AFRL-RY-WP-TR-2022-0210

LIFELONG LEARNING FORESTS

Joshua T. Vogelstein
John Hopkins University

SEPTEMBER 2022
Final Report

STINFO COPY

AIR FORCE RESEARCH LABORATORY
SENSORS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7320
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with The Under Secretary of Defense memorandum dated 24 May 2010 and AFRL/DSO policy clarification email dated 13 January 2020. This report is available to the general public, including foreign nationals.

Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RY-WP-TR-2022-0210 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

"//Signature//"

OLIVER A. NINA
Program Manager
Decision Sciences Branch
Multi-Domain Sensing Autonomy Division

"//Signature//"

OLGA L. MENDOZA-SCHROCK, Chief
Decision Sciences Branch
Multi-Domain Sensing Autonomy Division

"//Signature//"

ROY L. BALLARD
Division Chief
Multi-Domain Sensing Autonomy Division
Sensors Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

*Disseminated copies will show "//Signature//" stamped or typed above the signature

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
September 2022	Final	5 March 2018	31 March 2022

4. TITLE AND SUBTITLE
LIFELONG LEARNING FORESTS

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8650-18-2-7834	N/A	N/A
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
N/A	N/A	Y1SF

6. AUTHOR(S)
Joshua T. Vogelstein

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
John Hopkins University
3400 N. Charles St
Baltimore, MD 21218

8. PERFORMING ORGANIZATIONREPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Research Laboratory, Sensors Directorate Wright-Patterson Air Force Base, OH 45433-7320 Air Force Materiel Command, United States Air Forces	Defense Advanced Research Projects Agency (DARPA/MTO) 675 North Randolph Street Arlington, VA 22203	AFRL/RYAT	AFRL-RY-WP-TR-2022-0210

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
This material is based on research sponsored by the Air Force Research Lab (AFRL) and the Defense Advanced Research Projects Agency (DARPA) under agreement number FA8650-18-2-7834.The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Labs (AFRL), the Defense Advanced Research Projects Agency (DARPA) or the U.S. Government. Report contains color.

14. ABSTRACT
In biological learning, data are used to improve performance not only on the current task, but also on previously encountered, and as yet unencountered tasks. In contrast, classical machine learning which we define as starting from a blank slate, or tabula rasa, using data only for the single task at hand. While typical transfer learning algorithms can improve performance on future tasks, their performance on prior tasks degrades upon learning new tasks (called forgetting). Many recent approaches for continual or lifelong learning have attempted to maintain performance given new tasks. But striving to avoid forgetting sets the goal unnecessarily low: the goal of lifelong learning, whether biological or artificial, should be to improve performance on both past tasks (backward transfer) and future tasks forward transfer with any new data. Our key insight is that even though learners trained on other tasks often cannot make useful decisions on the current task, they may have learned representations that are useful for this task. Thus, although ensembling decisions is not possible, ensembling representations can be beneficial whenever the distributions across tasks are sufficiently similar. Moreover, we can ensemble representations learned independently across tasks in quasilinear space and time. We therefore propose two algorithms: representation ensembles of (1) trees and (2) networks. Both algorithms demonstrate forward and backward transfer in a variety of simulated and real data scenarios, including tabular, image, and spoken, and adversarial tasks. This is in stark contrast to the reference algorithms we compared to, all of which failed to transfer either forward or backward, or both, despite that many of them require quadratic space or time complexity.

15. SUBJECT TERMS
lifelong learning, transfer learning, out-of-distribution learning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
Unclassified	Unclassified	Unclassified	SAR	21

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Oliver Nina	

Table of Contents

List of Figures

List of Tables

1 High-Level Project Progress

Considering biological intelligence, an artificial intelligence agent should not only learn *tabula rasa* or tasks at hand, but also transfer knowledge from past and future or yet unseen tasks. In this vein, several state-of-the-art algorithms have tried to avoid catastrophic forgetting, an issue that has bugged the AI community for over 30 years. However, we have come up with a representation-ensembling approach that shows positive transfer for both past (forward) and future (backward) tasks. Our key insight is that even though learners trained on other tasks often cannot make useful decisions on the current task (the two tasks may have non-overlapping classes, for example), they may have learned representations that are useful for this task. Thus, although ensembling decisions is not possible, ensembling representations can be beneficial whenever the distributions across tasks are sufficiently similar. Moreover, we can ensemble representations learned independently across tasks in quasilinear space and time.

Despite the huge success in enabling forward and backward transfer using our representation-ensembling approach, we find the learning scenarios poorly defined in the literature. This makes solving lifelong learning problems difficult, especially in complicated learning scenarios where the agent must interact with the environment. In this vein, we make a small change to existing formal definitions of learnability by relaxing the assumption that the training data are sampled from the same distribution as the evaluation distribution. This change leads to the introduction of learning efficiency, which quantifies the amount a learner can leverage data for a given problem, regardless of whether it is an in- or out-of-distribution problem. We then prove the relationship between various generalized notions of learnability. We show that weak and strong OOD learnability are different, even though they are the same for in-distribution learnability. Finally, we show how this framework is sufficiently general to characterize transfer, multitask, meta, continual, and lifelong learning. We hope this unification helps bridge the gap between empirical practice and theoretical guidance in real world problems and provides insight into the gap between natural and machine learning.

Preventing catastrophic forgetting while continually learning new tasks is an essential problem in lifelong learning. Structural regularization (SR) refers to a family of algorithms that mitigate catastrophic forgetting by penalizing the network for changing its “critical parameters” from previous tasks while learning a new one. The penalty is often induced via a quadratic regularizer defined by an importance matrix, e.g., the (empirical) Fisher information matrix in the Elastic Weight Consolidation framework. In practice and due to computational constraints, most SR methods crudely approximate the importance matrix by its diagonal. In this study, we propose Sketched Structural Regularization (Sketched SR) as an alternative approach to compress the importance matrices used for regularizing in SR methods. Specifically, we apply linear sketching methods to better approximate the importance matrices in SR algorithms, which can improve the performance of SR algorithms with diagonal approximation.

1.1 Big Wins

We have achieved a significant improvement in backward and forward transfer. More specifically, the state of the art of CIFAR 10x10 **forward transfer was negative for all tasks**, **whereas for our algorithms, forward transfer was positive for all tasks.** Similarly, the previous state of the art for **backward transfer was negative for all tasks, whereas for our algorithms backwards transfer was positive for all tasks.**

Sketched structural regularization (sketched SR) methods achieve the following improvements. (1) Sketched SR consistently outperforms its diagonal counterpart on overcoming catastrophic forgetting, in both synthetic experiments and benchmark lifelong-learning tasks, including permuted-MNIST and CIFAR-100. (2) SCP ++, a special algorithm of the sketched SR methods, has been integrated into HRL CARLA and STELLAR systems, which significantly improves the performance maintenance of the systems. (3) We developed theoretical guarantees to the approximation ability of sketched SR.

1.2 What did not work

We have not deployed our algorithm in any real lifelong learning environment where the agent interacts continuously with the environment. More work is needed to provide theoretical guarantees and bounds to our proposed algorithm. Moreover, the tasks could be weighted based on task-similarity between the tasks that may dampen the effect of adversarial tasks on the current task performance. Again, there is a popular view that any amount, no matter the quantity, of source data can improve performance on the target data. Therefore, the standard practice of throwing all the source data available to a blackbox algorithm and expect to get better performance. However, our experiments show that we do not need infinite source data, in fact, more source data can impair, rather than improve, performance. Instead, we need a finite amount of good source data. Being able to intelligently choose source data to improve the target task may boost the performance of the lifelong learning algorithms more. This is a work in progress.

1.3 Lessons learned

Lifelong learning is an incredibly complicated scenario for learning machines. The L2M program catalyzed the global effort towards this goal, but the community remains at the early stages. To date, there are meager formal definitions of what lifelong learning is and is not. There are many papers that assert that they solve aspects of lifelong learning without clarifying the details of which kinds of problems they are even purporting to solve. Theory is nearly utterly lacking. And the empirical work often fails to demonstrate any transfer at all, which is the heart of lifelong learning problems. So, while the program really pushed the field forward, we have still only just begun.

2 Final Task Progress, Accomplishments, and Plans

*The Program Manager wants to receive this information in a **well-organized** and **systematic** manner. This should be a cumulative discussion of your project's progress to date. Back up all your claims with actual data.*

The overall technical area (TA) wins, and solutions can be summarized as:

- Proposed three key metrics to measure the performance of a lifelong learning agent.
- Implemented and demonstrated the efficacy of representation ensembling as a lifelong learning algorithm with quasilinear time and space complexity, the only method that exhibits both positive forward and backward transfer.
- Developed streaming forest to do streaming lifelong learning using random forest.
- Proposed weak and strong out-of-distribution learning (OOD) theorem and established learning efficiency as a generalized metrics to measure learnability for in-distribution, transfer, multitask, meta-, continual and lifelong learning.
- Demonstrated that for increasing source task sample size, target task risk does not decrease monotonically.
- Proposed a kernel density based generative model for better posterior calibration which eventually splits the task distribution into smaller parts for better transfer between tasks.

a. Lifelong Learning Metrics

Lifelong learning is a complicated learning setup where a learning agent constrained by resources sequentially learns multiple tasks and transfers knowledge between the tasks. Due to the dynamic nature of the learning environment, it is hard to ascertain whether a learner has learned all the tasks introduced so far or has forgotten some of the older tasks. We solved this problem by introducing three performance metrics characterizing three key aspects of lifelong learning, namely- improvement of performance on the current task by virtue of learning the previous tasks (forward transfer), improvement on past tasks by virtue of seeing the current task (backward transfer), and overall improvement on a particular task by virtue of having seen all the tasks (learning efficiency).

Learning efficiency is the ratio of the generalization error of an algorithm that has learned on one dataset, as compared to the generalization error of that same algorithm on a different dataset. Typically, we are interested in situations where the former dataset is a subset of the latter dataset. Let R^t be the risk associated with task t, and ${S^t}_n$ be the data from S_n that is specifically associated with task t, so $R^t(f({S^t}_n))$ is the risk on task t of the hypothesis learned by f only on task t data, and $R^t(f(S_n))$ denotes the risk on task t of the hypothesis learned on all the data.

Definition 1 (Learning Efficiency): The learning efficiency of algorithm f for given task t with sample size n is

$$ {LE_n}^t(f) = \frac{E\,[R^t(f({S_n}^t)]}{E[R^t(S_n)]} \qquad 1 $$

We say that algorithm f has learned all the tasks up to t with data S_n, if and only if ${LE_n}^t(f) > 1$ for all the tasks up to t.

Definition 2 (Forward Learning Efficiency): The forward learning efficiency of f for task t given n samples is

$$FLE_n{}^t(f) := \frac{E[R^t(f({S_n}^t))]}{E[R^t(f({S_n}^{\leq t}))]} \quad 2$$

We say an algorithm (positive) forward transfers for task t if and only if ${FLE_n}^t(f) > 1$. In other words, if ${FLE_n}^t(f) > 1$, then the algorithm has used data associated with past tasks to improve performance on task t.

Definition 3 (Backward Learning Efficiency): The backward learning efficiency of f for task t given n samples is

$$BLE_n{}^t(f) := \frac{E[R^t(f({S_n}^{\leq t}))]}{E[R^t(f(S_n))]} \quad 3$$

We say an algorithm (positive) backward learns task t if and only if ${BLE_n}^t(f) > 1$. In other words, if ${BLE_n}^t(f) > 1$, then the algorithm has used data associated with future tasks to improve performance on previous tasks.

Expected Progress: As lifelong learning is hard and there are many aspects of learning associated with a lifelong learning agent, we wanted to employ our performance metrics on a complicated streaming lifelong learning environment.

Actual Progress: We have used simplified learning scenarios where the data within each task arrives in batches to show the applicability of the proposed metrics. It establishes the foundation of evaluating performance in a more complicated environment as the proposed metrics are generalized enough to be adapted for a more complicated environment.

A major bottleneck while developing these performance indexes is that its application is closely tied with the learning system we have developed so far. As we have only established a foundational conjecture of representation ensembling to mitigate forgetting, the proposed metrics are not adapted for a more complicated setup.

b. Lifelong Learning through Representation Ensembling

Classical machine learning starts from a blank slate, or tabula rasa, using data only for the single task at hand. While typical transfer learning algorithms can improve performance on future tasks, their performance on prior tasks degrades upon learning new tasks (called forgetting). This has bugged the artificial intelligence (AI) community for over 30 years. However, trying not to forget the old tasks sets the bar unnecessarily low as biological agents not only learn the task at hand, but they also improve on all the

previous tasks by virtue of seeing all the tasks. In our work, we have introduced two representation-ensembling approaches: one is semiparametric synergistic forest (SynF) based on random forest and the other one is parametric synergistic network (SynN) based on deep-nets and demonstrated both positive forward and backward transfer over 10 tasks derived from CIFAR-100 dataset.

Our approach divides a learning model into three constituent parts: encoder (encodes input data to a representation space amenable to inference), channel (takes the representation from the encoder and estimates the distribution of the task) and decoder (does inference based on the output of the voter). A detailed structure of our proposed approach is shown in Figure 1.

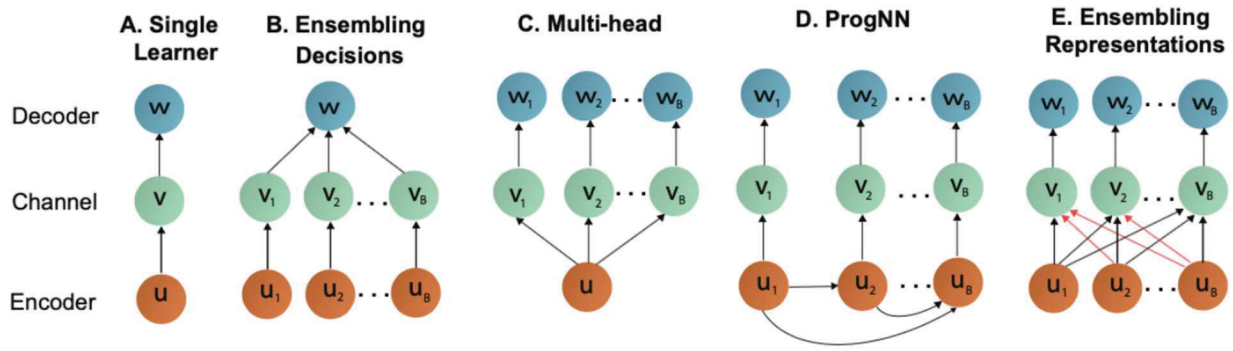

Figure 1: Schemas of Learning Models

We have demonstrated our lifelong learning capability on the CIFAR-10 X 10 dataset. It is a dataset with 10 tasks derived from the CIFAR-100 dataset. Each task consists of 10 classes each. Figure 2 shows positive forward and backward transfer for our proposed approach on CIFAR-10X10.

Expected Progress: We wanted to demonstrate we can not only overcome forgetting but also improve on the past, present and yet unseen tasks by virtue of learning independent representations for each task and ensembling them.
Actual Progress: One key thing to note is that our approach proposes a fundamental way to ensure transferability between tasks.

The insights achieved from this project led to the subsequent projects where we tried to enhance our implementation to work in streaming and challenging environments.

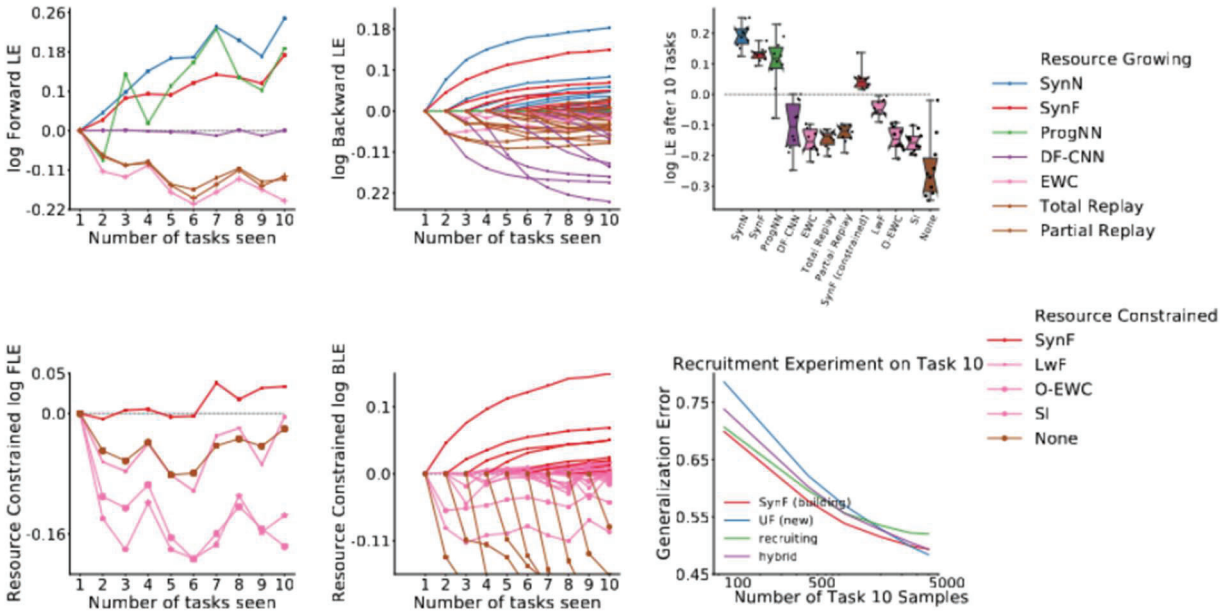

Figure 2: Performance of Different Algorithms on the CIFAR 10x10 Vision Experiments.

c. Streaming Decision Trees and Forests

Our proposed lifelong learning approach uses standalone learners (random forest or deep net) trained on individual tasks as the backbone of the learning system. A natural augmentation of our approach to streaming setup can be achieved by simply learning each task representation in a streaming manner. Learning representation based on deep nets in a streaming manner is straightforward whereas random forest does not have a rich literature of streaming approaches with efficient time and space complexity. We tried to solve this issue by proposing a simple streaming forest algorithm where we incrementally update the representation by splitting leaves when certain criteria are met. For a particular task, each update does not modify the existing partitions, and we assume that the task data distributions remain the same.

Expected Progress: We expected to adapt SynF to streaming classifications tasks while maintaining its forward and backward learning efficiencies. By adding new samples to a specific old task, streaming SynF would be able to improve the performance of all learned tasks. Please refer to the active PRs in section 5 for a streaming lifelong learning implementation on a simulation setup.

Actual Progress: We developed streaming tree algorithms for single tasks and started incorporating them with SynF. On three common datasets, we benchmarked streaming algorithms by incrementally updating them with 100-sample data batches. Our Stream Decision Forests (SDFs) and Stream Decision Trees (SDTs) consistently perform better than or close to existing methods, including Hoeffding Trees (HTs) and Mondrian Forests (MFs).

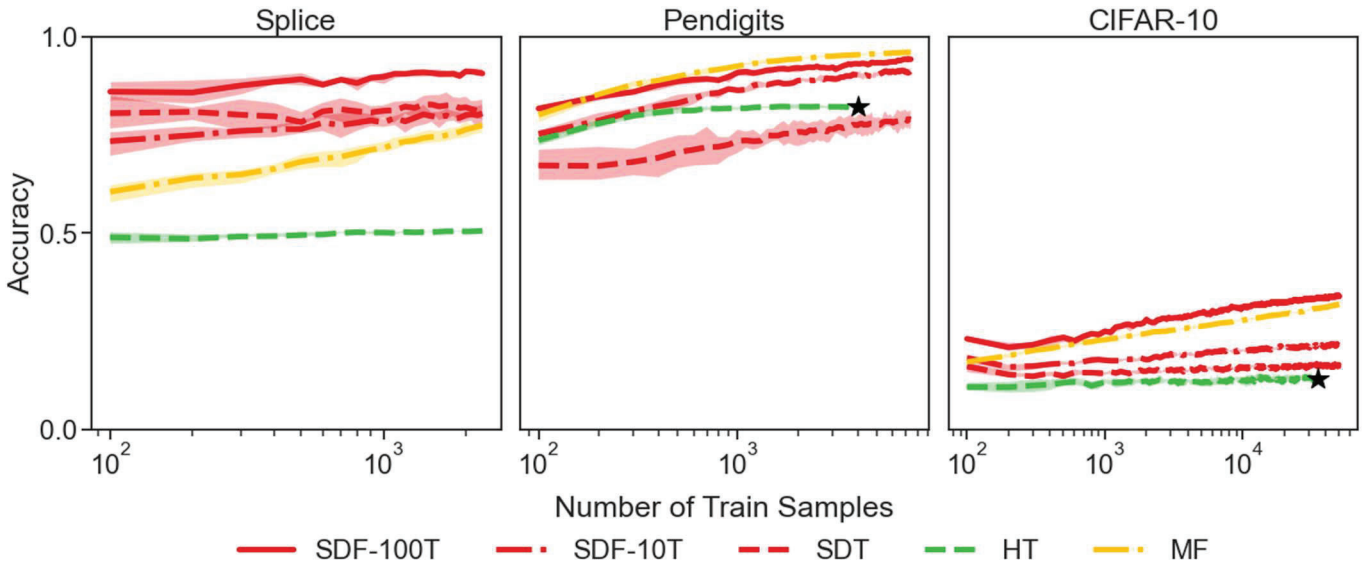

Figure 3: Multiclass Classifications on the Splice (left), Pendigits (center), and CIFAR-10 (right) Datasets.

Further steps include implementing resource constraints and adapting to domain shifts. Thus, our methods could complement our lifelong learning algorithms to address streaming data, rather than merely streaming tasks.

d. Out of distribution learning theory

There are many variations of out-of-distribution (OOD) learning, such as transfer learning, multitask learning, meta-learning, continual learning, etc. We have established a common framework that unites all the different variations under a single notation. Furthermore, we have proposed learning efficiency to quantify OOD learning. We know how to quantify in-distribution learning, nonetheless existing literature lacks in how to quantify OOD. There are many reasonable approaches to this. Finally, what can we prove about OOD learning? What guarantees do we have? What is out of reach? We seek to answer all these questions adequately with this subtask.

Expected Progress: Establish a framework which can formalize most, if not all, instances of OOD learning (such as transfer learning, multitask learning, etc). This framework should thus establish a common ground and notation for all these various scenarios. Furthermore, a few theorems about OOD learning, as well as its limits.

Actual Progress: We have successfully developed a framework that formalizes OOD learning under a common notation. We have enumerated many types of common OOD learning scenarios and have shown how they can be understood under our framework. We also successfully extended the concepts of strong and weak learning to the OOD case. We came up with a metric as well, learning efficiency, that has been successfully used to quantify learning progress by a learner, and for understanding how well a particular OOD learner performs. We proved a few very interesting results as well, such as the inequivalence of strong and weak learning in the OOD case (as opposed to their equivalence in-distribution).

In what follows we demonstrate some examples of theoretical results proved:

1. Weak OOD learning (performing better than chance) implies OOD learning (having learnt anything from the OOD data)
2. Strong (performing arbitrarily well) and weak learning are not equivalent in the OOD setting, unlike with in-distribution problems (due to boosting).

e. Non-Monotonic Out-of-Distribution Risk

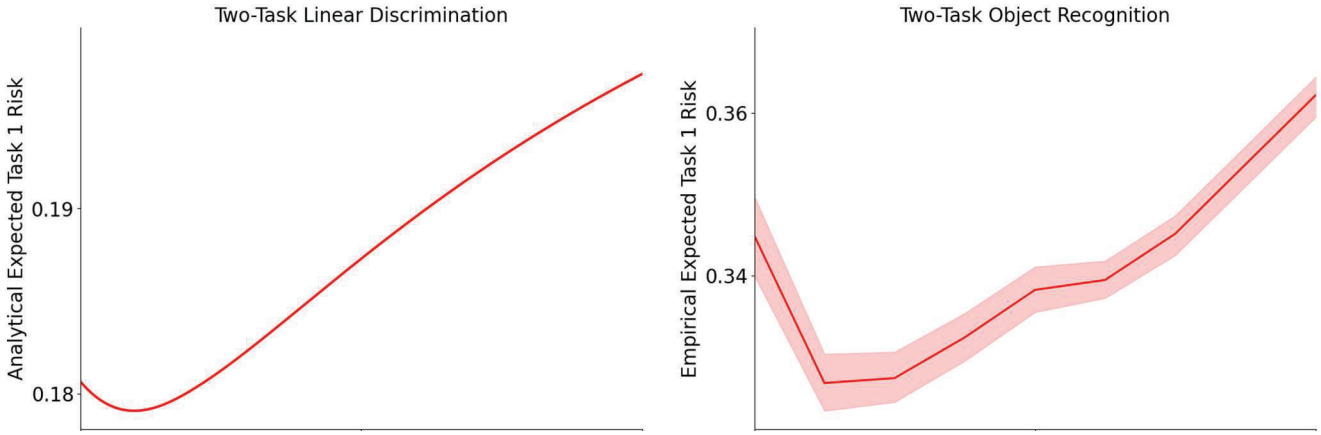

Figure 4: Non-monotonic Risk of Target Task as a Function of Source Task Sample Size.

We consider a learning agent that encounters two sequential tasks. Both are one-dimensional Gaussian classification tasks, so the only decision to be made is: where to draw a threshold. The only difference between the two tasks is that the categorization boundary in the second task is shifted slightly relative to the first. Importantly, the learning agent has not been overtrained on the first task. Rather it has learned something, but its estimate still has high variance, such that more related data could be helpful, or harmful. Our preliminary results illustrate a counter-intuitive result (**Fig 4, left**): *the performance on the first task is a non-monotonic function of the amount of data from the second task.* At first, performance on the first task gets better, but then, more data from the second task *impairs* performance. Is this merely a theoretical anomaly, or is non-monotonic performance something that really happens in the natural world in prospective learning settings? To test this, we then checked whether the same phenomenon occurs when using the classic machine learning benchmark CIFAR-10 (**Fig 4, right**) with a standard convolutional neural network. Increasing data from one task first leads to improved performance on a different task, but more data impairs performance. Note that this is in stark contrast to classical results in retrospective learning where more data *always* enhances performance. The idea that more data is always better is baked-in to the fundamental theorems of learning (which are typically stated in terms of: “if sample size is larger than N, then we converge to a desirable result”). This idea is also baked-in to the practice of machine learning, in which models are trained on more and more data, without considering the possibility that more data might actually hurt (because in retrospective learning, it does not). This simple scenario illustrates the inherent additional complexity of prospective learning including the potential cost of more data, which is heretofore unheard of.

f. Kernel Density Polytope:

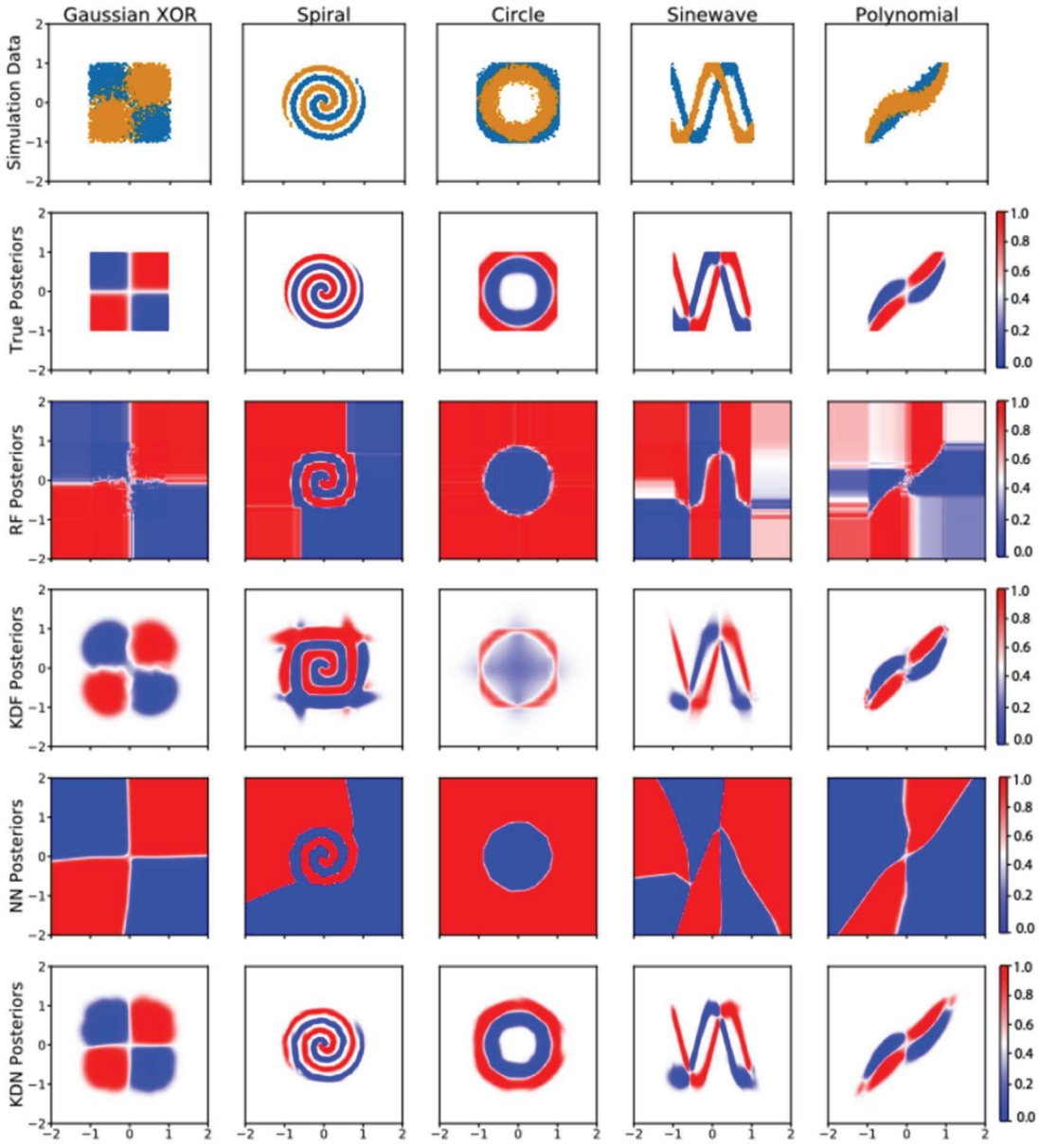

Figure 5: Simulation Distributions and Posterior Estimates by Different Algorithms.

The fight between discriminative versus generative goes deep, in both the study of artificial and natural intelligence. In our view, both camps have complementary value, so we sought to synergistically combine them. Here, we propose a methodology to convert deep discriminative networks to kernel generative networks. We leveraged the fact that deep models, including both random forests and deep networks, learn internal representations which are unions of polytopes with affine activation functions to conceptualize them both as generalized partitioning rules. We replace the affine activation within each polytope with a Gaussian kernel resulting in well-calibrated posteriors for in- and out-of-distribution regions and a huge compression in the number of parameters compared to those of their parent models.

In Figure 5, we have demonstrated the posteriors from our proposed approach for 5 different simulation datasets.

Expected Progress: Using kernel generative model we can split a task distribution into smaller constituent parts or polytopes. This will enable more efficient transfer learning between tasks by discarding the polytopes that are not useful for new tasks. Moreover, as we have a generative model, we do not need to save old task data.

Actual Progress: We have demonstrated the possibility of having more transfer between tasks using less samples. Please refer to the active PRs in section 5.1. This is a work under progress, and we are trying to prove theoretical properties and do elaborate experiments on image data using the proposed model.

3 Transition

PI Vogelstein worked in collaboration with Chris White at Microsoft Research (MSR). MSR cares deeply about lifelong learning and has provided datasets and new real-world challenges for us.

PI Vogelstein also recently began discussing lifelong learning with the founder of Sensie, a startup focusing on real-time delivery of wellness information based on body measurements.

We expect to transfer knowledge explicitly to both MSR and Sensie by way of code and other IP.

4 Publications and Presentations

Table 1: List of Publications

Title, Authors	Description/Type	Status
Representation Ensembling for Synergistic Lifelong Learning with Quasilinear Complexity, **Authors**:Jayanta Dey, Joshua Vogelstein, Hayden Helm, Will Levine, Ronak Mehta, Ali Geisa, Haoyin Xu, Gido van de Ven, Emily Chang, Chenyu Gao, Weiwei Yang, Bryan Tower, Jonathan Larson, Christopher White, Carey Priebe	arxiv preprint	Under Review by Transactions on Machine Learning Research
Towards a theory of out-of-distribution learning **Authors:** Ali Geisa, Ronak Mehta, Hayden S Helm, Jayanta Dey, Eric Eaton, Jeffery Dick, Carey E Priebe, Joshua T Vogelstein	arxiv preprint	Under Review by JMLR
Simplest Streaming Trees **Authors:** Haoyin Xu, Jayanta Dey, Sambit Panda, Joshua T Vogelstein	arxiv preprint	In preparation
Deep discriminative to kernel generative modeling **Authors:** Jayanta Dey, Ashwin De Silva, Will LeVine, Jong Shin, Haoyin Xu, Ali Geisa, Tiffany Chu, Leyla Isik, Joshua T Vogelstein	arxiv preprint	In preparation
When are Deep Networks really better than Decision Forests at small sample sizes, and how? **Authors**: Haoyin Xu, Kaleab A. Kinfu, Will LeVine, Sambit Panda, Jayanta Dey, Michael Ainsworth, Yu-Chung Peng, Madi Kusmanov, Florian Engert, Christopher M. White,	arxiv preprint	In preparation

Joshua T. Vogelstein, Carey E. Priebe		
Lifelong Learning with Sketched Structural Regularization **Authors:** Haoran Li, Aditya Krishnan, Jingfeng Wu, Soheil Kolouri, Praveen K. Pilly, Vladimir Braverman	conference paper	accepted at ACML 2021

5 Software packages / Code repos

Table 2: Code Repositories

Title	Description/Type	Link to Repository	Link to associated publication
ProgLearn	Software package to experiment lifelong learning in simulated and real data scenarios.	https://github.com/neurodata/ProgLearn	https://arxiv.org/abs/2004.12908
KDG	Code repos	https://github.com/neurodata/kdg	https://arxiv.org/abs/2201.13001
SDTF	Software package to use streaming decision tree.	https://github.com/neurodata/SDTF	https://arxiv.org/pdf/2110.08483
CoreL2M	Code repos	https://github.com/lihr04/corel2m	https://proceedings.mlr.press/v157/li21b/li21b.pdf

a. Active PRs:

i. **Segmentation of the ISIC dataset using proglearn.**

ii. **Streaming lifelong learning forest.**

iii. **Lifelong learning using kernel density graph.**

6. Appendix I – Publications

Please attach full copies of all relevant publications.

2. Li, Haoran, Aditya Krishnan, Jingfeng Wu, Soheil Kolouri, Praveen K. Pilly, and Vladimir Braverman. "Lifelong Learning with Sketched Structural Regularization." In *Asian Conference on Machine Learning*, pp. 985-1000. PMLR, 2021.
3. Vogelstein, Joshua T., Jayanta Dey, Hayden S. Helm, Will LeVine, Ronak D. Mehta, Ali Geisa, Haoyin Xu et al. "Representation Ensembling for Synergistic Lifelong Learning with Quasilinear Complexity." *arXiv preprint arXiv:2004.12908*(2020).
4. Geisa, Ali, Ronak Mehta, Hayden S. Helm, Jayanta Dey, Eric Eaton, Jeffery Dick, Carey E. Priebe, and Joshua T. Vogelstein. "Towards a theory of out-of-distribution learning." *arXiv preprint arXiv:2109.14501* (2021).
5. Dey, Jayanta, Ashwin De Silva, Will LeVine, Jong M. Shin, Haoyin Xu, Ali Geisa, Tiffany Chu, Leyla Isik, and Joshua T. Vogelstein. "Deep discriminative to kernel generative modeling." *arXiv preprint arXiv:2201.13001* (2022).
6. Xu, Haoyin, Jayanta Dey, Sambit Panda, and Joshua T. Vogelstein. "Simplest Streaming Trees." *arXiv preprint arXiv:2110.08483* (2022).
7. Xu, Haoyin, Michael Ainsworth, Yu-Chung Peng, Madi Kusmanov, Sambit Panda, and Joshua T. Vogelstein. "When are Deep Networks really better than Random Forests at small sample sizes?." *arXiv preprint arXiv:2108.13637* (2021).

Acronyms