the KL information $I(b_n : P)$ is found to be

$$I(b_n : P) = \sum_{x=0}^{n} b_n \ln \frac{b_n}{P} = \left(\frac{\lambda}{2n}\right)^2 + o\left(\frac{1}{n^2}\right).$$

Example 3.5: This example considers convergence of a sequence of negative binomial distributions to the logarithmic series distribution (see Johnson and Kotz [5, p. 170]). We wish to examine the rate of convergence, as ϵ approaches zero, of

$$I_\epsilon = \sum_{k=1}^{\infty} p_\epsilon(k) \ln \frac{p_\epsilon(k)}{q(k)},$$

where

$$p_\epsilon(k) = \frac{\Gamma(k+\epsilon)}{k!\Gamma(1+\epsilon)} \frac{\epsilon}{Q^\epsilon - 1}\left(\frac{P}{Q}\right), \qquad k = 1, 2, \cdots,$$

and

$$q(k) = \left(\frac{P}{Q}\right)^k / [k \ln Q], \qquad k = 1, 2, \cdots,$$

with $Q > 1$ and $P = Q - 1$. It can be shown that

$$I_\epsilon \leq \frac{\epsilon^2}{4} h(Q),$$

where

$$h(Q) = \frac{(Q-1)(2Q-1)}{Q \ln Q} - \ln^2 Q.$$

For moderate values of Q, say less than ten, the approximation

$$I_\epsilon \approx \frac{\epsilon^2}{4} \ln Q \tag{13}$$

works very well.

REFERENCES

[1] A. C. Berry, "The accuracy of the gaussian approximation to the sum of independent variates," *Trans. Amer. Math. Soc.*, vol. 49, pp. 122–136, 1941.

[2] H. Cramer, *Random Variables and Probability Distributions*, Cambridge Tracts in Mathematics No. 36. Cambridge: Cambridge Univ. Press, 1937.

[3] C. G. Essen, "Fourier analysis of distribution functions. A mathematical study of the Laplace–Gaussian law," *Acta Math.* vol. 77, pp. 1–125, 1945.

[4] B. W. Gnedenko and A. N. Kolomogorov, *Limit Distributions for Sums of Independent Random Variables*, (translated from the Russian). Cambridge: Addison–Wesley, 1968.

[5] N. L. Johnson and S. Kotz, *Discrete Distributions*. Boston,, MA: Houghton Mifflin, 1969.

[6] ——, *Continuous Univariate Distributions*. Boston, MA: Houghton Mifflin, 1970.

[7] J. H. B. Kemperman, "On the optimum rate of transmitting information," *Ann. Math. Statist.*, vol. 40, no. 6, pp. 2156–2177, 1969.

[8] S. Kullback, "A lower bound for discrimination information in terms of variation," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 126–127, 1966.

[9] ——, *Information Theory and Statistics*. New York: Dover, 1968.

[10] A. M. Lyapunov, "Nouvelle forme du theoreme sur la limite de probabilities," *Mem. Acad. Sci. St. Petersbourg*, vol. 12, no. 5, pp. 1–24, 1901.

[11] V. V. Petrov, *Sums of Independent Random Variables*. (translated from the Russian). Berlin: Springer-Verlag, 1975.

On Loss Functions Which Minimize to Conditional Expected Values and Posterior Probabilities

John W. Miller, *Member, IEEE*, Rod Goodman, *Member, IEEE*, and Padhraic Smyth, *Member, IEEE*

***Abstract*—A *loss function*, or *objective function*, is a function used to compare parameters when fitting a model to data. The loss function gives a distance between the model output and the desired output. Two common examples are the squared-error loss function and the cross entropy loss function. It is well known that minimizing the mean-square error loss function is equivalent to minimizing the mean square difference between the model output and the expected value of the output given a particular input. This property of minimization to the expected value is formalized as "*P*-admissibility." The necessary and sufficient conditions for *P*-admissibility, leading to a parametric description of all *P*-admissible loss functions are found. In particular, it is shown that two of the simplest members of this class of functions are the squared error and the cross entropy loss functions. One application of this work is in the choice of a loss function for training neural networks to provide probability estimates.**

***Index Terms*—Objective functions, loss functions, probability estimation.**

I. BACKGROUND

A *loss function*, or *objective function*, is a function used to compare parameters when fitting a mathematical model to data. For example, in linear regression, the problem is to find the line $f(x)$ which best "fits" a collection of data points x_i, y_i $(i = 1, \cdots, N)$. The line is a model M, which gives an estimate $\hat{y} = f(x)$ of the value y for each value x. The parameters to the model, $\underline{\theta}$, are the two values required to describe a line. Normally in linear regression the squared-error (SE) loss function $L(y, \hat{y}) = (y - \hat{y})^2$ is used, meaning that the "best fit" is considered to be the $\underline{\theta}$ that minimizes the average loss: $1/N \sum_{i=1}^{N} L(y, \hat{y})$. In the context of neural networks, the values $\underline{\theta}$ are the network's weights and thresholds. The estimate $\hat{y}$ is the network's output. We will call this method of finding the minimum average loss the "training algorithm," and the $\{\underline{x}_i, y_i\}$ values the "training data." Without loss of generality, y is assumed to be scalar. Table I summarizes our notation. Fig. 1 diagrams the relationship between the introduced symbols.

II. PROBABILITY EESTIMATION

Consider estimating the conditional mean, $E[y \mid \underline{x}]$, the mean value of y taken over all training samples with a given value for $\underline{x}$. For the case where y is a scalar binary value, $y \in \{0, 1\}$:

$$E[y \mid \underline{x}] = p(y = 1 \mid \underline{x}).$$

Manuscript received December 5, 1991. This work was supported in part by DARPA under Grant AFOSR-90-0199 and in part by NSF Grant ENG-8711673. This work was presented in part at the IEEE International Symposium on Information Theory, Budapest, Hungary, June 24–28, 1991. This work carried out in part by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

J. W. Miller is with Microsoft Research, 9S/1051, One Microsoft Way, Redmond, WA 98052.

R. Goodman is with the Department of Electrical Engineering, California Institute of Technology, 116-81, Pasadena, CA 91125.

P. Smyth is with the Communication Systems Research, Jet Propulsion Laboratory, 238-420, Pasadena, CA 91109.

IEEE Log Number 9209602.

TABLE I
SUMMARY OF NOTATION

Symbol	Explanation	Neural Network Example
θ	Parameters to a model	The weights and thresholds
$\underline{x}_i$, y_i	Training data: $\underline{x}_i$ is input, y_i is output, for sample i. The subscript is often omitted.	Training Data
$M(\theta)$	Model, a set of mappings $x \rightarrow \hat{y}$ parameterized by θ	The set of all possible functions that a network architecture can implement
$f(\underline{x})$	The mapping produced by the model with a given parameter set θ	A network with fixed weights and fixed thresholds
$\hat{y}_i$	For some given $\underline{x}_i$ the value $f(\underline{x}_i)$ This value is assumed to be scalar in this correspondence	The network output
$L(y_i, \hat{y}_i)$	The error between the desired and actual output	Squared Error (for example)

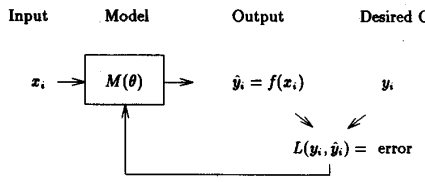

Fig. 1. Diagram for training—Setting θ to minimize error.

In classification problems $p(y = 1 \mid \underline{x})$ is sometimes called the *a posteriori* probability of y, the probability of $y = 1$ after the evidence $\underline{x}$ is known. In this correspondence, we deal with systems which estimate $E[y \mid \underline{x}]$, and include systems for estimating *a posteriori* probabilities as a special case. The training data is assumed to consist of independent samples from some underlying probability distribution over $\underline{x}$, y. There are two separate reasons why our estimate of $E[y \mid \underline{x}]$ will be different from the "true" probabilities that exist in the underlying distribution. The first reason is due to the limitation of the dataset. If the dataset is infinite, by the law of large numbers we can get an arbitrarily accurate estimate of the conditional probabilities $E[y \mid \underline{x}]$ by counting the frequency of occurrence in the training dataset. Since the dataset is finite, we only have a sampled estimate of this true underlying probability. This sampling error in the estimation is not considered here. Instead the probabilities referred to in the correspondence will be the relative frequencies found in the training set. The second reason for error in estimation is due to limitations of the model. A model may not have a parameter set $\underline{\theta}$ such that the function $f(\underline{x}) = E[y \mid \underline{x}]$ can be perfectly represented. A *sufficiently powerful* model $M(\underline{\theta})$ is one which, for some $\underline{\theta}$, is capable of producing $\hat{y}(\underline{\theta}, \underline{x}) = E[y \mid \underline{x}]$. For a sufficiently powerful model, the average loss must be minimized at $\hat{y} = E[y \mid \underline{x}]$:

$$\min_{\underline{\theta}} E[L(y, \hat{y}) \mid \underline{x}] \quad \text{is achieved when} \quad \hat{y}(\underline{\theta}, \underline{x}) = E[y \mid \underline{x}].$$

We will assume the values of the target y in the training sets are bounded such that $0 \leq y \leq 1$. This choice of upper and lower bounds for y is consistent with model outputs which represent probabilities. The results can be extended to problems with other upper and lower bounds by appropriately scaling and shifting the y values. The unbounded case is not studied here, but it can be shown that if no restriction is placed on the distribution of y, then no loss function can be guaranteed to minimize to the expected value of y.

It is useful to rewrite the "minimization at $\hat{y} = E[y \mid x]$" requirement in a way that explicitly shows the probability distribution used to calculate the expected value. Call this distribution $p(y)$. Until Section VII, all distributions and expected values will be "given $\underline{x}$." To simplify the notation, the explicit reference to the input $\underline{x}$ will be dropped in Sections II–VI. We can now replace the expected value with a definite integral, giving the formal definition:

$$\forall p(y) \qquad \text{s.t.} \int_0^1 p(y) \cdot dy = 1$$

$$\min_{0<\hat{y}<1} \int_0^1 p(y) \cdot L(y, \hat{y}) \cdot dy = \int_0^1 p(y) \cdot L(y, \overline{y}) \cdot dy, \quad \text{(c1)}$$

where $\overline{y} \equiv \int_0^1 p(y) \cdot y \cdot dy$. If $L(y, \hat{y})$ satisfies this property (c1), we call it "P-admissible," to indicate that the loss function is admissible for use in probability estimation or expected value estimation. A sufficiently powerful model trained using a P-admissible loss function will provide an estimate $\hat{y} = E[y \mid \underline{x}]$. In Sections III through VI, we study the set of loss functions which are P-admissible. In Section VII, the results are extended to models that may not be sufficiently powerful to approximate $E[y \mid \underline{x}]$—this is the more realistic case.

III. A Simple Example

Consider a dataset describing a set of patients. Let:

x = symptom of a disease,

y = presence of disease in patient (1 or 0).

Our training data is a set of medical records for 100 patients, all with symptom x. 90 of these patients have disease y, 10 do not, so from this sample:

$$E[y \mid x] = p(y = 1 \mid x) = \frac{90}{100} = 0.9.$$

Now, if we train a model on this sequence of zeros and ones, does the model output $E[y \mid x] = 0.9$? The answer is yes, if $L(y, \hat{y})$ is P-admissible.

IV. Two P-Admissible Loss Functions

The squared-error function is known to be P-admissible:

$$L_{se}(y, \hat{y}) = (\hat{y} - y)^2.$$

Least-squared minimization has the property that the derivative with respect to $\hat{y}$ is a linear function. This can be used to advantage in designing computationally efficient gradient based optimization schemes (training algorithms).

A loss function commonly used in neural network training algorithms is the cross entropy function [1]–[4]:

$$L_{ce}(y, \hat{y}) = y \cdot \log\left(\frac{y}{\hat{y}}\right) + (1 - y) \log\left(\frac{1 - y}{1 - \hat{y}}\right).$$

The cross entropy loss function is used for maximum likelihood estimation of model parameters when the training data consists of

classification labels [1]. L_{ce} is also P-admissible [2]. This is shown by finding the value of $\hat{y}$ such that $E[L_{ce}(y, \hat{y}) \mid \underline{x}]$ is minimized. Equation (c1) requires that the minimization solution be $\hat{y} = E[y]$. We can easily derive this by solving the minima problem:

$$\begin{aligned}\frac{\partial}{\partial \hat{y}} E[L_{ce}(y, \hat{y})] &= 0 \\ &= \frac{\partial}{\partial \hat{y}} \int_0^1 p(y) \cdot L_{ce}(y, \hat{y}) \cdot dy \\ &= \int_0^1 p(y) \cdot \frac{\partial}{\partial \hat{y}} L_{ce}(y, \hat{y}) \cdot dy \\ &= \int_0^1 p(y) \cdot \frac{\hat{y} - y}{\hat{y}(1-\hat{y})} \cdot dy \\ &= \frac{1}{\hat{y}(1-\hat{y})}(\hat{y} - E[y]).\end{aligned}$$

This partial derivative equals zero when $\hat{y} = E[y]$. Since the second partial derivative is positive at $\hat{y} = E[y]$, this extremum is the minimum. Thus, (c1) is satisfied and the cross entropy function is P-admissible.

V. Necessary and Sufficient Conditions for P-Admissibility

Define $h(\hat{y}) = L(0, \hat{y})$ to be the value of the loss function when the target output y is zero. In Appendix A, it is shown that P-admissibility is equivalent to the restrictions (r1) and (r2):

$$L(y, \hat{y}) = \int h'(\hat{y}) \cdot \frac{\hat{y} - y}{\hat{y}} \cdot d\hat{y} + C(y), \tag{r1}$$

$$h'(\hat{y}) > 0, \qquad \text{for } 0 < \hat{y} < 1. \tag{r2}$$

The prime indicates the derivative, $h'(\hat{y}) = d/d\hat{y}\, h(\hat{y})$. The value $C(y)$ is a constant with respect to $\hat{y}$. It has no effect on minimization and may be set to zero when defining a loss function. The restrictions (r1) and (r2) may be used to generate loss functions that minimize to the conditional expectation. Note that a P-admissible loss function may be described entirely by $h(\hat{y}) = L(0, \hat{y})$, the value of the loss function when the target $y = 0$. For example, the squared-error loss function has error $h(\hat{y}) = \hat{y}^2$ when the desired output is zero. Substitute this into (r1) with $C(y) = y^2$ to find $L(y, \hat{y}) = (y - \hat{y})^2$ as expected.

VI. A Symmetry Restriction

So far, we have used P-admissibility to restrict the class of loss functions under study. A second restriction on a loss function is the condition of logical symmetry:

$$L(0, \hat{y}) = L(1, 1 - \hat{y}). \tag{c2}$$

This condition is natural when the labels $y = 1$ and $y = 0$ are arbitrary. In the simple medical records example, we used $y = 1$ to indicate the presence of a disease. If instead we had used $y = 0$ to indicate the disease, would the results have been the same? The answer is true generally only if the loss function obeys the symmetry condition (c2).

Symmetry can be use to greatly simplify (r1). Define

$$k(\hat{y}) = \int \frac{h'(\hat{y})}{\hat{y}} \cdot d\hat{y}.$$

Then, (r1) may be written

$$L(y, \hat{y}) = h(\hat{y}) - y \cdot k(\hat{y}) + C(y). \tag{r3}$$

Thus, $L(1, 1 - \hat{y}) = h(1 - \hat{y}) - k(1 - \hat{y}) + C(1)$, and by definition $L(0, \hat{y}) = h(\hat{y})$. From (c2), find $k(\hat{y}) = h(\hat{y}) - h(1 - \hat{y}) - C(1)$. Finally, substitute back into (r3) to get the simple form:

$$L(y, \hat{y}) = h(\hat{y}) + y[h(1 - \hat{y}) - h(\hat{y})] + C_1(y).$$

We show in Appendix B that symmetry further restricts the form of a P-admissible loss function. Specifically $h(\hat{y})$ must satisfy (r4):

$$\frac{1 - \hat{y}}{\hat{y}} = \frac{h'(1 - \hat{y})}{h'(\hat{y})}. \tag{r4}$$

Hampshire and Pearlmutter [5] independently arrived at (r4) for the case where the targets are binary $\{0, 1\}$. In this correspondence, we show that this result applies to objective function analysis for more general distributions $p(y)$.

It follows from equation (r4) that at least one of the following cases must be true:

$h'(\hat{y})$ has a zero at $\hat{y} = 0$ or
$h'(\hat{y})$ has a pole at $\hat{y} = 1$.

Simple functions satisfying this restriction are:

$$h_1'(\hat{y}) = \hat{y},$$
$$h_2'(\hat{y}) = \frac{1}{1 - \hat{y}}.$$

By substitution into (r1), it is seen that h_1 defines the objective function:

$$L_1(y, \hat{y}) = 0.5\hat{y}^2 - \hat{y}y.$$

An objective function can be multiplied by a constant with respect to $\hat{y}$ or added to a constant w.r.t. $\hat{y}$ without changing the minimization, thus $L_1(y, \hat{y})$ is equivalent to $L_{se} = (\hat{y} - y)^2$. Similarly h_2 may be substituted into (r1) to generate the cross entropy objective function. Hence, by applying the result of Appendix B, we see that the well known loss functions L_{se} and L_{ce} are two of the most simple functions of a class that satisfy P-admissibility and the symmetry condition.

VII. Insufficiently Powerful Models

In Section II, we developed P-admissibility by considering that a sufficiently powerful model should produce conditional expected values. In practice, the model may not be able to produce these ideal outputs for every different input $\underline{x}$. What if the model is not sufficiently powerful?

Let $g(\underline{x}) = E[y \mid \underline{x}]$. We reintroduce the explicit reference to $\underline{x}$ in the notation. From Section II, we know $g(\underline{x})$ is the output of a sufficiently powerful model after training with a P-admissible loss function. Let $\hat{y}$ be the output of an "imperfect" model, one which may not be sufficiently powerful. In Appendix C, it is shown that if $L(y, \hat{y})$ is P-admissible then,

$$\min_{\underline{\theta}} E[L(y, \hat{y})] \text{ is equivalent to } \min_{\underline{\theta}} E[L(g(\underline{x}), \hat{y})]. \tag{r5}$$

Here, the expected value is taken over the joint distribution of $\underline{x}$, y in the training data. In words, (r5) says that the model which is found by minimizing the expected loss, produces outputs which come as close as possible to the output of an "ideal" model. Furthermore, this "closeness" is defined by the loss function. For instance if the loss function is squared error, then a real model trained to minimize the expected squared error between y and $\hat{y}$ will produce an output for which the expected square error between $\hat{y} = f(\underline{x})$ and $E[y \mid \underline{x}]$ is minimized. This was shown for the squared-error loss function by White [6]. Result (r5) generalizes this result to all possible loss functions that produce estimates of the conditional expectation.

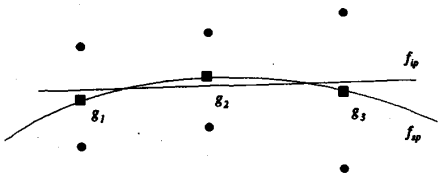

Fig. 2. Comparison of two models fitted to six points.

Fig. 2 shows an example graphically. Here we have 6 points of training data with three unique x. The conditional expected values of y for these three inputs are shown by g_1, g_2, g_3. The figure shows a curve f_{sp} from a sufficiently powerful model. A sufficiently powerful model in this case is any set of curves (parameterized by $\underline{\theta}$) that include a curve which passes through g_1, g_2, g_3. For example, second order polynomials are sufficiently powerful for this case since there are three points. If we set the $\underline{\theta}$ by minimizing the expected value of a P-admissible loss function, then the function $f_{sp}(x)$ will pass through these points. This is the result of Section II. Now suppose we fit these points with a line. A line is "insufficiently powerful" for this problem, because it does not have the 3 degrees of freedom necessary to pass through g_1, g_2, g_3. Result (r5) says that the line will come as close as possible to the three points. The measure of "closeness" is the same loss function which we used with the 6 training points. In other words, we would have found the same line if the training points had been g_1, g_2, g_3, rather than the 6 points. This would be false if we had selected $\underline{\theta}$ using a non-P-admissible loss functions. For instance, if the loss function were $L(y, \hat{y}) \equiv (y - \hat{y})^4$ then f_{sp} would not pass through the points g_1, g_2, g_3, and the line would not be a quartic error approximation of the outputs of the f_{sp} function.

VIII. Conclusion

We have generalized and extended previously known results on the topic of obtaining conditional estimates from a trained model. In particular, we derived necessary and sufficient conditions for an objective function to minimize to the expected value of the desired output y given an input $\underline{x}$. The objective function $L(y, \hat{y})$ was found to be uniquely specified by the function $L(0, \hat{y})$. This function $L(0, \hat{y})$ was found to satisfy further restrictions when a condition of logical symmetry is required. These restrictions and the relation between $L(y, \hat{y})$ and $L(0, \hat{y})$ define the class of all objective functions that minimize to the conditional expectation. This includes objective functions that minimize to a probability. Two of the simplest functions in this class were found to be the well-known squared error and cross entropy objective functions. When the model is incapable of mapping all inputs $\underline{x}$ to the ideal output $g(\underline{x}) = E[y \mid \underline{x}]$, it was found that after training to minimize $E[L(y, \hat{y})]$, the model minimizes the expected error in its approximation of $g(\underline{x})$ as measured by $L(g(\underline{x}), \hat{y})$.

IX. Appendix A
Conditions for Minimization to $E(y \mid \underline{x})$

Proof: (c1)$\Leftrightarrow$((r1)·(r2)), where (c1) (r1) and (r2) are defined as

$$\forall p(y) \quad \text{s.t.} \quad \int_0^1 p(y)\cdot dy = 1, \quad \overline{y} = \int_0^1 p(y)\cdot y \cdot dy, \quad \overline{y} \in (0, 1)$$

$$\min_{0<\hat{y}<1} \int_0^1 p(y)\cdot L(y, \hat{y})\cdot dy = \int_0^1 p(y)\cdot L(y, \overline{y})\cdot dy \qquad \text{(c1)}$$

$$L(y, \hat{y}) = \int h'(\hat{y})\cdot \frac{\hat{y}-y}{\hat{y}} \cdot d\hat{y} + C(y) \qquad \text{(r1)}$$

$$h'(\hat{y}) > 0, \qquad \text{for } 0 < \hat{y} < 1, \qquad \text{(r2)}$$

and where we assume $h'(\hat{y}) = d/d\hat{y}\, h(\hat{y})$ exists for $\hat{y} \in (0, 1)$.

First, it will be shown that ((r1)·(r2)) $\Rightarrow$ (c1). Consistent with the notation used in (c1), we will use an overbar to indicate the expected value of a function taken with respect to $p(y)$:

$$\overline{L}(\hat{y}) = \int_0^1 p(y) L(y, \hat{y})\cdot dy.$$

Substitute in (r1),

$$\overline{L}(\hat{y}) = \int_0^1 p(y)\left[\int h'(\hat{y})\frac{\hat{y}-y}{\hat{y}}\, d\hat{y} + C(y)\right] dy.$$

Take the derivative with respect to $\hat{y}$:

$$\frac{\partial}{\partial \hat{y}}\overline{L}(\hat{y}) = \int_0^1 p(y) h'(\hat{y})\left(\frac{\hat{y}-y}{\hat{y}}\right) dy. \qquad \text{(A.1)}$$

Here, we took the derivative under the integral. This equality holds if $h'(\hat{y})$ exists, as we have assumed for $\hat{y} \in (0, 1)$. A local minimum can occur for $x \in (0, 1)$, if and only if:

$$\frac{\partial}{\partial \hat{y}}\overline{L}(\hat{y}) = 0 \quad \text{and} \quad \frac{\partial^2}{\partial \hat{y}^2}\overline{L}(\hat{y}) > 0.$$

From (A.1):

$$\frac{\partial}{\partial \hat{y}}\overline{L}(\hat{y}) = h'(\hat{y})\int_0^1 p(y)\, dy - \frac{h'(\hat{y})}{\hat{y}}\int_0^1 y p(y)\, dy,$$
$$= h'(\hat{y})\left(1 - \frac{\overline{y}}{\hat{y}}\right). \qquad \text{(A.2)}$$

Given (r2) it is clear that $\partial/\partial\hat{y}\,\overline{L}(\hat{y}) = 0$ for $\hat{y} \in (0, 1)$, if and only if $\hat{y} = \overline{y}$. Taking derivatives again shows that this extremum represents a minimum:

$$\frac{\partial^2}{\partial \hat{y}^2}\overline{L}(\hat{y}) = h''(\hat{y}) - h''(\hat{y})\frac{\overline{y}}{\hat{y}} + \frac{\overline{y}}{\hat{y}^2} + h'(\hat{y})\frac{\overline{y}}{\hat{y}^2},$$
$$\frac{\partial^2}{\partial \hat{y}^2}\overline{L}(\overline{y}) = \frac{h'(\overline{y})}{\overline{y}} > 0.$$

Thus it has been shown ((r1)·(r2)) $\Rightarrow$ (c1).

It remains to be proven that (c1)$\Rightarrow$ ((r1)·(r2)). Condition (c1) certainly requires:

$$\int_0^1 p(y)\frac{\partial}{\partial \hat{y}} L(y, \hat{y})\, dy = 0 \qquad \text{at } \hat{y} = \overline{y}. \qquad \text{(A.3)}$$

Without loss of generality, as a change of notation, let

$$\frac{\partial}{\partial \hat{y}} L(y, \hat{y}) = G(y, \hat{y})\cdot(\hat{y} - y). \qquad \text{(A.4)}$$

Substituting this into (A.3):

$$\int_0^1 p(y) G(y, \hat{y})(\hat{y} - y)\, dy = 0 \quad \text{at } \hat{y} = \overline{y}. \qquad \text{(A.5)}$$

Now consider the distribution

$$p(y) = \rho\cdot\delta(y - y_1) + (1-\rho)\delta(y - y_2) \quad (0 \le y_1 < y_2 \le 1). \qquad \text{(A.6)}$$

The δ function is the unit impulse function. Condition (c1) requires that the minima be at $\overline{y}$ for all $p(y)$, therefore it must be true for the binary distribution given in (A.6). Notice $\overline{y} = \rho y_1 + (1-\rho)y_2$. Evaluating (A.5) with this distribution gives

$$\rho G(y_1, \hat{y})(\overline{y} - y_1) + (1-\rho)G(y_2, \hat{y})(\overline{y} - y_2) = 0 \quad \text{at } \hat{y} = \overline{y},$$

$$\rho G(y_1, \overline{y})[\rho y_1 + (1-\rho)y_2 - y_1]$$
$$+(1-\rho)G(y_2, \overline{y})[\rho y_1 + (1-\rho)y_2 - y_2] = 0,$$

$$\rho(1-\rho)(y_2-y_1)(G(y_1,\overline{y})-G(y_2,\overline{y}))=0.$$

Therefore, $G(y_1,\overline{y}) = G(y_2,\overline{y})$, where ρ can be chosen to set $\overline{y}$ arbitrarily in (y_1, y_2). Thus, $G(y_1,\hat{y}) = G(y_2,\hat{y}))$ for any $y_1 < \hat{y} < y_2$. Therefore, $G(y,\hat{y})$ is independent of y, so we may write

$$\frac{\partial}{\partial\hat{y}}L(y,\hat{y}) = G(\hat{y})\cdot(\hat{y}-y).$$

Evaluating at $y = 0$ and using the definition of $h(\hat{y})$ shows

$$G(\hat{y}) = \frac{h'(\hat{y})}{\hat{y}}.$$

Substituting this $G(\hat{y})$ back into (A.4) and then integrating both sides of the equation gives the desired result (r1). The result (c1)⇒(r2) follows easily from the requirement that the unique extremum found by (r1), be a minimum rather than a maximum. Since (c1) ⇒ ((r1) · (r2)) and ((r1) · (r2)) ⇒ (c1), it has been shown (c1) ⇔ ((r1) · (r2)). □

X. Appendix B
Restrictions on $h(\hat{y})$

Proof: ((c1) · (c2)) ⇒ (r4), where (c1) and (c2) are as given in Appendix A, and (r4) is

$$\frac{1-\hat{y}}{\hat{y}} = \frac{h'(1-\hat{y})}{h'(\hat{y})}. \tag{r4}$$

Given (c1), as in Appendix A, the equilibrium equation (A.3) must hold for distribution (A.6) with $y_1 = 0$ and $y_2 = 1$,

$$\rho\frac{\partial}{\partial\hat{y}}L(1,\hat{y}) + (1-\rho)\frac{\partial}{\partial\hat{y}}L(0,\hat{y}) = 0 \qquad \text{at } \hat{y}=\rho.$$

Using (c2):

$$-\rho\frac{\partial}{\partial\hat{y}}L(0,1-\hat{y}) + (1-\rho)\frac{\partial}{\partial\hat{y}}L(0,\hat{y}) = 0 \qquad \text{at } \hat{y}=\rho,$$

$$\frac{1-\rho}{\rho} = \frac{h'(1-\rho)}{h'(\rho)}.$$

Since ρ is arbitrary $\in (0, 1)$, (r4) is proven. □

XI. Appendix C
Minimization to a Probability with Insufficiently Powerful Models

Proof: (c1)⇒(r5). The condition for P-admissibility, (c1), is previously defined and (r5) is defined

$$\min_{\underline{\theta}} E[L(y,\hat{y})] \text{ is equivalent to } \min_{\underline{\theta}} E[L(g(\underline{x})),\hat{y}], \tag{r5}$$

where $g(\underline{x}) = \int p(y \mid \underline{x})\cdot y\,dy$ and $\hat{y} = f(\underline{x},\underline{\theta})$. Here the expected value symbol E refers to the expected value taken over a probability distribution on $\underline{x}$, y. So for any function $F(\underline{x}, y)$ the expected value $E[F(\underline{x}, y)]$ is shorthand for the equivalent integral forms

$$E[F(\underline{x},y)] = \int_{\underline{x}}\int_{y} p(\underline{x},y)F(\underline{x},y)\,dy\,d\underline{x}, \tag{A.7}$$

$$= \int_{\underline{x}} p(\underline{x})\left[\int_{y} p(y\mid\underline{x})F(\underline{x},y)\,dy\right]d\underline{x}. \tag{A.8}$$

We assume these integrals actually exist and that the distribution functions $p(\underline{x}, y)$, $p(\underline{x})$, and $p(y \mid \underline{x})$ all exist where they are evaluated within the integral. If $p(\underline{x}, y)$ is the probability of choosing a training pair $(\underline{x}, y)$ from a random sample from a finite training set then the integral corresponds to an equivalent finite sum. As shown in Appendix A, (c1)⇒(r1).

Let

$$k(\hat{y}) = \int\frac{h'(\hat{y})}{\hat{y}}\,d\hat{y}.$$

Then, (r1) may be written

$$L(y,\hat{y}) = h(\hat{y}) + C(y) - y\cdot k(\hat{y}). \tag{A.9}$$

Thus,

$$E[L(y,\hat{y})] = E[h(\hat{y}) + C(y) - y\cdot k(\hat{y})],$$
$$= E[h(\hat{y})] + E[C(y)] - E[y\cdot k(\hat{y})].$$

Using (A.8) with $F = y\cdot k(\hat{y})$:

$$E[L(y,\hat{y})] = E[h(\hat{y})] + E[C(y)] - \int_{\underline{x}} p(\underline{x})\left[\int_{y} p(y\mid\underline{x})y\cdot k(\hat{y})\,dy\right]d\underline{x}.$$

Note that given an input $\underline{x}$, the output $\hat{y}$ is constant. Thus, $k(\hat{y})$ may be factored out of the integral over y:

$$E[L(y,\hat{y})] = E[h(\hat{y})] + E[C(y)] - \int_{\underline{x}} p(\underline{x})\cdot k(\hat{y})\left[\int_{y} p(y\mid\underline{x})y\,dy\right]d\underline{x}.$$

Using the definition $g(\underline{x}) = \int p(y \mid \underline{x})\cdot y\,dy$,

$$E[L(y,\hat{y})] = E[h(\hat{y})] + E[C(y)] - \int_{\underline{x}} p(\underline{x})\cdot k(\hat{y})\cdot g(\underline{x})\,d\underline{x}.$$

Substitute the definition $p(\underline{x}) = \int_{y} p(\underline{x}, y)\,dy$,

$$E[L(y,\hat{y})] = E[h(\hat{y})] + E[C(y)] - \int_{\underline{x}}\left[\int_{y} p(\underline{x},y)'dy\right]\cdot k(\hat{y})\cdot g(\underline{x})\,d\underline{x}.$$

Now, use (A.7)

$$E[L(y,\hat{y})] = E[h(\hat{y})] + E[C(y)] - E[g(\underline{x})\cdot k(\hat{y})],$$
$$= E[h(\hat{y}) + g(\underline{x}) - g(\underline{x})\cdot k(\hat{y})] + E[C(y) - g(\underline{x})].$$

Use (A.9) once more:

$$E[L(y,\hat{y})] = E[L(g(\underline{x}),\hat{y})] + E[C(y) - g(\underline{x})]. \tag{A.10}$$

Since $E[C(y) - g(\underline{x})]$ is a constant with respect to $\underline{\theta}$, it follows from (A.10):

$$\min_{\underline{\theta}} E[L(y,\hat{y})] \text{ is equivalent to } \min_{\underline{\theta}} E[L(g(\underline{x}),\hat{y})].$$ □

References

[1] E. Baum and F. Wilczek, "Supervised learning of probability distributions by neural networks," in *Neural Information Processing Systems*, D. Anderson, Ed. New York: Amer. Inst. of Physics, 1988, pp. 52–61.

[2] A. El-Jaroudi and J. Makhoul, "A new error criterion for posterior probability estimation with neural nets," in *IEEE Proc. 1990 Int. Joint Conf. Neural Networks*, San Diego, CA, June 1990, pp. III-185–192.

[3] S. Solla, E. Levin, and M. Fleisher, "Accelerated learning in layered neural networks," *Complex Syst.*, vol. 2, no. 6, pp. 625–640, 1988.

[4] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *IEEE Proc. 1990 Conf. on Acoustics, Speech and Signal Processing*, Albuquerque, NM, 1990, pp. 1361–1364.

[5] J. Hampshire and B. Pearlmutter, "Equivalence proofs for multi-layer perceptron classifiers and the Bayesian discriminant function," in *Proceedings of the 1990 Connectionist Models Summer School*, D. Touretzky *et al.*, Eds. San Francisco, CA: Morgan Kaufmann, 1990, pp. 159–172.

[6] H. White, "Learning in artificial neural networks: A statistical perspective," *Neural Computation*, vol. 1, no. 4, pp. 425–464, 1990.