THE TAYLOR EXPANSION FOR DROPOUT IS DIVERGENT
(EXTENDED ABSTRACT)

Natalie Schluter
Department of Computer Science
IT University of Copenhagen
Copenhagen, Denmark
natschluter@itu.dk

ABSTRACT

This extended abstract examines the assumptions of the derived equivalence between dropout noise injection and $L_2$ regularisation for logistic regression with negative log loss [Wager et al. (2013)]. We show that the approximation method is based on a divergent Taylor expansion. Hence, subsequent work using this approximation to compare the dropout trained logistic regression model with standard regularisers, remains unfortunately ill-founded to date.

1 INTRODUCTION

This extended abstract examines the assumptions of the derived equivalence between dropout noise injection and $L_2$ regularisation for logistic regression with negative log loss [Wager et al. (2013)]. An important amount of empirical work has been devoted to trying to empirically extend this finding and simultaneously reconcile it with some seemingly contradictory behaviours of the noise regulariser term in question [Srivastava et al. (2014); Helmbold & Long (2015, 2016)]. For instance, Helmbold & Long (2016) remark that dropout training leads to negative parameters, even when the output is a positive multiple of the inputs, signaling co-adaptation of parameters. They also explain that for neural networks, the dropout penalty grows exponentially in the depth of the network in cases where the $L_2$ regulariser grows linearly. These are both behaviours that are in strict contrast with the behaviour of the $L_2$ regulariser.

Extending Bishop (1995)'s result to dropout noise injection for Generalised Linear Models (GLMs) [Wager et al. (2013)] applied the second order approximation technique for logistic regression models with negative log loss, trained with dropout noise injection. They conclude that dropout training is “first-order” equivalent to training with $L_2$ regularisation, after a special scaling of the features. We show that the approximation method is based on a divergent Taylor expansion. Hence, subsequent work using this approximation to compare the dropout trained logistic regression model with standard regularisers, remains unfortunately ill-founded to date.

2 PRELIMINARIES

Let $y \in \{0, 1\}$ be a response variable given a feature vector $x \in \mathbb{R}^d$ and $A$ is the log partition function, $A(z) = (1 + \exp(-z))^{-1}$. The logistic regression model for the distribution of $y$ is given by $p_\beta(y|x) = h(y) \exp(y \cdot \beta - A(x \cdot \beta))$, where $h(y)$ is some quantity independent of $x$ and $\beta \in \mathbb{R}^d$ is the natural parameter (weight vector). Correspondingly, the loss function, negative log likelihood, is given by $l_{x,y}(\beta) := -\log p_\beta(y|x)$. So finding optimal $\beta^*$ with respect to a dataset with $n$ examples, amounts to minimising expected loss over the dataset $\{(x_i, y_i)\}_{i=1}^n$: $\beta^* := \arg\min_{\beta \in \mathbb{R}^d} \mathbb{E}[l_{x_i,y_i}(\beta)]$.

A dropout noised example $x_i \in \mathbb{R}^d$ is defined $\tilde{x}_i := x_i \odot \xi_i$, where $\odot$ is the Hadamard product (i.e., element-wise vector product) and $\xi_i \in \{0, (1 - \delta)^{-1}\}^d$, such that $\xi_{ij} = 0$ with probability $\delta$ and otherwise $\xi_{ij} = (1 - \delta)^{-1}$. That is, components of $\tilde{x}_i$ are dropped to 0 following a Bernoulli($\delta$) distribution and are otherwise a scaled version of the respective components from $x_i$. 
Wager et al. (2013) use the approximation technique from Bishop (1995) to study dropout. They express GLM log loss as log loss with a sort of drop-out regulariser. That is, on expectation given the dropout noise, the loss can be expressed as

\[
E_\xi[l_{\xi,Y}(\beta)] = -(y x_1 \cdot \beta + A(x_1 \cdot \beta) - A(x_1 \cdot \beta) - E_\xi[A(x_1 \cdot \beta)]) = l_{\xi,Y}(\beta) + R(\beta)
\]

where \(R(\beta)\) can be seen as a regulariser to be minimised with loss in lieu of training with dropout.

Wager et al. (2013) derived an approximation for \(R(\beta)\) that revealed an equivalence with the \(L_2\) expression. In order to do this, they resort to a polynomial approximation of \(R(\beta)\) of the expected log partition function \(A\) by means of taking the first three terms (up to the second order) of the Taylor expansion of \(E_\xi[A(x_1 \cdot \beta)]\) on low noise levels, following Bishop (1995) for Gaussian noise. We will show here that the Taylor expansion in the case of dropout does not converge, making the approximation upon which this is based non-applicable.

With \(\bar{x}_1 \cdot \beta\) close to \(x_1 \cdot \beta\), i.e., on low noise levels, and writing \(R_2\) for the quadratic approximation, they obtain \(R_2(\beta) = \frac{1}{2}A''(\bar{x}_1 \cdot \beta)Var_\xi[\bar{x}_1 \cdot \beta]\).

We are studying dropout, at probability \(\delta\) and otherwise scaling by \(\frac{1}{1-\delta}\). To model this situation, we define the constants (with respect to dropout probability and \(i\)) if no parameters are dropped

\[
B_i := (\bar{x}_1 \cdot \beta - x_1 \cdot \beta) = \left(\frac{x_1 \cdot \beta}{1-\delta} - x_1 \cdot \beta\right) = (x_1 \cdot \beta) \left(\frac{\delta}{1-\delta}\right).
\]

To take into consideration a weight dropping to 0, we introduce the boolean random variable \(Y\) which follows a Bernoulli distribution with parameter \(E_\xi[Y] = (1 - \delta)\). We let \(Z_i := B_i \cdot Y\). Since \(B_i^n = (x_1 \cdot \beta)^n \left(\frac{\delta}{1-\delta}\right)^n\) in general, for \(E_\xi[Z^n]\), we have \(E_\xi[Z^n] = B_i^n E_\xi[Z^n] = B_i^n E_\xi[Y^n]
= (x_1 \cdot \beta)^n \left(\frac{\delta^n}{(1-\delta)^n}\right)\). So, in particular for \(n = 2\), we have

\[
E_\xi[(\bar{x}_1 \cdot \beta - x_1 \cdot \beta)^2] = (x_1 \cdot \beta)^2 \left(\frac{\delta^2}{1-\delta}\right)
\]

Where in the case of logistic regression, \(A''(x_1 \cdot \beta) = p_i(1-p_i)\) and \(p_i = (1 + e^{-x_1 \cdot \beta})^{-1}\).

3 K-TH ORDER POLYNOMIAL APPROXIMATIONS

An order \(k\) polynomial approximation of \(E_\xi[A(x_1 \cdot \beta)]\) is given by taking the first \(k\) terms of the Taylor expansion, with \(\bar{x}_1 \beta\) close to \(x_1 \beta\):

\[
E_\xi[A(x_1 \cdot \beta)] \approx \sum_{m=0}^{k} \frac{E_\xi[A^{(m)}(\bar{x}_1 \cdot \beta) \cdot (x_1 \cdot \beta - x_1 \cdot \beta)^m]}{m!}
\]

\[
= \sum_{m=0}^{k} \frac{A^{(m)}(x_1 \cdot \beta) \cdot E_\xi[(\bar{x}_1 \cdot \beta - x_1 \cdot \beta)^m]}{m!}
\]

\[
= \sum_{m=0}^{k} \frac{A^{(m)}(x_1 \cdot \beta) \cdot (x_1 \cdot \beta)^m \delta^m}{m!(1-\delta)^{m-1}}
\]

Let us denote \(A_k := A^{(k)}(x \cdot \beta)\). Let \(p := (1 + e^{-x \cdot \beta})^{-1}\). So, \(A_0 = p\) and \(A_1 = p(1-p) =: p'\), where \(p'\) is the derivative of \(p\). We now provide an explicit form for \(A_k\), in the case of logistic regression.

The Triangle of numbers, \(T(n, k)^{\dagger}\) is defined as

\[
T(n, k) = k! S_2(n, k)
\]

\(^{\dagger}\)Sequence number A019538 in the OEIS [oeis.org](https://oeis.org)
for \( n \geq 1 \) and \( 1 \leq k \leq n \), where \( S_2(n, k) \) are Stirling numbers of the second kind.

It is straightforward to show by induction the following result, which we include in the long version of this paper.

**Theorem 1.** For \( k \geq 1 \), \( A_k = p' \sum_{j=1}^{k} (-1)^{j-1} \cdot p^{j-1} \cdot T(k, j) \).

## 4 The appropriateness of the Taylor expansion

We note that each component \( x_{ij} \) of \( \tilde{x}_i \beta \) is in the set \( \{0, \frac{x_i \beta}{1-\rho} \} \). In other words, the perturbed value \( \tilde{x}_i \beta \) needs to be within the radius of convergence \( r \) of the Taylor expansion, \( |\tilde{x}_i \beta - x_i \beta| < r \). However, the value \( x_i \beta \) is completely unconstrained: the only assumption made thus far (following Wager et al. (2013)) being that the levels of noise are “low”. For additive Gaussian noise, noise levels can be adjusted so that this assumption is satisfied, so long as there is some \( r \) for which this the Taylor series expansion converges. However, dropout causes arbitrarily many values of \( x_i \) to be zero, thus \( \tilde{x}_i \beta \) can be very close to zero, even though \( x_i \beta \) can be arbitrarily large.

We now show that the radius of convergence is bounded, which excludes any meaningful study of dropout using Taylor series expansion of the expected log partition function. Recall \( R(\beta) = \mathbb{E}_x[A(\tilde{x}_i \cdot \beta)] - A(x_i \cdot \beta) \) and that \( p_i = (1 + \exp(x_i \beta))^{-1} \). Also recall the explicit expression for \( A_k \) from Theorem 1.

**Theorem 2.**

\[
R_\infty(\beta) := \lim_{k \to \infty} \sum_{j=2}^{k} A_j E[(\tilde{x}_i \cdot \beta - x_i \cdot \beta)^j] \] converges to \( R(\beta) \) if and only if \( |\tilde{x}_i \beta - x_i \beta| < 2\pi \).

That is, Theorem 2 states that the radius of convergence of \( R_\infty(\beta) \) is \( 2\pi \) for examples in the dataset that the un-noised model will classify as true. Since \( x_i \beta \) is unbounded, and certainly larger than \( 2\pi \) arbitrarily much of the time during training, we have a proof of Theorem 2.

**Proof.** We want to test for convergence of the ratio of factors, and if convergent, calculate the convergence ratio. For this we use the root test. Using the \( \limsup \) takes into account \( p \to 1 \).

\[
\limsup_{k \to \infty} \sqrt[k]{\frac{A_k}{k!}} = \limsup_{k \to \infty} 2k \sqrt[k]{\frac{A_{2k}}{(2k)!}} = \limsup_{k \to \infty} 2k \sqrt[k]{\frac{1}{(2k)!} \sum_{j=1}^{2k} (-p)^j \cdot T(2k, j)} \]

\[
= \limsup_{k \to \infty} 2k \sqrt[k]{\frac{1}{(2k)!} \sum_{j=1}^{2k} (-1)^j \cdot j \cdot S(2k, j)} \]

\[
= \limsup_{k \to \infty} 2k \sqrt[k]{\frac{1}{(2k)!} \sum_{j=1}^{2k} j \cdot S(2k, j)} \]

\[
= \limsup_{k \to \infty} 2k \sqrt[k]{\frac{1}{(2k)!} B_{2k}} = \limsup_{k \to \infty} 2k \sqrt[k]{\frac{1}{(2k)!} \frac{2 \cdot (2k)!}{(2\pi)^{2k}} \zeta(2k)} \]

\[
= \frac{1}{2\pi} \]

where is \( B_k \) is the \( k \)th Bell number. We used the well-established fact that

\[
B_{2k} \sim \frac{2 \cdot (2k)!}{(2\pi)^{2k}} \zeta(2k)
\]

asymptotically as \( k \) grows, where \( \zeta(2k) \) is the Riemann zeta function, which tends to 1 as \( k \to \infty \). This concludes the proof. \( \square \)
REFERENCES


