Some theoretical and experimental observerations on naïve discriminative learning

Stefan Evert
Friedrich-Alexander-Universität Erlangen-Nürnberg
Email: stefan.evert@fau.de

Antti Arppe
University of Alberta
Email: arppe@ualberta.ca

I. Introduction

Naïve Discrimative Learning (NDL) [5], [4], as implemented in [8], performs linguistic classification on the basis of direct associations between cues and outcomes, which are learned with the Rescorla-Wagner (R-W) equations. It is well known that the R-W model is related to neural networks (through the "delta rule" for gradient-descent training of a single-layer perceptron or SLP) and to linear least-squares regression, e.g. [6] and [5]. Furthermore, empirical observations have shown that, in comparison to other statistical classification techniques (logistic regression, support vector machines, random forrests, and memory-based learning), logistic regression appears in practice closest to NDL, in terms of overall prediction accuracy, agreement among individual predicted outcomes, and the distribution of estimated probabilities [3] [formally, does one need to bridge/fudge the empirical link here of NDL with logistic vs. the formal link of NDL with ordinary regression?]. However, most authors do not seem to be aware of the true depth of the afore-mentioned equivalence and of its implications.

Danks [6] argued that if the R-W equations successfully acquire the true associations between cues and outcomes, they should approximate an equilibrium state in which the expected change in associations $E[\Delta V] = 0$ if a cue-outcome event is randomly presented to the learner. This equilibrium state can be computed directly by solving a matrix equation, without carrying out many iterations of R-W updates. Note that – unless the learning rate is gradually decreased – the R-W model cannot converge to its equilibrium state (and will rather oscillate around the equilibrium).

In this paper, we show that the R-W equations are identical to gradient-descent training of a single-layer feed-forward neural network, which we refer to as a single layer perceptron (SLP)[1] here (Sec. III). Based on this result, we present a new, simpler derivation of the equilibrium conditions of [6] and prove that they correspond to the solution of a linear least-squares regression problem (Sec. IV). In Sec. V we discuss some consequences of these new insights.

[1]The term SLP is often reserved for a particular form of such a single-layer network using a Heavyside activation function. Here, we use it more generally to refer to any single-layer feed-forward network.

II. The Rescorla-Wagner equations

This section gives a mathematically precise definition of the R-W model, following the notation of [6]. The purpose of the R-W equations is to determine associations between a set of cues $C_1, \ldots, C_n$ and a single outcome O in a population of event tokens $(\mathbf{c}^{(t)}, o^{(t)})$ with

$$c_i^{(t)} = \begin{cases} 1 & \text{if } C_i \text{ is present} \\ 0 & \text{otherwise} \end{cases} \qquad o^{(t)} = \begin{cases} 1 & \text{if } O \text{ results} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

and $t = 1, \ldots, m$ (where $m = \infty$ can be allowed).

When presented with an event $(\mathbf{c}, o)$, the R-W equations update the associations V_i between cues and the outcome according to Eq. (2), which is a more formal notation of Eq. (1) from [6, 111].

$$\Delta V_i = \begin{cases} 0 & \text{if } c_i = 0 \\ \alpha_i \beta_1 \big(\lambda - \sum_{j=1}^n c_j V_j\big) & \text{if } c_i = 1 \wedge o = 1 \\ \alpha_i \beta_2 \big(0 - \sum_{j=1}^n c_j V_j\big) & \text{if } c_i = 1 \wedge o = 0 \end{cases} \tag{2}$$

Here, α_i is a measure of the salience of cue C_i, β_1 and β_2 are learning rates for positive ($o = 1$) and negative ($o = 0$) events, and λ is the maximal activation level of the outcome O. A simplified form of the R-W equations proposed by [9] assumes that $a_1 = \cdots = a_n = 1$, $\beta_1 = \beta_2 = \beta$ and $\lambda = 1$ (known as the W-H rule).

[6] argues that a successful R-W model should approach an equilibrium state of the association vector $\mathbf{V} = (V_1, \ldots, V_n)$ where the expected update $E[\Delta V_i] = 0$ if a random event token is sampled from the population. If we make the simplifying assumption that $\beta_1 = \beta_2 = \beta$, this condition corresponds to the equality

$$\lambda \frac{1}{m} \sum_{t=1}^{m} c_i^{(t)} o_i^{(t)} - \sum_{j=1}^{n} V_j \frac{1}{m} \sum_{t=1}^{m} c_i^{(t)} c_j^{(t)} = 0 \tag{3}$$

In Danks's notation, Eq. (3) can be written as

$$\lambda P(O, C_i) - \sum_{j=1}^{n} V_j P(C_i, C_j) = 0 \tag{4}$$

and is equal to his Eq. (11) [6, 113] multiplied by $P(C_i)$.

III. R-W and the single layer perceptron

We will now formulate a single-layer feed-forward neural network (SLP) whose learning behaviour – with gradient-descent training, which corresponds to the backprop algorithm for a SLP and is also known as the "delta rule" in this case – is identical to the R-W equations with equal positive and negative learning rates $\beta_1 = \beta_2$ (but no other restrictions). The SLP requires a slightly different representations of events as pairs $(\mathbf{x}^{(t)}, z^{(t)})$ with

$$x_i^{(t)} = \begin{cases} a_i & \text{if } C_i \text{ is present} \\ 0 & \text{otherwise} \end{cases} \qquad z^{(t)} = \begin{cases} \lambda & \text{if } O \text{ results} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Here, $a_i > 0$ is a (different) measure of the salience of cue C_i and $\lambda > 0$ the maximum activation of outcome O. Note that the event representation $(\mathbf{x}, z)$ is connected to the representation $(\mathbf{c}, o)$ through the equivalences $x_i = a_i c_i$ and $z = \lambda o$. In the W-H case, the two representations are identical.

The SLP uses a linear activation function $h(y) = y$ and Euclidean cost for the difference between y and the desired activation z. It computes the activation of the outcome as a linear combination $y = \sum_{i=1}^{n} w_i x_i$, where $\mathbf{w} = (w_1, \ldots, w_n)$ is the weight vector of the SLP. The cost associated with a given event token $(\mathbf{x}, z)$ is

$$E(\mathbf{w}, \mathbf{x}, z) = (z - y)^2 = \left(z - \sum_{i=1}^{n} w_i x_i \right)^2 \tag{6}$$

For batch updates based on the full population of event tokens, the corresponding batch cost is

$$E(\mathbf{w}) = \sum_t E(\mathbf{w}, \mathbf{x}^{(t)}, z^{(t)}) \tag{7}$$

If smaller batches are used, the sum j ranges over a subset of the population for each update step.

Given an event token $(\mathbf{x}, z)$, gradient-descent training of this SLP updates the weight vector by

$$\Delta w_i = -\delta \frac{\partial E(\mathbf{w}, \mathbf{x}, z)}{\partial w_i} \tag{8}$$

where $\delta > 0$ is the learning rate and the gradient $\partial E / \partial w_i$ is given by

$$\frac{\partial E(\mathbf{w}, \mathbf{x}, z)}{\partial w_i} = 2(z - y)(-x_i) = -2 \left(z - \sum_{j=1}^{n} w_j x_j \right) x_i \tag{9}$$

Inserting the definition of $\mathbf{x}$ and z from Eq. (5), we obtain

$$\Delta w_i = \begin{cases} 0 & \text{if } c_i = 0 \\ 2\delta a_i \big(\lambda - \sum_{j=1}^{n} c_j a_j w_j \big) & \text{if } c_i = 1 \wedge o = 1 \\ 2\delta a_i \big(0 - \sum_{j=1}^{n} c_j a_j w_j \big) & \text{if } c_i = 1 \wedge o = 0 \end{cases} \tag{10}$$

Comparing this with Eq. (2), we can set $V_j = a_j w_j$, i.e. we interpret the weight vector $\mathbf{w}$ of the SLP as salience-adjusted cue-outcome associations. With $\Delta V_i = a_i \Delta w_i$, we obtain

$$\Delta V_i = \begin{cases} 0 & \text{if } c_i = 0 \\ 2\delta a_i^2 \big(\lambda - \sum_{j=1}^{n} c_j V_j \big) & \text{if } c_i = 1 \wedge o = 1 \\ 2\delta a_i^2 \big(0 - \sum_{j=1}^{n} c_j V_j \big) & \text{if } c_i = 1 \wedge o = 0 \end{cases} \tag{11}$$

which is equal to the R-W equations for $\beta = 2\delta$ and $\alpha_i = a_i^2$.

The assumption $\beta_1 = \beta_2$ can be relaxed if we change the representation of events to

$$x_i^{(t)} = \begin{cases} a_i & \text{if } c_i = 1 \wedge o = 1 \\ a_i \sqrt{\frac{\beta_2}{\beta_1}} & \text{if } c_i = 1 \wedge o = 0 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

i.e. if we allow the salience of cues to differ between positive ($o = 1$) and negative ($o = 0$) events; the scaling facter β_2 / β_1 is the same for all cues C_i. We do not pursue this extension further here because it affects the equilibrium state in an unpredictable way. As [6] has already observed, the cue saliences α_i have no impact at all and the maximum activation level λ merely results in a linear scaling of the equilibrium state.

IV. R-W and least-squares regression

We have shown in Sec. III that the R-W equations describe the gradient-descent training of a SLP for the linear regression problem

$$\min_{\mathbf{w}} E(\mathbf{w}) = \min_{\mathbf{w}} \sum_t E(\mathbf{w}, \mathbf{x}^{(t)}, z^{(t)}) \tag{13}$$

This equivalence holds generally, not only in the case of the simplified W-H rule. Thus, both R-W and our SLP aim to solve the same regression problem.

If the training procedure is successful, the weight vector $\mathbf{w}$ should approach the least-squares solution. With single updates (corresponding to the R-W model), convergence cannot be achieved unless the learning rate is gradually reduced. With batch updates treating the entire population as a single batch, the cost $E(\mathbf{w})$ is a convex function of $\mathbf{w}$ and the gradient descent procedure converges to its unique minimum after a sufficient number of iterations.[2]. In order to keep the discussion straightforward, we assume the general case of a unique minimum in the present paper.

In order to express Eq. (13) more concisely, we define an $m \times n$ matrix $\mathbf{X} = (x_i^{(t)}) = (x_{ti})$ of cues for all event tokens in the population. The rows of this matrix correspond to event tokens t, the columns to cues i; i.e. row number t contains the input vector $\mathbf{x}^{(t)}$. We also define the column vector $\mathbf{z} = (z^{(1)}, \ldots, z^{(m)})$ of outcomes and recall that $\mathbf{w} = (w_1, \ldots, w_n)$ is a column vector of SLP weights. The batch cost can now be written as

$$E(\mathbf{w}) = \big(\mathbf{z} - \mathbf{X}\mathbf{w} \big)^2 \tag{14}$$

[2] In fact, the minimum of $E(\mathbf{w})$ might not be unique under certain circumstances, viz. if the correlation matrix of the cues is not positive definite, cf. [6, 115–116], who considers only the special case of "coextensive" cues

The least-squares solution must satisfy the condition $\nabla E(\mathbf{w}) = \mathbf{0}$, which leads to the so-called normal equations

$$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{z} \tag{15}$$

In the case of the W-H rule, $\mathbf{X}$ is a coincidence matrix between cues and events, with $x_{ti} \in \{0, 1\}$. A straightforward calculation shows that $\mathbf{X}^T\mathbf{X}$ is a square co-occurrence matrix with entries $f(C_i, C_j)$, and $\mathbf{X}^T\mathbf{z}$ is a vector of co-occurrence counts $f(C_i, O)$ between the cues and the outcome O. Dividing Eq. (15) by m, we obtain Eq. (4) with $\lambda = 1$, i.e.

$$P(O, C_i) - \sum_{j=1}^{n} V_j P(C_i, C_j) = 0$$

which is the same as Eq. (3) of [6] with rows multiplied by $P(C_i)$. Since linear regression is invariant wrt. the salience factors a_i (the weights are simply adjusted by reciprocal factors $1/a_i$ to achieve the same regression values) and scales linearly with λ, equivalence to the equilibrium conditions [6, 112–114] also holds for arbitrary values of a_i and λ.

In the general case where the regression problem has a unique solution, $\mathbf{X}^T\mathbf{X}$ is symmetric and positive definite. It can therefore be inverted and the least-squares solution is given by

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z} \tag{16}$$

Standard statistical software such as R [7] can be used to compute $\mathbf{w}^*$ reliably and efficiently. It is not necessary to carry out the iterative training procedure of the R-W model or the neural network, and there is no need to worry about convergence of the iterative training.

V. Consequences

We have shown that R-W association learning, a linear SLP neural network and linear regression are fully equivalent and should ideally lead to the same least-squares solution. As long as a researcher is only interested in the result of association learning, not in the iterative process itself, it is sufficient to calculate the least-squares solution directly from Eq. (16). Essentially, the R-W salience factors α_i have no effect on the learning result – because linear regression is not sensitive to such a scaling of the input variables – but only on the learning process: associations for cues with high salience α_i are learnt faster than for other cues. The parameter λ leads to a (trivial) linear scaling of the learning result, but has not effect on the learning process. Only different learning rates $\beta_1 \neq \beta_2$ affect the learning result, because they modify $\mathbf{X}^T\mathbf{X}$ in a complex way.

If R-W association learning or SLP training does not approximate the least-squares solution, it can arguably be considered to have failed. The only research question of interest that requires R-W iteration or application of the delta rule is thus: Under which circumstances and for which parameter settings does the R-W iteration converge or at least approximate the linear regression? This is particularly relevant for single-event updates (as specified for the R-W model), which are much less robust and lead

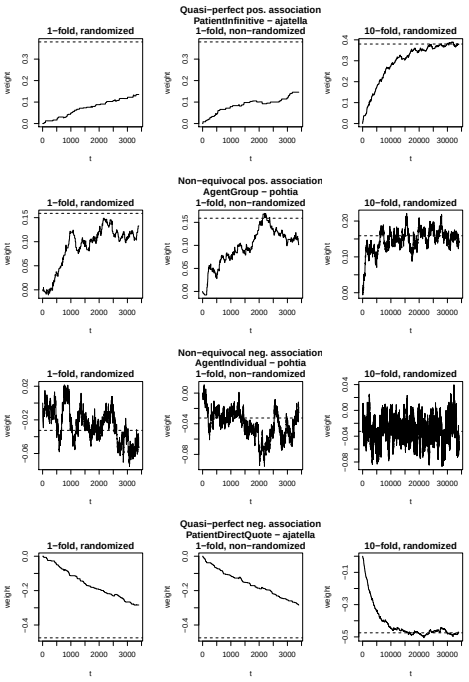

Fig. 1. Simulation results for a Finnish THINK dataset with selected linguistic contextual features (cues) and verbs (outcomes), using (i) R-W learning within a randomized version of the original dataset (3404 datapoints) (ii) R-W learning with a non-randomized version of the dataset (3404 datapoints), and (iii) R-W learning with a 10-fold, randomized multiple of the dataset (34040 datapoints). R-W cue-outcome association values presented with a solid line; corresponding equilibrium association values with a dotted line.

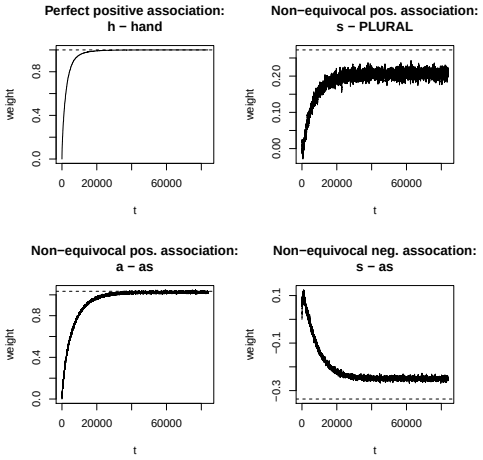

Fig. 2. Simulation results for a tiny, artificial dataset on English nouns and their pluralizations (10 types with 419 tokes), with selected orthographical unigraph features (cues) and lemmas (outcomes), using R-W learning with a 200-fold, randomized version of the dataset [5, Extension of Fig. 4]. R-W cue-outcome association values presented with a solid line; corresponding equilibrium association values with a dotted line.

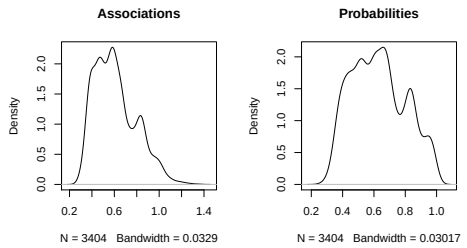

Fig. 3. Distributions of instance-wise estimated maximal associations (NDL) and probabilities (GLM) using the Finnish THINK dataset ($n = 3404$).

to larger fluctuations than batch updates. We plan to study these issues with the help of simulation experiments.

For instance, with a real dataset on the relationship between 46 linguistic contextual features as cues and 3404 occurrences of 4 near-synonymous Finnish THINK verbs as outcomes [1], [2], [8], the cue-outcome associations arising from a simulation of R-W learning (using the default parameters) do appear to (eventually) converge with the equilibrium values (Fig. 1). For the non-equivocal contextual features that have occurred, more or less, with multiple outcomes (e.g. AGENTGROUP and AGENTINDIVIDUAL with *pohtia*), this convergence seems to happen early, well within the course of the dataset. In contrast, with near-categorical features that in practice occur with only one of the four outcomes (e.g. the near-categorical co-occurrence of PATIENTINFINITIVE with *ajatella* vs. the categorical non-occurrence of PATIENTDIRECTQUOTE with this same verb), convergence appears to happen much slower, so that the entire dataset is not sufficient for this, requiring as many as approximately five or more times the extent of the original dataset. Furthermore, this simulation clearly shows, particularly in the case of AGENTINDIVIDUAL and *pohtia*, a remaining, sometimes quite substantial oscillation in the cue-outcome associations weights. In this respect, establishing how the learning parameters β_1 and β_2 might be adjusted in the course of the R-W learning process would be worthwhile. Interestingly, for the quasi-perfect cases, whether the order of datapoints in this dataset is randomized in the learning process, or not, does not appear to have an effect on the asymptotic result of R-W learning, in contrast to reservations in [6, 119] – but for the non-equivocal cases, the assumption of randomized order appears motivated. Of course, in all this one generally presumes that the proportions of the cue-outcome co-occurrences are the same in the overall population from which the dataset is a sample of.

In contrast, using the tiny, artificial dataset on English nouns and their pluralizations (10 types with 419 tokes), with orthographical unigraph features as cues and lemmas/features as outcomes ([5, Fig. 4]), in some but not all cases the cue-outcome association weights arising from R-W learning do not appear to converge with the equilibrium values, even with a 200-fold, randomized version of the dataset, e.g. '*s*' as cue and PLURAL as outcome, and '*s*' as cue and '*as*' as outcome (Fig. 2).[3]

Finally, having established NDL as linear regression with its well-known drawbacks (e.g. a propensity for overfitting the training data, especially if there is a large number n of cues), it will be interesting to contrast it with more state-of-the-art machine learning techniques, as a systematic follow-up and analysis of the empirical observations in [3].[4] We plan to carry out a mathematical analysis and empirical study of (i) logistic regression (a subtype of Generalized Linear Model [GLM]), which is more appropriate for categorical [Does it matter whether 'dichotomous' or 'polytomous'?] data than linear least-squares regression, and (ii) regularization techniques, which control overfitting and encourage sparse solutions.

REFERENCES

[1] Arppe, A. (2008). *Univariate, bivariate and multivariate methods in corpus-based lexicography – a study of synonymy*. Publications of the Department of General Linguistics, University of Helsinki, No. 44. URN: http://urn.fi/URN:ISBN:978-952-10-5175-3.

[2] Arppe, A. (2013). *polytomous: Polytomous logistic regression for fixed and mixed effects*. R package version 0.1.6. http://CRAN.R-project.org/package=polytomous

[3] Arppe, A. and Baayen, R. H. Statistical classication and principles of human learning. *Quantitative Investigations in Theorertical Linguistics, QITL-4*, Berlin, March 30, 2011.

[4] Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, 11, 295-328.

[5] Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P. and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118, 438-482.

[6] Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology*, 47, 109 –121. doi:10.1016/S0022- 2496(02)00016-0.

[7] R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

[8] Shaoul, C., Arppe, A., Hendrix, P., Milin, P. and Baayen, R. H. (2014). *ndl: Naive Discriminative Learning*. R package versions 0.1.6-0.2.16.

[9] Widrow, B., and Hoff, M. E. (1960). Adaptive switching circuits. *IRE WESCON convention record*, 96–104. New York: IRE. (Reprinted in Anderson, J. A., and Rosenfeld, E. (Eds.) (1988). *Neurocomputing: Foundations of research*, 123–134. Cambridge, MA: MIT Press.)

[3] Among various versions of the `ndl` package [8, v0.1.6 vs 0.2.16], there are several variants as how to calculate the equilibrium weights, but that would not appear to be the source of this observed non-convergence.

[4] Results using the Finnish THINK dataset [1], [2] indicated that (1) outcomes predicted by NDL and GLM (in the case of the 4-way *polytomous* outcome using the *one-vs-rest* technique implemented in the `polytomous` R package [2]) agree with a rate of 94.8%, that (2) the distributions of instance-wise maximum NDL cumulative association weights and corresponding maximum GLM expected probabilities have similar general distribution contours (cf. Fig. 3), and that (3) instance-wise maximum NDL association weights and corresponding maximum GLM expected probabilities correlate highly, with $r_{Spearman} = .950$, as do (4) individual NDL association weights and GLM log-odds, with $r_{Spearman} = .897$.