markov on [n].
- Then, we can use π to rewrite the objective as follows,

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n} F(\theta, i) = \mathbb{E}_{X\sim\pi}[G(\theta, X)],$$

where function $G(\theta, i) \triangleq \frac{1}{n\pi_i} F(\theta, i)$ for any $\theta \in \Theta, i \in [n]$.

Generalized SGD update rule

- The generalized update rule will then be,

$$\theta_{t+1} = Proj_{\Theta}(\theta_t - \gamma_{t+1} \nabla G(\theta_t, X_{t+1}))$$

- If π was an uniform distribution, $G(\theta, i) = F(\theta, i)$ for all $\theta \in \Theta$.
- This change of notation allows us to consider input sequences having possibly non-uniform stationary distributions, and is a version of importance sampling for RWSGD schemes.
- If the input sequence is instead a simple random walk on a connected graph $G(V, E)$ with $V = [n]$, we have $\pi \propto d$ (degree of nodes), and $G(\theta, i) = \frac{1^T d}{n d_i} F(\theta, i)$ for all $\theta \in \Theta, i \in V$.

Multivariate Estimators

- We would now, like to extend the results derived for scalar estimators to multivariate estimators.
- For any finite, irreducible Markov chain $\{X_t\}_{t\geq 0}$ with stationary distribution π, its estimator is defined as $\hat{\mu}_t(g) \triangleq \frac{1}{t}\sum_{i=1}^{t} g(X_i)$ for any vector-valued function $g : [n] \to \mathbb{R}^d$.
- The ergodic theorem states that for any initial distribution and any $g(.)$ such that $\mathbb{E}_\pi(g) = \sum_{i\in[n]} g(i)\pi_i < \infty$, we have

$$\hat{\mu}_t(g) \xrightarrow[t\to\infty]{a.s.} \mathbb{E}_\pi(g)$$

Asymptotic covariance matrix

- Similar to the asymptotic variance $\sigma_X^2(g)$ for a scalar-valued function g(·), we can also define the asymptotic covariance matrix $\Sigma_X(g)$ for vector-valued function $g(.)$,

$$\Sigma_X(g) \triangleq \lim_{t\to\infty} t.Var(\hat{\mu}_t(g)) = \lim_{t\to\infty} \frac{1}{t}.\mathbb{E}\{\Delta_t \Delta_t^T\}$$

  where $\Delta_t \triangleq \sum_{s=1}^{t}(g(X_s) - \mathbb{E}_\pi(g))$.
- Multivariate CLT: For any function $g : [n] \to \mathbb{R}^d$ that satisfies $\mathbb{E}_\pi(g^2) < \infty$, we have

$$\sqrt{t}[\hat{\mu}_t(g) - \mathbb{E}_\pi(g)] \xrightarrow[t\to\infty]{dist} \mathcal{N}(0, \Sigma_X(g))$$

Regularity conditions

- The step size is given by $\gamma_t = t^{-\alpha}$ for $\alpha \in (\frac{1}{2}, 1]$.
- There exists a unique minimizer θ^* in the interior of the compact set Θ with $\nabla f(\theta^*) = 0$, and matrix $\nabla^2 f(\theta^*)$ (resp. $\nabla^2 f(\theta^*) - \mathbf{I}/2$) is positive definite for $\alpha \in (\frac{1}{2}, 1]$ (resp. $\alpha = 1$).
- Gradients are bounded in the compact set Θ, that is, $sup_{\theta \in \Theta} sup_{i \in [n]} ||\nabla F(\theta, i)||_2 < \infty$.

Regularity conditions

- For every $z \in [n], \theta \in \mathbb{R}^d$, the solution $\tilde{F}(\theta, z) \in \mathbb{R}^d$ of the Poisson equation
  $\tilde{F}(\theta, z) - \mathbb{E}[\tilde{F}(\theta, X_{t+1})|X_t = z] = \nabla F(\theta, z) - \nabla f(\theta)$ exists, and $sup_{\theta \in \Theta, z \in [n]} ||\nabla \tilde{F}(\theta, z)||_2 < \infty$.
- The functions $F(\theta, i)$ are L-smooth for all $i \in [n]$, that is,
  $\forall \theta_1, \theta_2 \in \Theta, \forall i \in [n]$, we have
  $||\nabla F(\theta_1, i) - \nabla F(\theta_2, i)||_2 \leq L||\theta_1 - \theta_2||_2$.

CLT for generalized SGD

- For iterates $\{\theta_t\}_{t\geq 0}$ of the generalized SGD algorithm satisfying the regularity conditions, we have:

$$\theta_t \xrightarrow[t\to\infty]{a.s.} \theta^*, \text{ and } \frac{\theta_t - \theta^*}{\sqrt{\gamma_t}} \xrightarrow[t\to\infty]{dist} \mathcal{N}(0, \mathbf{V}_X)$$

where $\mathbf{V}_X$ is the solution to $\Sigma_X + \mathbf{K}\mathbf{V}_X + \mathbf{V}_X\mathbf{K}^T = \mathbf{0}$, $\Sigma_X = \Sigma_X(\nabla G(\theta^*, \cdot))$ is the asymptotic covariance matrix and $\mathbf{K}$ is the Hessian $(\nabla^2 f(\theta^*))$.

Efficiency ordering

- For two random walks $\{X_t\}_{t\geq 0}$ and $\{Y_t\}_{t\geq 0}$ with the same stationary distribution π, define $\{X_t\}_{t\geq 0}$ to be more efficient than $\{Y_t\}_{t\geq 0}$, written as $X \geq_E Y$, if and only if $\sigma^2_X(g) \leq \sigma^2_Y(g)$ for any $g : [n] \to R$.
- Assuming the regularity conditions, it is shown that $X \geq_E Y$ if and only if $\Sigma_X(g) \leq_L \Sigma_Y(g)$ for any vector-valued function g.

The big result

- After defining the notion of efficiency, they proceed to show that if $\sum_X(\nabla G(\theta^*, \cdot)) \leq_L \sum_Y(\nabla G(\theta^*, \cdot))$, then $V_X \leq_L V_Y$ where function $\nabla G(\theta^*, \cdot) : [n] \to \mathbb{R}^d$ is defined in the SGD iteration, V_X (resp. V_Y) is the covariance matrix corresponding to $\{X_t\}_{t\geq 0}$ (resp. $\{Y_t\}_{t\geq 0}$) as the stochastic input sequence.
- In other words, a Markov chain with a **smaller asymptotic variance has better convergence properties as an input to SGD**.

Up Next...

SRW v/s NBRW: Background

- Simple Random Walk (SRW) is a popular Markov chain. It is characterised by the fact that probability of transition from a node to any of its neighboring node is equal.
- A variant of SRW, termed Non-Backtracking Random Walk (NBRW) has gained traction recently. NBRW $\{Y_t\}_{t\geq 0}$ is a second-order non-reversible Markov chain characterised by:

$$P(Y_{t+1}=j|Y_t=i,Y_{t-1}=k)=\begin{cases}\frac{1}{d_i-1} & \text{if } i\neq j, j\in N(i), d_i>1,\\ 1 & \text{if } d_i=1, j\in N(i),\\ 0 & \text{otherwise}\end{cases}$$

SRW v/s NBRW: Which is better?

- Both SRW and NBRW share the same limiting (stationary) distribution of $\pi = \mathbf{d}/\mathbf{1}^T\mathbf{d}$.
- Previous works have shown that:

$$\mathrm{AV}_{NBRW} \leq \mathrm{AV}_{SRW}$$

- From the results proved in this paper, it follows that **NBRW is more efficient than SRW** as an SGD input sequence.

Shuffling vs IID sampling: Background

- Single and random shuffling methods sample all the samples non-repeatedly. In single shuffling, the same order is used for all epochs, whereas in random shuffling each epoch has a different random order.
- Though empirically shuffling methods have been better, no theoretical guarantees existed.
- By moving to a higher dimension, it is possible to formulate shuffling based methods as Markov chains with same limiting distribution as IID sampling.

Shuffling vs IID sampling: Formal proof

- It can then be shown that:

$$\mathbf{\Sigma}_X(\mathbf{g}) = 0$$

  i.e., asymptotic covariance matrix is zero for all vector valued functions $\mathbf{g}$ for shuffling methods.
- It then follows that shuffling based methods have a smaller covariance matrix in the Loewner ordering sense.
- Then using the key results, it is formally proven that shuffling based methods are superior to IID sampling as inputs to SGD.

Up Next...

Numerical Simulation: The setup

- Empirical demonstration of the theoretical results/conclusions.
- Considers two convex objective functions:

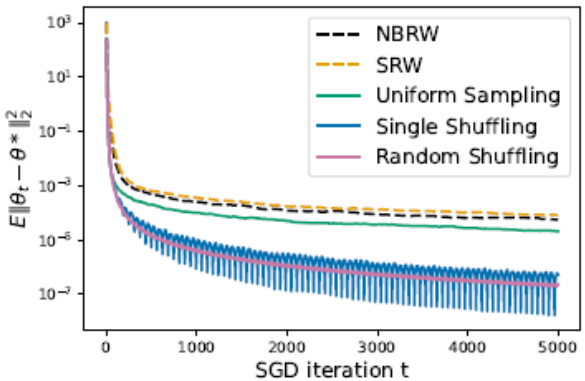
$$\tilde{f}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i \mathbf{x}_i^T \theta)) + \frac{1}{2}||\theta||_2^2$$

$$\hat{f}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \theta^T(\mathbf{a}_i\mathbf{a}_i^T + \mathbf{D}_i)\theta + \mathbf{b}^T\theta$$

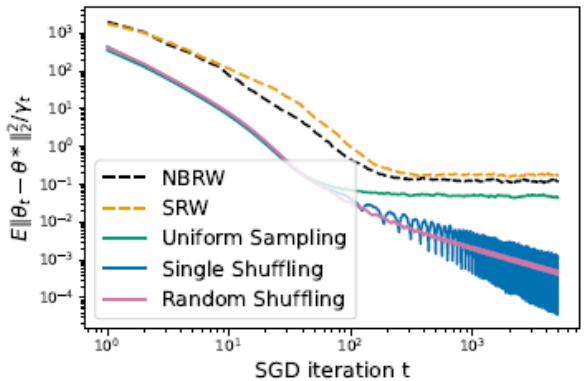
Numerical Simulation: The setup contd.

- Uses 6x6x3 cropped images from CIFAR-10 for $\tilde{f}$ and random vectors for $\mathbf{a}_i, \mathbf{b}$ and $\mathbf{D}_i$ for $\hat{f}$.
- The graph 'Dolphins' is considered.
- NBRW, SRW, single shuffling and random shuffling are compared.

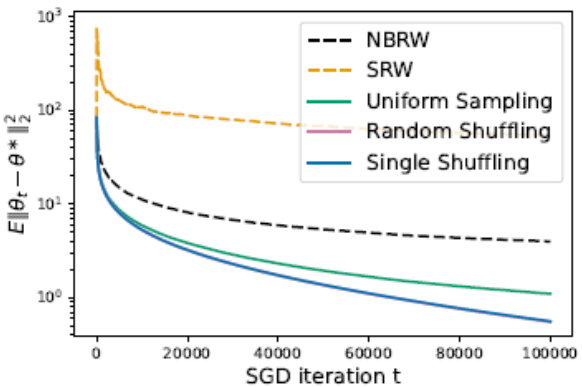
Numerical Simulation: Results



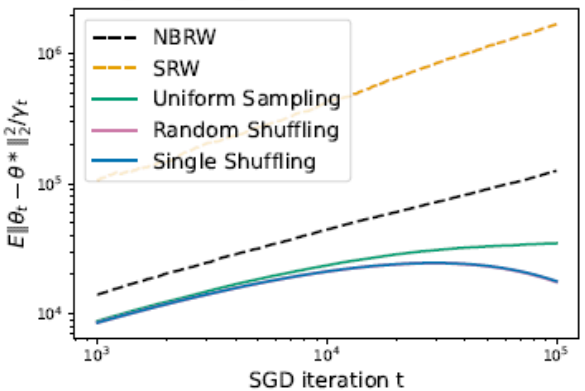
(a) Logistic regression; (MSE)



(b) Logistic regression; (scaled MSE)



(c) sum-non-convex fn.; (MSE)



(d) sum-non-convex fn.; (scaled MSE)

Figure: Results of simulations

What next?

- At this point, we believe we have a firm understanding of the theoretical nuances presented in the paper.
- Over the next couple of months, we hope to experimentally reconstruct the results presented in the paper.
- Once this is done, if time permits, we hope to extend the analysis to more datasets/ objective functions.

THANK YOU!