Shannon Inspired Approach to Limits of Learning

Contract No: W911NF-16-1-0561

Quarterly Report for Dec 1, 2016-Feb 28, 2017

PI: Vahid Tarokh

1 Introduction

Let us briefly recap the progress in the previous quarter. We considered the general problem of learning from data, where the objective of the learner is to build a generative/discriminative model based a set of historical data. The learner then would like to measure the performance of the inferred function with some objective in mind. In the previous quarter, the main focus was on the generalizability of the model class. Let $x^n = \{x_i\}_{i=1}^n$, where $x_i \in \mathcal{X}$, denote the data samples at the learner's disposal that are assumed to be drawn i.i.d. from an unknown joint density function $p(x)$. We assumed that the learner can express her objective in terms of minimizing a parametric loss function $\ell(x; \theta)$, which is a smooth function of the parameter vector θ. The learner solves for the unknown parameter vector $\theta \in \Theta$ where $\Theta \subseteq \mathbb{R}^k$. Further, let $L(\theta) := E\{\ell(x; \theta)\}$ be the risk associated with the parameter vector θ, where the expectation is with respect to the density $p(\cdot)$. Ideally, the goal of the learner is to choose the parameter vector θ^λ such that

$$\theta^* = \arg\min_{\theta \in \Theta} L(\theta) = \arg\min_{\theta \in \Theta} E\{\ell(x; \theta)\}. \tag{1}$$

Since the density function $p(\cdot)$ is unknown, the learner cannot compute θ^* and hence cannot achieve $L(\theta^*)$. Instead, she usually solves the empirical version of the problem and chooses $\hat{\theta}^\lambda(x^n)$ as follows:

$$\hat{\theta}^\lambda(x^n) = \arg\min_{\theta \in \Theta} \left\{ \hat{L}(\theta) \right\} = \arg\min_{\theta \in \Theta} \left\{ \hat{E}\{\ell(x_i; \theta)\} \right\}, \tag{2}$$

where $\hat{E}$ is the empirical mean operator with respect to the training samples x^n, i.e.,

$$\hat{E}\{f(x_i)\} = \frac{1}{n} \sum_{i=1}^{n} f(x_i), \tag{3}$$

and

$$\hat{L}(\theta) = \hat{E}\{\ell(x_i; \theta)\} = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i; \theta). \tag{4}$$

The question of whether a concept is asymptotically learnable in a hypothesis class (or a filtration of hypothesis classes) has been extensively studied and understood. On the other hand, the daunting task of machine learning is to understand how quickly this could be obtained and what is the best strategy of the learner given a finite data set. Most of the theoretical work on this front concerns worst-case and probabilistic upper bounds on the loss incurred when a new sample arrives. However, these bounds are usually pessimistic in nature and practitioners resort to methods such as cross validation for model selection and learning. We emphasize that the goal of this project is to devise a theoretical framework to obtain the limits of learning as new data arrives, which will in turn equip the learner with practical tools for finding the best learning machine within a given model class (limits of learning within each class) and also to be able to choose the optimal complexity within a filtration of model classes (limits of learning across classes)

In the previous quarter, we set out to answer this question and provided a framework inspired from information theory for estimating the out-of-sample prediction error from a given data set. In this quarter, we further provide a general definition of the limits of learning for vector-time series, and generalize the framework to include a general objective of the learner. Finally, given a data set x^n, we provide algorithms with practical complexity for achieving the limit of learning.

2 Definition of the Limit of Learning for Vector-Times Series

In this section, we provide the definition of the limit of learning.

Definition 1 (Limit of learning) *We define the limit of learning as*

$$LOL = I_U(X_{n+1}; X_1, \ldots, X_n, \Theta) - I_U(X_{n+1}; X_1, \ldots, X_n),$$

where $U(\cdot)$ can be thought of as the **measure** *of learning.*

Let us first provide two information theoretic interpretations for this definition by considering $I_U(\cdot;\cdot)$ to be the mutual information. In this case, X_{n+1} is distinguishable when $X_1, \ldots, X_n$ is at the learner's disposal if it is in the set of $2^{I(X_{n+1};X_1,\ldots,X_n)}$ jointly typical values (asymptotic equipartition property). Thus, the limit of learning as defined above is the amount of exponential loss in the error probability that is lost when the learner (who does not know Θ) predicts X_{n+1} from $X_1, \ldots, X_n$.

In the case where U is Shannon's entropy we have

$$I_U(X_{n+1}; X_1, \ldots, X_n, \Theta) - I_U(X_{n+1}; X_1, \ldots, X_n) = U(\Theta|X_1, \ldots, X_n) - U(\Theta|X_1, \ldots, X_{n+1}).$$

In this case, it is clear that $I_U(X_{n+1}; X_1, \ldots, X_n, \Theta) - I_U(X_{n+1}; X_1, \ldots, X_n) \geq 0$. In Section 3, we will discuss algorithms based sequential Monte Carlo methods to numerically compute the limit of learning.

Theorem 1 *Suppose that $X_1, \ldots, X_n$ are drawn i.i.d. from an unknown distribution p. Further assume that* $\dim(\Theta) = d$*. Then, under mild conditions, we have*

$$LOL = \frac{d}{2} \log \left(1 + \frac{1}{n}\right).$$

This result is reminiscent of Shannon's celebrated channel capacity formula, which is

$$C = \frac{K}{2} \log \left(1 + \frac{P}{\sigma^2}\right).$$

We believe that there is an analogous formula for the limit of learning for very general objectives of the learner when the data is drawn i.i.d. from an unknown distribution that does not necessarily belong to the model class. We also believe that this can be further extended to the dependent data under some assumptions. We are currently investigating these directions, and have derived some preliminary results.

3 Relationship with Model Selection

Next, we comment on the relation between the limit of learning (LOL) defined in the previous section and the task of model selection. Let us consider LOL again in the asymptotic limit as $n \to \infty$:

$$2\text{LOL} = d \log(1 + n) - d \log(n).$$

This resembles the Bayesian information criterion (BIC). Motivated by it, we propose to use the penalty function $-2U(\Theta|X_1, \ldots, X_n)$ instead of the BIC penalty function (which is reached in the asymptotic form of the LOL).

Lemma 1 *For a well-specified parametric case with finite dimensional parameters, the LOL-inspired model selection penalty is consistent, and achieves the same performance as BIC as $n \to \infty$.*

This result is expected given the asymptotic relation between LOL and BIC. Further, while the LOL-inspired criterion is asymptotically good, the question is how can one compute this criterion in practice.

While the new penalty function is analogous to BIC as n grows, the question is whether it would perform better than BIC for finite n. To answer this question, we present a simple experiment on the selecting the autoregressive (AR) order from AR class of orders $1, \ldots, 10$. In the experiment, we use the conditional mutual information as the LOL inspired model

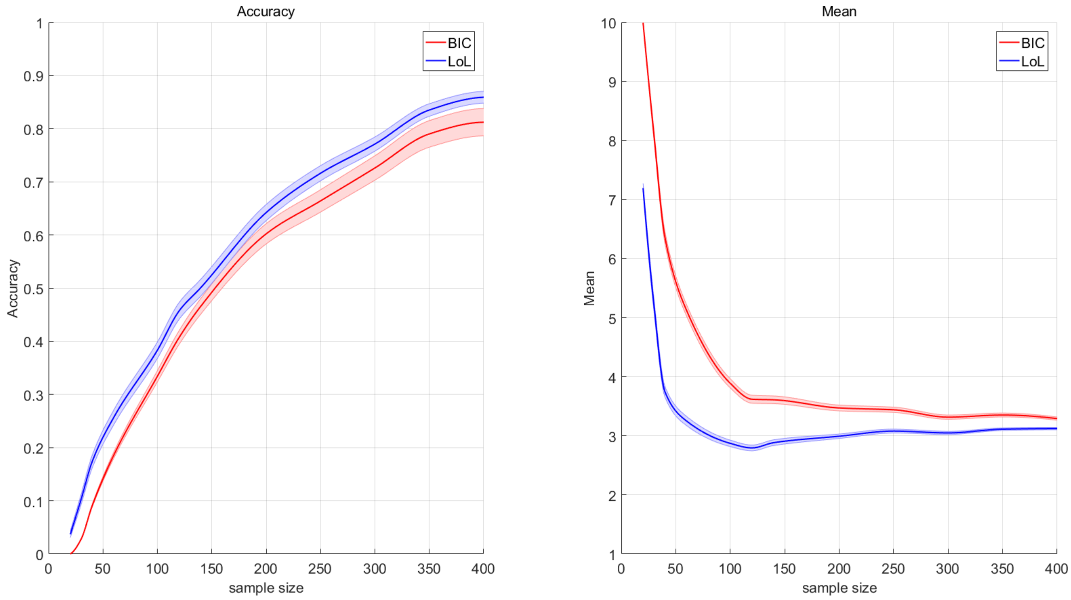

Figure 1: The left plot shows the empirical accuracy of selecting the true model using LOL and BIC. The right plot shows the average model order selected by LOL and BIC as a function of the sequence length. The noise is assumed to be i.i.d. normal Gaussian.

selection criterion and directly compare with BIC. We remark that the LOL-inspired criterion is calculated using the SMC that is discussed in Section 4. The true order of AR process used to generate the samples was 3, and the noise was generated as i.i.d. standard Gaussian. We sweep the experiment for varying sample sizes in the range from 20 to 400. The results for the error probability as well as the average selected model are plotted in Figure 1. As can be seen, the LOL-inspired model selection criterion provides superior performance to BIC for finite samples. We also observe that although both BIC and the conditional mutual information lead to consistent estimators for the order of the process, the latter converges faster and hence performs better in the finite sample case of interest in this example. In the second experiment, to show that this phenomenon is not just confined to the Gaussian noise process, we changed the generation noise to i.i.d. standard Laplacian process. As can be seen in Figure 2, similar trends as in the previous case could be observed in this case as well.

As a second well-studied example, we let $I_U(X_{n+1}; X_1, \ldots, X_n, \Theta)$ be the negative logarithm of the out-of-sample prediction error of X_{n+1} given $X_1, \ldots, X_n, \Theta$. In other words, $U(\Theta|X_1, \ldots, X_n)$ is supposed to be the squared error scaled by the Fisher information matrix. Under these assumptions the LOL-inspired model selection penalty gives back AIC (Akaike Information Criterion) and TIC (Takeuchi information criterion) in the well-specified and

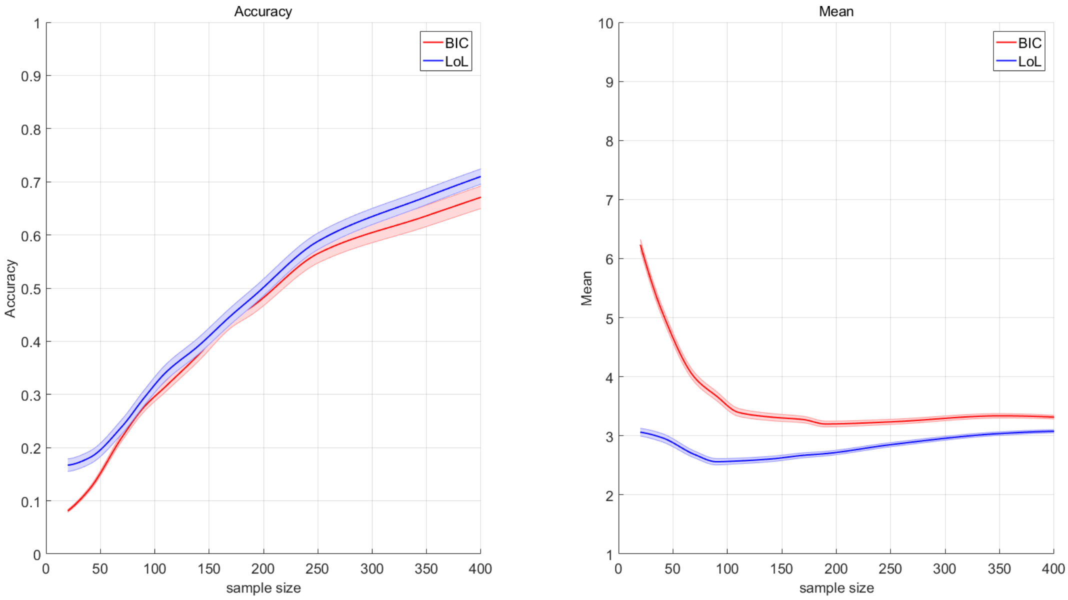

Figure 2: The left plot shows the empirical accuracy of selecting the true model using LOL and BIC. The right plot shows the average model order selected by LOL and BIC as a function of the sequence length. The noise is assumed to be i.i.d. normal Laplace.

mis-specified setting, respectively. In practice, one would mostly be concerned with the mis-specified setting as the real data almost never conforms to a true model.

Asymptotic Efficiency of TIC

Our next contribution concerns with the asymptotic efficiency of TIC under some regularity conditions. Before we state the result, we need to give a few definitions.

Let $\mathcal{A}_n$ be a set of candidate models and let $\mathcal{L}_n(\alpha) = E\{\ell(\boldsymbol{z}, \hat{\boldsymbol{\theta}}_n(\alpha))\}$ be the expected loss given by model α. Let α, $d(\alpha)$, $\mathcal{H}_n(\alpha) \subset \mathcal{R}_n^{d(\alpha)}$ be the candidate model, its dimension, and its associated parameter space, respectively. Define $\boldsymbol{\psi}_n(\boldsymbol{z}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{z}, \boldsymbol{\theta})$. Finally, let $V(\boldsymbol{\theta}) = E\{\nabla^2_{\boldsymbol{\theta}}\, \ell(\boldsymbol{z}, \boldsymbol{\theta})\}$ and $J(\boldsymbol{\theta}) = \mathrm{var}\{\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{z}, \boldsymbol{\theta})\}$.

Theorem 2 *Under mild assumptions, if $\tilde{\alpha}$ is chosen such that it minimizes*

$$\hat{\mathcal{L}}_n(\alpha) + n^{-1}\,\mathrm{tr}\left\{\hat{V}_n(\hat{\boldsymbol{\theta}}_n(\alpha))^{-1}\hat{J}_n(\hat{\boldsymbol{\theta}}_n(\alpha))\right\}$$

over $\alpha \in \mathcal{A}_n$, where $\boldsymbol{\theta}(\alpha)$ is the minimizer of $E_n\boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta})$ over $\mathcal{H}_n(\alpha)$ for each $\alpha \in \mathcal{A}_n$, and $\hat{V}_n(\cdot)$ and $\hat{J}_n(\cdot)$ are sample estimators of $V(\cdot)$ and $J(\cdot)$. Then we have

$$\frac{\mathcal{L}_n(\tilde{\alpha})}{\mathcal{L}_n(\alpha)} \to_p 1.$$

Inspired from this result in this case we define LOL as follows:

LOL = the reduction in the logarithm of the optimal predictive loss when revealing x_{n+1}.

We also have an algorithm based on this idea to perform model selection.

Algorithm 1 Penalized model selection

```
input A set of candidate models $\mathcal{A}_n$
output A model $\hat{\alpha} \in \mathcal{A}_n$
for $\alpha \in \mathcal{A}_n$ do
    Find the minimal loss estimator
        $\hat{\boldsymbol{\theta}}_n(\alpha) = \arg\min_{\boldsymbol{\theta} \in \alpha} \frac{1}{n} \sum_{i=1}^{n} l_n(\boldsymbol{z}_i, \boldsymbol{\theta})$    (5)
    and record the minimum as $\hat{\mathcal{L}}_n(\alpha)$
end for
Select
    $\hat{\alpha} = \arg\min_{\alpha \in \mathcal{A}_n} \underline{\mathcal{L}_n}(\alpha), \quad \text{where } \underline{\mathcal{L}_n}(\alpha) \triangleq \hat{\mathcal{L}}_n(\alpha) + \frac{1}{n} \operatorname{tr}\{\hat{V}(\alpha)^{-1} \hat{J}(\alpha)\}$    (6)
where
    $\hat{V}(\alpha) = -\frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}^2 \, l_n(\boldsymbol{z}_i, \hat{\boldsymbol{\theta}}_n(\alpha)), \quad \hat{J}(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \nabla_{\boldsymbol{\theta}} \, l_n(\boldsymbol{z}_i, \hat{\boldsymbol{\theta}}_n(\alpha)) \right\} \left\{ \nabla_{\boldsymbol{\theta}} \, l_n(\boldsymbol{z}_i, \hat{\boldsymbol{\theta}}_n(\alpha)) \right\}^{\mathrm{T}}$    (7)
```

There are clear paths for the extension of the framework to more practical settings. The first is to drop the assumption that the data of interest is generated by an i.i.d. process. In this case, the asymptotic performance of the popular criteria, such as AIC, TIC, and BIC are not justified. In order to move forward, we will assume a dependence structure between the random variables, and will incorporate existing knowledge of the learner.

Thus far, our focus was on learning using parametric models. The second area for extension is to consider non-parametric models, which is the subject of our current research. In fact, we have derived e a preliminary algorithm for the computation of the limit of learning in a non-parametric model setting as well. The algorithm is based on the ϵ-covering of the function space, and then designing appropriately vanishing rate of ϵ as the sample size increases. This includes learning from different experts and spaces.

4 Computation Algorithms

In this section, we focus on numerical methods for the calculation of the limits of learning. We discuss our methods for both well-specified and mis-specified cases.

4.1 Computation of LOL in the Well-Specified Case

In the well-specified case the posterior could be computed given the samples. In particular, we have used the SMC (sequential Monte Carlo) method to compute the posterior and consequently the limit of learning. This method was used to calculated the entropic LOL case was provided as an example in the previous section. We remark that the algorithm is general and could be used to calculate the limit of learning in general in the well-specified case. We have developed software for this purpose, which is posted along with this report.

4.2 Computation of LOL in the Mis-Specified Case

There are several applications where the learner receives the data points one sample at a time, and would like to start building a model for the learning process. In particular, the learner would like to start from simpler hypotheses and start expanding to more complicated hypotheses as more data arrives. If this is implemented naively, it is possible for the data arrival pattern to lead to fluctuations in the selected model in an arbitrary model. On the other hand, smooth model selection may be desired. In this case, the learner would increase the model complexity smoothly and monotonically as new data samples arrive.

Next, we will present our proposed sequential model expansion algorithm in the mis-specified case. It includes tracking several experts (candidate models), and maintaining an active set of experts at each time (to reduce computational cost). Then, the algorithm would perform switching of the active subset when new data samples arrive. The algorithm would output a set of weights associated with the active experts that will be used to predict the next data point at each time. The full description of an algorithm that performs based on the described method is presented next.

Algorithm 2 Sequential model expansion

```
input $\{(\boldsymbol{y}_t, \boldsymbol{x}_t : t = 1, \ldots, T\}$, $\eta > 0$, $\zeta \in [0, 1]$, $w_{0,1} = 1, w_{0,2} = \cdots = w_{0,K} = 0$, candidate models $\mathcal{A}_n = \{\alpha_1, \alpha_2, \ldots\}$, $s = 0$ ($\alpha_{s+1}, \ldots, \alpha_{s+K}$ are the maintained active subsets of models), $K \in \mathbb{N}$, threshold $\rho \in [0, 1]$
output $\boldsymbol{p}_t = [p_{t,1}, \ldots, p_{t,K}]^{\mathrm{T}}$ (predictive distribution over the active models) for each $t = 1, \ldots, T$
for $t = 1 \to T$ do
    Obtain $y_t$ and compute $v_{t,k} = w_{t-1,k} \exp\{-\eta \, \underline{\mathcal{L}_n}(\alpha_{s+k})\}$ for each $k = 1, \ldots, K$, where $\underline{\mathcal{L}_n}(\alpha)$ was defined using all the data up to time step $t$
    Let
        $w_{t,k} = \begin{cases} (1-\zeta) \; v_{t,k} & \text{if } k = 1 \\ (1-\zeta) \; v_{t,k} + \zeta \; v_{t,k-1} & \text{if } 1 < k < K \\ v_{t,k} + \zeta \; v_{t,k-1} & \text{if } k = K \end{cases}$
    Let $p_{t,k} = (\sum_{k=1}^{K} w_{t,k})^{-1} w_{t,k}$, $k = 1, \ldots, K$
    if $p_{t,1} \leq \rho$ and $p_{t,K} \geq 1 - \rho$ then
        Let $s = s + 1$
        Let $w_{t,k} = w_{t,k'}$, where $k = 1, \ldots, K$ and $k' = (k+1 \bmod K)$ (relabeling the active models)
    end if
end for
```

Finally, we remark that we can use surrogates for the limits of learning for dependent data as well. For example, cross-validation is used as a surrogate for the prediction loss. These are the subjects of our current investigation.

5 Non-identifiability

We are continuing to work on exploring multiple optima in non-negative matrix factorization. At the beginning of the last quarter, we were starting to use rapidly-exploring random trees to try to cover the posterior space. We have made several optimizations to our rapidly-exploring random trees to take into account the fact that they expand along narrow areas/manifolds, such as having a branch “hug” the edge of a constraint until it can make no farther progress toward the randomly sampled node. We also have it create “intermediate” nodes to help it fill interior spaces rather than always have points only at an edge. Finally, we have moved to Stein’s Discrepancy – a very new metric for evaluating whether a set of samples reflects a distribution – as our evaluation. While principled and more flexible than variational lower bounds, much of the work (some of it continuing) has been in understanding this metric and where it fails. That said, we are seeing that our approach generates samples that are spread farther apart than typical samplers and we outperform gradient-based optimizers. However, we perform as well or worse than random restarts, suggesting that in many of these problems, we need samples that are spaced even farther apart.

6 Limits of Learning in the Cognitive Setting

We are continuing to investigate neural circuits in the exacting setting that knowledge acquisition can occur from single interactions, the results of these acquisitions are rapidly evaluatable subcircuits, and recall in response to an external input can be in the form of a rapid evaluation of a composition of subcircuits that have been acquired at arbitrary different earlier times. The most basic computational task here is the cognitive function of association, which is defined here to mean that a subcircuit is to be set up so that the excitation of a set of neurons that represent A will in future cause the excitation of a set of neurons that represent B.

We note that this is a distinct and more onerous notion than the traditionally considered "auto-associative" networks where a set of memorized strings have to be retrieved. In particular the composability constraint adds a dimension that has not been considered before. The analysis is in terms of the three parameters: n the number of neurons, d the number of other neurons each neuron is connected to, and k, the inverse of the maximum synaptic strength when neurons have a firing threshold of unity.

One goal is to determine the capacity C of the system, namely the number of associations that can be memorized in succession so that they all, even the earliest ones memorized, remain functional after all the memorizations. In particular one wants to understand how C depends on $n, d,$and k. Our preliminary results study a simple mechanism, which we call the Basic Mechanism, that can work, in principle even with weak synapses. We show that the composability constraint is severe and limits the capacity to $O((d/k)^2)$, even independent of n, while an information theoretic estimate would be a higher $O(nd)$, linear in the number of synapses. For example for the realistic case of $d = n^{1/2}$, the capacity is linear in n, the number of neurons, and not the number of synapses. We have both upper and lower bounds: With the composability constraint capacity about $O((d/k)^2)$ is achievable, and that is close to the limit. In contrast, in the absence of the composability constraint, the higher information theoretic bound of $O(nd)$ (equal to $O(n^{3/2})$ in the case in question) can be approached.

7 Publications

[1] Jie Ding, Vahid Tarokh, and Yuhong Yang, "Bridging AIC and BIC: a new criterion for autoregression."

[2] Jie Ding, Vahid Tarokh, and Yuhong Yang, "Optimal variable selection in regression models."

[3] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez, "Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations," arXiv:1703.03717.