Towards Efficient Replay in Federated Incremental Learning

Supplementary Material

A. Dataset

Class-Incremental Task Dataset: New classes are incrementally introduced over time. The dataset starts with a subset of classes, and new classes are added in subsequent stages, allowing models to learn and adapt to an increasing number of classes.

- **CIFAR10:** A dataset with 10 object classes, including various common objects, animals, and vehicles. It consists of 50,000 training images and 10,000 test images.
- **CIFAR100:** Similar to CIFAR10, but with 100 fine-grained object classes. It has 50,000 training images and 10,000 test images.
- **Tiny-ImageNet:** A subset of the ImageNet dataset with 200 object classes. It contains 100,000 training images, 10,000 validation images, and 10,000 test images.

Domain-Incremental Task Dataset: New domains are introduced gradually. The dataset initially contains samples from a specific domain, and new domains are introduced at later stages, enabling models to adapt and generalize to new unseen domains.

- **Digit10:** Digit-10 dataset contains 10 digit categories in four domains: **MNIST**[19], **EMNIST**[3], **USPS**[13], **SVHN**[34].Each dataset is a digit image classification dataset of 10 classes in a specific domain, such as handwriting style.
  - **MNIST:** A dataset of handwritten digits with a training set of 60,000 examples and a test set of 10,000 examples.
  - **EMNIST:** An extended version of MNIST that includes handwritten characters (letters and digits) with a training set of 240,000 examples and a test set of 40,000 examples.
  - **USPS:** The United States Postal Service dataset consists of handwritten digits with a training set of 7,291 examples and a test set of 2,007 examples.
  - **SVHN:** The Street View House Numbers dataset contains images of house numbers captured from Google Street View, with a training set of 73,257 examples and a test set of 26,032 examples.
- **Office31:** A dataset with images from three different domains: Amazon, Webcam, and DSLR. It consists of 31 object categories, with each domain having around 4,100 images.
- **DomainNet:** A large-scale dataset with images from six different domains: Clipart, Painting, Real, Sketch, Quickdraw, and Infograph. It contains over 0.6 million images across 345 categories.

B. Baseline

- **Representative FL models:**
  - **FedAvg:** : It is a representative federated learning model, which aggregates client parameters in each communication. It is a simply yet effective model for federated learning.
  - **FedProx:** It is also a representative federated learning model, which is better at tackling heterogeneity in federated networks than FedAvg
- **Custom methods:**
  - **Fixed:** we train the model only from the first task and evaluate it for all the coming sequence of tasks.
  - **DANN+FL:** Here we adopt the robust adversarial-based method DANN[9]. This baseline mainly follows the domain adaptation paradigm which is different from the incremental learning setting and are often prone to issues like catastrophic forgetting.
  - **Shared:** Inspired by the multi-task learning scenario[40], we adopt all front layers before the last fully connected layer as shared layers, and use relevant different fully-connected layers to get outputs for different tasks.
- **Models for federated class-incremental learning:**
  - **FCIL:** This approach addresses the federated class-incremental learning and trains a global model by computing additional class-imbalance losses. A proxy server is introduced to reconstruct samples to help clients select the best old models for loss computation.
  - **FedCIL:** This approach employs the ACGAN backbone to generate synthetic samples to consolidate the global model and align sample features in the output layer. Authors conduct experiments in the FCIL scenario, and here we adopt it to our FDIL setting.

C. Configurations

For local training, the batch size is 64, learning rate for our models is 0.01/0.001 for {Office31, CIFAR10, CIFAR100}/{Digit10, DomainNet, Tiny-ImageNet}. For the update of the personalized informative model, the epoch is set to 40 for each client. For the multi-task learning structure in our approach, we treat all previous layers before the last fully-connected layer as share layers, and we use two different fully-connected layers to get outputs as the auxiliary classifier result and target classification result. We build the Virtual Machine(VM) to simulate the experiment environment and set up different processes to simulate different clients. The VM is configured with 8 RTX4090 and 6 2.3GHz Intel Xeon CPUs.

D. Detailed Re-Fed Framework with FedAvg

Algorithm 2: Re-Fed for FIL with FedAvg Algorithm

Input: T: communication round; K: number of clients; η: learning rate; $\{\mathcal{T}^t\}_{t=1}^n$: distributed dataset with n tasks; w: parameter of the model; v_k: personalized informative model in client k; λ: factor of information proportion.

Initialize the parameter w;
for $c = 1$ ***to*** T **do** // When the t-th new task arrives
  Server randomly selects a subset of devices S_t and send w^{t-1} to them;
  for *each selected client* $k \in S_t$ **in parallel do**
    Receive the distributed global model w^{t-1} and initializess the personalized informative model v_k^{t-1};
    Update v_k^{t-1}in s local iterations with previous local samples $\mathcal{T}_{k,local}^{t-1}$:
    $v_{k,s}^{t-1} = v_{k,s-1}^{t-1} - \eta\left(\sum_{i=1}^{M} \nabla l\left(f_{v_{k,s-1}^{t-1}}(\tilde{x}_{k,t-1}^{(i)}), \tilde{y}_{k,t-1}^{(i)}\right) + q(\lambda)(v_{k,s-1}^{t-1} - w^{t-1})\right), \; q(\lambda) = \frac{1-\lambda}{2\lambda}, \lambda \in (0, 1).$
    for *During the update of* v_k^{t-1} **do**
      Calculate the importance score for the sample $(\tilde{x}_{k,t-1}^{(i)}, \tilde{y}_{k,t-1}^{(i)})$ after total s iterations:
      $I(\tilde{x}_{k,t-1}^{(i)}) = \sum_{p=1}^{s} \frac{G^p(\tilde{x}_{k,t-1}^{(i)})}{p}.$
    end
    Cache previous samples with higher importance scores;
    Train the local model with cached samples and the new task $(\tilde{x}_{k,t}^{(i)}, \tilde{y}_{k,t}^{(i)})$ in s iterations:
    $w_{k,s}^t = w_{k,s-1}^t - \eta\left(\sum_{i=1}^{M} \nabla l\left(f_{w_{k,s-1}^t}(\tilde{x}_{k,t}^{(i)}), \tilde{y}_{k,t}^{(i)}\right)\right).$
    Send the model w_k^t back to the server.
  end
  The server aggregates the local models: $w^t = \sum_{k \in S_t} \frac{1}{|S_t|} w_k^t$.
end

E. Experimental Results

In this section, we further provide more details about the experiment results on the test accuracy and communication rounds. We record the test accuracy of the global model at training stage of each task and the communication rounds required to achieve the corresponding performance. Then, as we use a form of "Early-Emphasis" to accumulate the gradient norms and calculate the sample importance scores in Re-Fed, we compare and show results with two other methods of calculation of sample importance scores.

E.1. Detailed Results of Test Accuracy.

Table 5, 6, 7, 8 and 9 show the results of test accuracy on each incremental task in the **Acc** (Accuray) line, where "Δ" denotes the improvement of our method with other baselines. Here we measure average accuracy over all tasks on each client in the **Acc** line and highlight the best test accuracy in **bold**.

E.2. Detailed Results of Communication Round.

Table 5, 6, 7, 8 and 9 show the detailed results of communication round on each incremental task in the **CoR** (Communication Round) line and highlight the results of fewest number of communication rounds in underline.

E.3. Different Weighting Methods for Gradient Norms.

Table 10 shows the impact of using different methods to calculate the sample importance score with gradient norms in the update of personalized informative models. Here we adopt three methods: **Average Weighting:** we assign an equal weight to gradient norms from different iterations; **Early-Emphasis:** a higher weight to gradient norms in the early-training as adopted by Re-Fed; and **Late-Emphasis**: the sorting of samples with the sample importance score obtained by the method of **Early-Emphasis** is reversed.

Table 5. Performance comparisons of various methods on CIFAR10 with 5 incremental tasks (2 new classes for each task).

CIFAR10 (α = 1.0)								
Method	Target	2	4	6	8	10	Avg	$\Delta(\uparrow)$
FedAvg	Acc	92.65	76.67	42.90	40.46	26.73	55.88	2.57↑
	CoR	142	123	125	98	119	122	10↑
FedProx	Acc	92.39	74.18	39.84	37.55	25.87	53.97	4.48↑
	CoR	153	137	141	123	132	137	25↑
Fixed	Acc	92.65	62.48	36.54	24.20	19.21	47.02	11.43↑
	CoR	142	0	0	0	0	/	/
DANN+FL	Acc	93.07	77.81	44.32	36.98	24.86	55.41	3.04↑
	CoR	151	140	150	126	145	142	30↑
Shared	Acc	92.65	76.19	42.15	38.24	23.91	54.63	3.82↑
	CoR	142	117	116	83	118	115	3↑
FCIL	Acc	92.65	78.07	43.66	40.28	25.04	55.94	2.51↑
	CoR	142	125	108	92	121	118	6↑
FedCIL	Acc	**94.05**	**80.22**	46.19	35.50	27.35	56.66	1.79↑
	CoR	148	150	146	147	150	148	36↑
Re-Fed	Acc	92.65	79.23	**47.41**	**43.75**	**29.22**	**58.45**	/
	CoR	142	109	116	85	106	112	

Table 6. Performance comparisons of various methods on CIFAR100 with 10 incremental tasks (10 new classes for each task).

CIFAR100 (α = 5.0)													
Method	Target	10	20	30	40	50	60	70	80	90	100	Avg	$\Delta(\uparrow)$
FedAvg	Acc	58.70	43.72	48.69	38.28	30.81	26.16	24.90	20.72	18.97	17.21	32.82	5.57↑
	CoR	137	121	76	135	102	143	90	86	132	75	110	6↑
FedProx	Acc	56.51	42.02	48.03	39.11	32.33	27.24	26.50	20.88	19.67	18.03	33.03	5.36↑
	CoR	146	134	112	139	119	140	125	105	132	99	125	21↑
Fixed	Acc	58.70	34.52	35.09	30.37	27.01	23.96	18.18	14.78	11.47	9.27	26.34	12.05↑
	CoR	137	0	0	0	0	0	0	0	0	0	/	/
DANN+FL	Acc	58.82	44.12	46.84	39.66	31.54	27.93	24.21	24.03	21.32	19.73	33.82	4.57↑
	CoR	145	129	123	138	124	134	112	109	129	121	126	22↑
Shared	Acc	58.70	42.53	48.49	39.10	31.88	27.39	25.85	25.74	24.35	18.30	34.23	4.16↑
	CoR	137	117	82	137	113	137	103	97	135	89	115	11↑
FCIL	Acc	58.70	45.65	51.87	42.37	37.32	32.01	29.00	28.47	24.99	23.02	37.33	1.06↑
	CoR	137	123	77	134	105	140	96	88	130	73	110	6↑
FedCIL	Acc	**61.20**	**47.05**	49.66	38.14	32.69	24.11	23.90	23.99	19.89	17.98	33.86	4.53↑
	CoR	146	138	123	131	125	143	122	129	130	126	131	27↑
Re-Fed	Acc	58.70	43.66	**53.53**	**40.17**	**38.71**	**35.96**	**31.25**	**28.77**	**27.53**	**25.61**	**38.39**	/
	CoR	137	104	80	105	93	121	85	105	120	87	104	

Table 7. Performance comparisons of various methods on Tiny-ImageNet with 10 incremental tasks (20 new classes for each task).

		Tiny-ImageNet (α = 10.0)											
Method	Target	20	40	60	80	100	120	140	160	180	200	Avg	$\Delta(\uparrow)$
FedAvg	Acc	85.80	68.58	57.22	43.75	40.52	41.13	34.10	29.59	28.40	27.58	45.67	5↑
	CoR	132	143	139	125	107	97	128	121	109	98	120	7↑
FedProx	Acc	82.02	66.15	54.32	40.57	38.80	38.99	30.59	24.12	22.76	21.82	42.01	8.66↑
	CoR	127	140	142	134	120	113	114	121	110	108	123	10↑
Fixed	Acc	85.80	51.07	30.94	28.11	25.30	24.26	19.48	17.18	14.66	12.34	30.91	19.76↑
	CoR	132	0	0	0	0	0	0	0	0	0	/	/
DANN+FL	Acc	85.24	68.16	55.32	41.11	36.45	35.38	28.83	24.54	21.09	20.77	41.69	8.98↑
	CoR	138	140	141	131	124	126	137	128	121	123	131	18↑
Shared	Acc	85.80	67.21	56.49	42.05	40.17	37.59	28.61	25.90	23.89	22.19	42.99	7.68↑
	CoR	132	135	145	125	119	127	129	116	130	125	128	15↑
FCIL	Acc	85.80	71.94	61.02	50.73	44.25	**42.40**	36.96	34.51	31.36	29.58	48.86	1.81↑
	CoR	132	130	127	112	106	109	124	122	121	108	119	6↑
FedCIL	Acc	**86.43**	69.39	58.11	45.74	41.02	38.93	31.29	27.65	25.17	24.41	44.81	5.86↑
	CoR	146	144	137	121	117	126	132	140	124	129	132	19↑
Re-Fed	Acc	85.80	**72.06**	**65.29**	**52.39**	**45.93**	42.15	**38.88**	**36.95**	**35.19**	**32.07**	**50.67**	/
	CoR	132	120	126	121	91	103	110	114	112	92	113	

Table 8. Performance comparisons of various methods on Digit10 with 4 domains and Office-31 with 3 domains.

		Digit10 (α = 0.1)						**Office-31** (α = 1.0)				
Method	Target	MNIST	EMNIST	USPS	SVHN	Avg	$\Delta(\uparrow)$	Amazon	Dlsr	Webcam	Avg	$\Delta(\uparrow)$
FedAvg	Acc	92.82	88.62	84.02	77.59	85.76	3.99↑	58.08	31.62	39.25	42.98	8.76↑
	CoR	112	82	96	122	103	22↑	144	136	135	138	9↑
FedProx	Acc	93.07	87.43	85.67	79.09	86.32	3.43↑	58.69	34.25	43.01	45.32	6.42↑
	CoR	114	93	89	118	103	22↑	145	146	139	143	14↑
Fixed	Acc	92.82	85.35	82.11	71.26	82.48	7.27↑	58.08	24.56	37.44	40.03	11.71↑
	CoR	112	0	0	0	/	/	144	0	0	/	/
DANN+FL	Acc	**96.07**	87.30	82.81	76.44	85.66	4.09↑	**59.95**	42.21	45.21	49.12	2.62↑
	CoR	132	107	116	129	120	39↑	149	144	141	145	16↑
Shared	Acc	92.82	82.10	80.36	74.77	82.51	7.24↑	58.08	35.33	37.55	43.65	8.09↑
	CoR	112	76	84	103	93	12↑	144	122	124	130	1↑
FCIL	Acc	92.82	88.62	84.02	77.59	85.76	3.99↑	58.08	31.62	39.25	42.98	8.76↑
	CoR	112	82	96	122	103	22↑	144	136	135	138	9↑
FedCIL	Acc	94.61	90.24	87.55	83.85	89.06	0.69↑	59.37	45.91	46.26	50.51	1.23↑
	CoR	118	86	92	125	105	24↑	146	139	148	144	15↑
Re-Fed	Acc	92.82	**91.64**	**88.57**	**85.96**	**89.75**	/	58.08	**47.07**	**50.80**	**51.74**	/
	CoR	112	68	73	71	81		144	125	118	129	

Table 9. Performance comparisons of various methods on DomainNet with 6 domains.

DomainNet (α = 10)									
Method	Target	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg	$\Delta(\uparrow)$
FedAvg	Acc	52.07	36.22	45.09	46.59	49.36	51.73	46.84	3.39↑
	CoR	141	128	97	108	136	115	121	11↑
FedProx	Acc	50.31	33.64	41.77	45.04	47.44	49.12	44.55	5.68↑
	CoR	<u>136</u>	131	115	130	137	116	128	1↑
Fixed	Acc	52.07	29.58	32.24	38.91	40.09	46.30	39.87	10.36↑
	CoR	141	0	0	0	0	0	/	/
DANN+FL	Acc	**55.66**	36.44	42.02	38.84	45.89	50.01	44.81	5.42↑
	CoR	142	126	109	112	137	121	125	15↑
Shared	Acc	52.07	35.22	37.83	35.19	40.52	41.76	40.43	9.80↑
	CoR	141	113	98	125	120	96	116	6↑
FCIL	Acc	52.07	36.22	45.09	46.59	49.36	51.73	46.84	3.39↑
	CoR	141	128	97	<u>108</u>	136	115	121	11↑
FedCIL	Acc	54.52	38.98	40.45	41.77	45.09	47.28	44.68	5.55↑
	CoR	148	136	128	112	142	125	132	22↑
Re-Fed	Acc	52.07	**42.26**	**48.11**	**48.98**	**53.34**	**56.66**	**50.23**	/
	CoR	141	<u>103</u>	<u>97</u>	109	<u>118</u>	<u>91</u>	<u>110</u>	

Table 10. Performance comparisons of three weighting methods for gradient norms in two incremental scenarios.

Dataset	Class-Incremental Scenario			Domain-Incremental Scenario		
	CIFAR10	CIFAR100	Tiny-ImageNet	Digit10	Office31	DomainNet
Early-Emphasis	**29.22**	**25.61**	**32.07**	**85.96**	**50.80**	**56.66**
Average-Weighting	28.73	24.88	30.42	85.71	48.95	56.04
Late-Emphasis	26.57	22.18	28.08	84.36	47.29	53.90

F. Analysis of the Federated Incremental-Learning Framework: Re-Fed

In this section, we prove the convergence of personalized informative models. To simplify the notation, here we conduct an analysis on a fixed task while the convergence does not depend on the IL setting. We first define following standard assumptions.

Assumption 1 (L_2 Distance.) The L_2 distance between the optimal local models $\hat{w_k} := \arg\min_{w_k}\{f(w_k)\}$ and the optimal global model $\hat{w} := \arg\min_{w}\{\frac{1}{K}\sum_{k=1}^{K}\nabla f(w_k)\}$ is bounded by:

$$||\hat{w_k} - \hat{w}|| \leq M,\ \forall k \in [K]. \tag{7}$$

Assumption 2 (Gradient Variance.) The variance of stochastic gradients is finite and bounded at all clients by:

$$\mathbb{E}\Big[||\nabla f(\hat{w_k})||^2\Big] \leq \sigma^2,\ \forall k \in [K]. \tag{8}$$

Assumption 3 (Strong Convexity.) There exists $\mu_k \in \mathbb{R}_+$ and a unique solution $\hat{w_k}$:

$$f(w_k) - f(\hat{w_k}) \geq \langle \nabla f(\hat{w_k}), \hat{w_k} - w_k \rangle + \frac{\mu_k}{2}||w_k - \hat{w_k}||^2. \tag{9}$$

F.1. Proof of Theorem 3.1

Definition 1 (Personalized Informative Model Formulation.) Denote the objective of personalized informative model v_k on client k while $f(\cdot)$ is strongly convex as:

$$\begin{aligned} \hat{v_k}(\lambda) &:= \arg\min_{v_k}\left\{f(v_k) + \frac{q(\lambda)}{2}||v_k - \hat{w}||^2\right\} \\ q(\lambda) &:= \frac{1-\lambda}{2\lambda},\ \lambda \in (0,1) \end{aligned} \tag{10}$$

where $\hat{w}$ denotes the global model.

Lemma 1 (Proportion of Global and Local Information.) *For all $\lambda \in (0,1)$ and $\lambda \rightarrow f(\lambda_k)$ is non-increasing:*

$$\begin{aligned} &\frac{\partial f(\hat{v_k}(\lambda))}{\partial \lambda} \leq 0 \\ &\frac{\partial ||\hat{v_k}(\lambda) - \hat{w}||^2}{\partial \lambda} \geq 0. \end{aligned} \tag{11}$$

Then, for $k \in [K]$, we can get:

$$\lim_{\lambda \to 0} \hat{v_k}(\lambda) := \hat{w}. \tag{12}$$

Proof. The proof here directly follows the proof in Theorem 3.1 [10]. As λ declines and $q(\lambda)$ grows, the objective of Eq. 10 tends to optimize $||v_k - \hat{w}||^2$ and increase the local empirical training loss $f(v_k)$, leading to the convergence on the global model. Hence we can modify the λ value to adjust the optimization direction of our model v_k thus the dominance of local and global model information.

Theorem 3.1 (Personalized Informative Model.) *Assuming the global model w^t converges to the optimal model $\hat{w}$ with $g(t)$ for any client $k \in [K]$ at each communication round t: $\mathbb{E}\Big[||w^t - \hat{w}||^2\Big] \leq g(t)$ and $\lim_{t\to\infty} g(t) = 0$, then there exists a constant $C < \infty$ such that the personalized informative model v_k^t can converge to the optimal model $\hat{v_k}$ with $Cg(t)$.*

Proof. Here we first introduce the Lemma 2 here proved by [21] Lemma 13.

Lemma 2 ([21] Lemma 13.) Under assumptions above, $f(v_k)$ is μ_k-strongly convex at each communication round t, we have:

$$\begin{aligned} \mathbb{E}\Big[||v_k^{t+1} - \hat{v_k}||^2\Big] \leq{}& (1 - \eta(\mu_k + q(\lambda)))\,\mathbb{E}\Big[||v_k^t - \hat{v_k}||^2\Big] + \eta^2\left(\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k})\right)^2 + \eta^2 q(\lambda)^2\mathbb{E}\Big[||w^t - \hat{w}||^2\Big] \\ &+ 2\eta^2 q(\lambda)\left(\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k})\right)\sqrt{\mathbb{E}\Big[||w^t - \hat{w}||^2\Big]} + 2\eta q(\lambda)\sqrt{\mathbb{E}\Big[||v_k^t - \hat{v_k}||^2\Big]\mathbb{E}\Big[||w^t - \hat{w}||^2\Big]}. \end{aligned} \tag{13}$$

Assume $g(t+1) \leq g(t)$ and let positive number A be chosen such that $A(g(t) - g(t+1)) \leq g^2(t)$, and we arrive at $(1 - \frac{g(t)}{A})g(t) \leq g(t+1)$. Then, we prove the **Theorem 3.2** by induction. Assuming that $\mathbb{E}\Big[||v_k^t - \hat{v_k}||^2\Big] \leq Cg(t)$ where $C > 0$ and $C \geq \frac{\mathbb{E}\Big[||v_k^0 - \hat{v_k}||^2\Big]}{g(0)}$, the learning rate $\eta = \frac{2g(t)}{A(\mu_k + q(\lambda))}$, here we can continue with **Lemma 2**:

$$\begin{aligned}\mathbb{E}\Big[||v_k^{t+1} - \hat{v_k}||^2\Big] \leq & (1 - \frac{2g(t)}{A})Cg(t) + \frac{4q(\lambda)\sqrt{C}g(t)}{A(\mu_k + q(\lambda))} \\ & + \frac{4g(t)^2}{A^2(\mu_k + q(\lambda))^2}\left((\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k}))^2 + q(\lambda)^2 g(t) + 2q(\lambda)\sqrt{g(t)}(\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k}))\right).\end{aligned} \tag{14}$$

Therefore, if we let $C = \max\{\frac{\mathbb{E}\Big[||v_k^0 - \hat{v_k}||^2\Big]}{g(0)}, 16, \frac{4\left((\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k}))^2 + q(\lambda)^2 g(t) + 2q(\lambda)\sqrt{g(t)}(\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k}))\right)}{A(\mu_k + q(\lambda))^2(1 - \frac{1}{(1 + \frac{\mu_k}{q(\lambda)})})}\}$, then we have:

$$\begin{aligned}&\frac{4q(\lambda)\sqrt{C}g(t)^2}{A(\mu_k + q(\lambda))} + \frac{4g(t)^2}{A^2(\mu_k + q(\lambda))^2}\left((\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k}))^2 + q(\lambda)^2 g(t) + 2q(\lambda)\sqrt{g(t)}(\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k}))\right) \leq \\ &\frac{q(\lambda)Cg(t)^2}{A(\mu_k + q(\lambda))} + \frac{4g(t)^2}{A^2(\mu_k + q(\lambda))^2}\left((\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k}))^2 + q(\lambda)^2 g(t) + 2q(\lambda)\sqrt{g(t)}(\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k}))\right) = \\ &\frac{Cg(t)^2}{A} \cdot \frac{1}{(1 + \frac{\mu_k}{q(\lambda)})} + \frac{4g(t)^2}{A^2(\mu_k + q(\lambda))^2}\left((\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k}))^2 + q(\lambda)^2 g(t) + 2q(\lambda)\sqrt{g(t)}(\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k}))\right) \leq \\ &\frac{Cg(t)^2}{A} \cdot \frac{1}{(1 + \frac{\mu_k}{q(\lambda)})} + \frac{g(t)^2}{A^2} \cdot CA\left(1 - \frac{1}{(1 + \frac{\mu_k}{q(\lambda)})}\right) = \frac{Cg(t)^2}{A}.\end{aligned} \tag{15}$$

The first inequality uses the fact that $16 \leq C$ and consequently $4\sqrt{C} \leq C$. The second inequality results from the definition of C as $\frac{4\left((\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k}))^2 + q(\lambda)^2 g(t) + 2q(\lambda)\sqrt{g(t)}(\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k}))\right)}{A(\mu_k + q(\lambda))^2}\} \leq C(1 - \frac{1}{(1 + \frac{\mu_k}{q(\lambda)})})$. Hence, combining the results of 14 and 15 yields

$$\begin{aligned}\mathbb{E}\Big[||v_k^{t+1} - \hat{v_k}||^2\Big] &\leq (1 - \frac{2g(t)}{A})Cg(t) + \frac{Cg(t)^2}{A} \\ &= (1 - \frac{g(t)}{A})Cg(t) \\ &\leq Cg(t+1),\end{aligned} \tag{16}$$

and we have the desired result.