Appendix

In appendix, Sec. A provides supplementary information associated with experiments. In Sec. B, update rule derivation of ECL-ISVR is introduced. The proof of convergence analysis of proposed ECL-ISVR is shown in Sec. C.

A. Supplementary information of experiments

Supplementary information associated with experiments in Sec. 7 is provided.

Source code: To support experimental reproducibility, we provide a part of the latest version of the source code as supplementary material. Although the constructed software will run on multiple GPUs connected with the network, all computational experiments were conducted inside a single server with 8 GPUs. For communication between nodes, we used Gloo in PyTorch as a GPU communication library. In addition, ZeroMQ is also needed to realize dual variable swap operation, namely send/receive commands associated with dual variables are ought to be done at almost the same timing. For DNN optimization code construction, PyTorch (v1.6.0) with CUDA (v10.2) was used.

Training data division: In Tables 1 and 2, the number of training data samples held by each node for fashion-MNIST used in (T1), (T2) and CIFAR-10 used in (T3) is summarized. As explained in Subsec. 7.1, training samples are heterogeneously divided over $N = 8$ nodes. Each local node holds a different number of data and they are composed of 8 randomly selected classes out of a total of 10 classes.

Table 1. Number of training data samples held by each node when using fashion-MNIST in (T1) and (T2), where class index is associated with 0: T-shirt/top, 1: trouser, 2: pullover, 3: dress, 4: coat, 5: sandal, 6: shirt, 7: sneaker, 8: bag, 9: ankle boot.

Class index	0	1	2	3	4	5	6	7	8	9	Total
Node 1	0	1,100	969	947	798	862	890	911	888	0	7,365
Node 2	887	1092	0	940	879	842	0	860	801	1,007	7,308
Node 3	992	0	933	0	856	953	939	822	876	998	7,369
Node 4	1,092	936	1,068	1,085	835	807	1,048	0	0	974	7,845
Node 5	1,072	1,041	1,082	1,059	0	0	1,073	875	882	1,008	8,092
Node 6	0	903	975	934	826	867	1,049	934	919	0	7,407
Node 7	1,046	928	0	1,035	898	808	0	801	819	972	7,307
Node 8	911	0	973	0	908	861	1,001	797	815	1,041	7,307
Total	6,000	6,000	6,000	6,000	6,000	6,000	6,000	6,000	6,000	6,000	60,000

Table 2. Number of training data samples held by each node when using CIFAR-10 in (T3), where class index is associated with 0: airplane, 1: automobile, 2: bird, 3: cat, 4: deer, 5: dog, 6: frog, 7: horse, 8: ship, 9: truck.

Class index	0	1	2	3	4	5	6	7	8	9	Total
Node 1	0	916	807	789	665	718	742	759	740	0	6,136
Node 2	739	910	0	784	733	702	0	717	667	839	6,091
Node 3	826	0	778	0	713	794	782	685	730	831	6,139
Node 4	910	780	890	904	696	672	873	0	0	812	6,537
Node 5	893	867	901	883	0	0	894	729	735	840	6,742
Node 6	0	753	813	778	688	722	874	778	766	0	6,172
Node 7	871	774	0	862	748	674	0	668	683	810	6,090
Node 8	761	0	811	0	757	718	835	664	679	868	6,093
Total	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	50,000

Processing time: The breakdown of processing time T_{total}, which is composed of (i) local node computation time T_{local} for training and test data sets and (ii) communication time T_{com}, is shown. Note that we used Gloo in PyTorch as a GPU communication library because NCCL[6] cannot be applicable because "Send" and "Recv" commands are needed for our asynchronous decentralized communication shown in Fig. 1.

The table below summarizes T_{local} for training/test data sets and T_{com} of ECL-ISVR (PDMM-ISVR) separately measured for each round.

Table 3. Node-averaged processing time for each round when using (T1) fashion MNIST with convex logistic regression model trained with a part of optimization algorithms on (N1) multiplex ring network

	$T_{\text{local, train}}$ [min]	$T_{\text{local, test}}$ [min]	T_{com} [min]	T_{total} [min]
ECL-ISVR (PDMM-ISVR)	43.83	74.13	40.52	158.48
ECL-ISVR (ADMM-ISVR)	45.62	78.04	42.63	166.29
ECL (PDMM-SGD)	56.50	86.40	91.41	234.31
ECL (ADMM-SGD)	57.80	81.15	85.49	224.44
GT-SVR	82.34	72.56	61.90	216.80
FedProx	36.22	80.68	43.33	160.23
DSGD	34.12	81.17	44.73	160.02
D^2	47.30	90.21	48.32	185.83

Table 4. Node-averaged processing time for each round when using (T1) fashion MNIST with convex logistic regression model trained with a part of optimization algorithms on (N2) random network

	$T_{\text{local, train}}$ [min]	$T_{\text{local, test}}$ [min]	T_{com} [min]	T_{total} [min]
ECL-ISVR (PDMM-ISVR)	1.49	6.70	0.20	8.39
ECL-ISVR (ADMM-ISVR)	1.53	6.98	0.21	8.72
ECL (PDMM-SGD)	1.78	7.15	0.38	9.31
ECL (ADMM-SGD)	1.74	6.87	0.37	8.98
GT-SVR	2.81	7.20	0.21	10.22
FedProx	1.38	6.96	0.21	8.55
DSGD	1.27	6.82	0.21	8.30
D^2	1.40	6.62	0.20	8.22

[6] https://pytorch.org/docs/stable/distributed.html

Table 5. Node-averaged processing time for each round when using (T2) fashion MNIST with non-convex ResNet32 model trained with a part of optimization algorithms on (N1) multiplex ring network

	$T_{\text{local, train}}$ [min]	$T_{\text{local, test}}$ [min]	T_{com} [min]	T_{total} [min]
ECL-ISVR (PDMM-ISVR)	43.83	74.13	40.52	158.48
ECL-ISVR (ADMM-ISVR)	45.62	78.04	42.63	166.29
ECL (PDMM-SGD)	56.50	86.40	91.41	234.31
ECL (ADMM-SGD)	57.80	81.15	85.50	224.45
GT-SVR	82.34	72.56	61.90	216.80
FedProx	36.22	80.68	43.33	160.23
DSGD	34.12	81.17	44.73	160.02
D^2	47.30	90.21	48.32	185.83

Table 6. Node-averaged processing time for each round when using (T2) fashion MNIST with non-convex ResNet32 model trained with a part of optimization algorithms on (N2) random network

	$T_{\text{local, train}}$ [min]	$T_{\text{local, test}}$ [min]	T_{com} [min]	T_{total} [min]
ECL-ISVR (PDMM-ISVR)	43.83	74.13	40.51	158.48
ECL-ISVR (ADMM-ISVR)	45.62	78.04	42.63	166.29
ECL (PDMM-SGD)	56.50	86.40	91.41	234.31
ECL (ADMM-SGD)	57.80	81.15	85.50	224.45
GT-SVR	82.33	72.57	61.90	216.80
FedProx	36.22	80.68	43.33	160.23
DSGD	34.12	81.17	44.73	160.02
D^2	47.30	90.21	48.32	185.83

Table 7. Node-averaged processing time for each round when using (T3) CIFAR-10 with non-convex ResNet32 model trained with a part of optimization algorithms on (N1) multiplex ring network

	$T_{\text{local, train}}$ [min]	$T_{\text{local, test}}$ [min]	T_{com} [min]	T_{total} [min]
ECL-ISVR (PDMM-ISVR)	91.05	129.65	113.22	333.92
ECL-ISVR (ADMM-ISVR)	96.81	137.84	124.46	359.11
ECL (PDMM-SGD)	112.19	142.86	274.38	529.43
ECL (ADMM-SGD)	127.34	147.13	282.52	556.99
GT-SVR	129.64	113.48	156.78	399.80
FedProx	65.80	138.45	125.01	329.26
DSGD	62.40	141.22	127.43	331.05
D^2	80.33	141.31	121.50	343.14

Table 8. Node-averaged processing time for each round when using (T3) CIFAR-10 with non-convex ResNet32 model trained with a part of optimization algorithms on (N2) random network

	$T_{\text{local, train}}$ [min]	$T_{\text{local, test}}$ [min]	T_{com} [min]	T_{total} [min]
ECL-ISVR (PDMM-ISVR)	43.83	74.13	40.52	158.48
ECL-ISVR (ADMM-ISVR)	45.62	78.04	42.63	166.29
ECL (PDMM-SGD)	56.50	86.40	91.41	234.31
ECL (ADMM-SGD)	57.80	81.15	85.49	224.44
GT-SVR	32.34	72.56	61.90	216.80
FedProx	36.22	80.68	43.33	160.23
DSGD	34.12	81.27	44.73	160.02
D^2	47.30	90.21	48.32	185.83

Performance differences among nodes: Performance differences in the trained models optimized with ECL-ISVR (PDMM-ISVR/ADMM-ISVR) are summarized in Tables 9–14. As a result, performance differences among nodes are sufficiently small, i.e., almost equivalent models are obtained on $N = 8$ nodes.

Table 9. Performance differences among $N = 8$ nodes when using (T1) fashion MNIST with convex logistic regression model trained with the proposed ECL-ISVR (PDMM-ISVR/ADMM-ISVR) on (N1) multiplex ring network

Node index	1	2	3	4	5	6	7	8	Average [%]
Reference (1 node)	81.98	–	–	–	–	–	–	–	81.98
ECL-ISVR (PDMM-ISVR)	80.87	80.83	80.86	80.82	80.82	80.82	80.86	80.83	80.84
ECL-ISVR (ADMM-ISVR)	80.83	80.90	80.90	80.89	81.00	80.89	80.88	80.88	80.90
ECL (PDMM-SGD)	80.39	80.37	80.31	80.40	80.34	80.44	80.36	80.39	80.38
ECL (ADMM-SGD)	80.51	80.48	80.46	80.48	80.52	80.54	80.54	80.49	80.50
GT-SVR	81.51	81.34	81.46	81.50	81.49	81.50	81.48	81.53	81.48
FedProx	74.39	74.27	74.34	74.31	74.34	74.35	74.31	74.37	74.34
DSGD	65.44	65.37	65.42	65.42	65.41	65.44	65.36	65.40	65.41
D^2	70.33	70.50	70.35	70.35	70.36	70.37	70.18	70.28	70.34

Table 10. Performance differences among $N = 8$ nodes when using (T1) fashion MNIST with convex logistic regression model trained with the proposed ECL-ISVR (PDMM-ISVR/ADMM-ISVR) on (N2) ramdom network

Node index	1	2	3	4	5	6	7	8	Average [%]
Reference (1 node)	81.66	–	–	–	–	–	–	–	81.66
ECL-ISVR (PDMM-ISVR)	79.71	79.59	79.75	79.50	79.72	79.67	79.57	79.61	79.64
ECL-ISVR (ADMM-ISVR)	80.57	80.58	80.52	80.49	80.51	80.53	80.56	80.56	80.54
ECL (PDMM-SGD)	78.38	78.38	78.38	78.32	78.37	78.26	78.39	78.35	78.35
ECL (ADMM-SGD)	78.40	78.29	78.47	78.44	78.43	78.08	78.29	78.52	78.37
GT-SVR	81.03	79.99	80.73	80.94	80.54	80.52	80.17	80.96	80.61
FedProx	72.00	71.86	71.85	71.95	71.88	71.74	71.87	71.98	71.89
DSGD	65.47	65.34	65.61	65.35	65.56	65.60	65.34	65.35	65.45
D^2	67.42	67.58	66.69	67.50	67.27	67.17	67.58	67.39	67.33

Table 11. Performance differences among $N = 8$ nodes when using (T2) fashion MNIST with non-convex ResNet32 model trained with the proposed ECL-ISVR (PDMM-ISVR/ADMM-ISVR) on (N1) multiplex ring network

Node index	1	2	3	4	5	6	7	8	Average [%]
Reference (1 node)	91.40	–	–	–	–	–	–	–	91.40
ECL-ISVR (PDMM-ISVR)	91.01	90.83	91.29	90.98	89.87	90.66	91.38	91.16	91.01
ECL-ISVR (ADMM-ISVR)	91.21	91.25	91.17	91.09	90.14	91.01	90.87	91.34	91.08
ECL (PDMM-SGD)	91.39	91.44	91.10	90.93	91.18	91.46	90.77	91.29	91.20
ECL (ADMM-SGD)	91.11	91.15	90.82	91.08	91.21	91.41	91.30	91.25	91.17
GT-SVR	89.42	90.83	89.94	91.00	89.51	90.06	90.79	90.99	90.32
FedProx	89.88	90.05	89.35	89.50	89.53	88.49	89.89	90.03	89.59
DSGD	86.64	86.25	86.67	86.89	86.64	85.71	85.75	86.39	86.50
D^2	87.35	87.53	88.17	87.52	87.79	87.83	87.76	87.35	87.66

Table 12. Performance differences among $N = 8$ nodes when using (T2) fashion MNIST with non-convex ResNet32 model trained with the proposed ECL-ISVR (PDMM-ISVR/ADMM-ISVR) on (N2) multiplex random network

Node index	1	2	3	4	5	6	7	8	Average [%]
Reference (1 node)	90.71	–	–	–	–	–	–	–	90.71
ECL-ISVR (PDMM-ISVR)	90.79	91.09	90.22	90.39	90.92	90.45	91.00	90.20	90.63
ECL-ISVR (ADMM-ISVR)	91.00	90.64	90.63	90.41	90.78	90.72	90.78	88.78	90.47
ECL (PDMM-SGD)	90.89	90.50	90.18	90.05	90.84	89.79	90.46	89.19	90.24
ECL (ADMM-SGD)	91.54	91.23	90.21	91.23	91.34	91.06	91.20	90.42	91.03
GT-SVR	88.80	85.67	90.38	89.03	88.35	87.59	86.07	89.80	88.21
FedProx	86.17	85.90	87.94	85.90	88.09	86.94	86.71	87.45	86.89
DSGD	85.24	84.53	83.92	84.62	83.92	84.91	83.88	84.73	84.47
D^2	85.95	85.10	85.67	85.24	85.91	85.70	85.62	84.99	85.52

Table 13. Performance differences among $N = 8$ nodes when using (T3) CIFAR-10 with non-convex ResNet32 model trained with the proposed ECL-ISVR (PDMM-ISVR/ADMM-ISVR) on (N1) multiplex ring network

Node index	1	2	3	4	5	6	7	8	Average [%]
Reference (1 node)	73.69	–	–	–	–	–	–	–	73.69
ECL-ISVR (PDMM-ISVR)	73.90	74.02	72.59	71.62	73.49	74.40	73.90	73.60	73.44
ECL-ISVR (ADMM-ISVR)	73.36	73.73	72.62	72.10	73.59	73.03	71.79	72.96	72.90
ECL (PDMM-SGD)	71.53	73.21	73.58	73.95	73.88	72.84	72.50	72.60	73.01
ECL (ADMM-SGD)	74.20	74.13	73.65	74.01	74.05	73.92	74.46	73.83	74.03
GT-SVR	73.07	72.56	70.97	72.98	73.10	69.68	71.60	63.78	70.97
FedProx	67.52	68.34	67.07	68.73	67.55	68.88	61.16	68.09	67.17
DSGD	53.87	53.56	53.21	55.24	53.27	55.06	52.22	51.99	53.55
D^2	60.53	62.11	61.29	61.62	61.47	61.93	61.80	61.32	61.51

Table 14. Performance differences among $N = 8$ nodes when using (T3) CIFAR-10 with non-convex ResNet32 model trained with the proposed ECL-ISVR (PDMM-ISVR/ADMM-ISVR) on (N2) random network

Node index	1	2	3	4	5	6	7	8	Average [%]
Reference (1 node)	72.93	–	–	–	–	–	–	–	72.93
ECL-ISVR (PDMM-ISVR)	73.61	72.97	69.77	72.82	72.87	72.72	72.66	73.64	72.63
ECL-ISVR (ADMM-ISVR)	74.29	70.68	71.25	71.93	72.45	69.65	72.01	71.85	71.76
ECL (PDMM-SGD)	73.28	73.40	72.85	72.19	66.40	71.78	72.26	72.40	71.82
ECL (ADMM-SGD)	65.01	63.19	64.85	60.37	65.23	62.76	64.87	64.13	63.80
GT-SVR	69.40	70.06	64.71	68.79	69.26	64.89	68.86	64.41	67.55
FedProx	62.67	60.24	62.09	58.51	61.28	63.28	60.36	61.29	61.21
DSGD	50.67	44.78	47.41	48.35	50.24	46.98	47.57	49.84	48.23
D^2	53.91	49.99	53.16	54.23	50.87	51.29	51.87	48.51	51.73

B. Derivation of ECL-ISVR's update rule

We now provide the proposed ECL-ISVR's update rule in Alg. 2 by reformulating the problem (5). For a stationary point, the subdifferential of the cost in (5) must include zero:

$$\mathbf{0} \in \underbrace{\mathbf{A}\mathbf{J}\nabla q^\star(\mathbf{J}^\mathrm{T}\mathbf{A}^\mathrm{T}\boldsymbol{\lambda})}_{T_1(\boldsymbol{\lambda})} + \underbrace{\partial \iota_{\ker(\mathbf{I}-\mathbf{P})}(\boldsymbol{\lambda})}_{T_2(\boldsymbol{\lambda})}, \tag{26}$$

where the differential operator ∇ is used in T_1 while subdifferential operator ∂ is applied in T_2 since the indicator function $\iota_{\ker(\mathbf{I}-\mathbf{P})}(\boldsymbol{\lambda})$ includes discontinuous points, and the operator $\in$ reflects that the subdifferential of the indicator function can be multi-valued at discontinuous points. The operator splitting is an effective method for finding a stationary point for problems of the form of (26).

Before introducing operator splitting algorithms, we define several operators, as summarized in literatures (Bauschke et al., 2011; Ryu & Boyd, 2016). The resolvent operator R_{T_i} and the Cayley operator C_{T_i} are defined as

$$\begin{aligned} R_{T_i} &= (\mathrm{Id} + \eta T_i)^{-1}, \qquad (27)\\ C_{T_i} &= (\mathrm{Id} + \eta T_i)^{-1}(\mathrm{Id} - \eta T_i)\\ &= 2(\mathrm{Id} + \eta T_i)^{-1} - (\mathrm{Id} + \eta T_i)^{-1}(\mathrm{Id} + \eta T_i)\\ &= 2(\mathrm{Id} + \eta T_i)^{-1} - \mathrm{Id}\\ &= 2R_{T_i} - \mathrm{Id}, \qquad (28) \end{aligned}$$

where Id is the identity operator, $^{-1}$ is the inverse operator, and $\eta\,(>0)$ denotes the step-size.

In the proposed ECL-ISVR, Peaceman-Rachford Splitting (PRS) (Peaceman & Rachford, 1955) and Douglas-Rachford Splitting (DRS) (Douglas & Rachford, 1956) to obtain methods associated with PDMM (Zhang & Heusdens, 2017; Sherson et al., 2018) and ADMM (Gabay & Mercier, 1976). We first derive the PRS. A reformulation of (26) results in

$$\mathbf{0} \in (\mathrm{Id} + \eta T_2)(\boldsymbol{\lambda}) - (\mathrm{Id} - \eta T_1)(\boldsymbol{\lambda}). \tag{29}$$

Let an auxiliary variable $\boldsymbol{z}$ be associated with the lifted dual variable $\boldsymbol{\lambda}$ through the relation $\boldsymbol{\lambda} \in R_{T_1}(\boldsymbol{z})$. Then, (29) can be written as

$$\begin{aligned} &\mathbf{0} \in (\mathrm{Id} + \eta T_2)R_{T_1}(\boldsymbol{z}) - C_{T_1}(\boldsymbol{z}),\\ &\mathbf{0} \in R_{T_1}(\boldsymbol{z}) - R_{T_2}C_{T_1}(\boldsymbol{z}),\\ &\mathbf{0} \in \tfrac{1}{2}(C_{T_1} + \mathrm{Id})(\boldsymbol{z}) - \tfrac{1}{2}(C_{T_2} + \mathrm{Id})C_{T_1}(\boldsymbol{z}), \end{aligned} \tag{30}$$

which implies that the stationary point condition can be written as

$$\boldsymbol{z} \in C_{T_2}C_{T_1}(\boldsymbol{z}), \quad \boldsymbol{\lambda} \in R_{T_1}(\boldsymbol{z}), \qquad \text{(PRS)}. \tag{31}$$

This indicates that the lifted dual variables are recursively updated through two different Cayley operators C_{T_1} and C_{T_2}. An alternative operator splitting, DRS can also be used as a basis for solving (26). By applying the averaged operator to (31), DRS is obtained:

$$\boldsymbol{z} \in \tfrac{1}{2}C_{T_2}C_{T_1}(\boldsymbol{z}) + \tfrac{1}{2}\boldsymbol{z}, \quad \boldsymbol{\lambda} \in R_{T_1}(\boldsymbol{z}), \qquad \text{(DRS)}. \tag{32}$$

By introducing another auxiliary variable, $\boldsymbol{y}$ for $\boldsymbol{\lambda}$, the update rules for PRS (31) and DRS (32) can be decomposed into

$$\boldsymbol{\lambda} \in R_{T_1}(\boldsymbol{z}) = (Id + \eta T_1)^{-1}(\boldsymbol{z}), \tag{33}$$
$$\boldsymbol{y} \in C_{T_1}(\boldsymbol{z}) = (2R_{T_1} - Id)(\boldsymbol{z}) = 2\boldsymbol{\lambda} - \boldsymbol{z}, \tag{34}$$
$$\boldsymbol{\varsigma} \in R_{T_2}(\boldsymbol{y}) = (Id + \eta T_2)^{-1}(\boldsymbol{y}), \tag{35}$$
$$\boldsymbol{z} \in \tfrac{1}{2}C_{T_2}(\boldsymbol{y}) + \tfrac{1}{2}\boldsymbol{z} = \tfrac{1}{2}(2R_{T_2} - Id)(\boldsymbol{y}) + \tfrac{1}{2}\boldsymbol{z} = \tfrac{1}{2}(2\boldsymbol{\varsigma} - \boldsymbol{y}) + \tfrac{1}{2}\boldsymbol{z}. \tag{36}$$

First, update rule associated with (33) and (34) is derived. Since $T_1(\boldsymbol{\lambda}) = \mathbf{A}\mathbf{J}\nabla f^\star(\mathbf{J}^\mathrm{T}\mathbf{A}^\mathrm{T}\boldsymbol{\lambda})$, the update procedure using the resolvent operator is reformulated by

$$\begin{aligned}
&\boldsymbol{\lambda} \in (Id + \eta T_1)^{-1}(\boldsymbol{z}),\\
&\boldsymbol{z} \in (Id + \eta T_1)(\boldsymbol{\lambda}),\\
&\mathbf{0} \in \eta \mathbf{A}\mathbf{J}\nabla f^\star(\mathbf{J}^\mathrm{T}\mathbf{A}^\mathrm{T}\boldsymbol{\lambda}) + \boldsymbol{\lambda} - \boldsymbol{z}.
\end{aligned} \tag{37}$$

From the basic property of convex function (Rockafellar, 1970), primal and dual variables are associated with $\boldsymbol{w} \in \nabla q^\star(\mathbf{J}^\mathrm{T}\mathbf{A}^\mathrm{T}\boldsymbol{\lambda})$, and the subdifferential of convex conjugate function satisfies $\nabla q^\star = (\nabla q)^{-1}$. Then, we obtain

$$\boldsymbol{w} \in \nabla q^\star(\mathbf{J}^\mathrm{T}\mathbf{A}^\mathrm{T}\boldsymbol{\lambda}), \tag{38}$$

$$\nabla q(\boldsymbol{w}) \in \mathbf{J}^\mathrm{T}\mathbf{A}^\mathrm{T}\boldsymbol{\lambda}. \tag{39}$$

Combining (37) and (38) gives

$$\begin{aligned}
&\mathbf{0} \in \eta \mathbf{A}\mathbf{J}\boldsymbol{w} + \boldsymbol{\lambda} - \boldsymbol{z},\\
&\boldsymbol{\lambda} \in \boldsymbol{z} - \eta \mathbf{A}\mathbf{J}\boldsymbol{w}.
\end{aligned} \tag{40}$$

By placing $\tilde{\boldsymbol{\lambda}} = \frac{1}{\eta}\boldsymbol{\lambda}$, $\tilde{\boldsymbol{y}} = \frac{1}{\eta}\boldsymbol{y}$, $\tilde{\boldsymbol{z}} = \frac{1}{\eta}\boldsymbol{z}$, (40) is reformed by

$$\tilde{\boldsymbol{\lambda}} \in \tilde{\boldsymbol{z}} - \mathbf{A}\mathbf{J}\boldsymbol{w}. \tag{41}$$

Combining (39) and (41) gives

$$\begin{aligned}
&\mathbf{0} \in \nabla q(\boldsymbol{w}) - \mathbf{J}^\mathrm{T}\mathbf{A}^\mathrm{T}(\boldsymbol{z} - \eta \mathbf{A}\mathbf{J}\boldsymbol{w}),\\
&\mathbf{0} \in \nabla q(\boldsymbol{w}) + \eta \mathbf{J}^\mathrm{T}\mathbf{A}^\mathrm{T}(\mathbf{A}\mathbf{J}\boldsymbol{w} - \tilde{\boldsymbol{z}}).
\end{aligned} \tag{42}$$

If the minimum exists, the integral of (42) gives

$$\boldsymbol{w}^{k+1} = \arg\min_{\boldsymbol{w}} \big(q(\boldsymbol{w}) + \tfrac{\eta}{2}\|\mathbf{A}\mathbf{J}\boldsymbol{w} - \tilde{\boldsymbol{z}}^k\|^2\big). \tag{43}$$

Following (41), the $\tilde{\boldsymbol{\lambda}}$-update rule is given by

$$\tilde{\boldsymbol{\lambda}}^{k+1} = \tilde{\boldsymbol{z}}^k - \mathbf{A}\mathbf{J}\boldsymbol{w}^{k+1}. \tag{44}$$

The combination of (34) and (44) gives

$$\tilde{\boldsymbol{y}}^{k+1} = 2\tilde{\boldsymbol{\lambda}}^{k+1} - \tilde{\boldsymbol{z}}^k = \tilde{\boldsymbol{z}}^k - 2\mathbf{A}\mathbf{J}\boldsymbol{w}^{k+1}. \tag{45}$$

Next, update rule associated with (35) and (36) is derived. Note that derivation detail is shown in (Sherson et al., 2018), although the step-size is doubled. For a normal cone operator $T_2(\boldsymbol{\varsigma}) = \partial \iota_{\ker(\mathbf{I}-\mathbf{P})}(\boldsymbol{\varsigma})$, the update procedure (35) is reformulated as

$$\begin{aligned}
&\mathbf{0} \in \eta T_2(\boldsymbol{\varsigma}) + \boldsymbol{\varsigma} - \boldsymbol{y},\\
&\mathbf{0} \in \partial \iota_{\ker(\mathbf{I}-\mathbf{P})}(\boldsymbol{\varsigma}) + \tfrac{1}{\eta}(\boldsymbol{\varsigma} - \boldsymbol{y}).
\end{aligned} \tag{46}$$

The integral of (46) gives

$$\boldsymbol{\varsigma}^{k+1} = \arg\min_{\boldsymbol{\varsigma}} \big(\iota_{\ker(\mathbf{I}-\mathbf{P})}(\boldsymbol{\varsigma}) + \tfrac{1}{2\eta}\|\boldsymbol{\varsigma} - \boldsymbol{y}^{k+1}\|^2\big) = \arg\min_{\boldsymbol{\varsigma}=\mathbf{P}\boldsymbol{\varsigma}} \big(\|\boldsymbol{\varsigma} - \boldsymbol{y}^{k+1}\|^2\big). \tag{47}$$

As remarked in Lemma IV.2 in (Sherson et al., 2018), the solution of (47) is given by the projectiuon onto the set of feasible $\boldsymbol{\varsigma}$, i.e.,

$$\boldsymbol{\varsigma}^{k+1} = \mathbf{\Pi}_{\ker(\mathbf{I}-\mathbf{P})}(\boldsymbol{y}^{k+1}). \tag{48}$$

Then, update procedure using Cayley operator (36) can be computed as

$$\begin{aligned}
\boldsymbol{z}^{k+1} &= \begin{cases} (2\mathbf{\Pi}_{\ker(\mathbf{I}-\mathbf{P})} - Id)(\boldsymbol{y}^{k+1}) = \mathbf{P}\boldsymbol{y}^{k+1}, & \text{(PRS)}\\ \frac{1}{2}(2\mathbf{\Pi}_{\ker(\mathbf{I}-\mathbf{P})} - Id)(\boldsymbol{y}^{k+1}) + \frac{1}{2}\boldsymbol{z}^k = \frac{1}{2}\mathbf{P}\boldsymbol{y}^{k+1} + \frac{1}{2}\boldsymbol{z}^k, & \text{(DRS)} \end{cases}\\
\tilde{\boldsymbol{z}}^{k+1} &= \begin{cases} \mathbf{P}\tilde{\boldsymbol{y}}^{k+1}, & \text{(PRS)}\\ \frac{1}{2}\mathbf{P}\tilde{\boldsymbol{y}}^{k+1} + \frac{1}{2}\tilde{\boldsymbol{z}}^k, & \text{(DRS)} \end{cases}
\end{aligned} \tag{49}$$

By decomposing (43), (45), and (49) into each node/edge procedures with parameter selection (22), ECL-ISVR's update rule in Alg. 2 is obtained.

C. Convergence analysis of ECL-ISVR

Convergence analysis of ECL-ISVR is summarized in Theorem 1, where its proof for strongly convex and general convex functions is shown in Subsec. C.1, and for non-convex functions is shown in Subsec. C.2. Note that some technical lemmas summarized in Subsec. C.3 are applied to derive lemmas in Subsec. C.1 and C.2.

C.1. Proof for convex function

Proof summary: Through this subsection, the function $\{f_i\}$ satisfies (D1) β-Lipschitz smooth with α-convex ($\alpha \geq 0$). Update rule follows Alg. 2. Note that difference between PDMM-ISVR and ADMM-ISVR is not considered in our convergence analysis because they just differs in gradient expectation update computation, as in (11) and (13).

First, we define several new variables used in convergence analysis. In Lemma 1, the update variance of control variate, including communication lag, is bounded. In Lemma 2, the update variance of control variate is bounded. Lemma 3, local node drift is bounded. In Lemma 4, progress in one round can be bounded. By applying Lemma 12 to Lemma 4, the final convergence rate for convex functions is obtained. For general convex functions, Lemma 9 is applied to Lemma 4 to obtain the convergence rate.

Preliminaries: Some variables used throughout this subsection is introduced. In ECL-ISVR, control variates are obtained by (25) and (24). From (11) and (13), the expectations of these calculated control variates on round r are given by

$$\begin{aligned}
\mathbb{E}[\boldsymbol{c}_i^r] &= \mathbb{E}[\tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}} g_i(\boldsymbol{w}_i^{r,k-1})] = \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}} \nabla f_i(\boldsymbol{w}_i^{r,k-1}),\\
\mathbb{E}[\boldsymbol{c}_{i|j}^r] &= \mathbb{E}[\tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}} g_j(\boldsymbol{w}_i^{r,k-1})] = \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}} \nabla f_j(\boldsymbol{w}_i^{r,k-1}),\\
\mathbb{E}[\bar{\boldsymbol{c}}_i^r] &= \textstyle\sum_{j\in\mathcal{E}_i}\sum_{k\in\mathcal{K}} \frac{1}{E_i K}\nabla f_j(\boldsymbol{w}_i^{r,k-1}),
\end{aligned}$$

where node index is summarized by $\{1,\dots,N\} \in \mathcal{N}$, edge index connected to the i-th node is summarized by $\{\mathcal{E}_i(1),\dots,\mathcal{E}_i(E_i)\} \in \mathcal{E}_i$, and inner loop index set on each round is summarized by $\{1,\dots,K\} \in \mathcal{K}$. We define $\gamma_{i|j}^r$ and γ_i^r as,

$$\begin{aligned}
\gamma_{i|j}^r &= \|\nabla f_i(\boldsymbol{w}_i^*) - \mathbb{E}[\boldsymbol{c}_{i|j}^r]\|^2,\\
\gamma_i^r &= \|\nabla f_i(\boldsymbol{w}_i^*) - \mathbb{E}[\boldsymbol{c}_i^r]\|^2,
\end{aligned}$$

where $\{\boldsymbol{w}_i^*\}$ denotes stationary point of $\{\boldsymbol{w}_i\}$, and then define the average of $\gamma_{i|j}$ over cross nodes/edges Γ as

$$\Gamma^r = \tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}} \frac{1}{E_i}\sum_{j\in\mathcal{E}_i} \gamma_{i|j}^r.$$

The client drift to model how much clients move from the starting point is defined by

$$\Theta^r = \tfrac{1}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}} \mathbb{E}\|\boldsymbol{w}_i^{r,k} - \boldsymbol{w}_i^{r,0}\|^2.$$

For simple notation, the minimum number of edges associated with a node is defined by

$$E_{\min} = \min(E_i) \qquad (i \in \mathcal{N}).$$

We often use the step size scaled by inner loop iteration K, noted by

$$\tilde{\mu} = \mu K \geq 0.$$

For each round after K times inner loop iterations, variables are updated as

$$\begin{aligned}
\boldsymbol{w}_i^{r+1,0} &= \boldsymbol{w}_i^{r,K},\\
\boldsymbol{c}_i^{r+1,0} &= \boldsymbol{c}_i^{r,K},\\
\boldsymbol{c}_{i|j}^{r+1,0} &= \boldsymbol{c}_{i|j}^{r,K},\\
\bar{\boldsymbol{c}}_i^{r+1,0} &= \bar{\boldsymbol{c}}_i^{r,K}.
\end{aligned}$$

Update variance of control variate under asynchronous communication: We will bound variance of control variate by the gradient of stationary point for each round in asynchronous decentralized communication, where its communication schedule is shown in Fig. 1 and explained in Sec. 2.

Lemma 1. *For local node update using Alg. 2 with asynchronous communication, $\{\gamma_i^r, \gamma_{i|j}^r, \Gamma^r\}$ on round r are bounded by*

$$\begin{aligned}
\gamma_i^r &\leq 2\beta\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})] - f_i(\boldsymbol{w}_i^*),\\
\gamma_{i|j}^r &\leq \tfrac{1}{2}\gamma_{i|j}^{r-1} + \beta\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})] - f_i(\boldsymbol{w}_i^*),\\
\Gamma^r &\leq \tfrac{1}{2}\Gamma^{r-1} + \tfrac{\beta}{NE_{\min}}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})] - f_i(\boldsymbol{w}_i^*)).
\end{aligned}$$

Proof. First, γ_i^r is bounded by

$$\gamma_i^r = \|\nabla f_i(\boldsymbol{w}_i^*) - \mathbb{E}[\boldsymbol{c}_i^r]\|^2 = \|\nabla f_i(\boldsymbol{w}_i^*) - \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}}\nabla f_i(\boldsymbol{w}_i^{r,k-1})\|^2 \leq 2\beta(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})] - f_i(\boldsymbol{w}_i^*)), \tag{50}$$

where Lemma 13 is applied in inequality.

Since communication for each edge is conducted once per K times inner loop iterations at random timing on each round, the expectation of $\boldsymbol{c}_{i|j}^r$ under this asynchronous communication is represented in a recursive update manner as

$$\mathbb{E}[\boldsymbol{c}_{i|j}^r] = \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}}\{\tfrac{k-1}{K}\mathbb{E}[\boldsymbol{c}_{i|j}^{r-1}] + \tfrac{K-k+1}{K}\mathbb{E}[\nabla f_j(\boldsymbol{w}_i^{r,k-1})]\}, \tag{51}$$

where this indicates that expectation computation is conducted by varying communication timing $k \in \mathcal{K}$, where $\boldsymbol{c}_{i|j}$ is not updated for $(k-1)$ times, and then, the dual variable $\boldsymbol{y}_{i|j}$ is transmitted from the node j; and $\boldsymbol{c}_{i|j}$ is updated, and used for remaining $K-(k-1)$ times. By using (51), $\gamma_{i|j}^r$ is bounded by

$$\begin{aligned}
\gamma_{i|j}^r &= \|\nabla f_i(\boldsymbol{w}_i^*) - \mathbb{E}[\boldsymbol{c}_{i|j}^r]\|^2\\
&= \|\nabla f_i(\boldsymbol{w}_i^*) - \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}}\{\tfrac{k-1}{K}\mathbb{E}[\boldsymbol{c}_{i|j}^{r-1}] + \tfrac{K-k+1}{K}\mathbb{E}[\nabla f_j(\boldsymbol{w}_i^{r,k-1})]\}\|^2\\
&\overset{(a)}{\leq} \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}}\{\tfrac{k-1}{K}\|\nabla f_i(\boldsymbol{w}_i^*) - \mathbb{E}[\boldsymbol{c}_{i|j}^{r-1}]\|^2 + \tfrac{K-k+1}{K}\|\nabla f_i(\boldsymbol{w}_i^*) - \mathbb{E}[\nabla f_j(\boldsymbol{w}_i^{r,k-1})]\|^2\}\\
&= \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}}\{\tfrac{k-1}{K}\gamma_{i|j}^{r-1} + \tfrac{K-k+1}{K}\|\nabla f_i(\boldsymbol{w}_i^*) - \mathbb{E}[\nabla f_i(\boldsymbol{w}_i^{r,k-1})]\|^2\}\\
&\overset{(b)}{\leq} \psi\gamma_{i|j}^{r-1} + 2(1-\psi)\beta(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})] - f_i(\boldsymbol{w}_i^*)),
\end{aligned} \tag{52}$$

where Jensen's inequality is used to obtain inequality (a) shown in the third line of (52), Lemma 13 is used for (b), and ψ is given by

$$\psi = \begin{cases} \frac{K-1}{2K} & (K \text{ is even number})\\ \frac{1}{2} & (K \text{ is odd number}) \end{cases}.$$

For simple notation, we assume that K is odd number from here, then $\psi = \frac{1}{2}$.

Γ^r is bounded by

$$\begin{aligned}
\Gamma^r &= \tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\tfrac{1}{E_i}\sum_{j\in\mathcal{E}_i}\gamma_{i|j}^r\\
&\overset{(a)}{\leq} \tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\tfrac{1}{E_i}\sum_{j\in\mathcal{E}_i}\{\tfrac{1}{2}\gamma_{i|j}^{r-1} + \beta(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})] - f_i(\boldsymbol{w}_i^*))\}\\
&\overset{(b)}{\leq} \tfrac{1}{2}\Gamma^{r-1} + \tfrac{\beta}{NE_{\min}}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})] - f_i(\boldsymbol{w}_i^*)),
\end{aligned} \tag{53}$$

where (52) is used for (a), Lemma 13 and $\frac{1}{E_i} \leq \frac{1}{E_{\min}}$ is used for (b). □

Variance of one round: We will bound one round update variance of primal model variable.

Lemma 2. *We can bound the variance of one round local node update in any round r and any step-size $\tilde{\mu} = \mu K \geq 0$:*

$$\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r+1,0} - \boldsymbol{w}_i^{r,0}\|^2 \leq 4\beta^2\tilde{\mu}^2\Theta^r + 4\tilde{\mu}^2\Gamma^r + \tfrac{16\beta\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})] - f_i(\boldsymbol{w}_i^*)) + \tfrac{12\tilde{\mu}^2\sigma^2}{NK}. \tag{54}$$

Proof. Local node update using (23) for one round is bounded by

$$
\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\Delta \boldsymbol{w}_i^r\|^2\\
&=\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r+1,0}-\boldsymbol{w}_i^{r,0}\|^2\\
&=\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|-\tfrac{1}{K}\sum_{k\in\mathcal{K}}\tilde{\mu}(g_i(\boldsymbol{w}_i^{r,k-1})+\bar{\boldsymbol{c}}_i^{r,k}-\boldsymbol{c}_i^{r,k})\|^2\\
&\overset{(a)}{\leq}\tfrac{1}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbb{E}\|\tilde{\mu}(g_i(\boldsymbol{w}_i^{r,k-1})+\bar{\boldsymbol{c}}_i^{r,k}-\boldsymbol{c}_i^{r,k})\|^2\\
&\overset{(b)}{\leq}\tfrac{1}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}[4\tilde{\mu}^2\mathbb{E}\|g_i(\boldsymbol{w}_i^{r,k-1})-\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+4\tilde{\mu}^2\mathbb{E}\|\bar{\boldsymbol{c}}_i^{r,k}\|^2+4\tilde{\mu}^2\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^*)-\boldsymbol{c}_i^{r,k}\|^2\\
&\quad+4\tilde{\mu}^2\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})-\nabla f_i(\boldsymbol{w}_i^*)\|^2]\\
&\overset{(c)}{\leq}\tfrac{1}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}[4\tilde{\mu}^2\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,k-1})-\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+4\tilde{\mu}^2\|\mathbb{E}[\bar{\boldsymbol{c}}_i^r]\|^2+4\tilde{\mu}^2\|\nabla f_i(\boldsymbol{w}_i^*)-\mathbb{E}[\boldsymbol{c}_i^r]\|^2\\
&\quad+8\beta\tilde{\mu}^2(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))+12\tilde{\mu}^2\sigma^2]\\
&\leq\tfrac{4\beta^2\tilde{\mu}^2}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbb{E}\|\boldsymbol{w}_i^{r,k-1}-\boldsymbol{w}_i^{r,0}\|^2+\tfrac{4\tilde{\mu}^2}{N}\sum_{i\in\mathcal{N}}\|\nabla f_i(\boldsymbol{w}_i^*)-\mathbb{E}[\bar{\boldsymbol{c}}_i^r]\|^2\\
&\quad+\tfrac{4\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\|\nabla f_i(\boldsymbol{w}_i^*)-\mathbb{E}[\boldsymbol{c}_i^r]\|^2+\tfrac{8\beta\tilde{\mu}^2}{N}\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))+\tfrac{12\tilde{\mu}^2\sigma^2}{NK}\\
&\overset{(d)}{\leq}4\beta^2\tilde{\mu}^2\Theta^r+\tfrac{4\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\tfrac{1}{E_i}\sum_{j\in\mathcal{E}_i}\gamma_{i|j}^r+\tfrac{4\tilde{\mu}^2}{N}\sum_{i\in\mathcal{N}}\gamma_i^r+\tfrac{8\beta\tilde{\mu}^2}{N}\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))+\tfrac{12\tilde{\mu}^2\sigma^2}{NK}\\
&\leq 4\beta^2\tilde{\mu}^2\Theta^r+4\tilde{\mu}^2\Gamma^r+\tfrac{16\beta\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))+\tfrac{12\tilde{\mu}^2\sigma^2}{NK},
\end{aligned}
\tag{55}
$$

where Jensen's inequality is used for (a), Lemma 10 is used for (b), Lemma 11, Lemma 13, and (D1) β-Lipschitz smoothness are used for (c), Jensen's inequality and is used for (d). □

Bounding local node drift: We will bound local node drift Θ^r.

Lemma 3. *We can bound the local node drift Θ^r in any round r and any step-size $\tilde{\mu}=\mu K\geq 0$:*

$$
\Theta^r\leq 9\tilde{\mu}^2\Gamma^r+\tfrac{36\beta\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,k})]-f_i(\boldsymbol{w}_i^*))+\tfrac{21\tilde{\mu}^2\sigma^2}{K}.
$$

Proof. A recursive bound of local node drift is given by

$$
\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,k}-\boldsymbol{w}_i^{r,0}\|^2\\
&=\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,k-1}-\mu(g_i(\boldsymbol{w}_i^{r,k-1})+\bar{\boldsymbol{c}}_i^{r,k}-\boldsymbol{c}_i^{r,k})-\boldsymbol{w}_i^{r,0}\|^2\\
&\overset{(a)}{\leq}\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,k-1}-\mu(\nabla f_i(\boldsymbol{w}_i^{r,k-1})+\bar{\boldsymbol{c}}_i^{r,k}-\boldsymbol{c}_i^{r,k})-\boldsymbol{w}_i^{r,0}\|^2+\mu^2\sigma^2\\
&\overset{(b)}{\leq}(1+\tfrac{1}{K-1})\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,k-1}-\boldsymbol{w}_i^{r,0}\|^2+K\mu^2\underbrace{\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\bar{\boldsymbol{c}}_i^{r,k}-\boldsymbol{c}_i^{r,k}+\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2}_{\mathcal{T}_4}+\mu^2\sigma^2,
\end{aligned}
\tag{56}
$$

where Lemma 11 is used for (a), and Lemma 10 is used for (b). $\mathcal{T}_4$ is bounded by

$$
\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\bar{\boldsymbol{c}}_i^{r,k}-\boldsymbol{c}_i^{r,k}+\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2\\
&=\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\bar{\boldsymbol{c}}_i^{r,k}+(\nabla f_i(\boldsymbol{w}_i^*)-\boldsymbol{c}_i^{r,k})+(\nabla f_i(\boldsymbol{w}_i^{r,0})-\nabla f_i(\boldsymbol{w}_i^*))\|^2\\
&\overset{(a)}{\leq}\tfrac{3}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\bar{\boldsymbol{c}}_i^{r,k}\|^2+\tfrac{3}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{c}_i^{r,k}-\nabla f_i(\boldsymbol{w}_i^*)\|^2+\tfrac{3}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^*)-\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2\\
&\leq\tfrac{3}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\bar{\boldsymbol{c}}_i^{r,k}-\nabla f_i(\boldsymbol{w}_i^*)\|^2+\tfrac{3}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{c}_i^{r,k}-\nabla f_i(\boldsymbol{w}_i^*)\|^2+\tfrac{3}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^*)-\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2\\
&\overset{(b)}{\leq}\tfrac{3}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\bar{\boldsymbol{c}}_i^{r,k}-\nabla f_i(\boldsymbol{w}_i^*)\|^2+\tfrac{3}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{c}_i^{r,k}-\nabla f_i(\boldsymbol{w}_i^*)\|^2+\tfrac{6\beta}{N}\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*)),
\end{aligned}
\tag{57}
$$

where Lemma 10 is used for (a) and Lemma 13 is used for (b). By substituting (57) into (56), we obtain

$$
\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,k}-\boldsymbol{w}_i^{r,0}\|^2\\
&\le (1+\tfrac{1}{K-1})\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,k-1}-\boldsymbol{w}_i^{r,0}\|^2+\mu^2\sigma^2\\
&\quad+\tfrac{3K\mu^2}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\bar{\boldsymbol{c}}_i^{r,k}-\nabla f_i(\boldsymbol{w}_i^*)\|^2+\tfrac{3K\mu^2}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{c}_i^{r,k}-\nabla f_i(\boldsymbol{w}_i^*)\|^2+\tfrac{6K\beta\mu^2}{N}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))\\
&\overset{(a)}{\le}(1+\tfrac{1}{K-1})\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,k-1}-\boldsymbol{w}_i^{r,0}\|^2+7\mu^2\sigma^2\\
&\quad+\tfrac{3K\mu^2}{N}\textstyle\sum_{i\in\mathcal{N}}\|\mathbb{E}[\bar{\boldsymbol{c}}_i^{r}]-\nabla f_i(\boldsymbol{w}_i^*)\|^2+\tfrac{3K\mu^2}{N}\textstyle\sum_{i\in\mathcal{N}}\|\mathbb{E}[\boldsymbol{c}_i^{r}]-\nabla f_i(\boldsymbol{w}_i^*)\|^2+\tfrac{6K\beta\mu^2}{N}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))\\
&\overset{(b)}{\le}(1+\tfrac{1}{K-1})\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,k-1}-\boldsymbol{w}_i^{r,0}\|^2+7\mu^2\sigma^2+\tfrac{6K\beta\mu^2}{N}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))\\
&\quad+\tfrac{3K\mu^2}{N}\textstyle\sum_{i\in\mathcal{N}}\tfrac{1}{E_i}\sum_{j\in\mathcal{E}_i}\gamma_{i|j}^r+\tfrac{3K\mu^2}{N}\textstyle\sum_{i\in\mathcal{N}}\gamma_i^r\\
&\le(1+\tfrac{1}{K-1})\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,k-1}-\boldsymbol{w}_i^{r,0}\|^2+3K\mu^2\Gamma^r+\tfrac{12\beta K\mu^2}{N}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))+7\mu^2\sigma^2,
\end{aligned}
\tag{58}
$$

where the fact that variance of $\boldsymbol{c}_i$ is less than σ^2/K is used for (a), Jensen's inequality is used for (b). Unrolling (58), we obtain following bound:

$$
\begin{aligned}
\Theta^r&=\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,k}-\boldsymbol{w}_i^{r,0}\|^2\\
&\le(3K\mu^2\Gamma^r+\tfrac{12\beta K\mu^2}{N}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))+7\mu^2\sigma^2)(\textstyle\sum_{\tau=0}^{k-1}(1+\tfrac{1}{K-1})^\tau)\\
&\overset{(a)}{\le}(3K\mu^2\Gamma^r+\tfrac{36K\beta\mu^2}{N}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,k})]-f_i(\boldsymbol{w}_i^*))+7\mu^2\sigma^2)3K\\
&=9K^2\mu^2\Gamma^r+\tfrac{36K^2\beta\mu^2}{N}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,k})]-f_i(\boldsymbol{w}_i^*))+21K\mu^2\sigma^2\\
&=9\tilde{\mu}^2\Gamma^r+\tfrac{36\beta\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,k})]-f_i(\boldsymbol{w}_i^*))+\tfrac{21\tilde{\mu}^2\sigma^2}{K},
\end{aligned}
\tag{59}
$$

where the fact that $\sum_{\tau=0}^{k-1}(1+\frac{1}{K-1})^\tau<3K$ used for (a) is proofed in Lemma 13 of (Karimireddy et al., 2020). □

Progress in one round: We bound all errors in a round.

Lemma 4. *Following holds in any round r and any step-size $\tilde{\mu}=K\mu$ satisfying $\tilde{\mu}\in[0,\min(\frac{1}{27\beta},\frac{1}{3\alpha}))$,*

$$
\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r+1,0}-\boldsymbol{w}_i^*\|^2+9\tilde{\mu}^2\Gamma^r\\
&\quad\le(1-\tfrac{\alpha\tilde{\mu}}{2})\{\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,0}-\boldsymbol{w}_i^*\|^2+9\tilde{\mu}^2\Gamma^{r-1}\}-(1-\tfrac{1}{E_{\min}})\tfrac{\tilde{\mu}}{N}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))+\tfrac{\tilde{\mu}^2\sigma^2}{K}(\tfrac{12}{N}+3).
\end{aligned}
$$

Proof. The update difference for a round is given by

$$
\Delta\boldsymbol{w}_i^r=-\tfrac{\tilde{\mu}}{K}\textstyle\sum_{k\in\mathcal{K}}(g_i(\boldsymbol{w}_i^{r,k-1})+\bar{\boldsymbol{c}}_i^{r,k}-\boldsymbol{c}_i^{r,k}),
$$

and its expectation satisfies

$$
\mathbb{E}[\Delta\boldsymbol{w}_i^r]=-\tfrac{\tilde{\mu}}{K}\textstyle\sum_{k\in\mathcal{K}}\nabla f_i(\boldsymbol{w}_i^{r,k-1}).
$$

The second moment of updated local node variable and its stationary point for a round is bounded by

$$
\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r+1,0}-\boldsymbol{w}_i^*\|^2\\
&=\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,0}+\Delta\boldsymbol{w}_i^r-\boldsymbol{w}_i^*\|^2\\
&=\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,0}-\boldsymbol{w}_i^*\|^2-\tfrac{2\tilde{\mu}}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbb{E}\Big[\Big\langle\nabla f_i(\boldsymbol{w}_i^{r,k-1}),\boldsymbol{w}_i^{r,0}-\boldsymbol{w}_i^*\Big\rangle\Big]+\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\Delta\boldsymbol{w}_i^r\|^2\\
&\overset{(a)}{\le}\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,0}-\boldsymbol{w}_i^*\|^2+\underbrace{\tfrac{2\tilde{\mu}}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbb{E}\left[\Big\langle\nabla f_i(\boldsymbol{w}_i^{r,k-1}),\boldsymbol{w}_i^*-\boldsymbol{w}_i^{r,0}\Big\rangle\right]}_{\mathcal{T}_5}\\
&\quad+4\beta^2\tilde{\mu}^2\Theta^r+4\tilde{\mu}^2\Gamma^r+\tfrac{16\beta\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))+\tfrac{12\tilde{\mu}^2\sigma^2}{NK},
\end{aligned}
\tag{60}
$$

where Lemma 2 is used for (a). The term $\mathcal{T}_5$ is bounded by

$$\begin{aligned}\mathcal{T}_5 &= \tfrac{2\tilde{\mu}}{NK}\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbb{E}[\Big\langle\nabla f_i(\boldsymbol{w}_i^{r,k-1}),\boldsymbol{w}_i^*-\boldsymbol{w}_i^{r,0}\Big\rangle]\\ &\overset{(a)}{\leq}\tfrac{2\tilde{\mu}}{NK}\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbb{E}[f_i(\boldsymbol{w}_i^*)-f_i(\boldsymbol{w}_i^{r,0})+\beta\|\boldsymbol{w}_i^{r,k-1}-\boldsymbol{w}_i^{r,0}\|^2-\tfrac{\alpha}{4}\|\boldsymbol{w}_i^{r,0}-\boldsymbol{w}_i^*\|^2]\\ &\overset{(b)}{\leq}-\tfrac{2\tilde{\mu}}{N}\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*)+\tfrac{\alpha}{4}\|\boldsymbol{w}_i^{r,0}-\boldsymbol{w}_i^*\|^2)+2\beta\tilde{\mu}\Theta^r,\end{aligned}\tag{61}$$

where Lemma 14 is used for (a) and Lemma 13 is used for (b). By using (61), (60) is bounded by

$$\begin{aligned}&\tfrac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r+1,0}-\boldsymbol{w}_i^*\|^2\\ &\leq\tfrac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,0}-\boldsymbol{w}_i^*\|^2-\tfrac{2\tilde{\mu}}{N}\mathbb{E}[\sum_{i\in\mathcal{N}}\{f_i(\boldsymbol{w}_i^{r,0})-f_i(\boldsymbol{w}_i^*)+\tfrac{\alpha}{4}\|\boldsymbol{w}_i^{r,0}-\boldsymbol{w}_i^*\|^2\}]+2\beta\tilde{\mu}\Theta^r+4\beta^2\tilde{\mu}^2\Theta^r+4\tilde{\mu}^2\Gamma^r\\ &\quad+\tfrac{16\beta\tilde{\mu}^2}{N}\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))+\tfrac{12\tilde{\mu}^2\sigma^2}{NK}\\ &\leq(1-\tfrac{\alpha\tilde{\mu}}{2})\tfrac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,0}-\boldsymbol{w}_i^*\|^2+(2\beta\tilde{\mu}+4\beta^2\tilde{\mu}^2)\Theta^r+2\tilde{\mu}^2\Gamma^{r-1}\\ &\quad+((16+\tfrac{4}{E_{\min}})\beta\tilde{\mu}^2-2\tilde{\mu})\tfrac{1}{N}\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))+\tfrac{12\tilde{\mu}^2\sigma^2}{KN}.\end{aligned}\tag{62}$$

Multiplying $9\tilde{\mu}^2$ to the inequality w.r.t. Γ^r in Lemma 1 is resulted in

$$9\tilde{\mu}^2\Gamma^r\leq 9(1-\tfrac{\alpha\tilde{\mu}}{2})\tilde{\mu}^2\Gamma^{r-1}+\tfrac{9}{2}(\alpha\tilde{\mu}-1)\tilde{\mu}^2\Gamma^{r-1}+\tfrac{9\beta\tilde{\mu}^2}{NE_{\min}}\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*)).\tag{63}$$

Multiplying $3\beta\tilde{\mu}$ to Lemma 3 is resulted in

$$\begin{aligned}3\beta\tilde{\mu}\Theta^r&\leq 27\beta\tilde{\mu}^3\Gamma^r+\tfrac{108\beta^2\tilde{\mu}^3}{N}\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,k})]-f_i(\boldsymbol{w}_i^*))+\tfrac{63\beta\tilde{\mu}^3\sigma^2}{K}\\ &\overset{(a)}{\leq}\tfrac{27}{2}\beta\tilde{\mu}^3\Gamma^{r-1}+\tfrac{\beta^2\tilde{\mu}^3}{N}(108+\tfrac{27}{E_{\min}})\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,k})]-f_i(\boldsymbol{w}_i^*))+\tfrac{63\beta\tilde{\mu}^3\sigma^2}{K},\end{aligned}\tag{64}$$

where Lemma 1 is used for (a). Adding three inequalities (62), (63), (64), we get

$$\begin{aligned}&\tfrac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r+1,0}-\boldsymbol{w}_i^*\|^2+9\tilde{\mu}^2\Gamma^r\\ &\leq(1-\tfrac{\alpha\tilde{\mu}}{2})\{\tfrac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,0}-\boldsymbol{w}_i^*\|^2+9\tilde{\mu}^2\Gamma^{r-1}\}+\{\tfrac{9}{2}\alpha\tilde{\mu}+\tfrac{27}{2}\beta\tilde{\mu}-\tfrac{5}{2}\}\tilde{\mu}^2\Gamma^{r-1}+(-\beta\tilde{\mu}+4\beta^2\tilde{\mu}^2)\Theta^r\\ &+\{-2+16\beta\tilde{\mu}+108\beta^2\tilde{\mu}^2+\tfrac{1}{E_{\min}}(13\beta\tilde{\mu}+27\beta^2\tilde{\mu}^2)\}\tfrac{\tilde{\mu}}{N}\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))+\tfrac{\tilde{\mu}^2\sigma^2}{K}(\tfrac{12}{N}+63\beta\tilde{\mu}).\end{aligned}\tag{65}$$

When we select step-size following $\tilde{\mu}\in(0,\min(\frac{1}{27\beta},\frac{1}{3\alpha}))$, it is guaranteed that $\{\frac{9}{2}\alpha\tilde{\mu}+\frac{27}{2}\beta\tilde{\mu}-\frac{5}{2}\}<0$, $(-\beta\tilde{\mu}+4\beta^2\tilde{\mu}^2)<0$, $(-2+16\beta\tilde{\mu}+108\beta^2\tilde{\mu}^2)<-1$, $(13\beta\tilde{\mu}+27\beta^2\tilde{\mu}^2)<1$, and $(\frac{12}{N}+63\beta\tilde{\mu})<(\frac{12}{N}+3)$. Hence, (65) is bounded by

$$\begin{aligned}&\tfrac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r+1,0}-\boldsymbol{w}_i^*\|^2+9\tilde{\mu}^2\Gamma^r\\ &\leq(1-\tfrac{\alpha\tilde{\mu}}{2})\{\tfrac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r,0}-\boldsymbol{w}_i^*\|^2+9\tilde{\mu}^2\Gamma^{r-1}\}-(1-\tfrac{1}{E_{\min}})\tfrac{\tilde{\mu}}{N}\sum_{i\in\mathcal{N}}(\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0})]-f_i(\boldsymbol{w}_i^*))+\tfrac{\tilde{\mu}^2\sigma^2}{K}(\tfrac{12}{N}+3).\end{aligned}\tag{66}$$

□

By using Lemma 12 and Lemma 4 for strongly convex functions ($\alpha>0$), $\mu\in[0,\min(\frac{1}{27\beta K},\frac{1}{3\alpha K}))$, $R\geq\max(\frac{27\beta}{2\alpha},\frac{3}{2})$, and it is assumed to be $E_{\min}\geq 2$, we obtain convergence rate as

$$\mathbb{E}[\tfrac{1}{N}\sum_{i\in\mathcal{N}}(f_i(\boldsymbol{w}_i^R)-f_i(\boldsymbol{w}_i^*))]\leq\mathcal{O}\Big(\tfrac{E_{\min}}{E_{\min}-1}\{\alpha D_0^2\exp(-\min(\tfrac{\alpha}{27\beta K},\tfrac{1}{3K})R)+\tfrac{\sigma^2}{\alpha RK}(3+\tfrac{12}{N})\}\Big),\tag{67}$$

where $D_0^2=\frac{1}{N}\sum_{i\in\mathcal{N}}(\|f_i(\boldsymbol{w}_i^{1,0})-f_i(\boldsymbol{w}_i^*)\|^2+\|\nabla f_i(\boldsymbol{w}_i^*)-\mathbb{E}[\boldsymbol{c}_i^{1,0}]\|^2)$.

Meanwhile for general convex functions, integrating Lemma 9 and Lemma 4 with $\alpha=0$, $\mu\in[0,\frac{1}{27\beta K})$, $R\geq 1$, is resulted in

$$\mathbb{E}[\tfrac{1}{N}\sum_{i\in\mathcal{N}}(f_i(\boldsymbol{w}_i^R)-f_i(\boldsymbol{w}_i^*))]\leq\mathcal{O}\Big(\tfrac{E_{\min}}{E_{\min}-1}\{\tfrac{\sigma D_0}{\sqrt{RKN}}\sqrt{3+\tfrac{12}{N}}+\tfrac{27\beta D_0^2}{R}\}\Big).\tag{68}$$

As shown in this analysis, the convergence speed is regulated by a node with the smallest number of connecting nodes,$E_{\min}$. E.g., when a network has a topology with small $E_{\min}$, such that a line connects N nodes (then $E_{\min} = 1$), the convergence rate will be slow. As an effect of our algorithm including variance reduction, stable convergence is expected even if each node has the statistically biased data subset. Our convergence analysis does not make any specific assumption on the data distribution bias.

In next subsection, we show the convergence analysis for non-convex functions.

C.2. Proof for non-convex function

We now analyze the convergence rate of the ECL-ISVR (PDMM-ISVR/ADMM-ISVR) with a potentially non-convex cost function. First, we will bound the update variance in control variates by taking communication lag into account is bounded by Lemma 5, and variance of the variable update for each round is bounded in Lemma 6, and the local node drift is bounded in Lemma 7. Combining three lemmas gives us the progress made in one round in Lemma 8. The final convergence rate is obtained from this one round progress with Lemma 9.

Preliminaries: Similar to the convergence analysis for convex functions as shown in subsection C.1, some new variables used throughout this subsection are introduced. In ECL-ISVR (PDMM-ISVR/ADMM-ISVR), control variates are obtained by (25) and (24). From (11) and (13), expectation of these calculated control variates on round r are given by

$$\begin{aligned}
\mathbb{E}[\boldsymbol{c}_i^r] &= \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}} \mathbb{E}[g_i(\boldsymbol{w}_i^{r,k-1})] = \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}} \nabla f_i(\boldsymbol{w}_i^{r,k-1}),\\
\mathbb{E}[\boldsymbol{c}_{i|j}^r] &= \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}} \mathbb{E}[g_j(\boldsymbol{w}_i^{r,k-1})] = \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}} \nabla f_j(\boldsymbol{w}_i^{r,k-1}),\\
\mathbb{E}[\bar{\boldsymbol{c}}_i^r] &= \textstyle\sum_{j\in\mathcal{E}_i}\sum_{k\in\mathcal{K}} \frac{1}{E_i K}\nabla f_j(\boldsymbol{w}_i^{r,k-1}),
\end{aligned}$$

where node index is summarized by $\{1,\dots,N\}\in\mathcal{N}$, edge index connected to the i-th node is summarized by $\{\mathcal{E}_i(1),\dots,\mathcal{E}_i(E_i)\}\in\mathcal{E}_i$, and inner loop index set on each round is summarized by $\{1,\dots,K\}\in\mathcal{K}$. To represent control variate changes for each round update, we define $\xi_{i|j}^r$ and ξ_i^r as

$$\begin{aligned}
\xi_{i|j}^r &= \|\nabla f_i(\boldsymbol{w}_i^{r,0}) - \mathbb{E}[\boldsymbol{c}_{i|j}^r]\|^2,\\
\xi_i^r &= \|\nabla f_i(\boldsymbol{w}_i^{r,0}) - \mathbb{E}[\boldsymbol{c}_i^r]\|^2.
\end{aligned}$$

Average of $\xi_{i|j}^r$ over cross nodes/edges is defined by

$$\Xi^r = \tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}} \frac{1}{E_i}\sum_{j\in\mathcal{E}_i}\xi_{i|j}^r.$$

The client drift from the starting point is defined by

$$\Theta^r = \tfrac{1}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbb{E}\|\boldsymbol{w}_i^{r,k} - \boldsymbol{w}_i^{r,0}\|^2.$$

For simple notation, the minimum number of edges associated with a node is defined by

$$E_{\min} = \min(E_i) \qquad (i\in\mathcal{N}).$$

We often use the step size scaled by inner loop iteration K, noted by

$$\tilde{\mu} = \mu K \geq 0.$$

For each round after K times inner loop iterations, variables are updated as

$$\begin{aligned}
\boldsymbol{w}_i^{r+1,0} &= \boldsymbol{w}_i^{r,K},\\
\bar{\boldsymbol{c}}_i^{r+1,0} &= \bar{\boldsymbol{c}}_i^{r,K},\\
\boldsymbol{c}_{i|j}^{r+1,0} &= \boldsymbol{c}_{i|j}^{r,K}.\\
\boldsymbol{c}_i^{r+1,0} &= \boldsymbol{c}_i^{r,K},
\end{aligned} \tag{69}$$

Update variance of control variate under asynchronous communication: We will bound update variance of expected control variate for each round in asynchronous decentralized communication, where its communication schedule is shown in Fig. 1 and explained in section 2.

Lemma 5. *For local node update using Alg. 2 with asynchronous communication, $\{\xi_i^r, \xi_{i|j}^r, \Xi^r\}$ are bounded in any round r and any step-size $\tilde{\mu} = \mu K \geq 0$ as*

$$\begin{aligned}
\xi_i^r &\leq \tfrac{\beta^2}{K}\textstyle\sum_{k\in\mathcal{K}}\mathbb{E}\|\boldsymbol{w}_i^{r,k-1} - \boldsymbol{w}_i^{r,0}\|^2,\\
\xi_{i|j}^r &\leq \tfrac{1}{2}\xi_{i|j}^{r-1} + \tfrac{\beta^2}{2K}\textstyle\sum_{k\in\mathcal{K}}\mathbb{E}\|\boldsymbol{w}_i^{r,k-1} - \boldsymbol{w}_i^{r,0}\|^2,\\
\Xi^r &\leq \tfrac{1}{2}\Xi^{r-1} + \tfrac{\beta^2}{2E_{\min}}\Theta^r.
\end{aligned}$$

Proof. ξ_i^r is bounded by

$$\begin{aligned}\xi_i^r &= \mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0}) - \mathbb{E}[\boldsymbol{c}_i^r]\|^2 \\ &= \mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0}) - \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}} \nabla f_i(\boldsymbol{w}_i^{r,k-1})\|^2 \\ &\overset{(a)}{\leq} \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}} \mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0}) - \nabla f_i(\boldsymbol{w}_i^{r,k-1})\|^2 \\ &\overset{(b)}{\leq} \tfrac{\beta^2}{K}\textstyle\sum_{k\in\mathcal{K}} \mathbb{E}\|\boldsymbol{w}_i^{r,k-1} - \boldsymbol{w}_i^{r,0}\|^2, \end{aligned} \tag{70}$$

where Jensen's inequality is used for (a) and (D1) β-Lipschitz smoothness is used for (b). Since communication for each edge is conducted once per K inner iterations at random timing on each round, the expectation of $\boldsymbol{c}_{i|j}^r$ under this asynchronous communication is represented in a recursive update manner as

$$\mathbb{E}[\boldsymbol{c}_{i|j}^r] = \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}}\{\tfrac{k-1}{K}\mathbb{E}[\boldsymbol{c}_{i|j}^{r-1}] + \tfrac{K-k+1}{K}\nabla f_j(\boldsymbol{w}_i^{r,k-1})\}, \tag{71}$$

where this indicates that expectation computation is conducted by varying communication timing $k \in \mathcal{K}$, where $\boldsymbol{c}_{i|j}$ is not updated for $(k-1)$ times, and then the dual variable $\boldsymbol{y}_{i|j}$ is transmitted from the node j and $\boldsymbol{c}_{i|j}$ is updated, and it is used for remaining $K-(k-1)$ times. By using (71), $\xi_{i|j}^r$ is bounded by

$$\begin{aligned}\xi_{i|j}^r &= \mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0}) - \mathbb{E}[\boldsymbol{c}_{i|j}^r]\|^2 \\ &= \mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0}) - \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}}\{\tfrac{k-1}{K}\mathbb{E}[\boldsymbol{c}_{i|j}^{r-1}] + \tfrac{K-k+1}{K}\nabla f_j(\boldsymbol{w}_i^{r,k-1})\}\|^2 \\ &\overset{(a)}{\leq} \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}}\{\tfrac{k-1}{K}\|\nabla f_i(\boldsymbol{w}_i^{r,0}) - \mathbb{E}[\boldsymbol{c}_{i|j}^{r-1}]\|^2 + \tfrac{K-k+1}{K}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0}) - \nabla f_j(\boldsymbol{w}_i^{r,k-1})\|^2\} \\ &\overset{(b)}{\leq} \psi\xi_{i|j}^{r-1} + (1-\psi)\beta^2\tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}} \mathbb{E}\|\boldsymbol{w}_i^{r,k-1} - \boldsymbol{w}_i^{r,0}\|^2, \end{aligned} \tag{72}$$

where Jensen's inequality is used in (a), and (D1) β-Lipschitz smooth property and ψ in following is used in (b),

$$\psi = \begin{cases} \frac{K-1}{2K} & (K \text{ is even number}) \\ \frac{1}{2} & (K \text{ is odd number}) \end{cases}.$$

For simple notation, we assume that K is odd number from here, then $\psi = \frac{1}{2}$.

Then, Ξ^r is bounded by

$$\begin{aligned}\Xi^r &= \tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\tfrac{1}{E_i}\sum_{j\in\mathcal{E}_i}\xi_{i|j}^r \\ &\leq \tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\tfrac{1}{E_i}\sum_{j\in\mathcal{E}_i}\{\tfrac{1}{2}\xi_{i|j}^{r-1} + \tfrac{\beta^2}{2K}\textstyle\sum_{k\in\mathcal{K}}\mathbb{E}\|\boldsymbol{w}_i^{r,k-1} - \boldsymbol{w}_i^{r,0}\|^2\} \\ &= \tfrac{1}{2}\Xi^{r-1} + \tfrac{\beta^2}{2E_{\min}}\Theta^r. \end{aligned} \tag{73}$$

□

Variance of one round We will bound one round update variance of primal model variable.

Lemma 6. *We can bound the variance of one round local node update in any round r and any step-size $\tilde{\mu} = K\mu \geq 0$:*

$$\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r+1,0} - \boldsymbol{w}_i^{r,0}\|^2 \leq 8\tilde{\mu}^2\beta^2\Theta^r + 4\tilde{\mu}^2\Xi^r + \tfrac{4\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2 + \tfrac{9\tilde{\mu}^2\sigma^2}{NK}.$$

Proof. Local node update using (23) for one round is bounded by

$$
\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\Delta \boldsymbol{w}_i^r\|^2\\
&=\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\boldsymbol{w}_i^{r+1,0}-\boldsymbol{w}_i^{r,0}\|^2\\
&=\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|-\tfrac{\tilde{\mu}}{K}\textstyle\sum_{k\in\mathcal{K}}(g_i(\boldsymbol{w}_i^{r,k-1})+\bar{\boldsymbol{c}}_i^{r,k-1}-\boldsymbol{c}_i^{r,k-1})\|^2\\
&\overset{(a)}{\leq}\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|-\tfrac{\tilde{\mu}}{K}\textstyle\sum_{k\in\mathcal{K}}(\nabla f_i(\boldsymbol{w}_i^{r,k-1})+\mathbb{E}[\bar{\boldsymbol{c}}_i^r]-\mathbb{E}[\boldsymbol{c}_i^r])\|^2+\tfrac{9\tilde{\mu}^2\sigma^2}{NK}\\
&\overset{(b)}{\leq}\tfrac{\tilde{\mu}^2}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,k-1})+\mathbb{E}[\bar{\boldsymbol{c}}^r]-\mathbb{E}[\boldsymbol{c}_i^r]\|^2+\tfrac{9\tilde{\mu}^2\sigma^2}{NK}\\
&=\tfrac{\tilde{\mu}^2}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbb{E}\|(\nabla f_i(\boldsymbol{w}_i^{r,k-1})-\nabla f_i(\boldsymbol{w}_i^{r,0}))+(\mathbb{E}[\bar{\boldsymbol{c}}_i^r]-\nabla f_i(\boldsymbol{w}_i^{r,0}))+\nabla f_i(\boldsymbol{w}_i^{r,0})-(\mathbb{E}[\boldsymbol{c}_i^r]-\nabla f_i(\boldsymbol{w}_i^{r,0}))\|^2+\tfrac{9\tilde{\mu}^2\sigma^2}{NK}\\
&\overset{(c)}{\leq}\tfrac{4\tilde{\mu}^2}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,k-1})-\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{4\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\|\mathbb{E}[\bar{\boldsymbol{c}}_i^r]-\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{4\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\|\mathbb{E}[\boldsymbol{c}_i^r]-\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2\\
&\qquad+\tfrac{4\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{9\tilde{\mu}^2\sigma^2}{NK}\\
&\overset{(d)}{\leq}4\tilde{\mu}^2\beta^2\Theta^r+\tfrac{4\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\tfrac{1}{E_i}\sum_{j\in\mathcal{E}_i}\xi_{i|j}^r+\tfrac{4\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\xi_i^r+\tfrac{4\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{9\tilde{\mu}^2\sigma^2}{NK}\\
&\leq 4\tilde{\mu}^2\beta^2\Theta^r+4\tilde{\mu}^2\Xi^r+\tfrac{4\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\xi_i^r+\tfrac{4\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{9\tilde{\mu}^2\sigma^2}{NK}\\
&\overset{(e)}{\leq}8\tilde{\mu}^2\beta^2\Theta^r+4\tilde{\mu}^2\Xi^r+\tfrac{4\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{9\tilde{\mu}^2\sigma^2}{NK},
\end{aligned}
\tag{74}
$$

where Lemma 11 is used for (a), Jensen's inequality is used for (b), Lemma 10 is used for (c), Jensen's inequality is used for (d), and Lemma 5 is used for (e). □

Bounding the drift: We will bound local node drift Θ^r.

Lemma 7. *We can bound the local node drift Θ^r in any round r and any step-size $\tilde{\mu}=\mu K$ satisfying $\tilde{\mu}\in[0,\frac{1}{32\beta})$,*

$$
\Theta^r\leq\tfrac{3\tilde{\mu}^2\sigma^2}{K}+\tfrac{12\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+12\tilde{\mu}^2\Xi^r+12\tilde{\mu}^2\beta^2\Theta^r.
$$

Proof. A recursive bound of local node drift is given by

$$
\begin{aligned}
&\mathbb{E}\|\boldsymbol{w}_i^{r,k}-\boldsymbol{w}_i^{r,0}\|^2\\
&\leq\mathbb{E}\|\boldsymbol{w}_i^{r,k-1}-\mu(g_i(\boldsymbol{w}_i^{r,k-1})+\bar{\boldsymbol{c}}_i^{r,k-1}-\boldsymbol{c}_i^{r,k-1})-\boldsymbol{w}_i^{r,0}\|^2\\
&\overset{(a)}{\leq}\mathbb{E}\|\boldsymbol{w}_i^{r,k-1}-\mu(\nabla f_i(\boldsymbol{w}_i^{r,k-1})+\bar{\boldsymbol{c}}_i^{r,k-1}-\boldsymbol{c}_i^{r,k-1})-\boldsymbol{w}_i^{r,0}\|^2+\mu^2\sigma^2\\
&\overset{(b)}{\leq}(1+\tfrac{1}{K-1})\mathbb{E}\|\boldsymbol{w}_i^{r,k-1}-\boldsymbol{w}_i^{r,0}\|^2+K\mu^2\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,k-1})+\bar{\boldsymbol{c}}_i^{r,k-1}-\boldsymbol{c}_i^{r,k-1}\|^2+\mu^2\sigma^2\\
&=(1+\tfrac{1}{K-1})\mathbb{E}\|\boldsymbol{w}_i^{r,k-1}-\boldsymbol{w}_i^{r,0}\|^2+\mu^2\sigma^2\\
&\qquad+K\mu^2\mathbb{E}\|(\nabla f_i(\boldsymbol{w}_i^{r,k-1})-\nabla f_i(\boldsymbol{w}_i^{r,0}))+(\bar{\boldsymbol{c}}_i^{r,k-1}-\nabla f_i(\boldsymbol{w}_i^{r,0}))+\nabla f_i(\boldsymbol{w}_i^{r,0})-(\boldsymbol{c}_i^{r,k-1}-\nabla f_i(\boldsymbol{w}_i^{r,0}))\|^2\\
&\overset{(c)}{\leq}(1+\tfrac{1}{K-1}+4K\beta^2\mu^2)\mathbb{E}\|\boldsymbol{w}_i^{r,k-1}-\boldsymbol{w}_i^{r,0}\|^2+\mu^2\sigma^2+4K\mu^2\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2\\
&\qquad+4K\mu^2\mathbb{E}\|\bar{\boldsymbol{c}}_i^{r,k-1}-\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+4K\mu^2\mathbb{E}\|\boldsymbol{c}_i^{r,k-1}-\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2,
\end{aligned}
\tag{75}
$$

where Lemma (11) is used for (a), Lemma 10 is used for (b), and Lemma 10 is used for (c) again. Unrolling (75), we obtain

following bound:

$$\begin{aligned}
\Theta^r &= \tfrac{1}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbb{E}\|\boldsymbol{w}_i^{r,k}-\boldsymbol{w}_i^{r,0}\|^2 \\
&\le (\tfrac{\tilde{\mu}^2\sigma^2}{K^2}+\tfrac{4\tilde{\mu}^2}{NK}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{4\tilde{\mu}^2}{NK}\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbb{E}\|\bar{\boldsymbol{c}}_i^{r,k-1}-\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2 \\
&\quad+\tfrac{4\tilde{\mu}^2}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbb{E}\|\boldsymbol{c}_i^{r,k-1}-\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2)\cdot\{\sum_{\tau=0}^{k-1}(1+\tfrac{1}{K-1}+\tfrac{4\beta^2\tilde{\mu}^2}{K})^\tau\} \\
&\overset{(a)}{\le} (\tfrac{\tilde{\mu}^2\sigma^2}{K^2}+\tfrac{4\tilde{\mu}^2}{NK}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{4\tilde{\mu}^2}{NK}\sum_{i\in\mathcal{N}}\tfrac{1}{E_i}\sum_{j\in\mathcal{E}_i}\xi_{i|j}^r+\tfrac{4\tilde{\mu}^2}{NK}\sum_{i\in\mathcal{N}}\xi_i^r)3K \\
&\le (\tfrac{\tilde{\mu}^2\sigma^2}{K^2}+\tfrac{4\tilde{\mu}^2}{NK}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{4\tilde{\mu}^2}{K}\Xi^r+\tfrac{4\tilde{\mu}^2}{K}\tfrac{1}{N}\sum_{i\in\mathcal{N}}\xi_i^r)3K \\
&\overset{(b)}{\le} (\tfrac{\tilde{\mu}^2\sigma^2}{K^2}+\tfrac{4\tilde{\mu}^2}{NK}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{4\tilde{\mu}^2}{K}\Xi^r+\tfrac{4\tilde{\mu}^2\beta^2}{K}\Theta^r)3K \\
&= \tfrac{3\tilde{\mu}^2\sigma^2}{K}+\tfrac{12\tilde{\mu}^2}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+12\tilde{\mu}^2\Xi^r+12\tilde{\mu}^2\beta^2\Theta^r, \qquad (76)
\end{aligned}$$

where Jensen's inequality and the fact that $\sum_{\tau=0}^{k-1}(1+\frac{1}{K-1})^\tau < 3K$ when $\tilde{\mu}\in[0,\frac{1}{32\beta})$, which is applied in Lemma 17 in (Karimireddy et al., 2020), is used for (a), Lemma 5 is used for (b). □

Progress made in each round: We bound all errors in a round.

Lemma 8. *Following holds in any round r and any step-size $\tilde{\mu}=K\mu$ satisfying $\tilde{\mu}\in[0,\frac{1}{32\beta})$,*

$$\left(\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0}+\Delta\boldsymbol{w}_i^r)]+12\beta\tilde{\mu}^2\Xi^r\right)\le\left(\tfrac{1}{N}\sum_{i\in\mathcal{N}}f_i(\boldsymbol{w}_i^{r,0})+12\beta\tilde{\mu}^2\Xi^{r-1}\right)-\tfrac{\tilde{\mu}}{3N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{\tilde{\mu}\sigma^2}{4K}(1+\tfrac{18}{N}).$$

Proof. Staring from the Lipschitz smoothness of $\{f_i\}$ and taking expectation by $r-1$ round provides

$$\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0}+\Delta\boldsymbol{w}_i^r)]\le\tfrac{1}{N}\sum_{i\in\mathcal{N}}f_i(\boldsymbol{w}_i^{r,0})+\tfrac{1}{N}\sum_{i\in\mathcal{N}}\left\langle\nabla f_i(\boldsymbol{w}_i^{r,0}),\mathbb{E}[\Delta\boldsymbol{w}_i^r]\right\rangle+\tfrac{\beta}{2N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\Delta\boldsymbol{w}_i^r\|^2, \quad (77)$$

where $\Delta\boldsymbol{w}_i^r$ is given by

$$\Delta\boldsymbol{w}_i^r=-\tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}}\tilde{\mu}(g_i(\boldsymbol{w}_i^{r,k-1})+\bar{\boldsymbol{c}}_i^{r,k}-\boldsymbol{c}_i^{r,k}),$$

and its expectation satisfies

$$\mathbb{E}[\Delta\boldsymbol{w}_i^r]=-\tfrac{\tilde{\mu}}{K}\textstyle\sum_{k\in\mathcal{K}}g_i(\boldsymbol{w}_i^{r,k-1}).$$

From (77), update difference is bounded by

$$\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0}+\Delta\boldsymbol{w}_i^r)]-\tfrac{1}{N}\sum_{i\in\mathcal{N}}f_i(\boldsymbol{w}_i^{r,0}) \\
&\le-\tfrac{\tilde{\mu}}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\left\langle\nabla f_i(\boldsymbol{w}_i^{r,0}),\mathbb{E}[\nabla f_i(\boldsymbol{w}_i^{r,k-1})]\right\rangle+\tfrac{\beta}{2N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\Delta\boldsymbol{w}_i^r\|^2 \\
&\overset{(a)}{\le}-\tfrac{\tilde{\mu}}{NK}\textstyle\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\left\langle\nabla f_i(\boldsymbol{w}_i^{r,0}),\mathbb{E}[\nabla f_i(\boldsymbol{w}_i^{r,k-1})]\right\rangle+4\tilde{\mu}^2\beta^3\Theta^r+2\tilde{\mu}^2\beta\Xi^r \\
&\quad+\tfrac{2\tilde{\mu}^2\beta}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{9\tilde{\mu}^2\beta\sigma^2}{2NK} \\
&\overset{(b)}{\le}-\tfrac{\tilde{\mu}}{2N}\textstyle\sum_{i\in\mathcal{N}}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{\tilde{\mu}}{2}\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\|\tfrac{1}{NK}\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}(\nabla f_i(\boldsymbol{w}_i^{r,k-1})-\nabla f_i(\boldsymbol{w}_i^{r,0}))\|^2 \\
&\quad+4\tilde{\mu}^2\beta^3\Theta^r+2\tilde{\mu}^2\beta\Xi^r+\tfrac{2\tilde{\mu}^2\beta}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{9\tilde{\mu}^2\beta\sigma^2}{2NK} \\
&\overset{(c)}{\le}-(\tfrac{\tilde{\mu}}{2}-2\tilde{\mu}^2\beta)\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{\tilde{\mu}}{2NK}\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\|\nabla f_i(\boldsymbol{w}_i^{r,k-1})-\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2 \\
&\quad+4\tilde{\mu}^2\beta^3\Theta^r+\tilde{\mu}^2\beta\Xi^{r-1}+\tfrac{\tilde{\mu}^2\beta^3}{E_{\min}}\Theta^r+\tfrac{9\tilde{\mu}^2\beta\sigma^2}{2NK} \\
&\overset{(d)}{\le}-(\tfrac{\tilde{\mu}}{2}-2\tilde{\mu}^2\beta)\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tilde{\mu}^2\beta\Xi^{r-1}+(\tfrac{\tilde{\mu}\beta^2}{2}+4\tilde{\mu}^2\beta^3+\tfrac{\tilde{\mu}^2\beta^3}{E_{\min}})\Theta^r+\tfrac{9\tilde{\mu}^2\beta\sigma^2}{2NK}, \qquad (78)
\end{aligned}$$

where Lemma 6 is used in (a), inequality in (b) follows from observation that $-ab=\frac{1}{2}((b-a)^2-a^2)-\frac{1}{2}b^2\le\frac{1}{2}((b-a)^2-a^2)$ for any $a,b\in\mathbb{R}$, Jensen's inequality is used in (c), and (D1) β-Lipschitz smoothness is used in (d).

Multiplying $12\beta\tilde{\mu}^2$ to the Lemma 5 results in

$$12\beta\tilde{\mu}^2\Xi^r \leq 6\beta\tilde{\mu}^2\Xi^{r-1} + \tfrac{6\beta\tilde{\mu}^2\beta^2}{E_{\min}}\Theta^r, \tag{79}$$

Multiplying $2\tilde{\mu}\beta^2$ to Lemma 7 results in

$$\begin{aligned} 2\tilde{\mu}\beta^2\Theta^r &\leq \tfrac{6\tilde{\mu}^3\beta^2\sigma^2}{K} + \tfrac{24\tilde{\mu}^3\beta^2}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2 + 24\tilde{\mu}^3\beta^2\Xi^r + 24\tilde{\mu}^3\beta^4\Theta^r \\ &\overset{(a)}{\leq} \tfrac{6\tilde{\mu}^3\beta^2\sigma^2}{K} + \tfrac{24\tilde{\mu}^3\beta^2}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2 + 12\tilde{\mu}^3\beta^2\Xi^{r-1} + (24+\tfrac{12}{E_{\min}})\tilde{\mu}^3\beta^4\Theta^r, \end{aligned} \tag{80}$$

where Lemma 5 is used in (a). Adding (78), (79), and (80) is resulted in

$$\begin{aligned} &\Big(\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0}+\Delta\boldsymbol{w}_i^r)]+12\beta\tilde{\mu}^2\Xi^r\Big) \leq \Big(\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}} f_i(\boldsymbol{w}_i^{r,0})+12\beta\tilde{\mu}^2\Xi^{r-1}\Big)+(-5\tilde{\mu}^2\beta+12\tilde{\mu}^3\beta^2)\Xi^{r-1}+(6\tilde{\mu}\beta+\tfrac{9}{2N})\tfrac{\tilde{\mu}^2\beta\sigma^2}{K} \\ &+\{-\tfrac{3}{2}\tilde{\mu}\beta^2+4\tilde{\mu}^2\beta^3+24\tilde{\mu}^3\beta^4+\tfrac{1}{E_{\min}}(7\tilde{\mu}^2\beta^3+12\tilde{\mu}^3\beta^4)\}\Theta^r-(\tfrac{\tilde{\mu}}{2}-2\tilde{\mu}^2\beta-24\tilde{\mu}^3\beta^2)\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2. \end{aligned} \tag{81}$$

When selecting the step-size $\tilde{\mu} = \mu K \leq \frac{1}{32\beta}$, it is guaranteed that $(-5\tilde{\mu}^2\beta+12\tilde{\mu}^3\beta^2) < 0$, $\{-\frac{3\tilde{\mu}\beta^2}{2}+4\tilde{\mu}^2\beta^3+24\tilde{\mu}^3\beta^4+\frac{1}{E_{\min}}(7\tilde{\mu}^2\beta^3+12\tilde{\mu}^3\beta^4)\} < 0$, $\frac{\tilde{\mu}}{3} < (\frac{\tilde{\mu}}{2}-2\tilde{\mu}^2\beta-24\tilde{\mu}^3\beta^2) \leq \frac{9\tilde{\mu}}{24}$, and $(6\tilde{\mu}\beta+\frac{9}{2N}) \leq (\frac{1}{4}+\frac{9}{2N})$. Hence, (81) is bounded by

$$\Big(\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}[f_i(\boldsymbol{w}_i^{r,0}+\Delta\boldsymbol{w}_i^r)]+12\beta\tilde{\mu}^2\Xi^r\Big) \leq \Big(\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}} f_i(\boldsymbol{w}_i^{r,0})+12\beta\tilde{\mu}^2\Xi^{r-1}\Big)-\tfrac{\tilde{\mu}}{3N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2+\tfrac{\tilde{\mu}\sigma^2}{4K}(1+\tfrac{18}{N}). \tag{82}$$

□

By using Lemma 9 and Lemma 8 for β-Lipschitz smooth non-convex function, $\mu \in [0, \frac{1}{32\beta K})$, $R \geq 1$, we obtain convergence rate as

$$\tfrac{1}{N}\textstyle\sum_{i\in\mathcal{N}}\mathbb{E}\|\nabla f_i(\boldsymbol{w}_i^{r,0})\|^2 \leq \mathcal{O}\Big(\tfrac{3\sigma\sqrt{Q_0}}{2\sqrt{RKN}}\sqrt{1+\tfrac{18}{N}} + \tfrac{3\beta Q_0}{R}\}\Big), \tag{83}$$

where $Q_0 = \frac{1}{N}\sum_{i\in\mathcal{N}}(f_i(\boldsymbol{w}_i^{1,0})-f_i(\boldsymbol{w}_i^*))$.

C.3. Technical lemmas

We now summarize several technical lemmas needed for the convergence analysis. Since our convergence analysis strategy follows the approach used in the SCAFFOLD paper (Karimireddy et al., 2020), many of lemmas shown in this subsection are taken from (Karimireddy et al., 2020). Therefore, detailed proofs are omitted.

C.3.1. Technical lemmas used in both convex and non-convex functions

The following lemma is useful for unrolling recursions and deriving convergence rate for the general convex function and non-convex function.

Lemma 9 (Sub-linear convergence rate)**.** *For every non-negative sequence $\{d^{r-1}\}_{r\geq 1}$ and any parameters $\mu_{\max} \geq 0, c_1 \geq 0, c_2 \geq 0, R \geq 0$, there exists a constant step-size $\mu \leq \mu_{\max}$ such that satisfies*

$$\Psi_R = \frac{1}{R+1}\sum_{r=1}^{R+1}\left(\frac{d^{r-1}}{\mu} - \frac{d^r}{\mu} + c_1\mu + c_2\mu^2\right) \leq \frac{d^0}{\mu_{\max}(R+1)} + \frac{2\sqrt{c_1 d^0}}{\sqrt{R+1}} + 2\left(\frac{d^0}{R+1}\right)^{\frac{2}{3}} c_2^{\frac{1}{3}}.$$

Proof. See Lemma 2 in (Karimireddy et al., 2020). □

Next, a relaxed triangle inequality, which is true for the squared L_2 norm, is introduced.

Lemma 10 (Relaxed triangle inequality)**.** *Let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_\tau\}$ be set of τ vectors. Then, following inequality is true for any $a > 0$,*

1. $\|\boldsymbol{v}_i + \boldsymbol{v}_j\|^2 \leq (1+a)\|\boldsymbol{v}_i\|^2 + (1+\frac{1}{a})\|\boldsymbol{v}_i\|^2, \qquad (a > 0)$
2. $\|\sum_{i=1}^{\tau} \boldsymbol{v}_i\|^2 \leq \tau \sum_{i=1}^{\tau} \|\boldsymbol{v}_i\|^2.$

Proof. The first statement is derived from the following identity for any $a > 0$,

$$\|\boldsymbol{v}_i + \boldsymbol{v}_j\|^2 \leq (1+a)\|\boldsymbol{v}_i\|^2 + (1+\tfrac{1}{a})\|\boldsymbol{v}_i\|^2 - \|\sqrt{a}\boldsymbol{v}_i + \tfrac{1}{\sqrt{a}}\boldsymbol{v}_j\|^2.$$

For the second inequality, we use the convexity of $\boldsymbol{v} \to \|\boldsymbol{v}\|^2$ and Jensen's inequality as

$$\|\tfrac{1}{\tau}\sum_{i=1}^{\tau} \boldsymbol{v}_i\|^2 \leq \tfrac{1}{\tau}\sum_{i=1}^{\tau} \|\boldsymbol{v}_i\|^2.$$

□

The following lemma is an elementary lemma about expectations of the norm of random vectors.

Lemma 11 (Separating mean and variance)**.** *Let $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_\tau\}$ be set of τ random variables, which are not necessarily independent. First suppose that the variance of $\boldsymbol{\phi}_i$ from its mean $\mathbb{E}[\boldsymbol{\phi}_i] = \boldsymbol{\varphi}_i$ is bounded by $\mathbb{E}\|\boldsymbol{\phi}_i - \boldsymbol{\varphi}_i\|^2 \leq \sigma^2$. Then, following inequality satisfies,*

$$\mathbb{E}\|\sum_{i=1}^{\tau} \boldsymbol{\phi}_i\|^2 \leq \|\sum_{i=1}^{\tau} \boldsymbol{\varphi}_i\|^2 + \tau^2\sigma^2.$$

Instead, let assume that their conditional mean is $\mathbb{E}[\boldsymbol{\phi}_i|\boldsymbol{\phi}_{i-1}, \ldots, \boldsymbol{\phi}_1] = \boldsymbol{\varphi}_i$, i.e., the variable $\{\boldsymbol{\phi}_i - \boldsymbol{\varphi}_i\}$ form a martingale difference sequence, and the variable is bounded by $\mathbb{E}\|\boldsymbol{\phi}_i - \boldsymbol{\varphi}_i\|^2 \leq \sigma^2$ as before. Then, a tighter bound is obtained,

$$\mathbb{E}\|\sum_{i=1}^{\tau} \boldsymbol{\phi}_i\|^2 \leq 2\|\sum_{i=1}^{\tau} \boldsymbol{\varphi}_i\|^2 + 2\tau\sigma^2.$$

Proof. See Lemma 4 in (Karimireddy et al., 2020). □

C.3.2. Technical lemmas used in only convex functions

The following lemma is useful for unrolling recursions and deriving convergence rate for the strongly convex case ($\alpha > 0$).

Lemma 12 (Linear convergence rate)**.** *For every non-negative sequence $\{d^{r-1}\}_{r\geq 1}$ and any parameters $h \geq 0, \mu_{\max} \in (0, 1/h], c \geq 0, R \geq \frac{1}{2h\mu_{\max}}$, there exists a constant step-size $\mu \leq \mu_{\max}$ and weight $\omega^r = (1-h\mu)^{1-r}$ such that for $\Omega_R = \sum_{r=1}^{R+1} \omega^r$*

$$\Psi_R = \frac{1}{\Omega_R}\sum_{r=1}^{R+1}\left(\frac{\omega^r}{\mu}(1-h\mu)d^{r-1} - \frac{\omega^r}{\mu}d^r + c\mu\omega^r\right) \leq \mathcal{O}\left(hd^0\exp(-h\mu_{\max}R) + \frac{c}{hR}\right).$$

Proof. See Lemma 1 in (Karimireddy et al., 2020). □

Lemma 13 (Upper bound of smooth convex function to the stationary point)**.** *Suppose that the function $\{f_i\}$ satisfies (D1) β-Lipschitz smooth and (D2) α-convex ($\alpha \geq 0$). The output of f_i and its stationary point $\{\boldsymbol{w}_i^*\}$ imply following:*

$$\|\nabla f_i(\boldsymbol{w}_i) - \nabla f_i(\boldsymbol{w}_i^*)\|^2 \leq 2\beta(f(\boldsymbol{w}_i) - f(\boldsymbol{w}_i^*))$$

Proof. See Theorem 2.1.5 in (Nesterov et al., 2018). □

Lemma 14 (Perturbed strong convexity)**.** *Suppose that function $\{f_i\}$ satisfies (D1) β-Lipschitz smooth and (D2) α-strongly convex ($\alpha > 0$). The following inequality holds for any three points $\{\boldsymbol{r}_i, \boldsymbol{w}_i, \boldsymbol{u}_i\}$, in the domain of f_i:*

$$\langle \nabla f_i(\boldsymbol{r}_i), \boldsymbol{w}_i - \boldsymbol{u}_i \rangle \geq f_i(\boldsymbol{w}_i) - f_i(\boldsymbol{u}_i) + \tfrac{\alpha}{4}\|\boldsymbol{u}_i - \boldsymbol{w}_i\|^2 - \beta\|\boldsymbol{w}_i - \boldsymbol{r}_i\|^2.$$

Proof. See Lemma 5 in (Karimireddy et al., 2020). □