Generalization error and overtraining
a geometric and Bayesian understanding

Andrés
<ivan.a.davidovich@ntnu.no>

Luca
<piero.mana@ntnu.no>

Draft of 31 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 more thorough 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: $\mathrm{p}(\text{candidate}|\,\text{data}, \text{initial info})$. 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(data| 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

3 Learning models and extreme models

Central to our Bayesian understanding of the generalization test is the difference between a *learning* plausibility model and a non-learning plausibility model, also called *extremal* or independent, and the relationship between the two. This relationship, extensively studied by Lauritzen (1974a,b; 1988; 1984), is closely connected with the notion of sufficient statistics and exponential families of distributions [refs***].

With a learning model I_l, knowledge about output and input data $(x_1, x_2, \dots) =: \boldsymbol{x}$, $(t_1, t_2, \dots) =: \boldsymbol{t}$ modifies our degree of belief about a new output datum x conditional on a new input datum t:

$$\mathrm{p}(x|\, t,\, \boldsymbol{x}, \boldsymbol{t},\, I_l) \neq \mathrm{p}(x|\, t,\, I_l), \tag{1}$$

and will generally be different for different sets of input and output data. We can equivalently say that our degree of belief about a set of data

doesn't factor into a product of their individual degrees of belief:

$$\begin{aligned}&\mathrm{p}(x_1, x_2, x_3, \dots | \, t_1, t_2, t_3, \dots, \, I_{\mathrm{l}}) \\ &\quad = \mathrm{p}(x_1 | \, x_2, x_3, \dots, \, t_1, t_2, t_3, \dots, \, I_{\mathrm{l}}) \times \mathrm{p}(x_2 | \, x_3, \dots, \, t_2, t_3, \dots, \, I_{\mathrm{l}}) \times \cdots \\ &\quad \neq \mathrm{p}(x_1 | \, t_1, \, I_{\mathrm{l}}) \times \mathrm{p}(x_2 | \, t_2, \, I_{\mathrm{l}}) \times \cdots \,.\end{aligned} \tag{2}$$

With an extremal model I_{e}, knowledge about other data doesn't affect our degree of belief:

$$\mathrm{p}(x | \, t, \, \boldsymbol{x}, \boldsymbol{t}, \, I_{\mathrm{e}}) = \mathrm{p}(x | \, t, \, I_{\mathrm{e}}). \tag{3}$$

Equivalently, our degree of belief about a set of data factorizes into the product of the marginals:

$$\mathrm{p}(x_1, x_2, x_3, \dots | \, t_1, t_2, t_3, \dots, \, I_{\mathrm{e}}) = \mathrm{p}(x_1 | \, t_1, \, I_{\mathrm{e}}) \times \mathrm{p}(x_2 | \, t_2, \, I_{\mathrm{e}}) \times \cdots \,. \tag{4}$$

Why is the latter kind of model called 'extremal'?

Learning models are typically given in the following integral form

$$\mathrm{p}(\boldsymbol{x} | \, \boldsymbol{t}, \, I) = \int_W \mathrm{d}w \; \mathrm{p}(w | \, I) \; \mathrm{p}(\boldsymbol{x} | \, \boldsymbol{t}, \, w, I) \qquad \text{for all } (\boldsymbol{t}, \boldsymbol{x}) \tag{5a}$$

with

$$\mathrm{p}(\boldsymbol{x} | \, \boldsymbol{t}, \, w, I) = \prod_i \mathrm{p}(x_i | \, t_i, \, w, I), \tag{5b}$$

which is a *convex combination* of densities $\mathrm{p}(\boldsymbol{x} | \, \boldsymbol{t}, \, w, I)$ labelled by a parameter $w \in W$, the weights of the combination being $\mathrm{p}(w | \, I)$. Each labelled density $\mathrm{p}(\boldsymbol{x} | \, \boldsymbol{t}, \, w, I)$ is an extremal model (w, I), as clear from the comparison of formulae (5b) and (4). In other words, a learning model is given as a convex combination of extremal models. Most important, this convex combination is *unique*: if we change the family of extremal models or the weight density $\mathrm{p}(w | \, I)$ the resulting learning model will assign different plausibilities to some data sets $(\boldsymbol{t}, \boldsymbol{x})$. This means that a learning model uniquely selects a family of extremal, non-learning models and a weight density.

By keeping the family of extremal models $\{(w, I) \mid w \in W\}$ fixed and considering all possible weight densities we generate a space of plausibility models. By construction this is a *convex set*, and its extremal points (a point being a plausibility model) are non-learning models of

the form (5b). This is why they are called 'extremal'. The extremal model labelled by w^* corresponds to a singular convex combination with a delta weight density $\delta(w - w^*)$. [***refs for convex spaces]

When we update a learning model with training data $(\boldsymbol{t}', \boldsymbol{x}')$, the weight density in the convex combination (5) is updated:

$$\mathrm{p}(w|\, \boldsymbol{x}', \boldsymbol{t}',\, I) = \frac{\mathrm{p}(w|\, I)\ \prod_i \mathrm{p}(x'_i|\, t'_i,\, w, I)}{\int_W \mathrm{d}w\ \mathrm{p}(w|\, I)\ \prod_i \mathrm{p}(x'_i|\, t'_i,\, w, I)}, \tag{6}$$

yielding a new point in the convex set; but the extremal points are unchanged. As the training data accumulate, the weight density become very peaked on some value w^*, becoming a delta:

$$\mathrm{p}(w|\, \boldsymbol{x}', \boldsymbol{t}',\, I) \xrightarrow{\text{data } (\boldsymbol{t}', \boldsymbol{x}') \text{ accumulate}} \delta(w - w^*). \tag{7}$$

So upon training our learning model approaches the set of extremal models.

4 Bayesian interpretation of the generalization test

Consider sets of input data $\boldsymbol{t} \coloneqq (t_1, t_2, \dots)$ and corresponding output data $(x_1, x_2, \dots)$.

The evidence is our degree of belief about the output data, given the input data and our initial information and assumptions I:

$$\mathrm{p}(x_1, x_2, \dots |\, t_1, t_2, \dots,\, I). \tag{8}$$

It's important to note that our joint degree of belief about the outputs is not the product of our marginal degrees of belief:

$$\begin{aligned}
&\mathrm{p}(x_1, x_2, x_3 \dots |\, \boldsymbol{t},\, I) \\
&= \mathrm{p}(x_1|\, x_2, x_3, \dots,\, \boldsymbol{t},\, I) \times \mathrm{p}(x_2|\, x_3, \dots,\, \boldsymbol{t},\, I) \times \mathrm{p}(x_3|\, \dots,\, \boldsymbol{t},\, I) \times \cdots \\
&\neq \mathrm{p}(x_1|\, \boldsymbol{t},\, I) \times \mathrm{p}(x_2|\, \boldsymbol{t},\, I) \times \mathrm{p}(x_3|\, \boldsymbol{t},\, I) \times \cdots
\end{aligned} \tag{9}$$

*** Old draft ***

✠ to be modified Consider sets of input quantities $\boldsymbol{t} \coloneqq (t_1, t_2, \dots)$ and corresponding output quantities $(x_1, x_2, \dots)$. A partially exchangeable plausibility model is a degree of belief about the output data, conditional on the input data, that is invariant under every relabelling of those output data that have the same input values. For example,

$$\begin{aligned}
&\mathrm{p}(x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, \dots \mid t_1 = 5, t_2 = 6, t_3 = 5, t_4 = 6, \dots\ I) = \\
&\mathrm{p}(x_1 = 3, x_2 = 2, x_3 = 1, x_4 = 4, \dots \mid t_1 = 5, t_2 = 6, t_3 = 5, t_4 = 6, \dots\ I) = \\
&\qquad \mathrm{p}(x_1 = 3, x_2 = 4, x_3 = 1, x_4 = 2, \dots \mid t_1 = 5, t_2 = 6, t_3 = 5, t_4 = 6, \dots\ I),
\end{aligned} \tag{10}$$

that is, we have the same degree of belief about $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$ and about $x_1 = 3, x_2 = 2, x_3 = 1, x_4 = 4$, where the values of x_1 and x_3 have been exchanged, because these correspond to input data having the same value, $t_1 = t_3 = 5$. This degree of belief also equals the one about $x_1 = 3, x_2 = 4, x_3 = 1, x_4 = 2$, where the values of x_2 and x_4 have also been exchanged, because these correspond to input data having the same value, $t_2 = t_4 = 6$. And so on. However, this degree of belief can be different from the one about $x_1 = 2, x_2 = 1, x_3 = 3, x_4 = 4$, where the values of x_1 and x_2 have been exchanged, because these correspond to different input values: $t_1 \neq t_2$.

A partially exchangeable model is completely determined if we specify our degree of belief about the relative frequencies of x values for each distinct input value t, in an indefinitely long sequence. [continue explanation***] This degree of belief is therefore represented by a density $\mathrm{p}(\{f_t\} \mid I)$ on a space equivalent to a product of simplices, one for each different value that t can assume. When this density is positive over the whole space the model is called *non-parametric*. This space acquires a huge dimensionality as the numbers of possible t and x values increase. For this reason we often specify singular densities, which are positive only over a lower-dimensional subset of this space. The corresponding model is called *parametric*, because we introduce parameters as coordinates on the lower-dimensional subset.

When the model is updated or trained with data D, the corresponding density $\mathrm{p}(\{f_t\} \mid I)$ is updated to $\mathrm{p}(\{f_t\} \mid D, I)$, which yields different predictions. This is an example of a *learning* model. As the number

of training data increases the density concentrates on a point of the lower-dimensional subset.

5 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

$$\mathrm{p}(y^{(T)}, \dots, y^{(1)} \mid x^{(T)}, \dots, x^{(1)}, I) \tag{11}$$

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

$$\mathrm{p}(y^{(T)}, \dots, y^{(1)} \mid x^{(T)}, \dots, x^{(1)}, I) = \int\cdots\int \Bigl[\prod_{x=1}^{M} \prod_{t}^{x^{(t)}=x} f_{y^{(t)}|x}\Bigr] \mathrm{p}(\boldsymbol{f}_1, \dots, \boldsymbol{f}_M \mid I)\, \mathrm{d}\boldsymbol{f}_1 \cdots \mathrm{d}\boldsymbol{f}_M. \tag{12}$$

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, $(f_{1|x}, \dots, f_{N|x}) =: \boldsymbol{f}_x$, 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

$$\mathrm{p}(y^{(T)}, \dots, y^{(1)} | \text{ same input } x, \boldsymbol{f}_x, I) = f_{y^{(T)}|x} \times \dots \times f_{y^{(1)}|x}. \tag{13}$$

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

$$\prod_{t}^{x^{(t)}=x} f_{y^{(t)}|x} \tag{14}$$

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:

$$\mathrm{p}(\text{all outputs} | \text{ inputs}, \boldsymbol{f}_1, \dots, \boldsymbol{f}_M, I) = \prod_{x=1}^{M} \prod_{t}^{x^{(t)}=x} f_{y^{(t)}|x}. \tag{15}$$

Now what if we don't know the relative frequencies $\boldsymbol{f}_x$, for any input? Then we assign a probability distribution over their possible valuesand use the law of total probability:

$$\mathrm{p}(\text{all outputs} | \text{ inputs}, I) =$$
$$\int \cdots \int \mathrm{p}(\text{all outputs} | \text{ inputs}, \boldsymbol{f}_1, \dots, \boldsymbol{f}_M, I)\, \mathrm{p}(\boldsymbol{f}_1, \dots, \boldsymbol{f}_M | I)\, \mathrm{d}\boldsymbol{f}_1, \dots, \mathrm{d}\boldsymbol{f}_M. \tag{16}$$

Substituting the explicit expression (15) in this formula we obtain formula (16). In summary,

$$\int \cdots \int \Big[\prod_{x=1}^{M} \prod_{t}^{x^{(t)}=x} f_{y^{(t)}|x} \Big]\, \mathrm{p}(\boldsymbol{f}_1, \dots, \boldsymbol{f}_M | I)\, \mathrm{d}\boldsymbol{f}_1 \cdots \mathrm{d}\boldsymbol{f}_M \ . \tag{17}$$

product over all inputs

probability for outputs from same input

uncertainty over all frequencies

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

$$\Delta \coloneqq \{(f_1, \dots, f_N) \mid f_i \geqslant 0, \textstyle\sum_i f_i = 1\}, \tag{18}$$

and the collection of possible frequencies $(\boldsymbol{f}_1, \dots, \boldsymbol{f}_M)$ belongs to the M-fold Cartesian product Δ^M. From now on we denote $\boldsymbol{f} \coloneqq (\boldsymbol{f}_1, \dots, \boldsymbol{f}_M)$.

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:

$$\begin{aligned}\mathrm{p}(y^{(T')}, \dots, y^{(T+1)} | \, x^{(T')}, \dots, x^{(T+1)}, \; y^{(T)}, x^{(T)}, \dots, y^{(1)}, x^{(1)}, \; I) = \\ \frac{\int \Bigl[\prod_{x=1}^{M} \prod_{t=T+1,\dots,T'}^{x^{(t)}=x} f_{y^{(t)}|x}\Bigr] \mathrm{p}(\boldsymbol{f} | \, I) \, \mathrm{d}\boldsymbol{f}}{\int \Bigl[\prod_{x=1}^{M} \prod_{t=1,\dots,T}^{x^{(t)}=x} f_{y^{(t)}|x}\Bigr] \mathrm{p}(\boldsymbol{f} | \, I) \, \mathrm{d}\boldsymbol{f}}.\end{aligned} \tag{19}$$

This formula is equivalent to (12) with and updated distribution for the frequencies:

$$\begin{aligned}\mathrm{p}(\boldsymbol{f} | \, y^{(T)}, x^{(T)}, \dots, y^{(1)}, x^{(1)}, \; I) \, \mathrm{d}\boldsymbol{f} = \\ \frac{\bigl[\prod_{x=1}^{M} \prod_{t=1,\dots,T}^{x^{(t)}=x} f_{y^{(t)}|x}\bigr] \mathrm{p}(\boldsymbol{f} | \, I)}{\int \bigl[\prod_{x=1}^{M} \prod_{t=1,\dots,T}^{x^{(t)}=x} f_{y^{(t)}|x}\bigr] \mathrm{p}(\boldsymbol{f} | \, I) \, \mathrm{d}\boldsymbol{f}} \, \mathrm{d}\boldsymbol{f}.\end{aligned} \tag{20}$$

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 $(\boldsymbol{q}_1, \dots, \boldsymbol{q}_M)$ 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,

$$\mathrm{p}(\boldsymbol{f} | \, y^{(T)}, x^{(T)}, \dots, y^{(1)}, x^{(1)}, \; I) \propto \exp\bigl[-\textstyle\sum_x T_x D(\boldsymbol{q}_x | \, \boldsymbol{f}_x)\bigr] \, \mathrm{p}(\boldsymbol{f} | \, I), \tag{21}$$

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 $\mathrm{p}(\boldsymbol{f}|\, I)\,\mathrm{d}\boldsymbol{f}$ greatly influence our inference (19). This distribution determines two important ***

6 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: $\mathrm{P}(A|\, D, I)$. Optimal decision is B that maximizes

$$\sum_A c(B|\, A)\,\mathrm{P}(A|\, D, I). \tag{22}$$

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

$$c'(B|\, A) \coloneqq c(B|\, A)\,\frac{\mathrm{P}(A|\, D, I)}{\mathrm{P}(A|\, D, I')}. \tag{23}$$

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 LaTeX, Emacs, AUCTeX, 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.)

Barndorff-Nielsen, O. E., Blæsild, P., Schou, G., eds. (1974): *Proceedings of Conference on Foundational Questions in Statistical Inference: Aarhus, May 7–12, 1973*. (University of Aarhus, Aarhus).

Barndorff-Nielsen, O. E., Dawid, A. P., Diaconis, P., Johansen, S., Lauritzen, S. L. (1984): *Discussion of Steffen Lauritzen's paper ['Extreme Point Models in Statistics']*. Scand. J. Statist. **11**^{2}, 83–91. See Lauritzen (1984).

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.

Lauritzen, S. L. (1974a): *Sufficiency, prediction and extreme models*. In: Barndorff-Nielsen, Blæsild, Schou (1974), 249–269. With discussion. Repr. without discussion in Lauritzen (1974b).

— (1974b): *Sufficiency, prediction and extreme models*. Scand. J. Statist. **1**^{3}, 128–134.

— (1984): *Extreme point models in statistics*. Scand. J. Statist. **11**^{2}, 65–83. See also discussion and reply in Barndorff-Nielsen, Dawid, Diaconis, Johansen, Lauritzen (1984).

— (1988): *Extremal Families and Systems of Sufficient Statistics*. (Springer, Berlin). First publ. 1982.

Levin, E., Tishby, N., Solla, S. A. (1990): *A statistical approach to learning and generalization in layered neural networks*. Proc. IEEE **78**^{10}, 1568–1574.

MacKay, D. J. C. (1992a): *Bayesian interpolation*. Neural Comp. **4**^{3}, 415–447. http://www.inference.phy.cam.ac.uk/mackay/PhD.html.

— (1992b): *A practical Bayesian framework for backpropagation networks*. Neural Comp. **4**^{3}, 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.