IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 7, 2020, accepted December 29, 2020, date of publication January 8, 2021, date of current version January 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3050176

Homeostasis-Inspired Continual Learning: Learning to Control Structural Regularization

JOONYOUNG KIM[1], (Member, IEEE), HYOWOON SEO[2], (Member, IEEE), WAN CHOI[1,3], (Fellow, IEEE), AND KYOMIN JUNG[1], (Member, IEEE)

[1]Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea
[2]Centre for Wireless Communications, University of Oulu, 90570 Oulu, Finland
[3]Institute of New Media and Communications, Seoul National University, Seoul 08826, South Korea

Corresponding author: Kyomin Jung (kjung@snu.ac.kr)

This work was supported in part by the BK21 FOUR Program of the Education and Research Program for Future ICT Pioneers, Seoul National University, in 2020, and in part by the Automation and Systems Research Institute (ASRI), Seoul National University.

ABSTRACT Learning continually without forgetting might be one of the ultimate goals for building artificial intelligence (AI). However, unless there are enough resources equipped, forgetting knowledge acquired in the past is inevitable. Then, we can naturally pose a fundamental question about how to control what knowledge and how much of it to forget to improve the overall accuracy. To give a clear answer to it, we propose a novel trainable network termed *homeostatic meta-model*. The proposed neuromorphic framework is a natural extension of the conventional concept *Synaptic Plasticity* (SP) for further optimizing the accuracy of continual learning. In the preceding works on SP and its variations, though they seek important network parameters for structural regularization, they care less about the intensity of regularization (IoR). Per contra, this work reveals that a careful selection of IoR during continual training can remarkably improve the accuracy of tasks. The proposed method balances IoR between newly learned knowledge and the previously-acquired ones rather than biasing it to a specific task or evenly balancing. To obtain effective and optimal IoRs for the real-time continual learning circumstances, we propose a homeostasis-inspired meta learning architecture. The proposed meta-model automatically controls the IoRs by capturing important parameters from the previous tasks and the current learning direction. We provide experimental results considering various types of continual learning tasks showing that the proposed method notably outperforms the conventional methods in terms of learning accuracy and knowledge forgetting. We also show that the proposed method is relatively stable and robust compared to the existing SP-based methods. Furthermore, the IoR generated by our model interestingly appears to be proactively controlled within a specific range, which resembles a negative feedback mechanism of *homeostasis* in synapses.

INDEX TERMS Continual learning, meta learning, optimization.

I. INTRODUCTION

Human intelligence includes abilities to learn and memorize. Though there are thousands of valuable works on building intelligent machines that learn well, we usually rely on the capacity of storage devices installed on the machines when it comes to the memorizing capability. In this regard, if a machine carries out a sequence of tasks under a memory-limited environment with a single neural network, i.e., a *continual learning* (CL) scenario, it cannot memorize whole knowledge learned in the past and may incur unintended forgetting phenomenon, which we usually call *a catastrophic forgetting* [1]–[3]. Such CL phenomenon happens more under strict conditions, where the network topology cannot be expanded and theactual episodic real data cannot be stored, which is called a strict CL.

Elastic Weight Consolidation [4], which is recognized as pioneering research of the strict CL scenario, proposes a way of alleviating such forgetting inspired by a neuroscientific principle, a *Synaptic Plasticity* (SP). The principle teaches us that more frequently activated synapses tend to strengthen over time in human brains. Likewise, the EWC leverages a regularization method based on information obtained from previously-learned tasks when learning a new one. This can be done by finding important parameters of the network for each task, handing their list over to the next task, and using

The associate editor coordinating the review of this manuscript and approving it for publication was Bijoy Chand Chatterjee.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

this information to conduct elastic parameter-wise regularization when dealing with the current task.

Though the approach above is an innovative piece of work, it leaves the following fundamental questions:

Q1) *Does such regularization always bring positive effects to continual learning?*

Given a limited amount of resources to a learner, the regularization done for less forgetting the knowledge acquired may disturb learning new knowledge. Thus, it is necessary for the learner to control an *intensity of regularization (IoR)*, to gain high overall accuracy when learning continually.

Q2) *How can the learner control the IoR for less forgetting?*

Intuitively, we can think of a simple example that can give us a clue to addressing this question. Suppose (x_1, y_1) and (x_2, y_2) are samples, and l is the loss function for learning the samples. Following the principle on *transfer-interference trade-off* in [5], the IoR should be set small when

$$\text{(Weak interference)} \quad \frac{l(x_1, y_1)}{\partial\theta} \cdot \frac{l(x_2, y_2)}{\partial\theta} > 0,$$

where θ denotes the network parameters and $\cdot$ represents a dot-product, meaning the learning directions of the two samples are similar. Thus, with weak interference, choosing small IoR can help increasing the overall accuracy of the learning. On the other hand, with strong interference, the IoR should be set large for less forgetting the acquired knowledge when

$$\text{(Strong interference)} \quad \frac{l(x_1, y_1)}{\partial\theta} \cdot \frac{l(x_1, y_2)}{\partial\theta} < 0.$$

To minimize the amount of forgetting knowledge and increase overall accuracy, the IoR must be controlled according to the principle mentioned above. Throughout this paper, we answer **Q1** and **Q2** by proposing a novel trainable meta network termed *homeostatic meta-model* (HM), which effectively and automatically controls IoR in the main network (Section 3). The term homeostatic is brought from the neuroscientific mechanism, *Homeostatic Plasticity*, which prevents neural activities from being driven towards runaway or quiescence states [6], [7]. The proposed HM aims to balance the IoR to optimize the overall accuracy of continual learning by capturing important parameters from the previous tasks and the current learning direction. Specifically, our model batch-wisely generates an optimal IoR, depending on the degree of interference between the previously-learned tasks and the currently drawn batch, and the Euclidean distance between the current-state parameters and the optimal parameters learned from the previous tasks.

To control the amount of correlation between the tasks when training the HM, we propose a block-wise permutation-based (`BPERM`) continual learning tasks (Section 4). The proposed type of task is a generalization of the existing `MNIST-PERM` type tasks [8], where the size of permuting blocks can be manually controlled. In consequence, we can possibly train the HM with meta samples depicting various kinds of relations between the tasks.

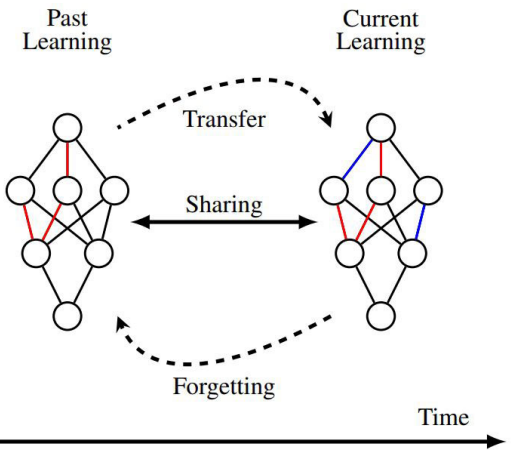


FIGURE 1. An illustration of the Stability-Plasticity.

To show the effectiveness and feasibility of the proposed HM, we conduct experiments on continual learning with a learner using various types of tasks. In the preliminary experiment, we found that using a fixed value of IoR degenerates the overall performance of continual learning in terms of accuracy and forgetting. This finding can give a clear answer to **Q1**. Moreover, in the main experiment (continual learning with 10 tasks), we found that the proposed method with IoR generated by the HM outperforms the conventional methods, i.e., EWC, Online-EWC [9], IMM [10]. Even in a long sequence of tasks (40 tasks), we showed that the proposed HM stably retains the previously-acquired knowledge compared to the other benchmark methods. Furthermore, we provide the trend of IoR over time during continual learning. The IoR generated by HM interestingly appears to be proactively controlled within a specific range, which resembles a negative feedback mechanism of *homeostasis* in synapses.

II. PRELIMINARIES

A. CONTINUAL LEARNING

As mentioned before, studies on continual learning consider a scenario of which a single machine carries out a series of tasks with a limited amount of resources, such as computational power, memory space, etc. Note that depending on the details, it is also termed as *lifelong learning* [11], endless *online learning*, the special case of *transfer learning* [12], and methods of *Overcoming Catastrophic Forgetting* [4], [10], [13], [14]. In the literature, there are several different approaches that alleviate the forgetting phenomena in continual learning, such as structural regularization [4], [9], [14], [15], adaptation with episodic memory [16]–[19], expansion of network topology [20], [21], and dedication of partial network to a specific task [13], [22], [23]. Among them, we are especially interested in the *structural regularization* based methods. In brief, the studies on such methods focus on finding *important network parameters* in a trained network that affects a lot on learning accuracy with a small difference

on their values. Having such information, we can regularize the network parameters when handling a series of tasks. Notable studies on the structural regularized based methods include the followings: [4] proposes the square of the gradient as diagonal Fisher information using Laplace approximation; [9] develops this method to accumulate previous information in an online manner; *Synaptic Intelligence* [15] saves the full trajectory of gradient steps rather than a few steps; *Structured Laplace* [14] computes sequentially block diagonal components of Hessian in the order through the layers; and IMM [10] suggests several transfer methods to make the search space of posterior weight smooth.

1) SYNAPTIC PLASTICITY

In order to understand the main idea proposed in this paper, we make a quick revisit to the previous works EWC [4] and Online-EWC [9] in advance. Conventionally in the related works on continual learning framework, it is considered that a set of I tasks

$$\mathcal{T} = \{T_1, \cdots, T_I\},$$

where its i-th element $T_i = \{(x_i^s, y_i^s, t_i^s)_{s=1}^{S_i}\}$ is S_i-triplet of an input $x_i^s \in \mathcal{X}$, a target $y_i^s \in \mathcal{Y}$ and a task descriptor $z_i^s \in \mathcal{Z}$ for $s \in \{1, \ldots, S_i\}$, is given to a learner to carry out a series of tasks in increasing order of task index i. Note that $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ respectively denote the set of inputs, targets, and task descriptors.

Now suppose the learner is training the sequence of tasks from T_1 to T_i, where $i \leq I$. Then, we can obtain the following negative log posterior of the network parameters θ in the approximated Bayesian perspective:

$$-\log p(T_i|\theta) + \frac{\alpha}{2}\sum_{j=1}^{i-1} ||\theta - \theta_j^*||_{F_j}^2, \tag{1}$$

where $\theta, \theta_j^*, \alpha$ and F_j represent the current-state network parameters of the learner, learned parameters after training task j, the IoR and Fisher information of task j, respectively. The norm ($||\theta||_x^2 = \theta^T \cdot diag(x) \cdot \theta$) describes Mahalanobis norm [24] and the exact posteriors are replaced as

$$\log p(\theta|T_j) \approx \mathcal{N}\left(\theta; \theta_j^*, F_j^{-1}\right), \quad {}^\forall j \in \{1 : i-1\} \tag{2}$$

using Laplace approximation [25]. Unfortunately, in the original EWC, there exists a problem that the second term, i.e., the regularization term, tends to increase linearly with respect to the increment of the number of tasks. To solve the problem, in [9], the authors propose Online-EWC, which re-centers the likelihood of the previous tasks with the optimal parameters of the $(i-1)$-th task, and as a consequence, we can have

$$-\log p(T_i|\theta) + \frac{\alpha}{2}||\theta - \theta_{i-1}^*||_{F_{1:i-1}}^2, \tag{3}$$

where $F_{1:i-1} := \sum_{j=1}^{i-1} F_j$ for brevity. This approach is quite effective for reducing the required memory space for handing over information from the previous tasks to the present and computational burden.

By replacing the negative log-likelihood with the loss function of the task T_i, $l_i(\theta) := -\log(p(\theta|T_i)$, then we have the EWC loss as

$$l_{i,EWC}(\theta) = l_i(\theta) + \frac{\alpha}{2}||\theta - \theta_{i-1}^*||_{F_{1:i-1}}^2. \tag{4}$$

B. META LEARNING

Meta-learning, also known as *learning to learn*, aims to design models that can learn new skills or adapt to new environments rapidly, allowing the algorithm to learn to exploit structures in the problems of interest in an automatic way. In the literature on the meta learning frameworks, there have been interesting works. A general framework for meta learning is proposed in [26] by mimicking the learning process with Recurrent Neural Network (RNN) and Model Agnostic Meta Learning (MAML) uses the original Stochastic Gradient Descent (SGD) as meta learner, but the initialization is learned via meta learning [27]. For an advance version of MAML, Meta-SGD [28] learns not only an initialization but also updating directions and the learning rates via meta learning.

1) META LEARNING FOR CONTINUAL LEARNING

Recently, the philosophy of meta learning is applied to continual learning framework to enhance its performance. Firstly, a simple optimizer which outputs the component-wise modified gradient for alleviating forgetting was proposed in [29]. On the other hand, MER [5] proposes a meta continual learning algorithm inspired by Reptile algorithm [30] which considers approximated second order correlation between parameters. However, the methods proposed in the preceding work lack scalability or incur high computing complexity. Unlike the previous studies, our approach is to design a relatively simple network for controlling the IoR (meta knowledge).

Towards applying the meta learning to continual learning, we must distinguish the tasks for meta training and testing. Therefore, we hereafter consider a pair of task sets,

$$\mathcal{T}^0 = \{T_1^0, \cdots, T_{I_0}^0\} \text{ and } \mathcal{T}^1 = \{T_1^1, \cdots, T_{I_1}^1\},$$

respectively denoting meta training and testing set, where I_0 and I_1 is positive integers such that $I_0 > I_1$. As done in the previous studies the continual learning scenario ([4], [19], [31]), $\mathcal{T}^0$ allows the learner to access all samples for the purposes of training its meta model which measures the importance of network parameters for new gradient update, while $\mathcal{T}^1$ is the actual sequence of tasks used for meta testing. Notice that the tasks in the meta test set $\mathcal{T}^1$ are only accessible during the meta testing process and cannot included in the meta training set for unbiased training.

III. HOMEOSTATIC META-MODEL

In this section, we propose a novel trainable network HM, as shown in Fig 2, which is well suited to the CL scenario to learn better and forget less. The role of the HM denoted by h_ϕ

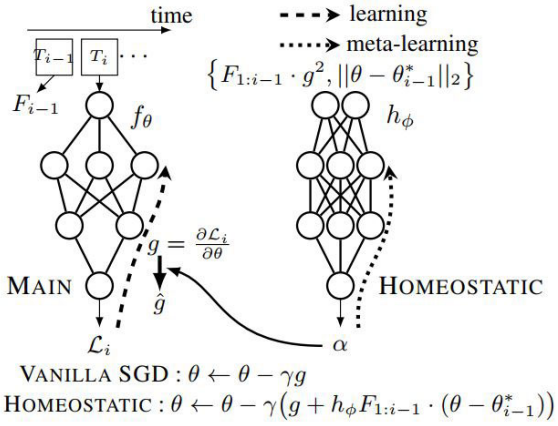


FIGURE 2. An illustration of the main network and the homeostatic meta-model (HM).

is to generate IoR, denoted as α. Recall that the fundamental meaning of regularization is to utilize side information for solving ill-posed problems, and correspondingly, the semantics of IoR indicates how large the optimization result is affected by constraints. Mathematically, the generation of the current-state IoR can be expressed as

$$\alpha = h_\phi\big(\underbrace{F_{1:i-1} \cdot g^2}_{(a)}, \underbrace{||\theta - \theta^*_{i-1}||_2}_{(b)}\big) \tag{5}$$

when the learner is training the task T_i. The proposed model h_ϕ is a neural network, which is lighter than the main network and takes two inputs: (a) a inner product of the element-wise square of current-state gradients of given batch from current task i, i.e., g^2 and accumulated Fisher information $F_{1:i}$, and (b) Euclidean distance between the current-state parameters θ and θ^*_{i-1} which is the learned parameters from the previous task. Note that (a) is to give the HM the information on the similarity between the current learning direction and the previously-accumulated importance of parameters, and (b) is to inform how far the current parameters deviates from the previously-learned parameters. In this way, our HM guarantees scalability, in other words, the computational burden of the homeostatic network does not depend on the number of parameters in main network. To put it precisely with mathematical expressions, we can derive the following loss function of the proposed structure

$$l_{i,main}(\theta) = l_i(\theta) + \frac{h_\phi}{2}||\theta - \theta^*_i||^2_{F_{1:i-1}}. \tag{6}$$

In the training phase, since the HM is allowed to access all the samples in the task set ($\mathcal{T}^0$), it arbitrarily draws $\mathcal{B}_{i-1}$ and $\mathcal{B}_i$ from $\mathcal{T}^0_{i-1}$ and $\mathcal{T}^0_i$ respectively, then takes their union $\mathcal{B}_{i-1} \cup \mathcal{B}_i$ as an input of the meta model h_ϕ. The gradient of the model f_θ taking the union batch, $g_\cup$ is denoted by

$$g_\cup = \nabla_\theta l\big(f_\theta(\mathcal{B}_{i-1} \cup \mathcal{B}_i)\big). \tag{7}$$

Note that the union gradient can play a role as a supervision for training h_ϕ with the absolute loss function,

$$l_{HM}(\hat{g}, g_\cup) = |\hat{g} - g_\cup|, \tag{8}$$

where

$$\hat{g} = g + h_\phi\big(F_{1:i-1} \cdot g^2, ||\theta - \theta^*_{i-1}||_2\big)F_{1:i-1} \cdot (\theta - \theta^*_{i-1}). \tag{9}$$

To stabilize the direction of the gradient $g_\cup$, we adopt larger batch and lower learning rate when training h_ϕ compared to main network training. For testing the trained meta model, we generate IoRs by $h_{\phi^*}((a),(b))$, where h_{ϕ^*} denotes the optimized meta-trained HM using Algorithm 1, and (a) and (b) are the inputs for the meta model as discussed above. Note that we freeze ϕ^* during the meta testing, while training the main network sequentially with tasks from $\mathcal{T}^1$.

Algorithm 1 Training Homeostatic Meta-Model h_ϕ

procedure Train (f_θ, h_ϕ, $\mathcal{T}^0$)
for all $\{T^0_{i-1}, T^0_i\}$ **in** T^0 **do**
  $\theta^*_{i-1} \leftarrow$ base train f_θ on T^0_{i-1} using SGD
  $\theta \leftarrow \theta^*_{i-1}$
  $F_{i-1} \leftarrow \mathrm{E}\big[\big(\nabla_\theta \mathcal{L}\big(f_\theta(T^0_{i-1})\big)\big)^2\big]$
  $F_{1:i-1} \leftarrow F_{1:i-2} + F_{i-1}$
  for batches $\mathcal{B}_{i-1}, \mathcal{B}_i$ **drawn respectively from** $T^0_{i-1}, \mathcal{T}^0_i$ **do**
    $g \leftarrow \nabla_\theta \mathcal{L}\big(f_\theta(\mathcal{B}_i)\big)$
    $g_\cup \leftarrow \nabla_\theta l\big(f_\theta(\mathcal{B}_{i-1} \cup \mathcal{B}_i)\big)$
    Update $\hat{g}$ using Eq (9) with h_ϕ
    $\theta \leftarrow \theta - \gamma\hat{g}$
    $\phi \leftarrow \phi - \eta\nabla_\phi l_{HM}(h_\phi, g, g_\cup, F_{1:i-1})$
  end for
end for
end procedure

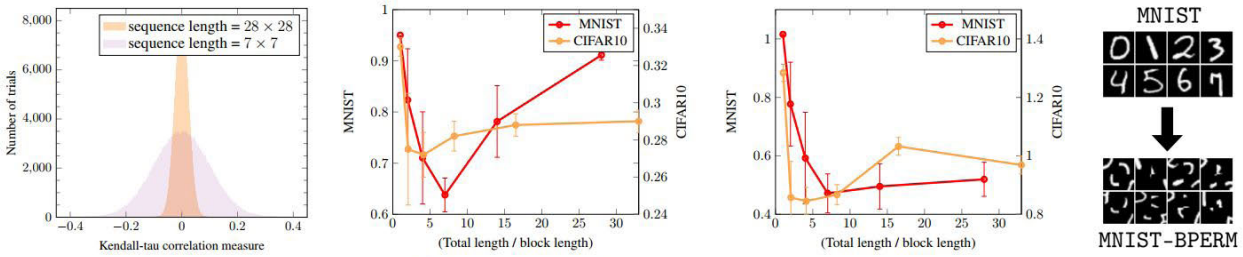
IV. PRELIMINARY EXPERIMENTS AND FINDINGS

Generally, `MNIST-PERM` [8], [32] is a widely-used set of tasks in continual learning literature, which is made by simply permuting the input pixels of `MNIST` dataset. Each permutation map is uniformly drawn and applied to generate a task. Since the number of pixels of a sample is generally a few hundreds or more, the Kendall tau correlation measure [33] between two randomly chosen permutation patterns is near zero, which means the correlation between them is very low. However, we can easily predict that the chance of drawing two permutation patterns with high correlation is larger in block-wise permutation tasks as shown in Fig. 2 (a).

A. BLOCK-WISE PERMUTATION

We term such block-wise permuted tasks as `BPERM`. The `BPERM` can be seen as a generalization of a pixel-wise permutation-based tasks, such as `MNIST-PERM` with block size of a single pixel. As shown in Fig 3 (d), the `MNIST-BPERM` looks like a Jigsaw puzzle, since it is made by dividing the input pixels into square blocks and permuting the blocks while the pixels inside the blocks are stay fixed.

To give more insights about the property of `BPERM`, we conducted experiments with `MNIST` and `CIFAR10`



(a) Correlation (b) Fully-Connected (c) CNN (d) BPERM

FIGURE 3. (a) A histogram of Kendall-tau correlation measure between randomly drawn two permutations (500, 000 trials). An illustration of the average accuracy of Task 1 (normal task) after training Task 2 (BPERM) with MNIST and CIFAR10 datasets with (b) Fully-connected networks and (c) 6 layers CNN. (d) An example of MNIST−BPERM.

datasets. For each random seed, we train a learner with a normal task and BPERM task consecutively, then evaluate the accuracy of the first task.

With two distinct architectures, both (b) and (c) consistently show the similar trends on the amount of information about the previous task that a learner forget, depending on the size of the blocks. Interestingly, the moderate-sized blocks cause more forgetting, which implies that the BPERM task gives larger interference than pixel-wise permutation task.

B. PERFORMANCE METRICS

To assess how a learner's performance evolves as it learns a sequence of tasks, we define average accuracy (A), average forgetting (F) (*backward transfer*) [17] and step forgetting (SF). Let $R_{i,j}$, for $j \leq i$, be the accuracy of the model on task T_j^1 after training the task T_i^1. Then we define a metric, *average accuracy*, defined as

$$A_k = \frac{1}{k}\sum_{j=1}^{k} R_{k,j},$$

which describes the average of the model accuracy on each task after training T_k^1. To measure the amount of knowledge forgot after the initial acquisition, we consider *average forgetting*, defined as

$$F_k = \frac{1}{(k-1)}\sum_{j=1}^{k-1}(R_{j,j} - R_{k,j}).$$

In addition, we also consider *step-forgetting*, defined as

$$SF_k = \frac{1}{(k-1)}\sum_{j=1}^{k-1}(R_{k-1,j} - R_{k,j}).$$

which indicates the amount of information loss owing to the acquisition of new knowledge by training T_k^1.

V. EXPERIMENT

To assess the performance of the HM-based continual learning, experimental results are provided in this section. In general, the performance of continual learning largely depends on how the sequence of tasks are chosen. To take care of this issue, we conducted the experiments with different random seeds and averaged the performance measures to avoid bias that comes from a naive choice of task sequence.

A. DATASETS

We consider three variants for each MNIST and CIFAR10 datasets: (1) MNIST−PERM and CIFAR−PERM permute input pixels of the MNIST and CIFAR10, respectively, through randomly drawn permutation maps, (2) MNIST−ROTA and CIFAR−ROTA rotate them by a randomly drawn integer angles in $[0, \pi]$ (3) MNIST−BPERM and CIFAR−BPERM block-wisely permute them through randomly drawn coarser permutation maps. Note that, for CIFAR10, each transformation is applied to all RGB dimensions.

B. METHODS

Here the proposed *homeostatic meta-model* (HM) is compared with six alternative baseline methods listed as follows.

- Single: a single learner based on SGD for a sequence of tasks.
- Independent: a dedicated (independent) learner based on SGD for each task.
- EWC [4]: regularized with the dedicated Fisher information F_i for each task i.
- OEWC [9]: regularized with the accumulated Fisher information $F_{1:i-1}$ as Eq. (4).
- IMM [10]: *Incremental Moment Matching* with a weight transfer method.
- Multi-task: allowed to access all the tasks (violation of strict CL scenario).

We adopt fully-connected networks with four hidden layers of 50 ReLU units for the main networks in each considering method and four additional hidden layers of 30 ReLU units are used in the HM. All the networks in baselines and the proposed HM is trained using SGD with mini-batch size of 100 samples.

C. OVERALL PERFORMANCE

In Fig. 4 and 6, the proposed HM consistently outperforms the other alternatives when training a sequence of tasks, in terms of average accuracy and forgetting, with equal or less number of model parameters (*resource-efficient*). Specifically

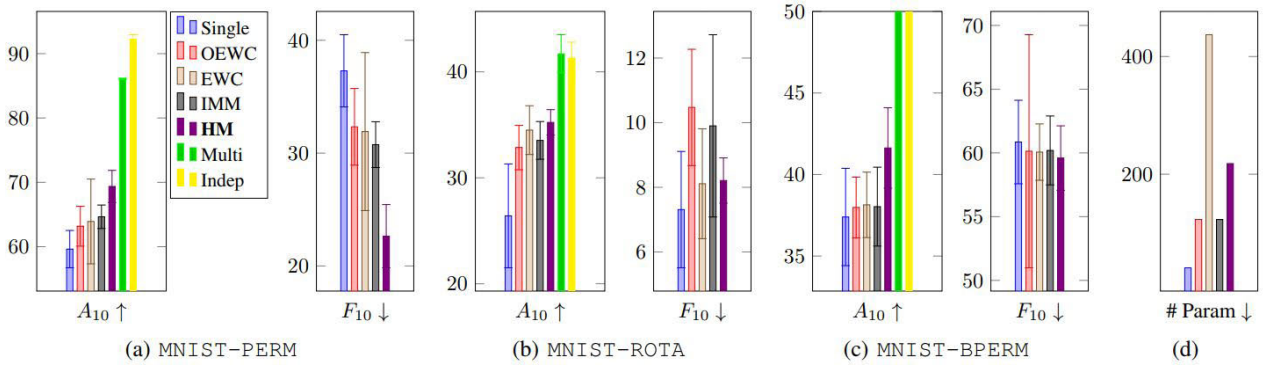


FIGURE 4. Comparisons on average accuracy (A_{10}) and forgetting (F_{10}), in percentage, of the considered learning methods including the HM, averaged over 10 random seeds on 10 tasks ((a) MNIST-PERM, (b) MNIST-ROTA, (c) MNIST-BPERM with block length = 4) (d) illustrates the number of model parameters employed in each method.) Note ↑ indicates that larger number is a better model, and ↓ means the opposite.

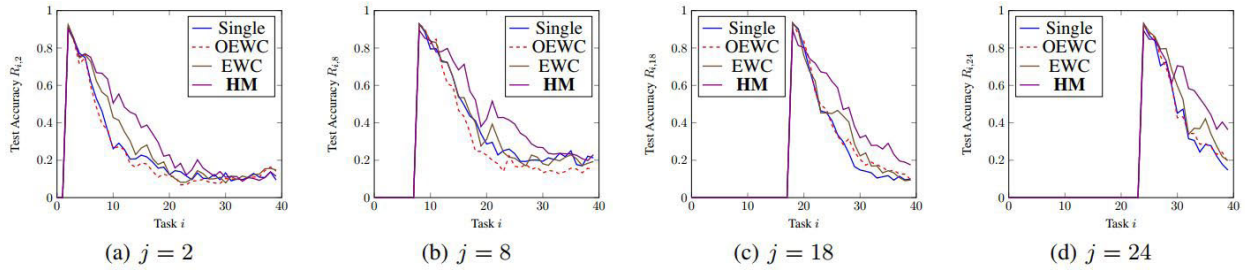


FIGURE 5. Evolution of test accuracy ($R_{i,j}$) of the main network with a longer sequence of MNIST-PERM tasks ($i \in [1, 40]$).

compared to EWC, the performance variance of the HM is relatively small, which implies that the proposed method is more robust than EWC (*robustness*).

To capture the performance of the HM with given a longer sequence of tasks, we conducted a experiment with 40 consecutive MNIST-PERM tasks. The experimental results are illustrated in Fig. 5. The proposed HM still outperforms the alternatives considerably. Fig. 5 (a) represents the evolution of the accuracy of the main network on the 2nd task. Likewise, each (b), (c) and (d) represents the accuracy of the main network on 8th, 18th and 24th task, respectively.

In Fig. 7, the trend of IoR and step-forgetting is provided during continual learning. The IoR generated by HM interestingly appears to be proactively controlled within a specific range. In most cases, IoR is small at each task boundary and becomes larger as learning progresses, until the next task boundary. By doing so, our model lead to increase degree of freedom in parameter space at each task boundary. In addition, the proposed HM shows smaller step-forgetting value compared to the alternatives, which implies the amount of information lost is the smallest among all the compared methods when the main model acquires new knowledge.

VI. RELATED LITERATURE

A. STRUCTURAL REGULARIZATION

Several studies focus on structural regularization based on measuring the importance of parameters. EWC [4] proposes the square of the gradient as diagonal Fisher information using Laplace approximation and OEWC [9] develop this method to accumulate previous information in online manner. Both of [9], [31] share a similar principle to divide whole learning phase into a training phase and the a importance update phase. Then, Synaptic Intelligence [15] saves the full trajectory of gradient steps for the importance of parameters rather than a single step, and Structured Laplace [14] computes sequentially block diagonal components of parameter correlation matrix in the order through the layers. Also, Incremental Moment Matching [10] suggests several transfer methods to make the search space of posterior parameter smooth and variational continual learning [34] generalizes various continual learning scenarios in *Bayesian* perspective using existing posterior approximation methods.

B. ADAPTATION WITH EPISODIC MEMORY

Other group of works assume a learner has fixed-size memory for episodic memory. In the beginning, iCarl [16] proposes a class incremental problem with a memory management method, and GEM [17] solves a continual learning task as an optimization problem with episodic memory constraints. Then MbPA [18] suggested that a random access method for episodic memory with large learning rate would be sufficient for the model adaptation.

C. EXPANSION OF NETWORK TOPOLOGY

If a learner can expand the size of the network, it can allocate a little portion of the whole network dedicated to each task

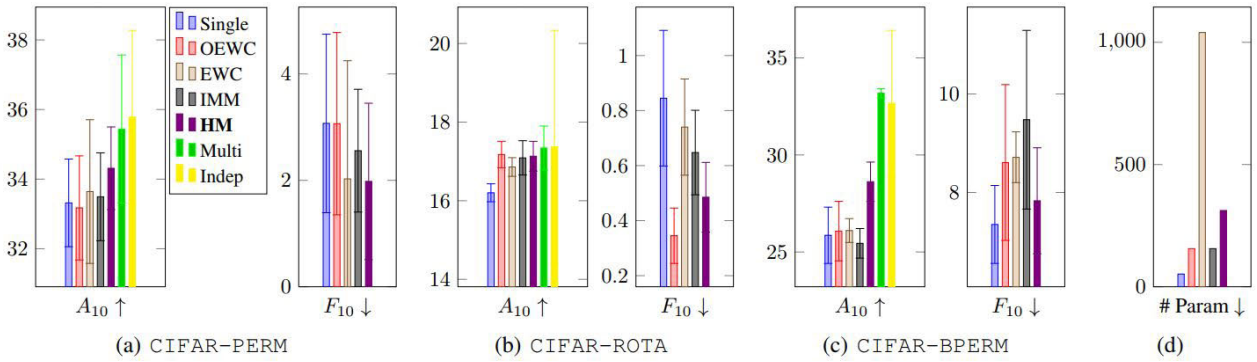


FIGURE 6. Comparisons on average accuracy (A_5) and forgetting (F_5), in percentage, of the considered learning methods including the HM, averaged over 4 random seeds on 5 tasks ((a) `CIFAR-PERM`, (b) `CIFAR-ROTA`, (c) `CIFAR-BPERM` with block length = 8, (d) figure illustrates the number of model parameters employed in each method.) Note ↑ indicates that larger number is a better model, and ↓ means the opposite.

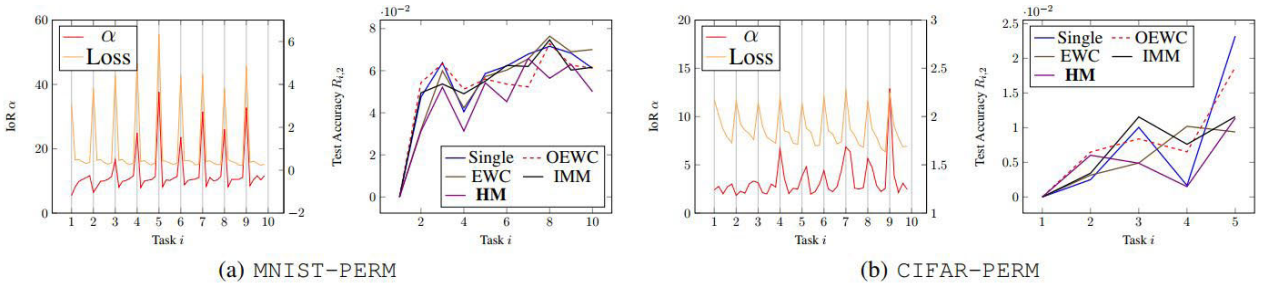


FIGURE 7. A trend of IoR generated by HM during continual learning and performance comparison of step-forgetting with (a) `MNIST-PERM` and (b) `CIFAR-PERM`.

for alleviating forgetting. [20] proposes the simple method of network expansion with an adaptor which summarizes previous information into the low dimensional vector. But their work has limitations in scalability. To overcome this limitation [21] suggests a gradual method of determining the number of neurons that expand according to specific thresholds.

D. DEDICATING A PARTIAL NETWORK TO A SPECIFIC TASK

In order to share a single network for multiple tasks, several studies softly divide the given network into some parts. Each part is dedicated to each task for avoiding interference. [22] defines a modular network to make a different pathway for each task and [23] proposes a network pruning method to help neurons to be uncorrelated for packing information in fixed network topology. Also, [13] proposes a hard attention method with annealing method for training, in order to dedicate the part of neural network for each task.

Algorithm 2 Testing Homeostatic Meta-Model h_{ϕ^*}

procedure Test $(f_\theta, h_{\phi^*}, \{\mathcal{T}_1^1, \mathcal{T}_2^1, \mathcal{T}_3^1, \ldots\})$
for all $\mathcal{T}_i^1$ **do**
  if $i = 1$ **then**
    $\theta_1^* \leftarrow$ train f_θ for one epoch on $\mathcal{T}_1^1$ using SGD
  else
    $\theta \leftarrow \theta_{i-1}^*$
    $F_{i-1} \leftarrow \mathbb{E}\left[\left(\nabla_\theta \mathcal{L}\left(f_\theta(\mathcal{T}_{i-1}^1)\right)\right)^2\right]$
    $F_{1:i-1} \leftarrow F_{1:i-2} + F_{i-1}$
    for each batch $\mathcal{B}_i$ **from** $\mathcal{T}_i^0$ **do**
      $g \leftarrow \nabla_\theta l\left(f_\theta(\mathcal{B}_i)\right)$
      Update θ using (9) with h_{ϕ^*}
    end for
    $\theta_i^* \leftarrow \theta$
  end if
end for
end procedure

VII. CONCLUSION

We proposed a *robust*, *resource-efficient*, and neuromorphic method, the HM, for strict CL by controlling the IoR. We showed that effective control of IoR via the HM can alleviate the amount of forgetting previously-acquired knowledge during the CL. Moreover, by considering a generalized task type, i.e., `BPEM`, for the CL, we highlighted the effectiveness of the proposed method. The experimental results demonstrate the proposed HM outperforms the state-of-the-art methods with less number of model parameters in terms of the learning accuracy and amount of information forgetting.

Until now, theoretical bounds on the capacity of continual learning are not yet found, and thus our ultimate goal is to find

TABLE 1. Dataset statistics.

Dataset	MNIST			CIFAR		
Type	PERM	ROTA	BPERM	PERM	ROTA	BPERM
number of tasks	10, 40	10		10, 40	10	
input size	28×28			32×32		
training smaples	10			10		
test images	10			10		
setup range	28×28	$[0, \pi]$	block 7×7	32×32	$[0, \pi]$	block 7×7

the bounds and algorithms that approach them. Moreover, the proposed HM in this paper is trained and tested with the same type of task, however, in the future work we will generalize the HM, to make it work for different types of tasks. For example, we can say that the HM is generalized if trained with `MNIST-BPERM` and work for `CIFAR-BPERM`. We believe that this will be possible, since the HM implicitly considers the correlation between the given sequence of tasks, while less caring about what the type of tasks are.

APPENDIX

Here we report the summary of datasets used for the experiments and describe the algorithm for testing our HM in detail.

META-TESTING ALGORITHM

Notice that the tasks in the meta test set $\mathcal{T}^1$ are only accessible during the meta testing process and cannot included in the meta training set for unbiased training.

REFERENCES

[1] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural networks for perception*. Amsterdam, The Netherlands: Elsevier, 1992, pp. 65–93.

[2] R. French, "Catastrophic forgetting in connectionist networks," *Trends Cognit. Sci.*, vol. 3, no. 4, pp. 128–135, Apr. 1999.

[3] R. Ratcliff, "Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions," *Psychol. Rev.*, vol. 97, no. 2, p. 285, 1990.

[4] J. Kirkpatrick and R. Pascanu, "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.

[5] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesauro, "Learning to learn without forgetting by maximizing transfer and minimizing interference," 2018, *arXiv:1810.11910*. [Online]. Available: http://arxiv.org/abs/1810.11910

[6] G. Turrigiano, "Homeostatic synaptic plasticity: Local and global mechanisms for stabilizing neuronal function," *Cold Spring Harbor Perspect. Biol.*, vol. 4, no. 1, Jan. 2012, Art. no. a005736.

[7] G. G. Turrigiano and S. B. Nelson, "Homeostatic plasticity in the developing nervous system," *Nature Rev. Neurosci.*, vol. 5, no. 2, pp. 97–107, Feb. 2004.

[8] R. K. Srivastava, J. Masci, S. Kazerounian, F. Gomez, and J. Schmidhuber, "Compete to compute," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2310–2318.

[9] J. Schwarz, J. Luketina, W. M. Czarnecki, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell, "Progress & compress: A scalable framework for continual learning," 2018, *arXiv:1805.06370*. [Online]. Available: http://arxiv.org/abs/1805.06370

[10] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang, "Overcoming catastrophic forgetting by incremental moment matching," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4652–4662.

[11] S. Thrun, "Lifelong learning algorithms," in *Learning to Learn*. Cham, Switzerland: Springer, 1998, pp. 181–209.

[12] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proc. ICML Workshop Unsupervised Transf. Learn.*, 2012, pp. 17–36.

[13] J. Serra, D. Suris, M. Miron, and A. Karatzoglou, "Overcoming catastrophic forgetting with hard attention to the task," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4555–4564.

[14] H. Ritter, A. Botev, and D. Barber, "Online structured laplace approximations for overcoming catastrophic forgetting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3738–3748.

[15] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3987–3995.

[16] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "ICaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2001–2010.

[17] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6467–6476.

[18] P. Sprechmann, S. M. Jayakumar, J. W. Rae, A. Pritzel, A. Puigdomènech Badia, B. Uria, O. Vinyals, D. Hassabis, R. Pascanu, and C. Blundell, "Memory-based parameter adaptation," 2018, *arXiv:1802.10542*. [Online]. Available: http://arxiv.org/abs/1802.10542

[19] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with A-GEM," 2018, *arXiv:1812.00420*. [Online]. Available: http://arxiv.org/abs/1812.00420

[20] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," 2016, *arXiv:1606.04671*. [Online]. Available: http://arxiv.org/abs/1606.04671

[21] J. Yoon, E. Yang, J. Lee, and S. Ju Hwang, "Lifelong learning with dynamically expandable networks," 2017, *arXiv:1708.01547*. [Online]. Available: http://arxiv.org/abs/1708.01547

[22] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra, "PathNet: Evolution channels gradient descent in super neural networks," 2017, *arXiv:1701.08734*. [Online]. Available: http://arxiv.org/abs/1701.08734

[23] A. Mallya and S. Lazebnik, "PackNet: Adding multiple tasks to a single network by iterative pruning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7765–7773.

[24] P. C. Mahalanobis, *On the Generalized Distance in Statistics*. Bengaluru, Karnataka: National Institute of Science of India, 1936.

[25] D. J. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[26] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. De Freitas, "Learning to learn by gradient descent by gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3981–3989.

[27] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[28] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-SGD: Learning to learn quickly for few-shot learning," 2017, *arXiv:1707.09835*. [Online]. Available: http://arxiv.org/abs/1707.09835

[29] R. Vuorio, D.-Y. Cho, D. Kim, and J. Kim, "Meta continual learning," 2018, *arXiv:1806.06928*. [Online]. Available: http://arxiv.org/abs/1806.06928

[30] A. Nichol and J. Schulman, "Reptile: A scalable metalearning algorithm," 2018, *arXiv:1803.02999*. [Online]. Available: http://arxiv.org/abs/1803.02999

[31] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 139–154.

[32] I. J. Goodfellow *et al.*, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," 2013, *arXiv:1312.6211*. [Online]. Available: http://arxiv.org/abs/1312.6211

[33] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, pp. 81–93, Jun. 1938.

[34] C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner, "Variational continual learning," 2017, *arXiv:1710.10628*. [Online]. Available: http://arxiv.org/abs/1710.10628



JOONYOUNG KIM (Member, IEEE) received the B.S. degree in electrical engineering from KAIST. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Seoul National University.

In 2009, he worked with LG Electronics, as a Research Internship. His research interests include crowdsourcing, continual and meta learning, and data analysis.



HYOWOON SEO (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Electrical Engineering (EE), Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012, 2014, and 2020, respectively. In 2018, he was a Visiting Scholar with the Centre for Wireless Communications (CWC), University of Oulu, Finland. He was a Postdoctoral Research Fellow with Seoul National University (SNU). He is currently a Postdoctoral Research Fellow with the CWC, University of Oulu. His research interests include wireless communication, the Internet-of-Things, physical layer security and privacy, vehicle-to-everything (V2X) communications, and deep learning. He received the Gold Prize for the 24th Samsung Humantech Paper Award, in 2018.



WAN CHOI (Fellow, IEEE) received the B.Sc. and M.Sc. degrees from the School of Electrical Engineering and Computer Science (EECS), Seoul National University (SNU), Seoul, South Korea, in 1996 and 1998, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, The University of Texas at Austin, in 2006.

From 1998 to 2003, he was a Senior Member of the Technical Staff with the Research and Development Division, KT, South Korea, where he researched 3G CDMA systems. From February 2007 to February 2020, he was a Professor with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. He is currently a Professor with the Department of Electrical and Computer Engineering, SNU. He is also a member of the Fellow Evaluation Committee and the IEEE Vehicular Technology Society and a fellow of the IEEE (recommended by IEEE Communication Society) for the contribution to "the analysis and design of multi-cell communication systems". He was a recipient of the Jack Neubauer Memorial Award (Best System Paper Award) from the IEEE Vehicular Technology Society, in 2002. He received the Dan Noble Fellowship Award from the IEEE Vehicular Technology Society, in 2006, the Asia Pacific Young Researcher Award from the IEEE Communication Society, in 2007, the Haedong Young Scholar Award from KICS, in 2012, and the Irwin-Jacobs Award from Qualcomm and KICS, in 2015. While at The University of Texas at Austin, he was also a recipient of the William S. Livingston Graduate Fellowship and the Information and Telecommunication Fellowship from the Ministry of Information and Communication (MIC), South Korea. He was an Executive Editor, from December 2014 to June 2019. He is currently the Executive Editor Chair of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, from July 2019. He is an Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, from 2009 to 2014, the IEEE WIRELESS COMMUNICATIONS LETTER, from 2012 to 2017, and a Guest Editor for the 5G *Wireless Communication System* Special Issue of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATION.



KYOMIN JUNG (Member, IEEE) graduated from the Seoul Science High School, in February 1996. He received the B.S. degree in mathematics from Seoul National University, in August 2003, and the Ph.D. degree from the MIT Mathematics Department, in June 2009. During his Ph.D., he worked with Microsoft Research Cambridge, in 2008, the IBM T.J. Watson Research, in 2007, and Bell Labs, in 2006 as Summer Research Internships. He was a Visiting Professor with MIT, in summer 2013, and with Microsoft Research Cambridge, in summer 2009 and summer 2010. He was a member of Laboratories for Information and Decision Systems (LIDS) with MIT Electrical Engineering and Computer Science Department. He worked with the Department of Computer Science, KAIST, from 2009 to 2013, where he had joint appointments with the Department of Electrical Engineering and the Department of Mathematics. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Seoul National University. He is also an Adjunct Professor with the Department of Mathematical Sciences, SNU. He was a recipient of the Samsung Lee Kun Hee Scholarship for graduate study. He was also a Gold Medalist from the International Mathematical Olympiad (IMO), in 1995. He has been a recipient of the Excellent New Faculty Funding from NRF Korea, since 2012.