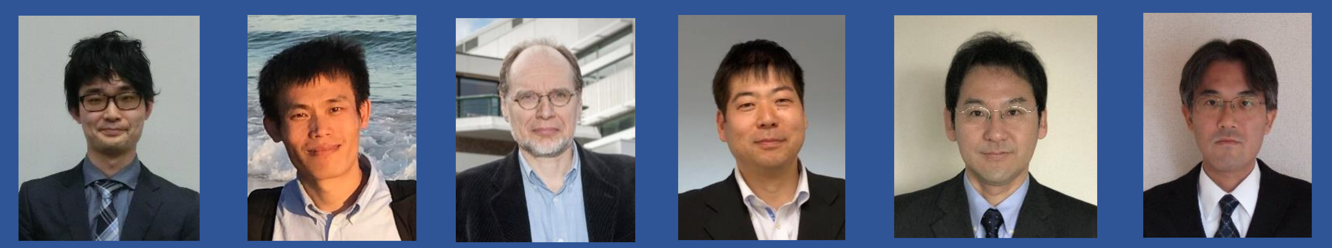
Asynchronous Decentralized Optimization with Implicit Stochastic Variance Reduction

Kenta Niwa[1,2], Guoqiang Zhang[3], W. Bastiaan Kleijn[4], Noboru Harada[1,2], Hiroshi Sawada[1], and Akinori Fujino[1]

1: NTT Communication Science Laboratories, 2: NTT Media Intelligence Laboratories, 3: University Technology of Sydney, 4: Victoria University of Wellington



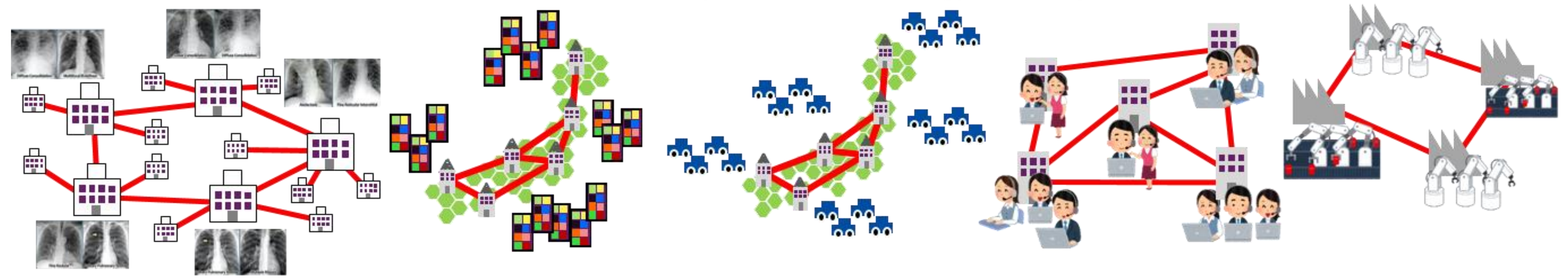
Goal

Background: We are entering an era of distributed data processing due to data volume, privacy-aware issues, and legal regulations, e.g., GDPR.

Goal: To train ML models without aggregating data to a central cloud.

1. Nodes are connected by decentralized network for flexible scalability.
2. Asynchronous communication among nodes are allowed.
3. Data subsets held on local nodes are statistically heterogeneous (non-IID).

Application Examples:



Conventional studies

Independently of network/communication configurations, many distributed training algorithms can be categorized into three trends.

(1) Average Consensus: SGD + Average

Weak robustness to non-IID data subsets

- FedAvg [McMahan *et al*., 2017]
- DSGD [Chen & Sayed, 2012]
- Gossip SGD [Ormandi *et al*. 2013]
- FedProx [Li *et al*., 2019]

(2) Stochastic Variance Reduction (SVR):

Stochastic gradient modification using global/local control variates.

Global control variate / Local control variate

$$\boldsymbol{w}_i^{k+1} = \boldsymbol{w}_i^k - \mu(g_i(\boldsymbol{w}_i^k) + \bar{\boldsymbol{c}}_i^k - \boldsymbol{c}_i^k),$$

- SVRG [Johnson & Zhang, 2013]
- SAGA [Defazio *et al*., 2014]
- SCAFFOLD [Karimireddy *et al*., 2020]
- GT-SVR [Xin *et al*., 2020]

(3) Primal-dual formalism:

Solve model matching **constraint** cost-sum **minimization** problem

$$\inf_{\boldsymbol{w}_1,..,\boldsymbol{w}_N} \frac{1}{N}\sum_{i\in\mathcal{N}} f_i(\boldsymbol{w}_i) \quad \text{s.t. } \mathbf{A}_{i|j}\boldsymbol{w}_i + \mathbf{A}_{j|i}\boldsymbol{w}_j = \mathbf{0}, \quad (\forall i\in\mathcal{N}, j\in\mathcal{E}_i),$$

$$\{\mathbf{A}_{i|j}, \mathbf{A}_{j|i}\} = \{\mathbf{I}, -\mathbf{I}\}$$

- Distributed ADMM [W. Shi *et al*., 2014]
- Primal-Dual Method of Multiplier (PDMM) [G. Zhang *et al*., 2017, T. Sherson *et al*., 2018]
- ECL [Niwa *et al*., 2020]
- FedSplit [Pathak & Wainwright, 2020]

E.g., Update rule of Edge-Consensus Learning (ECL) [Niwa *et al*., 2020]

$$\boldsymbol{w}_i^{k+1} = (\boldsymbol{w}_i^{r,k} - \mu g_i(\boldsymbol{w}_i^k) + \mu\eta_i \textstyle\sum_{j\in\mathcal{E}_i} \mathbf{A}_{i|j}^{\mathrm{T}} \boldsymbol{z}_{i|j}^k)/(1+\mu\eta_i E_i),$$

Stochastic gradient modification using dual variables $\boldsymbol{z}_{i|j}$

Note the similarity of the adjustment of stochastic gradient descent for SVR and for the primal dual formalism (e.g., ECL).

Main contribution

Key idea: Primal-dual formalism (e.g., ECL) may have an optimal condition where it matches SVR.

Reformulating ECL update rule:

$$\begin{aligned}\boldsymbol{w}_i^{k+1} &= (\boldsymbol{w}_i^k - \mu g_i(\boldsymbol{w}_i^k) + \mu\eta_i \textstyle\sum_{j\in\mathcal{E}_i} \mathbf{A}_{i|j}^{\mathrm{T}} \boldsymbol{z}_{i|j}^k)/(1+\mu\eta_i E_i) \\ &= \boldsymbol{w}_i^k - \mu[g_i(\boldsymbol{w}_i^k) + \underline{\tfrac{\eta_i}{1+\mu\eta_i E_i}\{\textstyle\sum_{j\in\mathcal{E}_i}(\boldsymbol{w}_i^k - \mathbf{A}_{i|j}^{\mathrm{T}}\boldsymbol{z}_{i|j}^k) - \mu E_i g_i(\boldsymbol{w}_i^k)\}}]\end{aligned}$$

ECL matches with SVR $\boldsymbol{w}_i^{k+1} = \boldsymbol{w}_i^k - \mu[g_i(\boldsymbol{w}_i^k) + \bar{\boldsymbol{c}}_i^k - \boldsymbol{c}_i^k]$ if underlined terms are modification using global/local control variates $\bar{\boldsymbol{c}}_i^k - \boldsymbol{c}_i^k$.

Contribution: We optimally select parameter η_i such that ECL matches with SVR.

- Investigating physical meaning of affine dual variables $\mathbf{A}_{i|j}^{\mathrm{T}}\boldsymbol{z}_{i|j}$, it is proportion to the sum of update difference between nodes. (-> a part of global control variate)
- By reformulating $\boldsymbol{w}$-update rule, we can optimally set η_i to follow SVR. $\eta_i = 1/(\mu E_i(K-1))$

Proposed method (ECL with Implicit SVR: ECL-ISVR)

- Optimal η_i is selected for the previous ECL [Niwa et al., 2020].

Cost function (with quadratic approximation)

$$\inf_{\boldsymbol{w}_1,..,\boldsymbol{w}_N} \tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}} f_i(\boldsymbol{w}_i)$$
$$\text{s.t. } \mathbf{A}_{i|j}\boldsymbol{w}_i + \mathbf{A}_{j|i}\boldsymbol{w}_j = \mathbf{0}, \quad (\forall i\in\mathcal{N}, j\in\mathcal{E}_i),$$

$$q_i(\boldsymbol{w}_i) = f_i(\boldsymbol{w}_i^{r,k}) + \langle g_i(\boldsymbol{w}_i^{r,k}), \boldsymbol{w}_i - \boldsymbol{w}_i^{r,k}\rangle + \tfrac{1}{2\mu}\|\boldsymbol{w}_i - \boldsymbol{w}_i^{r,k}\|^2.$$

Previous ECL

$$\boldsymbol{w}_i^{r,k+1} = \arg\min_{\boldsymbol{w}_i}(q_i(\boldsymbol{w}_i) + \textstyle\sum_{j\in\mathcal{N}_i} \tfrac{\eta_i}{2}\|\mathbf{A}_{i|j}\boldsymbol{w}_i - \boldsymbol{z}_{i|j}^{r,k}\|^2 + \textstyle\sum_{j\in\mathcal{N}_i}\tfrac{\rho_i}{2}\|\boldsymbol{w}_i - \boldsymbol{w}_j^{r,k}\|^2),$$

$$\boldsymbol{y}_{i|j}^{r,k+1} = \boldsymbol{z}_{i|j}^{r,k} - 2\mathbf{A}_{i|j}\boldsymbol{w}_i^{r,k+1}$$

$$\boldsymbol{z}_{i|j}^{r,k+1} = \begin{cases}\boldsymbol{y}_{j|i}^{r,k+1} & \text{(PDMM-SGD)} \\ \tfrac{1}{2}\boldsymbol{y}_{j|i}^{r,k+1} + \tfrac{1}{2}\boldsymbol{z}_{i|j}^{r,k} & \text{(ADMM-SGD)}\end{cases}$$

Affined dual variable is a part of global control variate. We select η_i such that ECL matches with SVR.

$$\eta_i = 1/(\mu E_i(K-1))$$
$$\rho_i = 0$$

Proposed ECL-ISVR

$$\boldsymbol{w}_i^{r,k+1} = \boldsymbol{w}_i^{r,k} - \mu\{g_i(\boldsymbol{w}_i^{r,k}) + \bar{\boldsymbol{c}}_i^{r,k} - \boldsymbol{c}_i^{r,k}\}$$

$$\bar{\boldsymbol{c}}_i^{r,k} = -\tfrac{1}{\mu K E_i}\textstyle\sum_{j\in\mathcal{E}_i}\{\mathbf{A}_{i|j}^{\mathrm{T}}\boldsymbol{z}_{i|j}^{r,k} - \tfrac{1}{K}(\mathbf{A}_{i|j}^{\mathrm{T}}\boldsymbol{z}_{i|j}^{r,k-1} + (1-\tfrac{1}{K})\mathbf{A}_{i|j}^{\mathrm{T}}\boldsymbol{z}_{i|j}^{r,k-2} + ,.., + (1-\tfrac{1}{K})^{(r-1)K+k-1}\mathbf{A}_{i|j}^{\mathrm{T}}\boldsymbol{z}_{i|j}^{1,0})\},$$

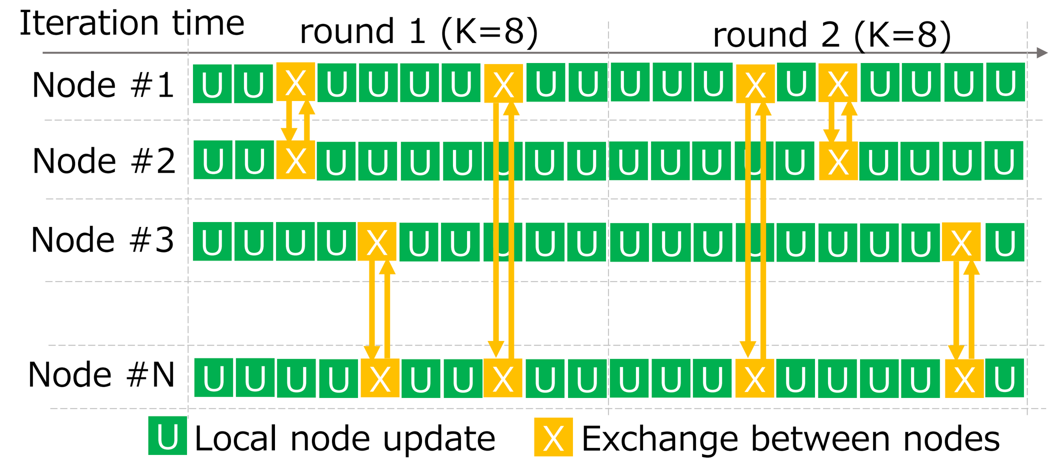
$$\boldsymbol{c}_i^{r,k} = \tfrac{1}{K}\{g_i(\boldsymbol{w}_i^{r,k}) + (1-\tfrac{1}{K})g_i(\boldsymbol{w}_i^{r,k-1}) + (1-\tfrac{1}{K})^2 g_i(\boldsymbol{w}_i^{r,k-2}) + ,.., + (1-\tfrac{1}{K})^{(r-1)K+k}g_i(\boldsymbol{w}_i^{1,0})\}.$$

$$\boldsymbol{y}_{i|j}^{r,k+1} = \boldsymbol{z}_{i|j}^{r,k} - 2\mathbf{A}_{i|j}\boldsymbol{w}_i^{r,k+1}$$

$$\boldsymbol{z}_{i|j}^{r,k+1} = \begin{cases}\boldsymbol{y}_{j|i}^{r,k+1} & \text{(PDMM-ISVR)} \\ \tfrac{1}{2}\boldsymbol{y}_{j|i}^{r,k+1} + \tfrac{1}{2}\boldsymbol{z}_{i|j}^{r,k} & \text{(ADMM-ISVR)}\end{cases}$$

- 2 algorithm flavors are existed.
  1: PDMM-ISVR: Peaceman-Rachford Splitting is applied.
  2: ADMM-ISVR: Douglas-Rachford Splitting is applied.

- Procedure is composed of alternatingly repeating
  U local node model updates and
  X asynchronously exchange dual variables.

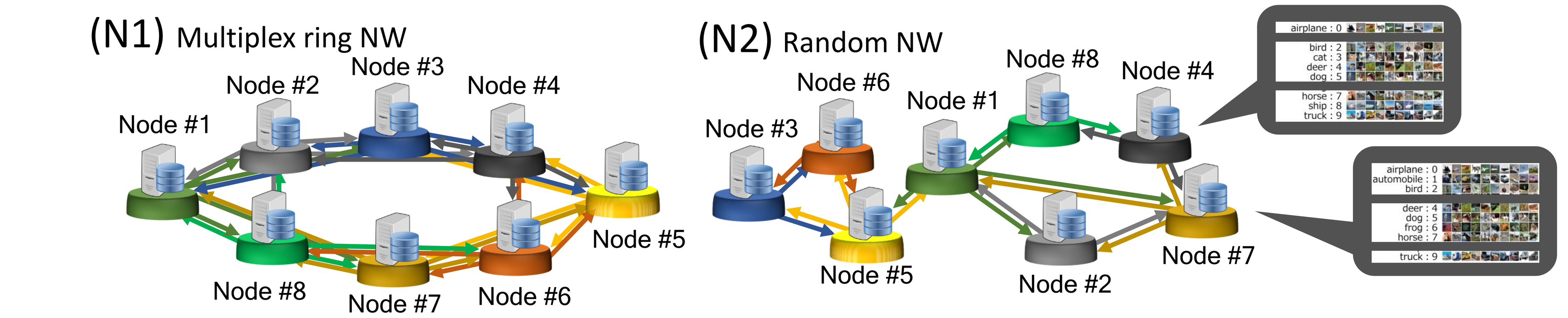


Algorithm 1 Proposed ECL-ISVR

```
▷ Set w_i = w_j (~Norm), z_{i|j} = 0, μ, x_i, A_{i|j}
for r ∈ {1, ..., R} (Outer loop round) do
    for i ∈ N do
        for k ∈ {1, ..., K} (Inner loop iteration) do                 } U
            ▷ Stochastic gradient calculation
            g_i(w_i) ← ∇f_i(w_i, χ_i)
            ▷ Update local primal and lifted dual variables
            w_i ← w_i − μ[g_i(w_i) + 1/(μKE_i) {Σ_{j∈E_i}
                    (w_i − A^T_{i|j} z_{i|j}) − μE_i g_i(w_i)}]
            for j ∈ E_i do
                y_{i|j} ← z_{i|j} − 2A_{i|j} w_i
            end for
            ▷ Procedure when communicated with j-th node            } X
            for j ∈ E_i^{r,k} (at random time) do
                communicate_{j→i}(y_{j|i})
                z_{i|j} ← { y_{j|i}                          (PDMM-ISVR)
                          { 1/2 z_{i|j} + 1/2 y_{j|i}        (ADMM-ISVR)
            end for
        end for
    end for
end for
```

Numerical experiments

- Aim of experiments is to identify algorithms that nearly reach performance of **reference** case where all data are available on a single node.
- Decentralized networks: (N1) multiplex ring and (N2) random topologies
- Asynchronous communication: Once per K=8 inner iterations on average
- Heterogeneous data: (T1) fashion MNIST and (T2) CIFAR-10 is divided to N=8 nodes where each node has 8 classes out of a total of 10 classes.



- Proposed methods (**PDMM-ISVR** and **ADMM-ISVR**) performed closest to the single-node **reference** scores with fast processing time.
- Previous ECL (**PDMM-SGD**) was next best. However, **ADMM-SGD** was unstable with long processing time due to doubled communication requirement.