

IEEE

CDS NEWSLETTERS

Volume 18, Number 1

ISSN 1550-1914

February 2024

Development of Natural and Artificial Intelligence

Contents

1	[Deep Learning] Dialogue: My Understanding of Post-Selection Misconduct in Deep Learning	2
2	[Deep Learning] Dialogue: Validation Error with Post-Selection Present is Downward Biased for Test Error	4
3	[Deep Learning] Dialogue: The Luckiest Network on Validation Performs Average During Tests	8
4	[Deep Learning] Dialogue Summary: Is “Deep Learning” Misconduct and What Should Researchers Do?	12
5	[Post-Selection] Dialogue Initiation: Is Post-Selection Generalizable?	16
6	Technical News: AI and Brain-Computer Interfaces	18
7	IEEE TCDS Table of Contents	19
	Supplement to “Dialogue: Validation Error with Post-Selection Present is Downward Biased for Test Error”	25
	An Intuitive View of Hongxiang Qiu’s Dialogue: “Validation Error with Post-Selection Present is Downward Biased for Test Error”	32

1 [Deep Learning] Dialogue: My Understanding of Post-Selection Misconduct in Deep Learning



*Soham Sarode, Indian Institute of Information Technology, Jabalpur, India
Email: 21bcs203@iiitdmj.ac.in*

I have done some deep learning research. The following is my understanding after reading J. Weng [1], On Deep Learning Misconduct, in the reference list of the Dialogue Initiation [2] in the last issue of the Newsletter.

The initial misconduct in Deep Learning involves authors intentionally concealing unfavorable data during training, possibly due to concerns about diminished accuracy. The second misconduct is characterized as cheating, where authors wrongly label and report training data as independent test data, undermining the integrity of the results. These misdeeds undermine transparency, mislead the AI community, and cast doubt on the credibility of achievements in Deep Learning research.

The “Pure-Guess Nearest Neighbors” (PGNN) is to play a devil’s role so that laymen can see the “devil” Deep Learning better. PGNN considers a classification system, which has a Training stage but without a Test stage. At the Training stage, PGNN stores the entire fitting set F . If the query is in the fitting set F , PGNN simply outputs the stored label of the query. Otherwise, PGNN needs to guess a label. Because the time to create PGNN versions is finite but not bounded, the author can simply keep generating PGNN versions till one perfect version comes up. There is always a finite time during which the process of randomly and uniformly guessing all labels for the validation set V will come up with a lucky classifier whose guessed labels for samples in the V set are all correct.

The four properties of this PGNN algorithm are:

i) PGNN classifier using Post-Selection achieves a zero error rate on the validation set V (and also on the test T if the author has T); the author needs a finite storage and a finite time, but unbounded.

ii) PGNN may also post-select multiple m luckiest networks in the training stage.

iii) Like Deep Learning, PGNN used a test set T in the Post-Selection (training) stage, and therefore its reported performance is not trustworthy.

iv) Like Deep Learning, PGNN lacks an independent test and therefore it is not trustworthy.

This PGNN algorithm will take a long time to train as it is trial and error guesswork, like almost all Deep Learning networks in Post-Selection which hand-tunes parameters, but lastly, PGNN will take a finite amount of time.

This was the problem found. For any research, the diagnosis of the problem is very important, and then

finding the solution to it is the second step. For example, the Developmental Networks from Dr. Weng's group provides a solution to the local minima problem faced by Post-Selection.

This is my understanding. I am very curious about how to avoid this misconduct in deep learning by Developmental Networks. Having done deep learning research, I suddenly realized that "Deep Learning Misconduct" paper. I have read the paper but I may or may not be correct here.

References

- [1] J. Weng. On "deep learning" misconduct. In *Proc. 2022 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 1–8, Beijing, China, Dec. 9-11 2022. SciTePress. arXiv:2211.16350.
- [2] J. Weng. Dialogue initiation: Is post-selection in deep learning fatal to deep learning? *IEEE CDS Newsletters*, 17(1):725–728, 2023.

2 [Deep Learning] Dialogue: Validation Error with Post-Selection Present is Downward Biased for Test Error



Hongxiang Qiu, Michigan State University, Michigan, USA
Email: qiuhongx@msu.edu

In this Dialogue, I will first formulate a sensible model describing investigators' post-selection behavior and the two issues Juyang (John) Weng stated in the Dialogue initiation. While I try to restrict the use of mathematics, I must introduce mathematical notations to define this model precisely. After that, I will provide intuitive interpretations of the main theoretical result whose proofs will appear elsewhere. To illustrate ideas, I will focus on supervised learning with independent and identically distributed (i.i.d.) data, perhaps the simplest case of machine learning.

I first introduce one sensible way to model the two issues about Deep Learning that John raised. I use O to denote a generic data point drawn from a distribution P . In the supervised learning setting, typically $O = (X, Y)$ where X is the independent variable and Y is the label or outcome. Consider $F + V + T$ i.i.d. data O_1, \dots, O_{F+V+T} drawn from a distribution P . Here, $(O_i)_{i=1}^F$, $(O_i)_{i=F+1}^{F+V}$ and $(O_i)_{i=F+V+1}^{F+V+T}$ are the fitting, validation and testing data sets, respectively. I will describe their usage in more detail in the next paragraph. I treat F, V, T as deterministic (i.e., given) and so the only randomness in these data sets comes from the random data points. Let $L(g, O)$ be a fixed loss or prediction error of a prediction model g evaluated at a data point O . The simplest example is the squared error loss $L(g, O) = L(g, (X, Y)) = (Y - g(X))^2$. Let $E(g, I) = \frac{1}{|I|} \sum_{i \in I} L(g, O_i)$ be the prediction error $L(g, \cdot)$ of a prediction model g averaged over data points O_i with i in an index set $I \subseteq \{1, \dots, F + V + T\}$. When L is the squared error loss, E is a sample average squared error. I use $[a : b]$ to denote the index set $\{a, a + 1, \dots, b\}$ for any positive integers a and b with $a \leq b$, and use O_I to denote $\{O_i : i \in I\}$ for an index set I .

The investigator considers an algorithm N that takes as input a hyperparameter θ and a data index set I , and outputs a prediction model such as a neural network. For example, N might minimize the prediction error on the fitting data, i.e. search for $\operatorname{argmin}_{g \in \mathcal{G}} E(g, [1 : F])$ (potentially with regularization) by (stochastic) gradient descent within a function class \mathcal{G} . The hyperparameter θ might include the structure of the neural network, the initial weights, the step size of (stochastic) gradient descent, etc. The investigator considers $K \geq 2$ hyperparameters $\theta_1, \dots, \theta_K$ and obtain K prediction models $(g_k = N(\theta_k, [1 : F]))_{k=1}^K$ using the fitting data $O_{[1:F]}$.

Next, the investigator peaks validation errors $E(g_k, [F + 1, F + V])$ and picks the model index $\hat{k} =$

$\operatorname{argmin}_k E(g_k, [F + 1 : F + V])$ with the smallest validation error¹. This step is called *post-selection* by Weng [4] because \hat{k} is selected after looking at the validation error. The investigator then only reports $E(g_{\hat{k}}, [F + 1, F + V])$ as a test error, without reporting the total number K of candidate prediction models considered or the performance $E(g_k, [F + 1, F + V])$ of all K prediction models. The testing set $O_{[F+V+1:F+V+T]}$ is never used by the investigator, for example, because the investigator only splits the data into two subsets of sizes F and V without having a separate testing data set. If the investigator manually tunes the hyperparameter, we may conceive K as an unknown large number and treat the above model as an approximation.

It is important to note that $\hat{k} = \hat{k}(O_{[F+1:F+V]}, g_1, \dots, g_K)$ is a random variable depending on the validation data $O_{[F+1:F+V]}$ and the K prediction models g_1, \dots, g_K ; it does not only depend on the fitting data $O_{[1:F]}$. Hereafter, I omit the dependence of \hat{k} on $(O_{[F+1:F+V]}, g_1, \dots, g_K)$ from notations for conciseness.

Formally, in the above procedure, the issue of hiding “bad-looking” data raised by John is the failure to report K and $E(g_k, [F + 1, F + V])$ for all k . The issue of absence of test is that $g_{\hat{k}}$ is obtained using both fitting data $O_{[1:F]}$ and validation data $O_{[F+1:F+V]}$, and its sample prediction error $E(g_{\hat{k}}, [F + V + 1 : F + V + T])$ on the testing set $O_{[F+V+1:F+V+T]}$ —or the population version $\mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_{\hat{k}}, [F + V + 1, F + V + T])$ —is never evaluated.

Hereafter, I assume that all expectations exist and use \mathbb{E}_A to denote the expectation over a generic random element A . For example, the aforementioned population test error $\mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_{\hat{k}}, [F + V + 1 : F + V + T])$ is the prediction error of $g_{\hat{k}}$ with the randomness in the independent test set $O_{[F+V+1:F+V+T]}$ integrated out and the random prediction model $g_{\hat{k}}$ fixed. This expectation can be interpreted in the following way: obtain $g_{\hat{k}}$ using the fitting and validation data $O_{[1:F+V]}$, collect many test data sets $O_{[F+V+1:F+V+T]}$, compute the test error $E(g_{\hat{k}}, [F + V + 1, F + V + T])$ for all these data sets, and the expectation $\mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_{\hat{k}}, [F + V + 1 : F + V + T])$ is the probabilistic limit of the average over these test data sets when the number of test data sets goes to infinity (thus the term “population version”). Because $g_{\hat{k}}$ is independent of the testing data and E is a sample average, this expectation can also be viewed as the probabilistic limit of $E(g_{\hat{k}}, [F + V + 1, F + V + T])$ as the size T goes to infinity by law of large numbers.

In the following, I will interpret the main theoretical result. The formal statements and proof can be found [in the supplement](#). Consider the bias of the validation error $E(g_{\hat{k}}, [F + 1 : F + V])$ for estimating even the smallest population test error among all K prediction models,

$$\mathbb{E}_{O_{[F+1:F+V]}} E(g_{\hat{k}}, [F + 1 : F + V]) - \min_{k \in \{1, \dots, K\}} \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F + V + 1 : F + V + T]). \quad (1)$$

Here, the first expectation $\mathbb{E}_{O_{[F+1:F+V]}} E(g_{\hat{k}}, [F + 1 : F + V])$ can be interpreted as follows: collect many validation data sets $O_{[F+1:F+V]}$, compute the smallest validation error $E(g_{\hat{k}}, [F + 1 : F + V]) = \min_k E(g_k, [F + 1 : F + V])$, take the average over all validation data sets, and the expectation $\mathbb{E}_{O_{[F+1:F+V]}} E(g_{\hat{k}}, [F + 1 : F + V])$ is the probabilistic limit of this average as the number of validation data sets—instead of the validation data size V —goes to infinity. This procedure to interpret the expectation disagrees with common practice in machine learning where the validation data size V is large, but the expectation is still informative of the behavior of the smallest validation error $E(g_{\hat{k}}, [F + 1 : F + V])$ on a typical validation data set. Note that $\mathbb{E}_{O_{[F+1:F+V]}} E(g_{\hat{k}}, [F + 1 : F + V])$ cannot be interpreted as the probabilistic limit

¹If more than one prediction model achieves the minimum, \hat{k} is an arbitrary one of them.

of $E(g_{\hat{k}}, [F + 1 : F + V])$ as the validation data set size V goes to infinity. In contrast, the other expectation, $\mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F + V + 1 : F + V + T])$, may still be interpreted as the probabilistic limit of the sample average test error $E(g_k, [F + V + 1 : F + V + T])$ of model g_k as the test data set size T goes to infinity. It is important to note that the post-selected model $g_{\hat{k}}$ may or may not achieve the smallest population test error $\min_k \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F + V + 1 : F + V + T])$, because, conditioning on the fitting set, \hat{k} is a random variable depending on the validation set while $\operatorname{argmin}_k \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F + V + 1 : F + V + T])$ is a constant.

I have shown that, the bias of the validation error $E(g_{\hat{k}}, [F + 1 : F + V])$ for estimating even the smallest population test error among all K prediction models, the term in (1), is never positive and often negative—that is, $E(g_{\hat{k}}, [F + 1 : F + V])$ is overly optimistic on average for even the best candidate prediction model's population test error. Since the population test error of the post-selected model cannot be smaller than the best candidate prediction model's population test error, the $E(g_{\hat{k}}, [F + 1 : F + V])$ is overly optimistic on average for the population test error of $g_{\hat{k}}$.

I have further proven that the bias in (1) is zero if and only if, for some fixed model index $k^* \in \{1, \dots, K\}$, it holds that $L(g_{k^*}, O) \leq L(g_k, O)$ for P -almost every O and all $k = 1, \dots, K$; that is, there exists an optimal candidate model g_{k^*} with the smallest prediction error for every data point. Here, k^* is a fixed index (conditioning on the fitting data), whereas \hat{k} is a random index depending on the validation data. In this case, the randomness in \hat{k} through post-selection on the validation data degenerates, and $E(g_{\hat{k}}, [F + 1 : F + V])$ is unbiased for the population test error of $g_{\hat{k}}$.

This unbiasedness condition is extremely stringent and often not satisfied. One example of this unbiasedness condition is when all K candidate prediction models are identical, in which case considering K prediction models is redundant. For instance, the fitting algorithm N solves a convex optimization problem from an initial point θ , so every local minimum is also a global minimum. In the metaphor of lottery tickets in Weng [4], this corresponds to the case where all tickets are identical. As another example, the K hyperparameters θ_k differ dramatically so that only one leads to a reasonable prediction model while all others always produce useless, improbable predictions, regardless of the fitting data. In the metaphor of lottery tickets in Weng [4], this corresponds to a rigged lottery where the outcome of every lottery draw is pre-determined by a fixed scheme, the post-selected model. This might happen when the hyperparameter θ includes the step size for (stochastic) gradient descent. Only one candidate step size (e.g., 0.01) converges to a reasonable prediction model. In contrast, all other candidate step sizes are huge (e.g., 1000), always leading to divergence in the fitting algorithm N and insane prediction models after reaching a pre-specified maximum number of iterations. In the metaphor of lottery tickets in Weng [4], this example corresponds to the case where all other purchased lottery tickets must be disqualified to enter the draw. I anticipate all the above examples rarely happening in practice.

The above unbiasedness condition can also be empirically falsified with the data by inspecting $L(g_k, O_i)$ for each $i \in [F + 1 : F + V]$: if $L(g_{\hat{k}}, O_i) > L(g_k, O_i)$ for some $k \neq \hat{k}$ and some $i \in [F + 1 : F + V]$, then the unbiasedness condition fails, i.e., the bias cannot be zero. This is because, when the unbiasedness condition holds, $g_{\hat{k}}$ must achieve the smallest population test error, and thus $\operatorname{Prob}(L(g_{\hat{k}}, O) \leq L(g_k, O) \text{ for all } k) = 1$.

Except on rare occasions where this unbiasedness condition holds, reporting $E(g_{\hat{k}}, [F + 1 : F + V])$ alone is not only biased, but even worse, downward biased. It tends to provide an overly optimistic estimate

of the population test error of any candidate prediction model g_k . We should consider reporting a better estimator of a test error. Reporting the total number K of candidate models and all their validation errors $E(g_k, [F + 1 : F + V])$ is more informative and transparent, even for manually tuning hyperparameters.

The main concern in this Dialogue, the bias, is a restrictive concept. In particular, I do not show any result about the magnitude of the bias in (1) or the distance between the smallest validation error $E(g_{\hat{k}}, [F + 1 : F + V])$ and the smallest population test error $\min_{k \in \{1, \dots, K\}} \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F + V + 1 : F + V + T])$, and my result does not preclude the possibility (i) that the bias in (1) tends to zero, or (ii) that the smallest validation error $E(g_{\hat{k}}, [F + 1 : F + V])$ converges to the smallest population test error $\min_{k \in \{1, \dots, K\}} \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F + V + 1 : F + V + T])$ in probability, as the number of data points grows to infinity.

There is also substantial literature on post-model-selection inference in statistics; for example, Berk et al. [1], Javanmard and Montanari [2], Kuchibhotla et al. [3], among others. I conjecture that similar ideas could be applied to estimating and making inferences about the smallest population test error $\min_k \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F + V + 1 : F + V + T])$, the population test error $\mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_{\hat{k}}, [F + V + 1 : F + V + T])$ of the post-selected model $g_{\hat{k}}$ conditional on the fitting and validation sets $O_{[1:F+V]}$ (thus conditional on $g_{\hat{k}}$), or other test errors.

References

- [1] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, pages 802–837, 2013.
- [2] Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.
- [3] Arun K Kuchibhotla, John E Kolassa, and Todd A Kuffner. Post-selection inference. *Annual Review of Statistics and Its Application*, 9:505–527, 2022.
- [4] Juyang Weng. On “deep learning” misconduct. In *Proc. 2022 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 1–8, Beijing, China, Dec. 9-11 2022. SciTePress. arXiv:2211.16350.

3 [Deep Learning] Dialogue: The Luckiest Network on Validation Performs Average During Tests



Xiang Wu, Nanjing University of Science and Technology, Nanjing, China
 Email: wux0213@hotmail.com

Inspired by Weng’s Dialogue and his papers reporting the misconducts in deep learning [2], we conducted some visual navigation experiments in realistic synthetic environments to test a deep learning model that consists of a Convolution Neural Network (CNN) and a Long Short-Term Memory (LSTM).

A simulated visual navigation task, as a simplified version of Autonomous Driver Assistance Systems (ADAS) is conducted. For the ADAS, some great challenges beyond ADAS Level 2 need to be addressed: (1) the GPS information conflicts with vision (e.g., obstacles, traffic lights, and detours); (2) visual objects have not been seen before; (3) resolving conflicts requires a trade-off between short-term goal and long-term goal.

We created a simulated environment for teaching and testing an agent to walk along a sideway with the GPS signal from Unity 3D to indicate the long-term destination, as shown in Fig. 1 (a). In the meanwhile, the agent needs to learn the basic traffic rules with supervision: go at the green traffic light; stop at the red traffic light; and walk around obstacles. The agent’s head is equipped with two uniform-pixel cameras. Each camera has a horizontal view field of 135 degrees to the front. The dimension of each binocular image captured by these cameras is $w \times h \times d = 256 \times 72 \times 3$, where d represents red-green-blue. All the spatio-temporal data were divided into three disjoint sets, the fit set F , the validation set V , and the test set T . For resources of the networks, see [4].

This Dialogue experimentally demonstrates the statement in the title for a set of 30 CNN-LSTM networks that are fairly designed and have more networks than the 20 CNN-LSTM networks reported in [4].

In our experiments, the 30 CNN-LSTMs have randomly initialized neuronal weights but the same set of hyper-parameters. We traced the identity of these networks over 20 epoch batch-training where each epoch corresponds to going through all the spatiotemporal sequences once in the fit data set F . In contrast, the Developmental Network 2 (DN-2) learns incrementally so it does not need iterations at all. Therefore, it is not fair to compare the *batch* learning of 30 CNN-SLTMs with the *incremental* learning of the sole DN-2 here, because the latter does not need multiple iterations at all and its solution to all network parameters is in a closed form at each time t . Note that the frame time t is different from the epoch.

Fig. 2 reports how these CNN-LSTM networks reduce their fit error on the fit set F , through multiple epochs through the batch data on fit set F .

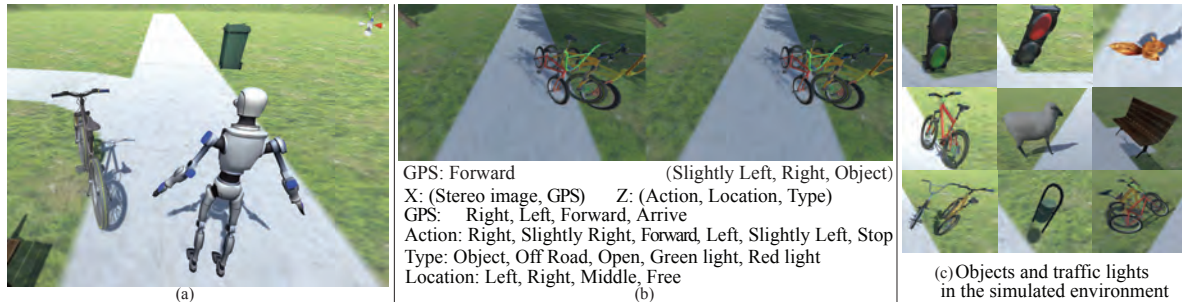


Figure 1: The settings of the simulated environment. In (a), the agent follows the GPS instructions to walk along the sideway. The simulated environment includes different lighting and vegetation conditions. The agent is equipped with two uniform-pixel cameras to capture binocular images. In (b), a binocular image sample with GPS instruction and corresponding motor supervision is listed. The supervision is formed by three motor concept zones. In (c), some objects in the simulated environment are shown. Some of these objects are used in the training scenes, while others are used in the disjoint test scenes. Namely, testing contains new objects that are not present in training. Adapted from [4].

At the end of 20 epochs, we find the luckiest network on the validation set V , as shown in Fig. 3. From the figure, we can see that the luckiest CNN-LSTM network indeed has the minimum validation error at epoch 20, but as we expected, it should not always be the luckiest network on other epochs 1 to 19.

We recorded the performances of all (30) CNN-LSTM networks on the test set T , as shown in Fig. 4. Note that epoch for CNN-LSTM in Fig. 4 is the iteration number through fit set F , but that for incremental learning DN-2 is the discrete time frame. They have very different meanings.

In summary, CNN-LSTM's performance is extremely sensitive to the luck of initial weights, which defeats the popular misunderstanding that Deep Learning is not sensitive to local minima. *What is importantly new is that the luckiest CNN-LSTM on validation only performs around the average of all (30) trained networks in a new test.* See in Fig. 4 the CNN-LSTM-L curve is around the CNN-LSTM-AVE curve (above and below, but near.) This experimental result confirms why all the "Deep Learning" methods should at least report the average validation error of all trained networks, as argued intuitively in [2] and mathematically proven by Weng in [3]. The luckiest error on the validation set V is expected to give only an *average* error in a disjoint test set T !

In contrast, the sole DN-2 reached the minimum error as early as epoch 1 by going through all time frames only *once* and only *incrementally*, much faster than any of these 30 CNN-LSTM networks. This is because DN-2 is optimal in the sense of maximum likelihood (ML) which is sufficient to reach the optimal solution within the first epoch! During 20 epochs, the performances of DN-2 are slightly across epochs different due to the effect of the synaptic maintenance mechanism [1].

References

- [1] Y. Wang, X. Wu, and J. Weng. Synapse maintenance in the where-what network. In *Proc. Int'l Joint Conference on Neural Networks*, pages 2823–2829, San Jose, CA, July 31 - August 5, 2011. NJ: IEEE Press.

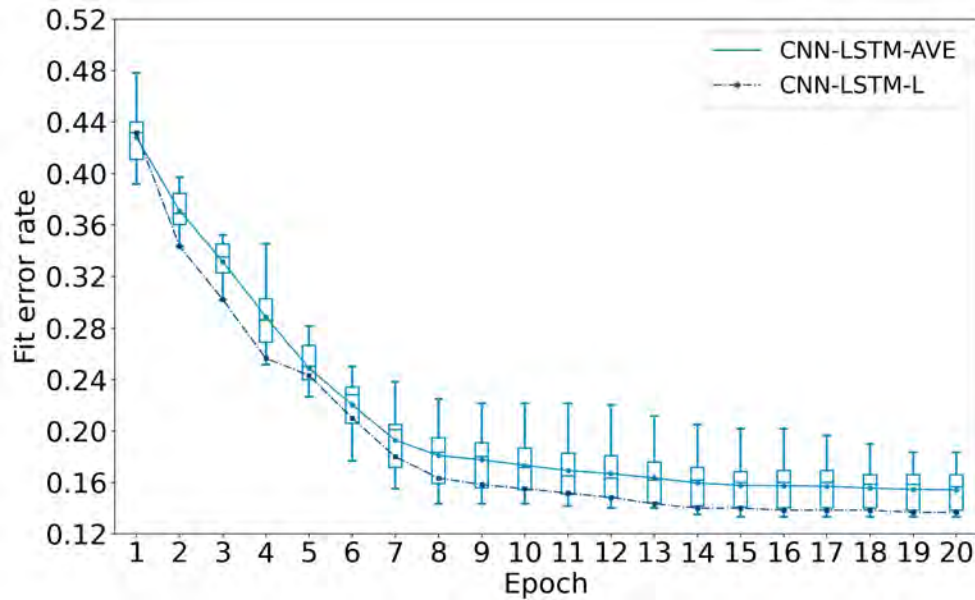


Figure 2: The distribution of the fit errors of 30 CNN-LSTM networks on fit set F . The boxplot (like candles) marks 5 values of the distribution of the ranked errors of 30 networks, namely the minimum (0%), 25%, median (50%), 75%, and maximum (100%). CNN-LSTM-AVE: average fit errors of 30 CNN-LSTM networks. CNN-LSTM-L: the fit error of the luckiest CNN-LSTM post-selected on the validation set V .

- [2] J. Weng. On “deep learning” misconduct. In *Proc. 2022 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 1–8, Beijing, China, Dec. 9-11 2022. SciTePress. arXiv:2211.16350.
- [3] J. Weng. Is ‘deep learning’ fraudulent in statistics? In *Proc. The 5th International Conference on Artificial Intelligence in Electronics Engineering (AIEE 2024)*, pages 1–8, Bangkok, Thailand, January 15-17 2024. NY: ACM Press.
- [4] X. Wu and J. Weng. The luckiest network gives the average error on disjoint tests: Experiments. In *Proc. The 5th International Conference on Artificial Intelligence in Electronics Engineering (AIEE 2024)*, pages 1–10, Bangkok, Thailand, January 15-17 2024. NY: ACM Press.

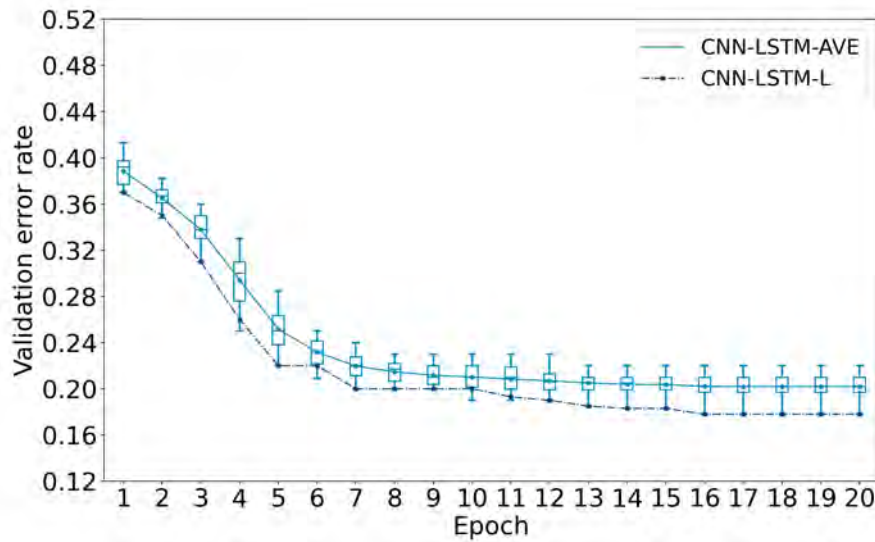


Figure 3: The distribution of the *validation* errors of 30 CNN-LSTM networks on validation set V . CNN-LSTM-AVE: average validation errors of 30 CNN-LSTM networks. CNN-LSTM-L: the validation error of the luckiest CNN-LSTM post-selected on the validation set V at epoch 20.

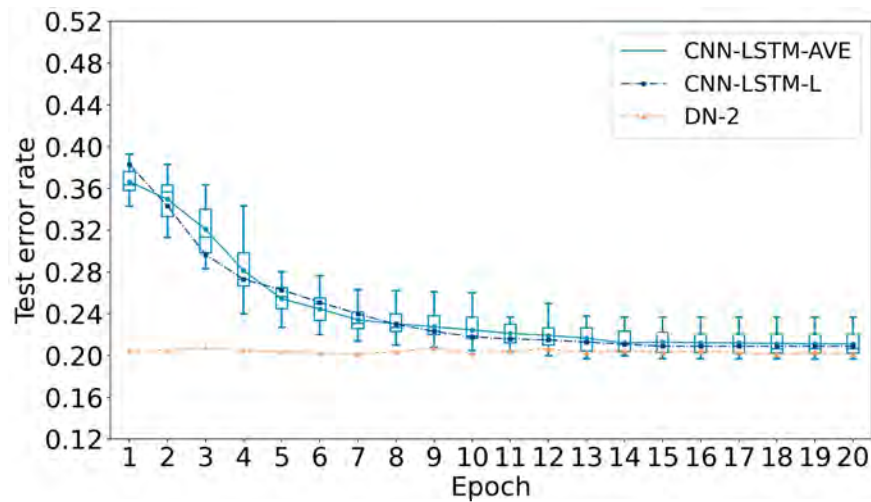


Figure 4: The distribution of the *test* errors of 30 CNN-LSTM networks on the set set T vs DN-2. CNN-LSTM-AVE: average test errors of 30 CNN-LSTM networks. CNN-LSTM-L: the test error of the luckiest CNN-LSTM post-selected on the validation set V at epoch 20. DN-2: The sole Developmental Network 2 trained by the fit set F .

4 [Deep Learning] Dialogue Summary: Is “Deep Learning” Misconduct and What Should Researchers Do?



*Juyang Weng, Brain-Mind Institute and GENISAMA, USA
Email: juyang.weng@gmail.com*

After my Dialogue Initiation was published in the last issue of the IEEE CDS Newsletter, I invited several authors of Deep Learning and LLMs to provide their positions and views on the Dialogue. These invited authors include Yann LeCun, Yoshua Bengio, Geoffrey Hinton, and Alec Radford. Unfortunately, none of them responded.

This issue of the Newsletter published three Dialogue responses, from Soham Sorode, Hongxiang Qiu, and Xiang Wu, respectively, as the readers can see above.

Let me first summarize responses to the six (6) questions that I raised in the Dialogue Initiation.

1. “It is simply a lack of perfection in the development of science. We should continue along this path.”

Among the three respondents, none expressed that we should continue along this path. All three respondents criticized the Deep Learning practice directly or indirectly. I would like to mention also some responses in a context other than this Dialogue. A colleague, also editor-in-chief of a journal, wrote informally to me that these Deep Learning articles are bad papers or bad reviews, but he did not let me cite his name. In response to my letter of complaint about Deep Learning misconduct sent to the IEEE Computational Intelligence Society, the 2023 president of the IEEE Computational Intelligence Society suggested “education”, but I did not see that he initiated any programs so far in his “education” direction. As the President of the IEEE Computational Society, he is responsible for establishing programs to correct the society-wide misconduct that I call cheating and hiding. Considering the current prevailing misconduct in almost all areas of the Society, has the society president done his job well?

2. “We need more transparency in the successes and failures of these tools.”

Soham Sorode wrote, “These misconducts undermine transparency, mislead the AI community, and cast doubt on the credibility of achievements in Deep Learning research.” Hongxiang Qiu wrote, “Reporting the total number of K of candidate models and all their validation errors . . . is more informative and transparent, even for manually tuning hyperparameters.” Xiang Wu’s experiment revealed how much the inflated performance is if one only reports the luckiest error on a validation set. He wrote, “*What is importantly new is that the luckiest CNN-LSTM on validation only performs around the average of all (30) trained networks in a new test.*” “This experimental result confirms why all the “Deep Learning” methods should at least report the average validation error of all trained networks.” From Xiang Wu’s response, the reader can see that

the luckiest CNN-LSTM (on the validation set) in his Fig. 3 (marked as CNN-LSTM-L) gives a test error in his Fig. 4 (marked as CNN-LSTM-L) that is around the mean error of 30 trained networks (marked as CNN-LSTM-AVE). The following is my hypothesis: If my PGNN model [2] were among Wu's candidate networks, it would have given a zero validation error at every epoch! This is drastically lower than all the curves in Wu's validation plot Fig. 3. However, PGNN's test error at all epochs in Wu's test error plot Fig. 4 would be only around the value of CNN-LSTM-AVE at Epoch 20. Sorode was correct in saying PGNN plays a devil's role. PGNN shows how bad the Post-Selection devil is!

3. "It is not true that Deep Learning amounts to cheating in the absence of a test if the paper is a willful repetition after being alerted of an early naïve mistake."

Qiu assumed this practice, "The testing set ... is never used by the investigator." After a company's upper-level management was alerted by me, the company should not willfully publish the same Post-Selection data in the absence of a test, even if we assume that the first Post-Selection paper from the same company was originally a naïve mistake. Such a willful publication amounts to willful cheating.

4. "It is not true that Deep Learning amounts to hiding bad-looking data if the paper is a willful repetition after being alerted for an early naïve mistake."

Under Qiu's model, "The issue of hiding 'bad-looking' data raised by John is the failure to report K and $E(g_k, [F + 1, F + V])$ for all k ." After the company's upper-level management was alerted by me, the company should not willfully publish Post-Selection papers that hide the performance data of all other trained networks, even if we assume that the first Post-Selection paper from the company was a naïve mistake. Such a willful publication amounts to the willful hiding of bad-looking data.

5. "Technically Deep Learning is correct because its reported error of the luckiest network on a validation set is a good estimate of the expected error on a future test."

To avoid ambiguity that is intrinsic in English, by "luckiest network", we always mean the luckiest network that is chosen based on its error on a validation set V , unless explicitly stated otherwise. Qiu wrote that the luckiest network is "downward biased" when it is compared with the luckiest network in a future test T , let alone that of the same network on test T . Wu's experimentally showed that Deep Learning is invalid because its reported error of the luckiest network on a validation set is inflated. In AIEE 2024 [2] I have theoretically proven (see Eq. (7) there) the theory behind Wu's experimental data. Namely, the minimum mean square error (MSE) of the luckiest network on a future test is the average error of all trained networks. This minimum MSE should be reported as the best estimate for a future test. However, Eq. (7) in [2] is only for randomly initialized parameters such as weights, not including hand-tuned hyper-parameters yet. Below, I will discuss my latest proof that all hand-tuning parameters are random since they depend on the random fit set and the random validation set. Qiu's Dialogue also stated that the performances should be reported for "manually tuning hyperparameters".

6. "Ethically Deep Learning is incorrect because only the average error of all trained networks on a validation set is a good estimate of the expected error on a future test."

Wu's data experimentally supports this statement. As I mentioned above, my AIEE 2024 paper [2] theoretically proved Wu's experimental data for the case of randomly initialized parameters such as weights. See below for hand-tuned hyperparameters.

In the following, I comment on each of the three Dialogues.

The author of the first Dialogue response, Soham Sorode, has done some deep-learning research. However, he seems to be more open and honest than many researchers who have done the same alleged misconduct. Honesty should be a minimum standard for a scholar.

The author of the second Dialogue response, Hongxiang Qiu, is a biostatistician. He seems to have the required statistical background to see through the weakness of Post-Selection. As far as I am aware, his result about the downward bias of the luckiest network is new. Although his proof is about expectation, the expectation tells the average behavior of each random sample from experiments. Qiu's derivation has many mathematical details that are not very accessible to some of our AI developers. My 37 rounds of reviews and interactions with Qiu mainly focused on his presentation about intuition. I thank Qiu for promptly giving me 37 rounds of responses that considerably improved the presentation of intuition. *In the supplement*, I provide an intuitive view to assist those who feel that Qiu's mathematical details are challenging.

The author of the third Dialogue response, Xiang Wu, presented a very important experimental result. As far as I know, he produced the first experiment in AI where a post-selected agent was subject to a new test! All Deep Learning results, as well as many other methods that use Post-Selections, typically report only the luckiest validation error. Few of them reported the *average* validation error, e.g., see [1]. Let us view the three data plots Figs. 2 to 4 in Wu conjunctively to see the following six points.

(i) The fit errors of 30 CNN-LSTM networks in Fig. 2 fluctuate, but have a trend to slowly reduce when the networks go through the same set of fit data set F as epochs. This is reasonable because the fit error on F is the objective function to be minimized by a gradient-based learning method.

(ii) The validation errors in Fig. 3 are typically larger than the corresponding fit errors in Fig. 2. This is also reasonable because the validation errors are not directly minimized by the gradient-based method. In particular, for the luckiest CNN-LSTM-L network on validation, its validation errors are typically larger than its corresponding fit errors.

(iii) Likewise, the test errors in Fig. 4 are typically also larger than the corresponding fit errors in Fig. 2.

(iv) The validation errors in Fig. 3 and test errors in Fig. 4 tend to have similar distribution ranges at each Epoch. Of course, the similarity depends on the similarity of distributions between the validation set V and test set T .

(v) I have proven using a statistical framework in [3, Theorem 3] and [4, Theorem 7], that every trained network has the same expected error in a future test, regardless of its luck on the validation set. This is true with not only randomly generated weight parameters but also hand-tuned hyperparameters. This conclusion is intuitive in lottery analogy, regardless of whether the lottery ticket numbers are randomly generated or hand-written. I hypothesize here that if the sizes of the fit set F and the test set T both go to infinity, the test errors of all 30 trained CNN-LSTM networks in Wu's experiment will approach the same value—their average, regardless of their ranks on validation V . In other words, the variance of the test errors in Fig. 4 will diminish with expanding F and T . We can see that the curves in Fig. 4 are already considerably *squashed* compared to Fig. 2. Likewise, if we repeatedly test 30 lottery tickets without bound, the average winning rates of the 30 lottery tickets will approach the same value—their average.

(vi) Whether or not all 30 CNN-LSTM networks have the same expected error in a future test depends also on whether any of them have a capability of abstraction. Unfortunately, this power is lacking with all the 30 CNN-STLMs because they lack an architecture of abstraction (e.g., emergent Turing machine or the

like). If a network has internal power to conduct abstraction, e.g., to learn invariance from sensorimotor samples, and to transfer the invariance to new but applicable contexts, such a network will have a smaller expected error in a test. From the motor actions in Fig. 1 of Wu, we can see that the abstract concepts, such as what, where, scale, orientation, and relation, have not sufficiently trained for DN-2. For example, the concept values for location are coarse (left, right, middle, free) and the concept for scale is absent. The resources for the DN-2 network are also very limited, which does not enable DN-2 to reach a drastically lower test error after the first epoch than what is shown. DN-2 needs more space to store weights than the CNN-LSTMs because it does not use convolution on purpose. Thus, the DN-2 network can abstract in its internal representations. However convolution cannot abstract any location-related concepts because convolution is shift-invariant.

It is worth noting that although both Hongxiang Qiu and Xiang Wu dealt with the performance of the luckiest network on the validation set, their comparisons are for different targets. Qiu's target is the expected error of the luckiest network on a disjoint test which could be a different network from the luckiest network on validation. Wu's target is the error of the same network when it is measured on a disjoint test. Qiu's target is typically smaller than Wu's target but never larger.

I hope that this Dialogue will inspire further discussions on this important issue that has plagued the AI discipline on a worldwide scale for many years. We added a short title [Deep Learning] for this Dialogue series so that this Dialogue can continue to accept new responses. If you like to submit a response to this Dialogue [Deep Learning], send your response to me and EIC Dongshu Wang by starting with the short title [Deep Learning] followed by your response title. Two different Dialogues, e.g., [Deep Learning] and [Post-Selection], can run in parallel in a future issue.

References

- [1] Q. Gao, G. A. Ascoli, and L. Zhao. BEAN: Interpretable and efficient learning with biologically-enhanced artificial neuronal assembly regularization. *Front. Neurorobot*, 15:1–13, June 1 2021.
- [2] J. Weng. On “deep learning” misconduct. In *Proc. 2022 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 1–8, Beijing, China, Dec. 9-11 2022. SciTePress. arXiv:2211.16350.
- [3] J. Weng. Misconduct in post-selection and deep learning. In *Proc. the 8th International Conf. on Control, Robotics and Sybernetics*, pages 1–9, Changsha, China, Dec. 22-24 2023. NJ: IEEE Press.
- [4] J. Weng. Conscious learning without post-selection misconduct. *International Journal of Humanoid Robotics*, 21(1):1–41, 2024. accepted and to appear.

5 [Post-Selection] Dialogue Initiation: Is Post-Selection Generalizable?



Xiang Wu, Nanjing University of Science and Technology, Nanjing, China
Email: wux0213@hotmail.com

In the last Dialogue labelled [Deep Learning] the main issue is not limited to so-called “Deep Learning”, but points to a technique called Post-Selection. Let me initiate a new Dialogue that covers a larger topic “Post-Selection”. Therefore, this new Dialogue is labeled [Post-Selection].

Post-Selection is a term originated in statistics, involving terms like Post-Selection Inference [1]. However, the post-selected network in Deep Learning has hardly been used for inference (i.e., lack of a test as Weng [2] argued).

The Post-Selection technique has been used in some other AI areas, such as evolutionary computation, genetic algorithms, and other areas that use randomly initialized trials, such as simulated annealing, random forests, swarm intelligence, and Extreme Learning Machines, to name a few.

Weng [3] raised: “Namely, it is a severe technical and protocol flaw in reporting only the luckiest network, regardless of the post-selection uses validation sets or test sets. At least the average error over Post-Selections must be reported. This conclusion has a great impact on evolutionary methods that often report only the luckiest network, instead of those of all networks in a population. Namely, the performances of all individual networks in an evolutionary generation should be reported. Furthermore, a reasonably disjoint test set must be used to evaluate the generalization of the luckiest network.”

I would like to raise the following topic for this Dialogue: Is the post-selected predictor useful for inference? If it is, in what sense? Is post-selected predictor generalizable for AI problems?

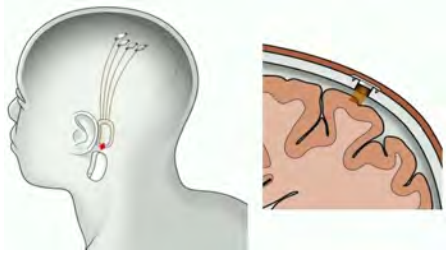
For possible publication in the next issue of the Newsletter, send your Dialogue response to me at wux0213@hotmail.com and EIC Dongshu Wang at wangdongshu@zzu.edu.cn by March 15, 2024. The size of your response is limited to 1 to 2 Newsletter pages.

References

- [1] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.
- [2] J. Weng. On “deep learning” misconduct. In *Proc. 2022 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 1–8, Beijing, China, Dec. 9-11 2022. SciTePress. arXiv:2211.16350.

- [3] J. Weng. Why deep learning's performance data are misleading. In *2023 4th Int'l Conf. on Artificial Intelligence in Electronics Engineering*, pages 1–10, Haikou, China, Jan. 6-8 2023. NY: ACM Press. arXiv:2208.11228.

6 Technical News: AI and Brain-Computer Interfaces



Elon Musk's Neuralink problems. Courtesy of medium.com

Neuralink founder Elon Musk announced on social media platform X that the first human patient implanted with the device from the brain chip startup company recovered well, with preliminary results indicating promising neural spike detection.

Similar experiments have been published before in research papers. Neuralink made a considerable step toward applications.

What is Neuralink for? In the short term, it's for helping people with paralysis. But that's not the whole answer. The idea is that these threads will read signals from a paralyzed patient's brain and transmit that data to an iPhone or computer, enabling the patient to control it with just their thoughts — no need to tap, type, or swipe.

But it's important to understand that this technology comes with staggering risks. Former Neuralink employees as well as experts in the field alleged that the company pushed for an unnecessarily invasive, potentially dangerous approach to the implants that can damage the brain (and apparently has done so in animal test subjects) to advance Musk's goal of merging with AI.

There are also ethical risks for society at large that go beyond just Neuralink. Several companies are developing tech that plugs into human brains, which can decode what's going on in our minds and has the potential to erode mental privacy and supercharge authoritarian surveillance. We have to prepare ourselves for what's coming.

7 IEEE TCDS Table of Contents

Volume 15, Issue 4, December 2023

Guest Editorial Special Issue on Hybrid Brain–Computer Collaborative Intelligent System

E. Q. Wu, P. Xiong, A. Song and P. X. Liu

3-D Tactile-Based Object Recognition for Robot Hands Using Force-Sensitive and Bend Sensor Arrays

X. Lu, D. Sun, H. Yin, et al

EEG-Based Emotion Recognition Using Trainable Adjacency Relation Driven Graph Convolutional Network

W. Li, M. Wang, J. Zhu and A. Song

Residual GCB-Net: Residual Graph Convolutional Broad Network on Emotion Recognition

Q. Li, T. Zhang, C. L. P. Chen, K. Yi and L. Chen

Causal Graph Convolutional Neural Network for Emotion Recognition

W. Kong, M. Qiu, M. Li, X. Jin and L. Zhu

Brain Biometrics of Steady-State Visual Evoked Potential Functional Networks

Y. Zhang, H. Shen, M. Li and D. Hu

Brain–Computer Interface Integrated With Augmented Reality for Human–Robot Interaction

B. Fang et al

A Novel Multiscale Dilated Convolution Neural Network With Gating Mechanism for Decoding Driving Intentions Based on EEGs

J. Sun, Y. Liu, Z. Ye and D. Hu

Graph Learning With Co-Teaching for EEG-Based Motor Imagery Recognition

Y. Zhang et al.

Exploratory Cross-Frequency Coupling and Scaling Analysis of Neuronal Oscillations Stimulated by Emotional Images: An Evidence From EEG

J. Chao, S. Zheng, C. Lei, H. Peng and B. Hu

Enhancing Visual Coding Through Collaborative Perception

L. An, Z. Yan, W. Wang, J. K. Liu and K. Yu

Clustering Based on Eye Tracking Data for Depression Recognition

M. Yang, C. Cai and B. Hu

Short-Interval Priming Effects: An EEG Study of Action Observation on Motor Imagery

Z. Sun, Y. -C. Jiang, Y. Li, J. Song and M. Zhang

Individual-Level fMRI Segmentation Based on Graphs

K. W. Tong, X. -Y. Zhao, Y. -X. Li and P. Li

A Spatiotemporal Channel Attention Residual Network With Extended Series Mean Amplitude Spectrum for Epilepsy Detection

Q. Wang, C. Huang, Q. Zeng, C. Li and T. Shu

Guest Editorial Special Issue on Emerging Topics on Development and Learning

D. Luo, A. Cangelosi, A. Sciutti, W. Wan and A. Tanevska

Interpretable Learned Emergent Communication for Human–Agent Teams

S. Karten, M. Tucker, H. Li, S. Kailas, M. Lewis and K. Sycara

Language-Model-Based Paired Variational Autoencoders for Robotic Language Learning

O. Özdemir, M. Kerzel, C. Weber, J. Hee Lee and S. Wermter

Unsupervised Multimodal Word Discovery Based on Double Articulation Analysis With Co-Occurrence Cues

A. Taniguchi, H. Murakami, R. Ozaki and T. Taniguchi

Trust in Robot–Robot Scaffolding

M. Kirtay, V. V. Hafner, M. Asada and E. Oztop

Efficient and Collision-Free Human–Robot Collaboration Based on Intention and Trajectory Prediction

J. Lyu, P. Ruppel, N. Hendrich, S. Li, M. Görner and J. Zhang

From State Transitions to Sensory Regularity: Structuring Uninterpreted Sensory Signals From Naive Sensorimotor Experiences

L. Goasguen, J. -M. Godon and S. Argentiari

Concurrent Skill Composition Using Ensemble of Primitive Skills

P. Dhakan, K. Kasmarik, P. Vance, I. Rañó and N. Siddique

An Ontology to Formalize a Creative Problem Solving Activity

C. Mercier

DisTop: Discovering a Topological Representation to Learn Diverse and Rewarding Skills

A. Aubret, L. Matignon and S. Hassas

A Platform for Holistic Embodied Models of Infant Cognition, and Its Use in a Model of Event Processing

M. Sagar et al.

Pick the Right Co-Worker: Online Assessment of Cognitive Ergonomics in Human–Robot Collaborative Assembly

M. Lagomarsino, M. Lorenzini, P. Balatti, E. D. Momi and A. Ajoudani

Interactive Robot Task Learning: Human Teaching Proficiency With Different Feedback Approaches

L. Hindemith, O. Bruns, A. M. Noller, N. Hemion, S. Schneider and A. -L. Vollmer

The Role of Object Physical Properties in Human Handover Actions: Applications in Robotics

N. F. Duarte, A. Billard and J. Santos-Victor

Visual Navigation Subject to Embodied Mismatch

X. Liu, D. Guo, H. Liu, X. Zhang and F. Sun

Do You Want to Make Your Robot Warmer? Make it More Reactive!

A. B. Giménez, E. Fernández-Rodicio, Á. Castro-González and M. A. Salichs

Kinematic Primitives in Action Similarity Judgments: A Human-Centered Computational Model

V. Nair et al.

Speakers Raise Their Hands and Head During Self-Repairs in Dyadic Conversations

E. E. Özkan, P. G. T. Healey, T. Gurion, J. Hough and L. Jamone

CBCL-PR: A Cognitively Inspired Model for Class-Incremental Learning in Robotics

A. Ayub and A. R. Wagner

REAL-X—Robot Open-Ended Autonomous Learning Architecture: Building Truly End-to-End Sensorimotor Autonomous Learning Systems

E. Cartoni, D. Montella, J. Triesch and G. Baldassarre

Performance-Based Iterative Learning Control for Task-Oriented Rehabilitation: A Pilot Study in Robot-Assisted Bilateral Training

Q. Miao et al.

A Deep Reinforcement Learning Algorithm Suitable for Autonomous Vehicles: Double Bootstrapped Soft-Actor–Critic-Discrete

J. Yang, J. Zhang, M. Xi, Y. Lei and Y. Sun

Machine Learning in Robot-Assisted Upper Limb Rehabilitation: A Focused Review

Q. Ai, Z. Liu, W. Meng, Q. Liu and S. Q. Xie

Leveraging Kernelized Synergies on Shared Subspace for Precision Grasping and Dexterous Manipulation

S. Katyara, F. Ficuciello, D. G. Caldwell, B. Siciliano and F. Chen

SOZIL: Self-Optimal Zero-Shot Imitation Learning

P. Hao, T. Lu, S. Cui, J. Wei, Y. Cai and S. Wang

Enable Fully Customized Assistance: A Novel IMU-Based Motor Intent Decoding Scheme

C. Yi et al.

Multidimensional Time-Series Life Cycle Costs Analysis of Intelligent Substation

Y. Jia and L. Ying

DeepCPG Policies for Robot Locomotion

A. M. Deshpande, E. Hurd, A. A. Minai and M. Kumar

Federated Reinforcement Learning for Collective Navigation of Robotic Swarms

S. Na et al.

YOLO-MS: Multispectral Object Detection via Feature Interaction and Self-Attention Guided Fusion

Y. Xie, L. Zhang, X. Yu and W. Xie

A Sim-to-Real Learning-Based Framework for Contact-Rich Assembly by Utilizing CycleGAN and Force Control

Y. Shi et al.

Low Resource-Reallocation Defense Strategies for Repeated Security Games With No Prior Knowledge and Limited Observability

J. Zhu, J. Zhang, Q. Ling and G. E. Dullerud

Obstacle Avoidance Learning for Robot Motion Planning in Human-Robot Integration Environments

Y. Hong, Z. Ding, Y. Yuan, W. Chi and L. Sun

An Efficient Graph Convolution Network for Skeleton-Based Dynamic Hand Gesture Recognition

S. -H. Peng and P. -H. Tsai

Alternated Greedy-Step Deterministic Policy Gradient

X. Wang, J. Zhang, Y. Gu, L. Huang, K. Yu and Y. Cheng

Event-Related Potential-Based Collaborative Brain-Computer Interface for Augmenting Human Performance Using a Low-Cost, Custom Electroencephalogram Hyperscanning Infrastructure

W. -J. Chen and Y. -P. Lin

A Bilateral Teleoperation System With Learning-Based Cognitive Guiding Force

Z. Ma, D. Shi, Z. Liu, J. Yu and P. Huang

Network Analysis on Cortical Morphometry in First-Episode Schizophrenia

M. Yin, W. Huang, Z. Liang, Q. Liu and X. Tang

Contour and Enclosed Region Refining for Contour-Based Instance Segmentation

W. Gu and S. Bai

Spatiotemporal Relationship Cognitive Learning for Multirobot Air Combat

H. Piao et al.

D2IFLN: Disentangled Domain-Invariant Feature Learning Networks for Domain Generalization

Z. Liu, G. CHen, Z. Li, S. Qu, A. Knoll and C. Jiang

EMG-Based Cross-Subject Silent Speech Recognition Using Conditional Domain Adversarial Network

Y. Zhang et al.

Supplement to “Dialogue: Validation Error with Post-Selection Present is Downward Biased for Test Error”

Hongxiang Qiu

Department of Epidemiology and Biostatistics, Michigan State University

In this supplement, I formally state the main theoretical results and provide elementary proof. I use the notations from the Dialogue.

Recall that O stands for a generic data point drawn from a distribution P . One might think that, because (i) both the validation data $O_{[F+1, F+V]}$ and the testing data $O_{[F+V+1, F+V+T]}$ are independent of the fitting data $O_{[1:F]}$, and (ii) $E(g, [F+1 : F+V])$, as a sample mean of i.i.d. random variables $L(g, O)$, is unbiased for the expectation $\mathbb{E}_O L(g, O)$, it must hold that $E(g, [F+1 : F+V])$ is unbiased for $\mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g, [F+V+1, F+V+T])$ for any given g obtained from the fitting data $O_{[1:F]}$, namely the bias $\mathbb{E}_{O_{[F+1:F+V]}} E(g, [F+1 : F+V]) - \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g, [F+V+1, F+V+T])$ is zero. Mathematically, since T is deterministic and the randomness in the testing data comes from its data points $O_{[F+V+1:F+V+T]}$, it holds that

$$\begin{aligned} & \mathbb{E}_{O_{[F+V+1, F+V+T]}} E(g, [F+V+1 : F+V+T]) \\ &= \mathbb{E}_{O_{[F+V+1, F+V+T]}} \frac{1}{T} \sum_{i=F+V+1}^{F+V+T} L(g, O_i) = \frac{1}{T} \sum_{i=F+V+1}^{F+V+T} \mathbb{E}_{O_i} L(g, O_i). \end{aligned}$$

Since each O_i is identically distributed, each $\mathbb{E}_{O_i} L(g, O_i)$ equals $\mathbb{E}_{O'} L(g, O')$ with O' denoting a data point independent of the fitting data $O_{[1:F]}$. Thus,

$$\begin{aligned} & \mathbb{E}_{O_{[F+V+1, F+V+T]}} E(g, [F+V+1 : F+V+T]) \\ &= \frac{1}{T} \sum_{i=F+V+1}^{F+V+T} \mathbb{E}_{O_i} L(g, O_i) = \frac{1}{T} \sum_{i=F+V+1}^{F+V+T} \mathbb{E}_{O'} L(g, O') = \mathbb{E}_{O'} L(g, O'). \end{aligned}$$

Similarly, $\mathbb{E}_{O_{[F+1:F+V]}} E(g, [F+1 : F+V]) = \frac{1}{V} \sum_{i=F+1}^{F+V} \mathbb{E}_{O_i} L(g, O_i) = \mathbb{E}_{O'} L(g, O')$ since the above argument holds with the testing data $O_{[F+V+1:F+V+T]}$ replaced by the validation data $O_{[F+1:F+V]}$. These equalities imply that

$$\mathbb{E}_{O_{[F+1:F+V]}} E(g, [F+1 : F+V]) = \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g, [F+V+1, F+V+T]) \quad (1)$$

for any given g that is obtained from the fitting data $O_{[1:F]}$ but does not depend on the validation data.

One might then make the following wrong conclusion based on a seemingly valid argument:

Invalid argument. By (1), $E(g_{\hat{k}}, [F+1 : F+V])$ must also be unbiased for $\mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_{\hat{k}}, [F+V+1, F+V+T])$:

$$\mathbb{E}_{O_{[F+1:F+V]}} E(g_{\hat{k}}, [F+1 : F+V]) - \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_{\hat{k}}, [F+V+1, F+V+T]) = 0, \quad (2)$$

because $g_{\hat{k}}$ is just a prediction modeled trained using $O_{[1:F]}$.

(1) is indeed correct with g independent of validation and testing data, but the fallacy in (2) arises from the fact that $g_{\hat{k}}$ further depends on the validation data $O_{[F+1:F+V]}$ as \hat{k} is a random variable whose randomness comes from both the fitting data $O_{[1:F]}$ and the validation data $O_{[F+1:F+V]}$. In other words, $g_{\hat{k}}$ is trained using both fitting data $O_{[1:F]}$ and validation data $O_{[F+1:F+V]}$ and depends on all K models, but (1) only holds for g trained using fitting data $O_{[1:F]}$ only, so the argument for (1) fails for $g_{\hat{k}}$.

I next show that the smallest validation error $E(g_{\hat{k}}, [F+1:F+V])$ is often downward biased for even the smallest population test error $\min_k \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F+V+1, F+V+T])$ among K prediction models. As mentioned in the Dialogue, it is important to note that the post-selected model $g_{\hat{k}}$ may or may not achieve the smallest population test error $\min_k \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F+V+1, F+V+T])$ among the K prediction models, because, conditioning on the fitting set, \hat{k} is a random variable depending on the validation set while $\operatorname{argmin}_k \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F+V+1:F+V+T])$ is constant and does not depend on any random data. There may be nonzero probability that the random variable \hat{k} is not a minimizer of $k \mapsto \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F+V+1:F+V+T])$.

Proposition 1. *Conditioning on the fitting set $O_{[1:F]}$, the bias of $E(g_{\hat{k}}, [F+1:F+V])$ as an estimator of $\min_k \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F+V+1, F+V+T])$,*

$$\mathbb{E}_{O_{[F+1:F+V]}} E(g_{\hat{k}}, [F+1:F+V]) - \min_k \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F+V+1, F+V+T]) \leq 0. \quad (3)$$

Proof. Proposition 1 follows immediately from Jensen's inequality. I provide an alternative elementary proof below.

For any fixed prediction model g_j ($j \in \{1, \dots, K\}$), we have

$$\begin{aligned} & \mathbb{E}_{O_{[F+1:F+V]}} E(g_{\hat{k}}, [F+1:F+V]) \\ &= \mathbb{E}_{O_{[F+1:F+V]}} \min_k E(g_k, [F+1:F+V]) && \text{(def of } \hat{k}) \\ &\leq \mathbb{E}_{O_{[F+1:F+V]}} E(g_j, [F+1:F+V]) && \text{(def of min)} \\ &= \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_j, [F+V+1, F+V+T]) && \text{(by (1)).} \end{aligned}$$

This inequality holds for all $j \in \{1, \dots, K\}$, and hence

$$\mathbb{E}_{O_{[F+1:F+V]}} E(g_{\hat{k}}, [F+1:F+V]) \leq \min_{j \in \{1, \dots, K\}} \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_j, [F+V+1, F+V+T]),$$

and (3) holds by replacing the dummy variable j in min with k . □

Thus, the bias of $E(g_{\hat{k}}, [F+1:F+V])$ as an estimator of the smallest population test error

$$\min_j \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_j, [F+V+1, F+V+T])$$

is never positive. It is important to note that the bias in (3) is not the bias for estimating the population test error $\mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_{\hat{k}}, [F+V+1, F+V+T])$ of the post-selected model $g_{\hat{k}}$ (conditional on the fitting and validation sets), because $g_{\hat{k}}$ might not achieve the smallest population test error among the K prediction

models.

One might argue that (3) is not a strict inequality, and so $E(g_{\hat{k}}, [F+1 : F+V])$ is still unbiased for $\min_j \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_j, [F+V+1, F+V+T])$ in some cases. I will next argue that

$$\mathbb{E}_{O_{[F+1:F+V]}} E(g_{\hat{k}}, [F+1 : F+V]) - \min_{k \in \{1, \dots, K\}} \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F+V+1, F+V+T]) = 0, \quad (4)$$

namely the equality in (3) holds so that $E(g_{\hat{k}}, [F+1 : F+V])$ is unbiased for estimating

$$\min_j \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_j, [F+V+1, F+V+T]),$$

only in special cases that I will describe next.

I have the following result on an equivalent condition for unbiasedness.²

Proposition 2. *Conditioning on the fitting set $O_{[1:F]}$, with O denoting a generic random data point drawn from P independent of the fitting set $O_{[1:F]}$, (4) holds if and only if*

$$\text{there exists } k^* \in \{1, \dots, K\} \text{ such that } \text{Prob}(O \in A) = 1, \quad (5)$$

where $A := \{o : L(g_{k^*}, o) \leq L(g_k, o) \text{ for all } k \in \{1, \dots, K\}\}$. Here, in contrast to \hat{k} , a random variable depending on the K prediction models and both fitting and validation data, k^* may depend on the fitting data but not the validation data.

Moreover, when (5) holds, the model g_{k^*} achieves the smallest population test error, namely

$$\mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_{k^*}, [F+V+1, F+V+T]) = \min_k \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F+V+1, F+V+T]), \quad (6)$$

and $\hat{k} \in \text{argmin}_k \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F+V+1, F+V+T])$ with probability one; that is, the distribution of \hat{k} is degenerate at k^* (if, in addition, the index set for optimal candidate models

$$\text{argmin}_k \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_k, [F+V+1, F+V+T])$$

contains only one element k^*).

The losses $L(g_{k^*}, O)$ and $L(g_k, O)$ are both random variables. They are the prediction errors of models g_{k^*} and g_k , respectively, on a data point O randomly drawn from P (independent of the fitting data $O_{[1:F]}$). For example, O may be a random data point in the validation or testing sets. The set $A = \{o : L(g_{k^*}, o) \leq L(g_k, o) \text{ for all } k \in \{1, \dots, K\}\}$ appearing in Proposition 2 may be uncountable, especially when the distribution P of O is continuous. The optimal candidate model index k^* is fixed conditioning on the fitting data; this is in sharp contrast with the post-selected model index \hat{k} , a random variable depending on the validation data conditioning on the fitting data.

²In general, if applied to a strictly convex/concave function, Jensen's inequality takes equality only when the random variable is a constant. The equality condition for the special case of Jensen's inequality in (3), as stated in Proposition 2, is less stringent, because min is not strictly concave.

Intuition. I next describe intuition for Proposition 2 before presenting the proof. I momentarily assume that P is discrete for simplicity. A rigorous argument can be made for an arbitrary distribution P , such as a continuous distribution, based on this intuition.

By the definition of \hat{k} , we have that the left-hand side of (4) equals

$$\mathbb{E}_{O_{[F+1:F+V]}} \min_k E(g_k, [F+1 : F+V]) - \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_{k^*}, [F+V+1, F+V+T]), \quad (7)$$

with a fixed (possibly unknown) model index k^* satisfying (6). The term in (7) further equals

$$\mathbb{E}_{O_{[F+1:F+V]}} \min_k E(g_k, [F+1 : F+V]) - \mathbb{E}_{O_{[F+1:F+V]}} E(g_{k^*}, [F+1, F+V]) \quad (8)$$

by (1), because k^* is fixed (even though k^* may be unknown) and g_{k^*} is trained using the fitting data only. Since both expectations are over the validation data $O_{[F+1:F+V]}$, we have that (8) equals

$$\mathbb{E}_{O_{[F+1:F+V]}} \left[\min_k E(g_k, [F+1 : F+V]) - E(g_{k^*}, [F+1 : F+V]) \right]. \quad (9)$$

In other words, the left-hand side of (4) equals (9). Note that the order of \min and \mathbb{E} has not been exchanged in the above derivation, even if the order might appear to have been exchanged when comparing (4) and (9).

Consider the random variable $Z := Z(O_{[F+1:F+V]}, g_1, \dots, g_K) := \min_k E(g_k, [F+1 : F+V]) - E(g_{k^*}, [F+1 : F+V])$ in the above expectation. Z may be interpreted as the gap between the smallest sample validation error and the best candidate prediction model's sample validation error, and depends on the validation data $O_{[F+1:F+V]}$ as well as the K prediction models. Since (i) (4) is equivalent to $\mathbb{E}Z = 0$, and (ii) $\min_k E(g_k, [F+1 : F+V]) \leq E(g_{k^*}, [F+1 : F+V])$ by definition so that $Z \leq 0$, we have that (4) is equivalent to $Z = 0$, that is, the smallest sample validation error always equals the sample validation error of g_{k^*} .

Moreover, $Z = \min_k E(g_k, [F+1 : F+V]) - E(g_{k^*}, [F+1 : F+V]) = 0$ is equivalent to $L(g_{k^*}, O) \leq L(g_k, O)$ for every k and every O : If $L(g_{k^*}, O) \leq L(g_k, O)$ for all k , it follows from the definition of E that $Z = 0$. The other direction can be shown by contradiction. Suppose that the opposite is true, namely $L(g_{k^*}, o) > L(g_{k'}, o)$ for some k' and some realized data point o . Consider the realization of the validation data being V copies of o . For this validation data, we have that $E(g_{k^*}, [F+1 : F+V]) > E(g_{k'}, [F+1 : F+V])$, that is, the validation error of g_{k^*} is not the smallest and thus $Z < 0$. In other words, Z does not equal zero for all validation data, a contradiction.

I will rely on the following rephrasing of a theorem in measure theory for probability measures, for example, Theorem 4.4.7 in Leadbetter et al. [1].

Theorem 1. Let G be a random variable. If $\mathbb{E}G = 0$ and $\text{Prob}(G \geq 0) = 1$, then $\text{Prob}(G = 0) = 1$.

I next prove Proposition 2.

Proof of Proposition 2. Proposition 2 is the equivalence between (4) and (5).

- (4) \Leftarrow (5): Suppose that (5) holds. Let $I \subseteq [F+1 : F+V+T]$ be any index set for data independent of the fitting set. Note that

$$\begin{aligned} B_I &:= \{o_I : L(g_{k^*}, o_i) \leq L(g_k, o_i) \text{ for all } k \in \{1, \dots, K\} \text{ and all } i \in I\} \\ &\subseteq C_I := \{o_I : E(g_{k^*}, I) \leq E(g_k, I)\}. \end{aligned}$$

Thus,

$$1 \geq \text{Prob}(O_I \in C_I) \geq \text{Prob}(O_I \in B_I) = 1 - \text{Prob}(O_I \in B_I^C)$$

where the complement of B_I is

$$B_I^C = \{o_I : L(g_{k^*}, o_i) > L(g_k, o_i) \text{ for some } k \in \{1, \dots, K\} \text{ and some } i \in I\}$$

With $D_i := \{o_I : L(g_{k^*}, o_i) > L(g_k, o_i) \text{ for some } k \in \{1, \dots, K\}\}$, we have $B_I^C = \cup_{i \in I} D_i$ and thus $0 \leq \text{Prob}(O_I \in B_I^C) \leq \sum_{i \in I} \text{Prob}(O_I \in D_i)$. Since $(O_i)_{i \in I}$ are i.i.d. copies of O distributed as P , (5) implies that $\text{Prob}(O_I \in D_i) = \text{Prob}(O \in A^C) = 1 - \text{Prob}(O \in A) = 0$ and therefore $0 \leq \text{Prob}(O_I \in B_I^C) \leq 0$. Thus, $1 \geq \text{Prob}(O_I \in C_I) \geq 1$ and $\text{Prob}(O_I \in C_I) = 1$. Taking I to be the testing data index $[F+V+1 : F+V+T]$, we have that $\text{Prob}(E(g_{k^*}, [F+V+1 : F+V+T]) = \min_k E(g_k, [F+V+1 : F+V+T])) = 1$ and thus (6) holds. Similarly, taking I to be the validation data index $[F+1 : F+V]$, we have that

$$\text{Prob}\left(\min_k E(g_k, [F+1 : F+V]) = E(g_{k^*}, [F+1 : F+V])\right) = 1, \quad (10)$$

that is, $\text{Prob}(Z = 0) = 1$. Therefore,

$$\begin{aligned} &\mathbb{E}_{O_{[F+1:F+V]}} E(g_{\hat{k}}, [F+1 : F+V]) \\ &= \mathbb{E}_{O_{[F+1:F+V]}} \min_k E(g_k, [F+1 : F+V]) && \text{(def of } \hat{k}) \\ &= \mathbb{E}_{O_{[F+1:F+V]}} E(g_{k^*}, [F+1 : F+V]) && \text{(by (10))} \\ &= \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_{k^*}, [F+V+1, F+V+T]) && \text{(by (1))} \\ &= \min_j \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_j, [F+V+1, F+V+T]), && \text{(by (6))} \end{aligned}$$

that is, (4) holds, namely (4) \Leftarrow (5).

- (4) \Rightarrow (5): I prove this direction by contradiction. Suppose that (5) does not hold. Consider k^* satisfying (6).

With O being a generic random data point distributed as P independent of the fitting data $O_{[1:F]}$ and $H_k := \{o : L(g_{k^*}, o) > L(g_k, o)\}$, I first show that there exists an index $k' \neq k^*$ such that

$$\text{Prob}(H_{k'}) \geq \varepsilon \quad (11)$$

for some constant $\varepsilon > 0$. The event in (5), $A = \{o : L(g_{k^*}, o) \leq L(g_k, o) \text{ for all } k \in \{1, \dots, K\}\}$, is equivalent to the following intersection of events

$$\bigcap_{k=1}^K H_k^C. \quad (12)$$

Since (5) does not hold by assumption, the probability of the complement event of (12) has non-zero probability, that is,

$$\text{Prob}\left(O \in \bigcup_{k=1}^K H_k\right) = \int \mathbb{1}_{\bigcup_{k=1}^K H_k}(o) dP(o) > 0.$$

Since

$$\text{Prob}\left(O \in \bigcup_{k=1}^K H_k\right) \leq \sum_{k=1}^K \text{Prob}(O \in H_k),$$

we have that

$$\sum_{k=1}^K \text{Prob}(H_k) > 0.$$

Therefore, there exists an index $k' \neq k^*$ such that $\text{Prob}(O \in H_{k'}) > 0$, and so there exists some constant $\varepsilon > 0$ such that $\text{Prob}(O \in H_{k'}) \geq \varepsilon$, i.e. (11) holds.

I next show that

$$\text{Prob}(Z < 0) > 0. \quad (13)$$

Because

$$\begin{aligned} J &:= \{o_{[F+1:F+V]} : L(g_{k^*}, o_i) > L(g_{k'}, o_i) \text{ for all } i \in [F+1 : F+V]\} \\ &\subseteq M := \{o_{[F+1:F+V]} : E(g_{k^*}, [F+1 : F+V]) - \min_k E(g_k, [F+1 : F+V]) > 0\}, \end{aligned}$$

where the latter set corresponds to the event $\{o_{[F+1:F+V]} : Z(o_{[F+1:F+V]}, g_1, \dots, g_K) < 0\}$, we have that

$$\text{Prob}(Z < 0) = \text{Prob}(O_{[F+1:F+V]} \in M) \geq \text{Prob}(O_{[F+1:F+V]} \in J).$$

Since $O_{[F+1:F+V]}$ are i.i.d. copies of O distributed as P independent of the fitting data $O_{[1:F]}$, by (11), $\text{Prob}(O_{[F+1:F+V]} \in J) = \prod_{i=F+1}^{F+V} \text{Prob}(H_{k'}) \geq \prod_{i=F+1}^{F+V} \varepsilon = \varepsilon^V > 0$, namely (13) holds.

I next show that $\text{Prob}(Z < 0) = 0$, a contradiction with (13). The left-hand side of (4) equals

$$\begin{aligned} &\mathbb{E}_{O_{[F+1:F+V]}} \min_k E(g_k, [F+1 : F+V]) - \mathbb{E}_{O_{[F+V+1:F+V+T]}} E(g_{k^*}, [F+V+1, F+V+T]) \\ &\hspace{20em} \text{(def of } \hat{k}) \\ &= \mathbb{E}_{O_{[F+1:F+V]}} \min_k E(g_k, [F+1 : F+V]) - \mathbb{E}_{O_{[F+1:F+V]}} E(g_{k^*}, [F+1, F+V]) \quad \text{(by (1))} \\ &= \mathbb{E}_{O_{[F+1:F+V]}} \left[\min_k E(g_k, [F+1 : F+V]) - E(g_{k^*}, [F+1 : F+V]) \right] \\ &= \mathbb{E}Z \quad \text{(def of } Z). \end{aligned}$$

Therefore, (4) is equivalent to $\mathbb{E}Z = 0$. By the definition of Z , $Z \leq 0$. Thus, with $G := -Z$, we have shown that $\mathbb{E}G = 0$ and $\text{Prob}(G \geq 0) = 1$. By Theorem 1, $\text{Prob}(G = 0) = \text{Prob}(Z = 0) = 1$, and so $\text{Prob}(Z < 0) = \text{Prob}(Z \leq 0) - \text{Prob}(Z = 0) = 1 - 1 = 0$. This is a contradiction with (13).

□

Intuitively, the above equivalent condition (5) for (4) means that the bias is zero if and only if (i) all truly best prediction models among the K models are identical, and (ii) the randomness in post-selection of \hat{k} degenerates because a truly best prediction model performs better than suboptimal ones on (almost) every data point so that $g_{\hat{k}}$ must be an optimal model among the K models.

References

- [1] Ross Leadbetter, Stamatis Cambanis, and Vladas Pipiras. *A basic course in measure and probability: Theory for applications*. Cambridge University Press, 2013. ISBN 9781139103947. doi: 10.1017/CBO9781139103947.

An Intuitive View of Hongxiang Qiu’s Dialogue: “Validation Error with Post-Selection Present is Downward Biased for Test Error”

Juyang Weng

Brain-Mind Institute and GENISAMA, USA

Hongxiang Qiu’s Dialogue deals with the expected error of the post-selected (the luckiest on the validation set V) network $g_{\hat{k}}$ among K trained networks $\{g_1, \dots, g_K\}$. Although the expected error is different from a sample error, we expect that if we do many experiments each of which gives a sample error, the average of sample errors approaches the expected error, according to the Law of Large Numbers.

Qiu has proven that if one reports the error of the luckiest network $g_{\hat{k}}$ on the validation set, the reported error is downward biased not only for the expected error of $g_{\hat{k}}$ in a future test but also for the expected error of the luckiest network g_{k^*} in a future test. The bias is zero if and only if the uncertainty from validation V to test T disappears in probability 1. If we use the lottery as an analogy, the luckiest ticket in the last lottery draw V will have worse expected luck in the future lottery draw T .

Therefore, Qiu’s result is very important. It says that post-selection is overly optimistic even for the luckiest predictor in a future test.

To assist those who feel the mathematics from Hongxiang Qiu is challenging, I slightly modify the notation below Eq. (3) in Qiu. Denote the fitting set, the validation set, and the test set as F , V , and T , respectively, but with their sizes all given (fixed). Note that $g_{\hat{k}}$ is a function of g_1, \dots, g_K and V , denoted as $g_{\hat{k}}(g_1, \dots, g_K, V)$. Then, conditioned on F , the four display expressions in Qiu below his Eq. (3) are rewritten below, using a slightly modified notation:

$$\begin{aligned}
 & \mathbb{E}_V E(g_{\hat{k}}(g_1, \dots, g_K, V)) \\
 &= \mathbb{E}_V \min_{g_j \in \{g_1, \dots, g_K\}} E(g_j, V) && \text{(def of } \hat{k} \text{)} \\
 &\leq \mathbb{E}_V E(g_j, V) && \text{(def of min)} \\
 &= \mathbb{E}_T E(g_j, T) && \text{(by i.i.d. in } V \cup T \text{).}
 \end{aligned}$$

The above inequality holds for all $g_j \in \{g_1, \dots, g_K\}$ and all $o \in V$, and hence

$$\mathbb{E}_V E(g_{\hat{k}}(g_1, \dots, g_K, V)) \leq \min_{g_j \in \{g_1, \dots, g_K\}} \mathbb{E}_T E(g_j, T),$$

and Eq. (3) in Qiu holds.

The iff (if and only if) condition is for the *sole* inequality (def of min) above to become equality. This observation might be intuitive for the reader to understand how to prove the iff condition.

Below, I slightly modify the iff condition in Proposition 2 of Qiu by introducing a concept called universe. In mathematics, a universe is a collection that contains all the entities one wishes to consider in a given situation. The universe U of a set $S = \{x \in U : g(x) = T\}$ is the domain of the proposition $g : U \mapsto \{T, F\}$. Outside the universe U , the proposition g is not defined. Since g is defined only in its domain, the universe

is necessary for any set.

Here, the universe $U = \mathcal{V}$, denoting the sample space of V , should be explicitly defined for the set A in Qiu. For any $g_{k^*} \in \{g_1, \dots, g_K\}$, define

$$A := \{o \in \mathcal{V} : L(g_{k^*}, o) \leq L(g_j, o) \text{ for all } g_j \in \{g_1, \dots, g_K\}\} \text{ with universe } \mathcal{V}. \quad (1)$$

Note that the set A is defined only on the universe \mathcal{V} . Conditioning on the fitting set F , the equality holds iff there exists $g_{k^*} \in \{g_1, \dots, g_K\}$ such that

$$\text{Prob}(o \in A) = 1 \quad (2)$$

for all $o \in \mathcal{V}$, or $P(\mathcal{V} - A) = P(\mathcal{V}) - P(A) = 1 - 1 = 0$. Namely, the luckiest g_{k^*} in our $\{g_1, \dots, g_K\}$ is the best for any $o \in \mathcal{V}$ with probability 1.

We rewrite the expression under the expectation in Eq. (9) of Qiu's supplement as

$$d(V) = \min_{g_k \in \{g_1, \dots, g_K\}} E(g_k, V) - E(g_{k^*}(g_1, \dots, g_K), V). \quad (3)$$

The above Eq. (2) means that the non-positive variable $d(V)$ becomes zero $d(V) \equiv 0$ (i.e., degenerated) with probability 1 in \mathcal{V} , because its expectation $\mathbb{E}_V d(V)$ is zero (Theorem 1 in Qiu's supplement). Therefore, due to the degeneracy in Eq. (2), the luckiest $g_{\hat{k}}$ on V must be the luckiest g_{k^*} on T . This is like a rigged lottery!

Specifically, the iff condition deals with two random variables, $E(g_{\hat{k}}(g_1, \dots, g_K, V))$ that denotes the first term in Eq. (3) and the second term in Eq. (3). Both terms depend on the random fitting set F , the set of K prediction models $\{g_1, \dots, g_K\}$, and the validation set V . However, the set A defined for the second term should use the universe \mathcal{V} to be tight for the only-if condition. This is because under expectation $\mathbb{E}_V d(V)$ the variable V may vary only in \mathcal{V} , which affects $d(V)$ in Eq. (3).