Online Supplementary Material for "(Machine) Learning Parameter Regions"

José Luis Montiel Olea[*] and James Nesbit[†]

Appendix B

In this appendix we consider the possibility of computing the misclassification error when the probability measure Q (the measure used by the oracle to compute misclassification error) differs from P (the measure used by the econometrician to generate random draws). Given an algorithm $\widehat{\lambda}_M$, the misclassification error of learning a concept λ thus becomes $\mathcal{L}(\widehat{\lambda}_M; \lambda, Q)$.

A concept $\lambda \in \Lambda$ is (Q, P) learnable if there exists an algorithm $\widehat{\lambda}_M$ and a function $m(\epsilon, \delta)$ such that for any $0 < \epsilon$ and $\delta < 1$:

$$P\left(\mathcal{L}(\widehat{\lambda}_M; \lambda, Q) < \epsilon\right) \geqslant 1 - \delta, \tag{1}$$

for all distributions P on Θ and for any $\lambda \in \Lambda$; provided $M \geqslant m(\epsilon, \delta)$.

We establish the following results

1. We provide a simple example, where θ has dimension $d = 1$, that shows that learning in the sense of (1) is impossible even if Λ has finite VC dimension. The example shows that when Q and P are different, learning becomes complicated because there is a lot of flexibility in the choice of P.

2. We also show that, not surprisingly, if we restrict P to belong to a class P_Q

[*] Columbia University, Department of Economics, 420 West 118th St., New York, NY 10027. (jm4474@columbia.edu).

[†] New York University, Department of Economics, NYU, 19 W.4th St., 6th Floor, New York, NY 10012. (jmn425@nyu.edu).

such that

$$\sup_{A \in \text{ continuity sets of } Q} |P(A) - Q(A)| \leqslant \eta,$$

for sufficiently small η, then learning is possible (for a fixed ϵ and δ) and the sufficient number of draws becomes

$$\ln\left(\frac{1}{\delta}\right)\frac{1}{2\epsilon - \eta}, \quad \eta < 2\epsilon.$$

which is larger than $\ln\left(\frac{1}{\delta}\right)\frac{1}{2\epsilon}$; the number of draws that would be required if P were equal to Q.

The example suggests that allowing P and Q to differ does not add much to our previous results.

Example:

Suppose that the parameter of interest lives in the real line, so that $d = 1$. Suppose that the concept class contains elements of the form $[a, \infty)$. The class has VC dimension 1.[1]

For notational simplicity, we identify sets of the form $[\lambda, \infty)$, $[\widehat{\lambda}, \infty)$ by the scalars λ, $\widehat{\lambda}$. Algebra shows that

$$P\left(\mathcal{L}(\widehat{\lambda}; \lambda, Q)\right) = |Q(\lambda) - Q(\widehat{\lambda})|.$$

Assume that Q is absolutely continuous with respect to the Lebesgue measure.

We show that in this example, learning is not possible. It is sufficient to show that for **any** algorithm $\widehat{\lambda}_M$, there exists ϵ, δ and λ such that for some P

$$P\left(\mathcal{L}(\widehat{\lambda}_M; \lambda, Q) \geqslant \epsilon\right) \geqslant \delta.$$

regardless of the sample size.

Fix $\lambda \in \mathbb{R}$ and let $\widehat{\lambda}_M$ be an arbitrary algorithm. Let M be an arbitrary sample size.

[1] Suppose we have 1 point, then λ can label it either 0 or 1, implying one point can be shattered. Suppose there are 2 points. We can generate labels $(0, 0)$, $(1, 1)$ and $(0, 1)$, but can't generate $(1, 0)$ labels. 2 points cannot be shattered, and thus the VC dimension (the largest number of points that can be shattered) of Λ is 1.

Without loss of generality,[2] consider algorithms $\widehat{\lambda}_M : (x_1, \ldots, x_m) \to \mathbb{R}$ such that for any set $(a, b) \subset \mathbb{R}$,

$$\widehat{\lambda}_M^{-1}(a, b) \neq \varnothing. \tag{2}$$

Take an arbitrary value λ^*, and an arbitrary set $(\underline{\lambda}^*, \overline{\lambda}^*)$, such that $\lambda^* \in (\underline{\lambda}^*, \overline{\lambda}^*)$. $\epsilon^* = \min\{Q(\overline{\lambda}^*) - Q(\lambda^*), Q(\lambda^*) - Q(\underline{\lambda}^*)\} > 0$. Such a set exists as Q is absolutely continuous w.r.t. to the Lebesgue measure. For any algorithm satisfying (2) we have

$$P\left(\mathcal{L}(\widehat{\lambda}_M; \lambda, Q) \geqslant \epsilon^*\right) \geqslant 1 - P(\widehat{\lambda}_M \in (\underline{\lambda}^*, \overline{\lambda}^*)).$$

For any sample size—and given that P is unrestricted—there is a P such that $P(\widehat{\lambda}_M \in (\underline{\lambda}^*, \overline{\lambda}^*))$ can be made arbitrary small. The example shows learning is impossible, even if the concept has finite VC dimension.

Now we show that if we allow for probability distributions P that are close to Q, learning is still possible. The result is not surprising at all, and all we need is to use the right definition of "closeness". Let

$$P_Q^\eta \equiv \left\{ P \;\middle|\; \sup_{A \in \text{ cont sets of } Q} |P(A) - Q(A)| \leqslant \eta \right\}.$$

We argue that the algorithm that sets $\widehat{\lambda}_M = \min\{x_i | x_i = 1\}$ or $\widehat{\lambda}_M = \max\{x_i | x_i = 0\}$ learns uniformly, for a fixed pair (ϵ, δ), where $\epsilon \geqslant \eta/2$.

The proof goes as follows. Fix $\lambda \in \mathbb{R}$. Find $\underline{\lambda}(\lambda) < \overline{\lambda}(\lambda)$ such that $Q(\overline{\lambda}(\lambda)) - Q(\lambda) = \epsilon = Q(\lambda) - Q(\underline{\lambda}(\lambda))$. Define the set $A(\lambda) = [\underline{\lambda}(\lambda), \overline{\lambda}(\lambda)]$. Then

$$\begin{aligned} P\left(\mathcal{L}(\widehat{\lambda}_M; \lambda, Q) \geqslant \epsilon\right) &= P(x_i \notin [\underline{\lambda}(\lambda), \overline{\lambda}(\lambda)], \quad \forall i) \\ &= (1 - P(A(\lambda)))^M. \end{aligned}$$

[2] If this were not the case, consider any (a, b) for which $\widehat{\lambda}_M^{-1}(a, b) = \varnothing$. Then we could pick $\lambda \in (a, b)$ and set $\epsilon^* = \min\{Q(a) - Q(\lambda), Q(\lambda) - Q(a)\}$. In this case, we have that for any P:

$$\begin{aligned} P\left(\mathcal{L}(\widehat{\lambda}_M, \lambda, Q) \geqslant \epsilon^*\right) &\geqslant P(\widehat{\lambda}_M \geqslant b) + P(\widehat{\lambda}_M \leqslant a) \\ &= 1 - P(\widehat{\lambda}_M \in (a, b)) = 1 - P(\varnothing) = 1. \end{aligned}$$

Note by definition $Q(A(\lambda)) = 2\epsilon$, which makes the line above equal to

$$(1 - [P(A(\lambda)) - Q(A(\lambda))] - 2\epsilon)^M,$$

implying

$$P(\mathcal{L}(\widehat{\lambda}_M; \lambda, Q) \geqslant \epsilon) \leqslant (1 - (2\epsilon - \eta))^M,$$

as for any $P \in P_Q^\eta$, we have $-\eta \leqslant P(A) - Q(A) \leqslant \eta$. Therefore for a fixed (ϵ, δ)

$$M \geqslant \ln\left(\frac{1}{\delta}\right)\frac{1}{2\epsilon - \eta}$$

suffices to learn the concept class. This requires more draws than when $Q = P$, which would be exactly

$$\ln\left(\frac{1}{\delta}\right)\frac{1}{2\epsilon}.$$

This formalizes the result that, if P is required to be sufficiently close to Q, then learning is indeed possible (but the number of draws required to learn is practically the same as when $P = Q$).

Appendix C

In this section we describe how to (machine) learn parameter regions that arise in the Latent Dirichlet Allocation (LDA) model of Blei et al. (2003). For more details, see Ke et al. (2019) (henceforth, KMN).

The LDA is a popular machine learning algorithm for text analysis. The LDA model assumes that there are K latent topics; a topic is a distribution over the V terms in the vocabulary, $\beta_k \in \Delta^{V-1}$. Each document d is characterized by a document-specific distribution over the K topics, $\theta_d \in \Delta^{K-1}$. The topic distributions $B \equiv (\beta_1, \ldots, \beta_K)$ and the topic compositions $\Theta \equiv [\theta_1, \ldots \theta_D]$ determine the mixture model for each word in document d.

Let $\mathbb{P}_d(t|B, \theta_d)$ denote the probability that a term $t \in \{1, \ldots, V\}$ appears in document d. The model assumes that

$$\mathbb{P}_d(t|B,\theta_d) = \sum_{k=1}^{K} \beta_{t,k}\theta_{k,d}.$$

The likelihood of corpus C is thus parameterized by (B,Θ) and given by

$$\begin{aligned}\mathbb{P}(C|B,\Theta) &= \prod_{d=1}^{D}\prod_{t=1}^{V}(\mathbb{P}_d(t|B,\Theta))^{n_{t,d}} \\ &= \prod_{d=1}^{D}\prod_{t=1}^{V}(B\Theta)_{t,d}^{n_{t,d}}, \qquad (3)\end{aligned}$$

where $n_{t,d}$ is count of the number of times term t appears in document d. We can collect the terms $\mathbb{P}_d(t)$ in the $V \times D$ matrix P and use (3) to write

$$P = B\Theta. \qquad (4)$$

Theorem 1 of KMN shows that the parameters of the likelihood, B and Θ in (4) are set identified and thus the choice of prior matters. They show that the range of posterior means can be described using solutions to the rank K *Non-negative Matrix Factorization* (NMF) of the term-document frequency matrix with weight $W_{t,d}$, $\widehat{P}$—the non-negative matrices (B,Θ) that solve:

$$\min_{B\in\mathbb{R}_+^{V\times K},\Theta\in\mathbb{R}_+^{K\times D}} \sum_{i=1}^{D}\sum_{t=1}^{V} W_{t,d}\left[\hat{P}_{t,d}\log\left(\frac{\hat{P}_{t,d}}{(B\Theta)_{t,d}}\right) - \hat{P}_{t,d} + (B\Theta)_{t,d}\right]. \qquad (5)$$

Let $\text{NMF}(\widehat{P}, W_{t,d})$, denote the set rank K NMF's. The set S is:

$$S \equiv \left\{(B,\Theta) \mid (B,\Theta) \in \text{NMF}(\widehat{P}, W_{t,d})\right\},$$

where (B,Θ) are assumed to be matrices whose columns are probability distributions.

The algorithm to solve for a solution of (5) is initialized randomly, and thus the algorithm induces a distribution over S. Thus we can compute the tightest bands that contain the set $\lambda(S)$ for some functional using random sampling.

KMN revisit the work of Hansen and McMahon (2016), studying the effects of increased transparency on the discussions of Federal Open Market Committee (FOMC). Let $\theta_{i,t}$ be the weight of i^{th} topic in meeting at time t, the Herfindahl index for the topic distribution is given by

$$H_t \equiv \sum_{i=1}^{K} \theta_{i,t}^2.$$

The specific functional of interest is the 'transparency coefficient' (λ) in the regression of the concentration measure on a dummy for the date in which the Federal Reserve changed its transparency policy (October 1993) and controls

$$H_t = \alpha + \lambda D(Trans)_t + \gamma X_t + \epsilon_t.$$

Note that λ is one dimensional, hence $d = 1$. The data are the FOMC transcripts from August 1987–January 2006 which have been extensively preprocessed. The meetings are broken into two sections FOMC1 and FOMC2 and the regression is run separately on each. The resulting dimensions of the term-document matrices are 9000×148 and 6000×148 for FOMC1 and FOMC2 respectively. The number of topics is set to $K = 40$. The resulting number of parameters estimated is large—B is 9000×40 and Θ is 40×148, a total of $365,920$.

KMN take $M = 120$ draws from $\lambda(S)$, corresponding to a misclassification error of at most 5.91% with probability at least 94.09% ($\epsilon = \delta = 0.0591$) and the iso-draw curve presented in Figure 1.

The tightest band containing $\lambda(S)$ for FOMC1 is $[-0.0380, 0.0466]$, and for FOMC2 is $[-0.0615, 0.0350]$.

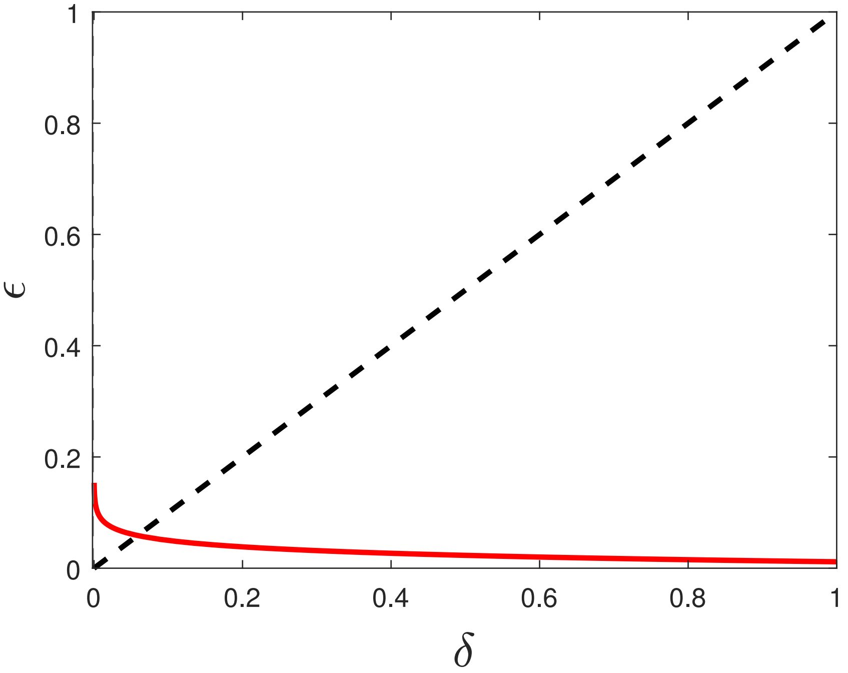

Figure 1: 'Iso-draw' curve for $M = 120$: the values of ϵ and δ that can be supported with 120 draws.

References

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Hansen, S. and M. McMahon (2016): "Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication," *Journal of International Economics*, 99, S114–S133.

Ke, S., J. L. Montiel Olea, and J. Nesbit (2019): "A Robust Machine Learning Algorithm for Text Analysis," *Working Paper*.