Permutation Compressors

Our team:

- Vyacheslav Naumov
- Anastasia Yagnych
- Mmesomachi Nwachukwu
- Alina Nurysheva

What is PermK?

For $d \geq n$ assume $d = qn$. Let $(\pi_1, \pi_2, \ldots, \pi_d)$ be a random permutation of $(1, 2, \ldots, d)$ then for $i \in \{1, \ldots, n\}$

$$\mathcal{C}_i(x) := n \cdot \sum_{j=q(i-1)+1}^{qi} x_{\pi_j} e_{\pi_j}$$

for $d < n$ and $n = qd$ let $(\pi_1, \pi_2, \ldots, \pi_n)$ be a random permutation of $(1, \ldots, 1, 2, \ldots, 2, \ldots, d, \ldots, d)$ where each appears q times, then

$$\mathcal{C}_i(x) := d x_{\pi_j} e_{\pi_j}$$

CGD

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta C_k(\nabla f(\mathbf{x}_k)),$$

where:

- $\mathbf{x}_k$ — parameter vector at iteration k,
- $\eta > 0$ — step size,
- C_k — compression operator,
- $\nabla f(\mathbf{x}_k)$ — gradient of the objective function f.

Classes of compressing operators

1. **Class** $B_1(\alpha, \beta)$:

$$\alpha\|\mathbf{x}\|^2 \leq \mathbb{E}[\|C(\mathbf{x})\|^2] \leq \beta\langle\mathbb{E}[C(\mathbf{x})], \mathbf{x}\rangle.$$

2. **Class** $B_2(\gamma, \beta)$:

$$\max\left\{\gamma\|\mathbf{x}\|^2, \frac{1}{\beta}\mathbb{E}[\|C(\mathbf{x})\|^2]\right\} \leq \langle\mathbb{E}[C(\mathbf{x})], \mathbf{x}\rangle.$$

3. **Class** $B_3(\delta)$:

$$\mathbb{E}[\|C(\mathbf{x}) - \mathbf{x}\|^2] \leq \left(1 - \frac{1}{\delta}\right)\|\mathbf{x}\|^2.$$

Complexity of the algorithm

1. For $C \in B_1(\alpha, \beta)$:

$$E_k \leq \left(1 - \frac{\alpha\mu}{\beta^2 L}\right)^k E_0,$$

   where $E_k = \mathbb{E}[f(\mathbf{x}_k)] - f^*$, L is the smoothness constant, μ is the strong convexity constant.

2. For $C \in B_2(\gamma, \beta)$:

$$E_k \leq \left(1 - \frac{\gamma\mu}{\beta L}\right)^k E_0.$$

3. For $C \in B_3(\delta)$:

$$E_k \leq \left(1 - \frac{\mu}{\delta L}\right)^k E_0.$$

Consider the following task

$$\min_{x\in\mathbb{R}^d}\left[f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right]$$

where:

- n is the number of workers/nodes,
- $f_i(x)$ is the loss function for the data on worker i,
- $x \in \mathbb{R}^d$ represents the model parameters.

Marina algorithm

$$x^{k+1} = x^k - \gamma g^k, \qquad g^k = \frac{1}{n}\sum_{i=1}^{n} g_i^k,$$

$$g_i^{k+1} = \begin{cases} \nabla f_i(x^{k+1}) & \text{if} \quad \theta_k = 1 \\ g^k + \mathcal{C}_i^k(\nabla f_i(x^{k+1}) - \nabla f_i(x^k)) & \text{if} \quad \theta_k = 0 \end{cases}$$

Input: starting point x^0, stepsize γ, probability $p \in (0,1]$, number of iterations T
Initialize $g^0 = \nabla f(x^0)$
for $k = 0, 1, \ldots, T-1$ **do**
    Sample $\theta_t \sim \mathrm{Be}(p)$
    Broadcast g^t to all workers
    for $i = 1, \ldots, n$ in parallel **do**
        $x^{t+1} = x^t - \gamma g^t$
        Set $g_i^{t+1} = \nabla f_i(x^{t+1})$ if $\theta_t = 1$, and $g_i^{t+1} = g^t + \mathcal{C}_i\left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right)$ otherwise
    end for
    $g^{t+1} = \frac{1}{n}\sum_{i=1}^{n} g_i^{t+1}$
end for
Output: $\hat{x}^T$ chosen uniformly at random from $\{x^t\}_{k=0}^{T-1}$

Compression Variance

$$\mathbb{E}[C(v)] = v, \quad \mathbb{E}[\|C(v) - v\|^2] \leq \omega \|v\|^2.$$

Communication Complexity

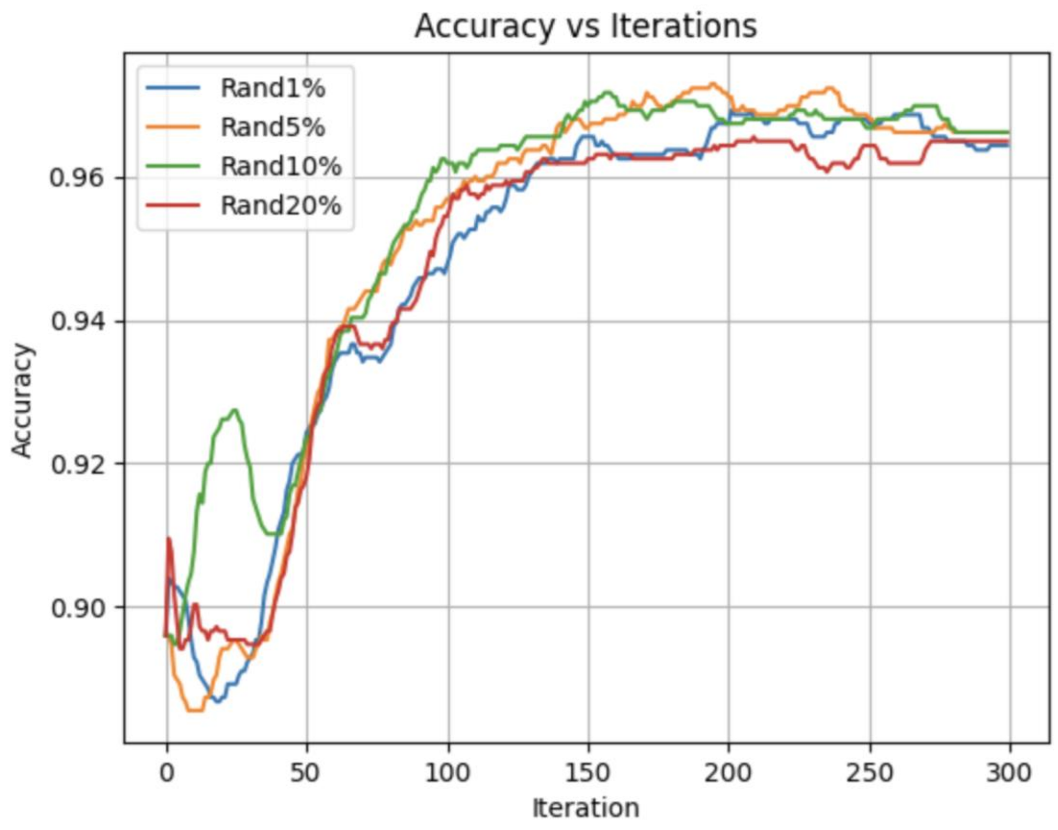
$$T = \mathcal{O}\left(\frac{\Delta_0}{\epsilon}\left(L_- + L_+\sqrt{\frac{1-p}{p} \cdot \frac{\omega}{n}}\right)\right),$$

where:

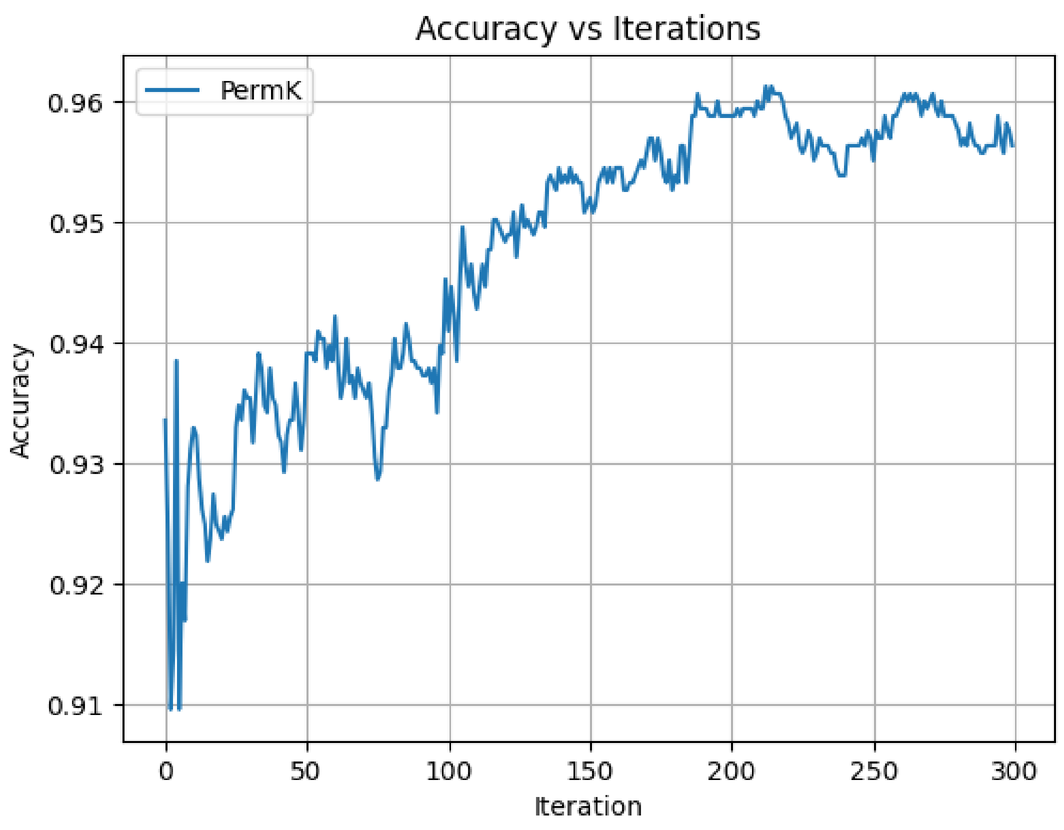
- $\Delta_0 = f(x_0) - f^*$,
- L_- and L_+ are gradient smoothness constants.

MARINA logistic regression

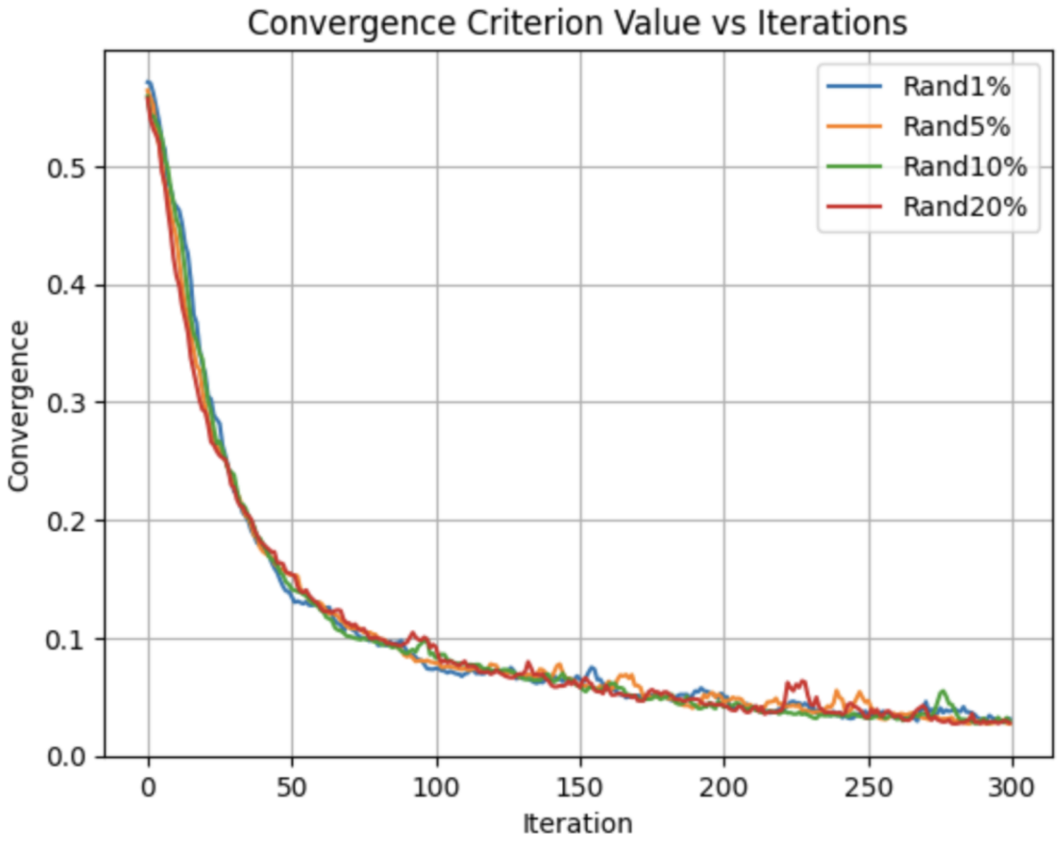
Rand



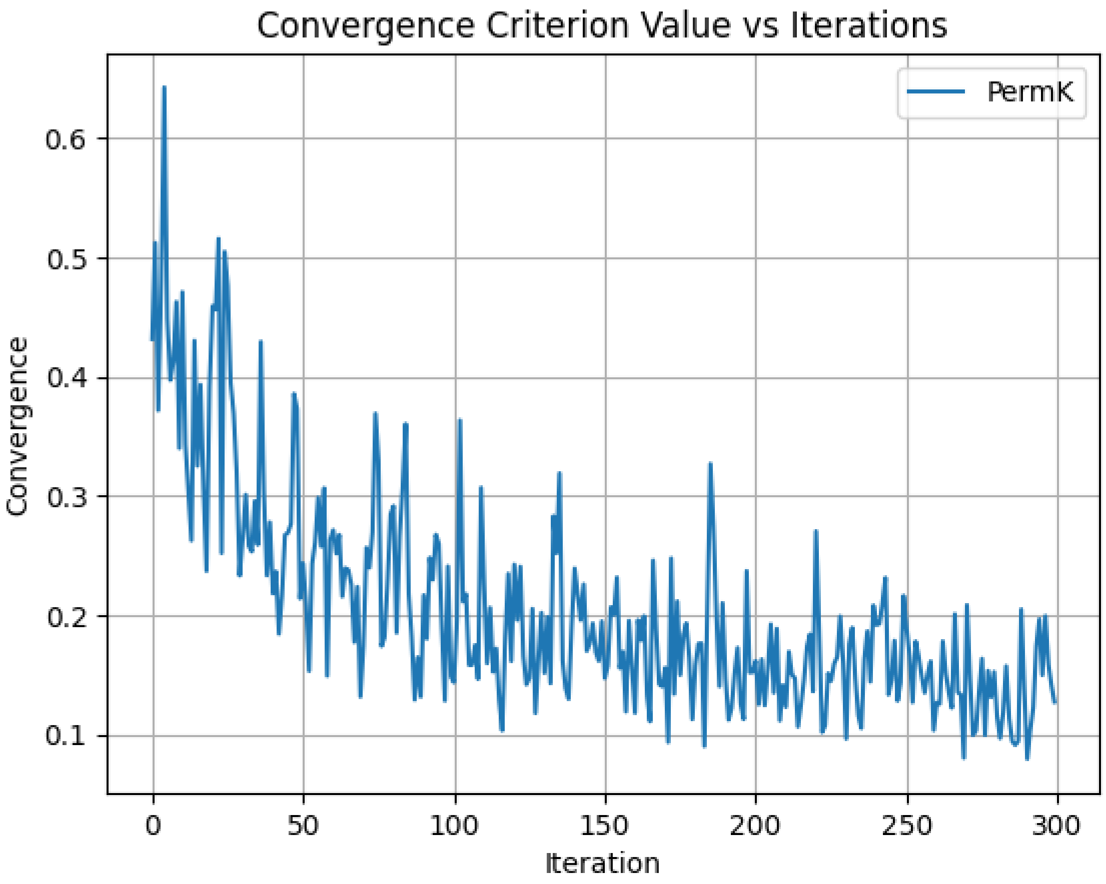
Perm

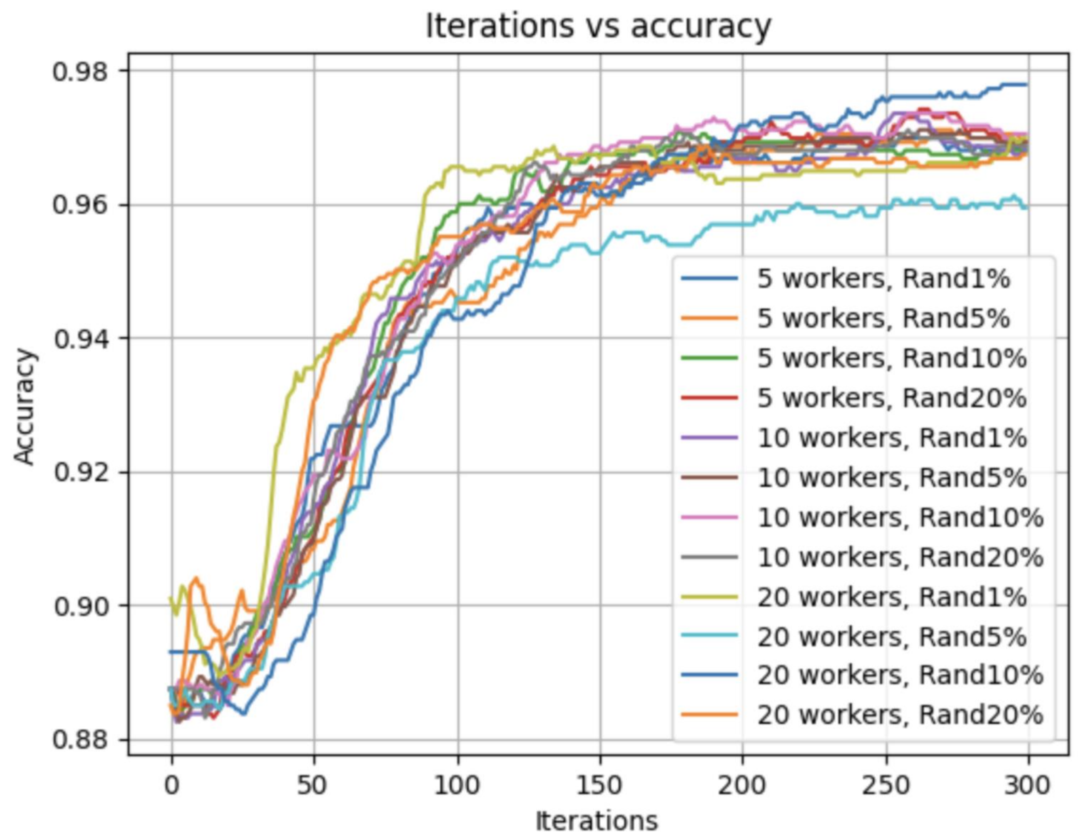


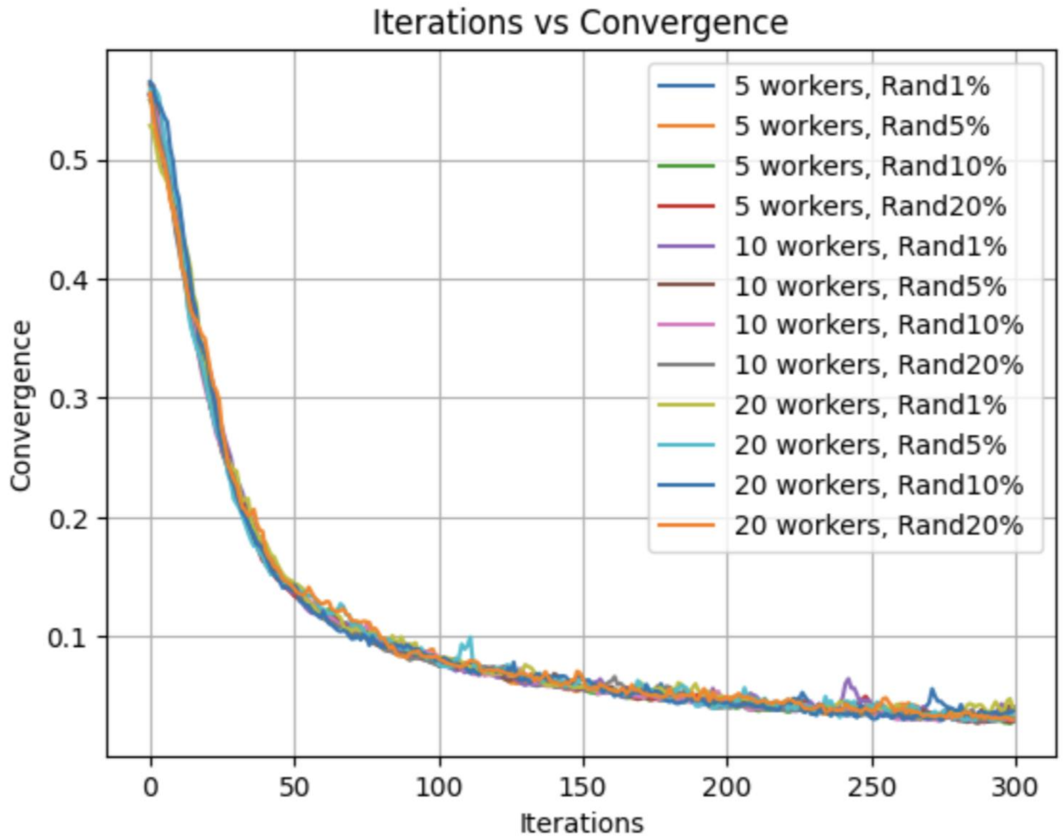
Rand


Convergence Criterion Value vs Iterations
Rand1%
Rand5%
Rand10%
Rand20%
Convergence
0.0
0.1
0.2
0.3
0.4
0.5
0
50
100
150
200
250
300
Iteration

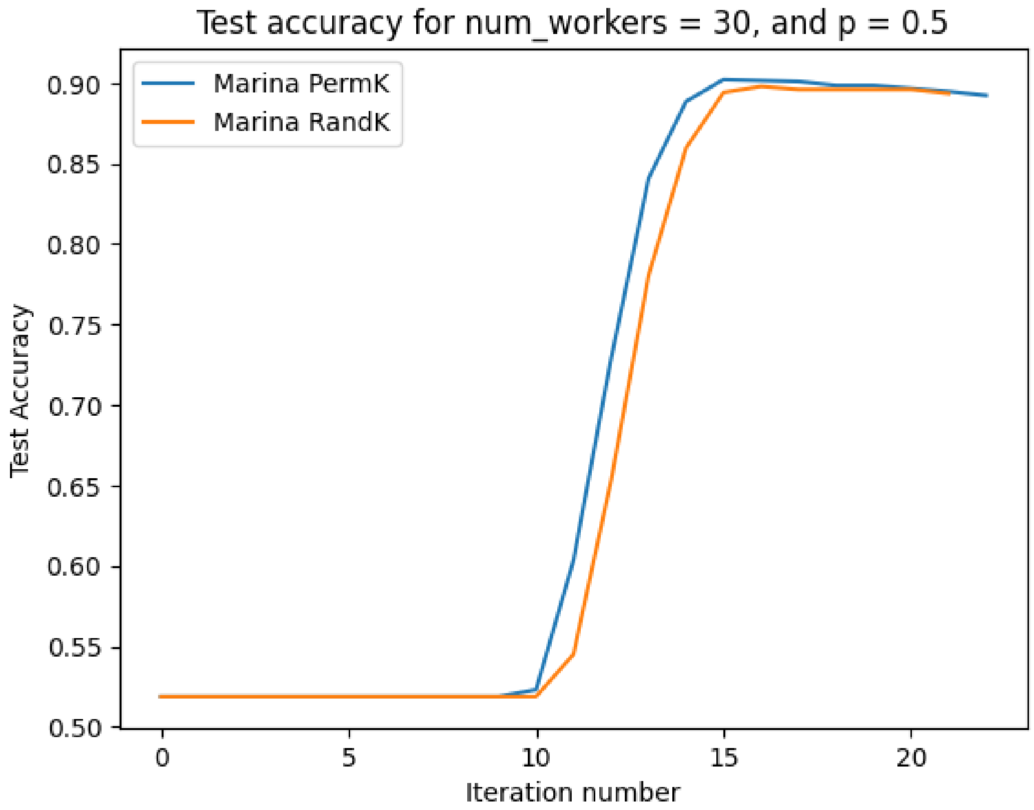
Perm

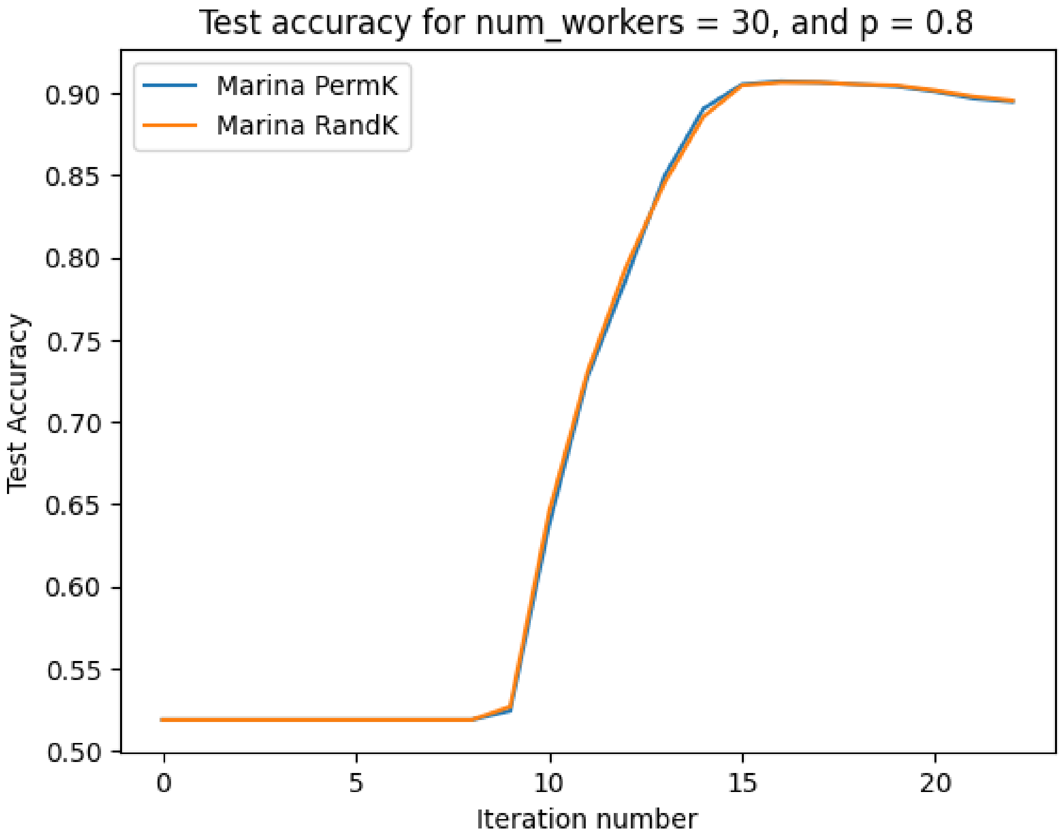

Convergence Criterion Value vs Iterations
PermK
Convergence
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200
250
300
Iteration


Iterations vs accuracy
Accuracy
Iterations
0.98
0.96
0.94
0.92
0.90
0.88
0
50
100
150
200
250
300
5 workers, Rand1%
5 workers, Rand5%
5 workers, Rand10%
5 workers, Rand20%
10 workers, Rand1%
10 workers, Rand5%
10 workers, Rand10%
10 workers, Rand20%
20 workers, Rand1%
20 workers, Rand5%
20 workers, Rand10%
20 workers, Rand20%

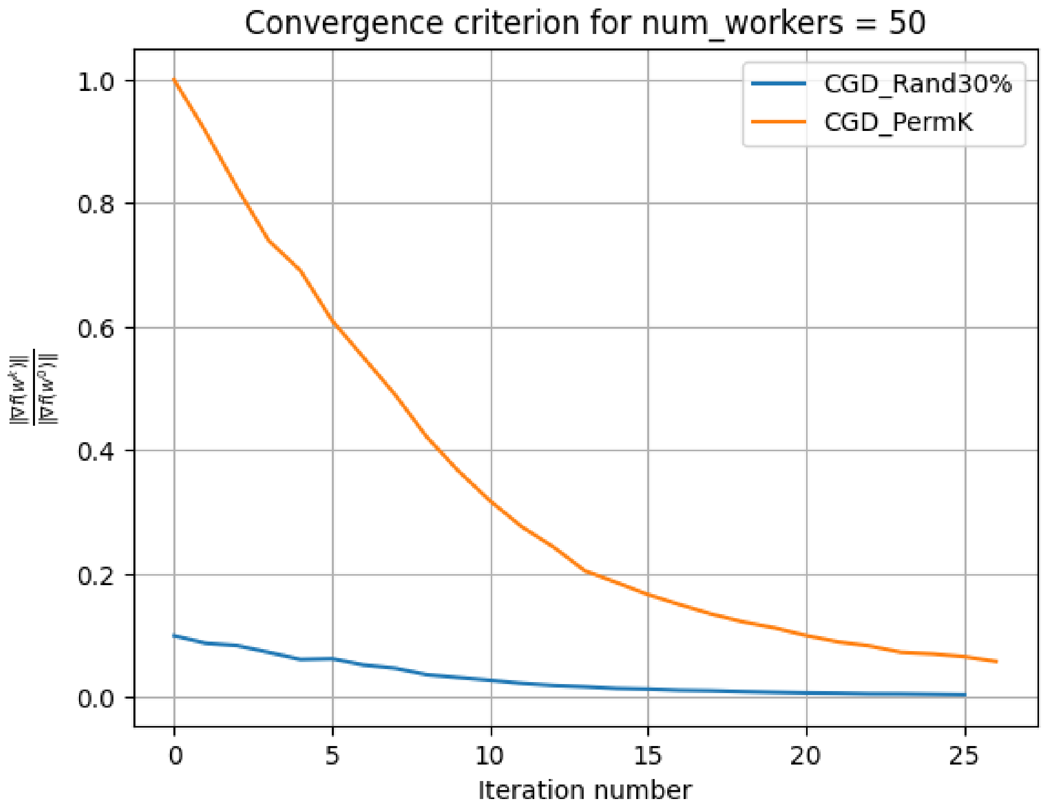

Iterations vs Convergence
Convergence
Iterations
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
300
5 workers, Rand1%
5 workers, Rand5%
5 workers, Rand10%
5 workers, Rand20%
10 workers, Rand1%
10 workers, Rand5%
10 workers, Rand10%
10 workers, Rand20%
20 workers, Rand1%
20 workers, Rand5%
20 workers, Rand10%
20 workers, Rand20%

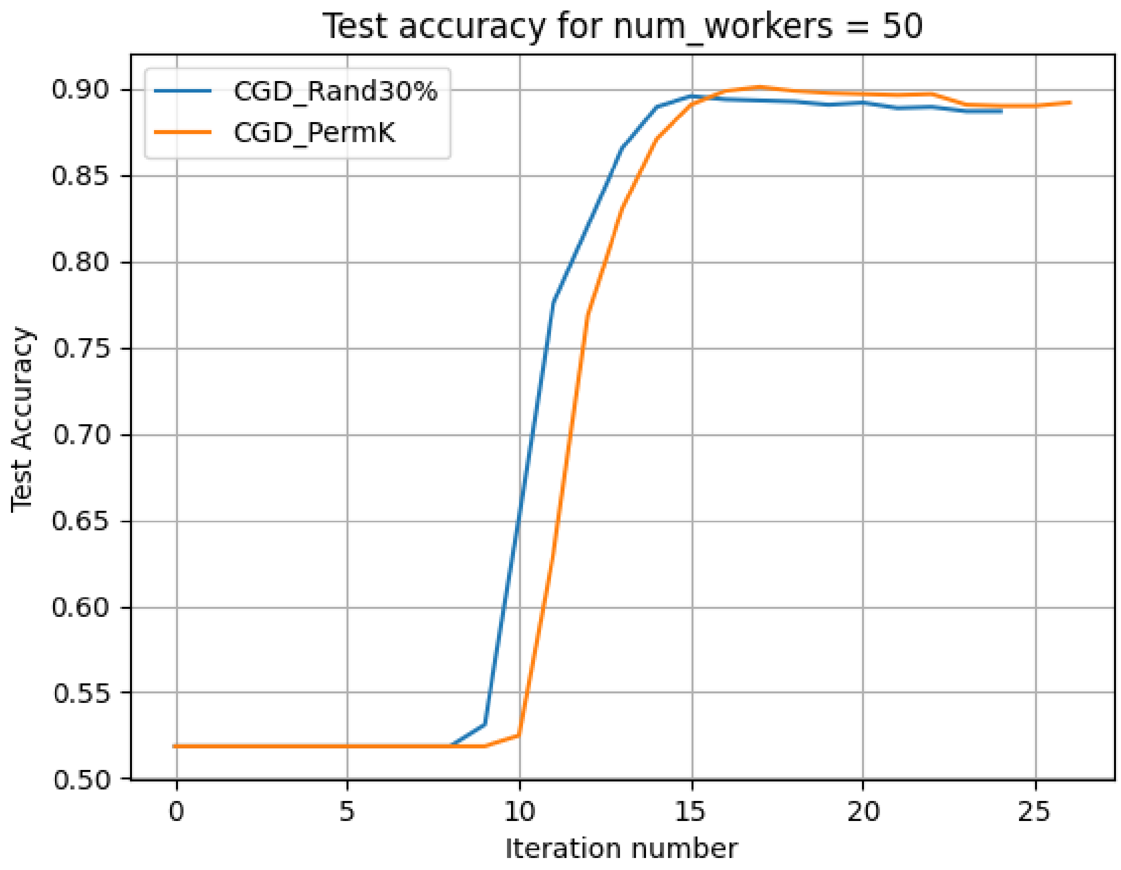
Tuning the parameter p, in Marina





PermK vs RandK in usual CGD





Thank you!