

# Generalization error and overtraining

## a geometric and Bayesian understanding

Andrés

<[ivan.a.davidovich@ntnu.no](mailto:ivan.a.davidovich@ntnu.no)>

Luca

<[piero.mana@ntnu.no](mailto:piero.mana@ntnu.no)>

Draft of 29 December 2018 (first drafted 6 August 2018)

The use of generalization error for testing and the phenomenon of overtraining are explained geometrically and given a Bayesian interpretation.

*Note: Dear Reader & Peer, this manuscript is being peer-reviewed by you. Thank you.*

## 1 Generalization tests, overtraining, probability calculus

In neural-network learning, and more generally in machine learning, it's customary to choose among candidate neural nets by testing them against a *generalization* data set separate from the *training* data set used to find the network parameters. Let's call this a *generalization test*; it may assume different forms, including cross-validation. The candidate neural nets differ in their architecture – number of nodes, weights, non-linear functions – or in the error function and weight regularizer used to train them.

The generalization test can also be used during the learning phase: it is sometimes observed that the generalization error reaches a minimum before the learning phase – that is, the minimization of the training error – is complete. This phenomenon is called *overtraining*. In this case the learning phase is stopped when the generalization error reaches a minimum.

MacKay (1992a,b; see also Bishop 2006), combining the analyses of Tishby, Levin, Solla (Tishby et al. 1989; Levin et al. 1990) and Gull (1989), gives an insightful interpretation of machine and neural-nets learning from the point of view of the probability calculus. For a discussion of the statements in the next two paragraphs, see MacKay (1992a,b) and Bishop (2006 esp. §§ 1.2–3, 3.3–4, 5.7).

From the point of view of the probability calculus it makes sense to choose among candidates by comparing their plausibilities conditional on the training data set:  $p(\text{candidate}|\text{data}, \text{initial info})$ . If we initially judge the candidates to be equally plausible, their plausibilities conditional on the data are proportional to the plausibility of the data

conditional on the candidates, that is, to the likelihoods of the candidates:  $p(\text{data} | \text{candidate, initial info})$ , also called the *evidence* for the candidate.

The evidence and the generalization test often lead to the same choice of candidate, and the evidence has several advantages compared to the generalization test. Yet, they aren't equivalent (for a thorough discussion of their differences see MacKay 1992a esp. § 6.6; 1992b esp. § 4.1). In particular it may happen that a candidate with poor evidence yields a good generalization test. This often indicates that the whole set of candidates represents our belief about the particular data poorly. So, even from the point of view of the probability calculus, the generalization test can be very useful.

The generalization test is not derived by an application of the probability calculus, however, unlike the evidence. Its idea is nevertheless intuitive and sound.

In this note we show that the generalization test can in fact be derived from the rules of probability. The derivation also shows why this test can indicate that the whole sets of candidates is a poor representation of our beliefs about the data.

## 2 Geometric preliminaries

---

\*\*\* Old draft \*\*\*

---

### 3 Synopsis

The literature usually associates overtraining with overfitting and too large number of parameters.

The relation between these aspects is more subtle than that. In fact, overtraining can be a consequence of *underfitting*. We have a clearer view approaching this question from the point of view of probability theory.

Consider a classification problem: given input  $x \in \{1, \dots, M\}$ , we want to predict output  $y \in \{1, \dots, N\}$ . Let's consider both discrete.

Two cases must be distinguished: whether every input has several possible outputs, or just one. The second case is a special instance of the first, and could be simply treated as the prediction of a 'function'  $x \mapsto y$ . But both cases are instances of inference with a *partially exchangeable* model.

Our assumptions are summarized in the proposition I. We assume that the probability of  $y$ , for each kind of input  $x$ , is exchangeable. Thus the probability for several pairs

$$p(y^{(T)}, \dots, y^{(1)} | x^{(T)}, \dots, x^{(1)}, I) \quad (1)$$

must have a partially exchangeable form [refs]: for every kind of input  $x$  we consider the  $N$  relative frequencies  $f_x := (f_{1|x}, \dots, f_{N|x})$  of the  $N$  possible outputs. The probability above is given by this integral:

$$p(y^{(T)}, \dots, y^{(1)} | x^{(T)}, \dots, x^{(1)}, I) = \int \cdots \int \left[ \prod_{x=1}^M \prod_t^{x^{(t)}=x} f_{y^{(t)}|x} \right] p(f_1, \dots, f_M | I) df_1 \cdots df_M. \quad (2)$$

The complicated look of this formula doesn't do justice to its simple and intuitive interpretation:

Choose a kind of input  $x$ . We're judging that the probability for a very large sequence of outputs generated with this input doesn't depend on their order. Suppose we knew the relative frequencies of each kind of output in this sequence,  $(f_{1|x}, \dots, f_{N|x}) =: f_x$ , and knew nothing else. Then out of symmetry reasons we should give a probability  $f_{y|x}$  of observing outcome  $y$  at the next observation. This is just a 'drawing

without replacement' problem. At the subsequent observations with the same input we give again the same probabilities to each  $y$ , because the sequence is so large that we approximate 'drawing without replacement' with 'drawing with replacement'. The probability for a sequence of outputs  $y^{(1)}, \dots, y^{(T)}$  from the same input  $x$  is then

$$p(y^{(1)}, \dots, y^{(T)} | \text{same input } x, f_x, I) = f_{y^{(T)}|x} \times \dots \times f_{y^{(1)}|x}. \quad (3)$$

Within a larger sequence of outputs from all possible inputs, the outputs  $y^{(0)t}$  coming from input  $x$  are those for which  $x^{(0)t} = x$ . Thus we can write the formula above more generally as

$$\prod_t^{x^{(t)}=x} f_{y^{(t)}|x} \quad (4)$$

The above reasoning applies for each kind of input  $x$ . Thus the probability for a sequence of outputs coming from different inputs is the product of the probabilities above for all different  $x$ :

$$p(\text{all outputs} | \text{inputs}, f_1, \dots, f_M, I) = \prod_{x=1}^M \prod_t^{x^{(t)}=x} f_{y^{(t)}|x}. \quad (5)$$

Now what if we don't know the relative frequencies  $f_x$ , for any input? Then we assign a probability distribution over their possible values and use the law of total probability:

$$p(\text{all outputs} | \text{inputs}, I) =$$

$$\int \dots \int p(\text{all outputs} | \text{inputs}, f_1, \dots, f_M, I) p(f_1, \dots, f_M | I) df_1, \dots, df_M. \quad (6)$$

Substituting the explicit expression (5) in this formula we obtain formula (6). In summary,

$$\int \dots \int \left[ \prod_{x=1}^M \prod_t^{x^{(t)}=x} f_{y^{(t)}|x} \right] p(f_1, \dots, f_M | I) df_1 \dots df_M. \quad (7)$$

probability for outputs  
from same input

product over all inputs

uncertainty over all frequencies

The possible frequencies give one input,  $f_x \equiv (f_{1|x}, \dots, f_{N|x})$ , belong to the  $(N - 1)$ -dimensional simplex

$$\Delta := \{(f_1, \dots, f_N) \mid f_i \geq 0, \sum_i f_i = 1\}, \quad (8)$$

and the collection of possible frequencies  $(f_1, \dots, f_M)$  belongs to the  $M$ -fold Cartesian product  $\Delta^M$ . From now on we denote  $f := (f_1, \dots, f_M)$ .

The probability for a new sequence of  $T'$  outputs given their inputs and given that we've learned a previous sequence of  $T$  input-output pairs is determined by Bayes's theorem:

$$\begin{aligned} p(y^{(T')}, \dots, y^{(T+1)} \mid x^{(T')}, \dots, x^{(T+1)}, y^{(T)}, x^{(T)}, \dots, y^{(1)}, x^{(1)}, I) = \\ \frac{\int \left[ \prod_{x=1}^M \prod_{t=T+1, \dots, T'}^{x^{(t)}=x} f_{y^{(t)}|x} \right] p(f \mid I) df}{\int \left[ \prod_{x=1}^M \prod_{t=1, \dots, T}^{x^{(t)}=x} f_{y^{(t)}|x} \right] p(f \mid I) df}. \end{aligned} \quad (9)$$

This formula is equivalent to (2) with an updated distribution for the frequencies:

$$\begin{aligned} p(f \mid y^{(T)}, x^{(T)}, \dots, y^{(1)}, x^{(1)}, I) df = \\ \frac{\left[ \prod_{x=1}^M \prod_{t=1, \dots, T}^{x^{(t)}=x} f_{y^{(t)}|x} \right] p(f \mid I)}{\int \left[ \prod_{x=1}^M \prod_{t=1, \dots, T}^{x^{(t)}=x} f_{y^{(t)}|x} \right] p(f \mid I) df} df. \end{aligned} \quad (10)$$

The form of this updated distribution has important consequences for our learning process.

If the number of learned data is enough large compared with the numbers  $M, N$  of possible inputs and outputs and with the magnitude of the initial distribution for the frequencies, and if the latter is strictly positive, then the updated distribution becomes very peaked on the collection of relative frequencies  $(q_1, \dots, q_M)$  of the learned outputs for all input values. This can be seen from the asymptotic expression in terms of the relative entropy (Kullback-Leibler divergence)  $D$ ,

$$p(f \mid y^{(T)}, x^{(T)}, \dots, y^{(1)}, x^{(1)}, I) \propto \exp[-\sum_x T_x D(q_x \mid f_x)] p(f \mid I), \quad (11)$$

where  $T_x$  is the number of observations with input  $x$ , with  $\sum_x T_x = T$ .

If the number of learned data is small compared with the dimensions of the input and output spaces, then the initial distribution for the frequencies  $p(f|I) df$  greatly influence our inference (9). This distribution determines two important \*\*\*

#### 4 Utility functions and probabilities

Utility of behaving as if proposition  $B \in X$  is true given that proposition  $A \in X$  is true:  $c(B|A)$ . Probability for  $A$  given  $D, I$ :  $P(A|D, I)$ . Optimal decision is  $B$  that maximizes

$$\sum_A c(B|A) P(A|D, I). \quad (12)$$

Now consider different probabilities for all  $A$  given the same data  $D$ :  $P(A|D, I')$ . The decision  $B$  will still be the same if we use a new utility

$$c'(B|A) := c(B|A) \frac{P(A|D, I)}{P(A|D, I')}. \quad (13)$$

So the same choice can be made with a different probability, if the utility is appropriately changed, provided  $p(A|D, I') > 0$  for all  $A$ .

This leads to a slightly more general view than Tishby et al.'s (1989; 1990) and Mackay's (1992a,b)

## Thanks

... to Mari & Miri for continuous encouragement and affection, and to Buster Keaton and Saitama for filling life with awe and inspiration. To the developers and maintainers of L<sup>A</sup>T<sub>E</sub>X, Emacs, AUC<sub>T</sub>E<sub>X</sub>, Open Science Framework, Python, Inkscape, Sci-Hub for making a free and unfiltered scientific exchange possible. ☒

## Bibliography

(‘de *X*’ is listed under *D*, ‘van *X*’ under *V*, and so on, regardless of national conventions.)

- Bishop, C. M. (2006): *Pattern Recognition and Machine Learning*. (Springer, New York).
- Gull, S. F. (1989): *Developments in maximum entropy data analysis*. In: Skilling (1989), 53–71.
- Levin, E., Tishby, N., Solla, S. A. (1990): *A statistical approach to learning and generalization in layered neural networks*. Proc. IEEE **78**<sup>10</sup>, 1568–1574.
- MacKay, D. J. C. (1992a): *Bayesian interpolation*. Neural Comp. **4**<sup>3</sup>, 415–447. <http://www.inference.phy.cam.ac.uk/mackay/PhD.html>.
- (1992b): *A practical Bayesian framework for backpropagation networks*. Neural Comp. **4**<sup>3</sup>, 448–472. <http://www.inference.phy.cam.ac.uk/mackay/PhD.html>.
- Skilling, J., ed. (1989): *Maximum Entropy and Bayesian Methods: Cambridge, England, 1988*. (Kluwer, Dordrecht).
- Tishby, N., Levin, E., Solla, S. A. (1989): *Consistent inference of probabilities in layered networks: predictions and generalizations*. Int. Joint Conf. Neural Networks **1989**, II-403–II-409.