Building Cost Functions Minimizing to Some Summary Statistics

Marco Saerens, *Member, IEEE*

***Abstract*—A learning machine—or a model—is usually trained by minimizing a given criterion (the expectation of the cost function), measuring the discrepancy between the model output and the desired output. As is already well known, the choice of the cost function has a profound impact on the probabilistic interpretation of the output of the model, after training. In this work, we use the calculus of variations in order to tackle this problem. In particular, we derive necessary and sufficient conditions on the cost function ensuring that the output of the trained model approximates 1) the conditional expectation of the desired output given the explanatory variables; 2) the conditional median (and, more generally, the q-quantile); 3) the conditional geometric mean; and 4) the conditional variance. The same method could be applied to the estimation of other summary statistics as well. We also argue that the least absolute deviations criterion could, in some cases, act as an alternative to the ordinary least squares criterion for nonlinear regression. In the same vein, the concept of "regression quantile" is briefly discussed.**

***Index Terms*—Absolute deviations, a posteriori probabilities estimation, conditional expectation estimation, cost function, L_1 approximation, loss function, median, median estimation, penalty function, performance criterion, q-quantile, quasi-likelihood.**

I. Introduction

AN IMPORTANT problem concerns the probabilistic interpretation to be given to the output of a learning machine or, more generally, a model, after training. It appears that this probabilistic interpretation depends on the cost function used for training. Artificial neural networks are almost always trained by minimizing a given criterion—the expectation of the cost function. It is therefore of fundamental importance to have a precise idea of what can be achieved with the choice of this criterion.

Consequently, there has been considerable interest in analyzing the properties of the mean square error criterion—the most commonly used criterion. It is, for instance, well known that artificial neural nets (or more generally any model), when trained using the mean square error criterion, produce as output an approximation of the expected value of the desired output conditional on the input—the explanatory variables—if "perfect training" is achieved (see for instance [45]). We say that perfect training is achieved if 1) the global minimum of the criterion is indeed reached after training and 2) the neural network is a "sufficiently powerful model" that is able to approximate the optimal estimator to any degree of accuracy (perfect model matching property).

It has also been shown that other cost functions, for instance the cross-entropy between the desired output and the model output in the case of pattern classification, lead to the same property of approximating the conditional expectation of the desired output as well. We may therefore wonder what conditions a cost function should satisfy in order that the model output has this property. In 1991, following the results of Hampshire and Pearlmutter [22], Miller *et al*. [37], [38] answered to this question by providing conditions on the cost function ensuring that the output of the model approximates the conditional expectation of the desired output given the input, in the case of perfect training.

In this work, we extend these results to the conditional **median** (and, more generally, the **q-quantile**), as well as other summary statistics (the conditional geometric mean and variance), by applying the calculus of variations. This is the main original contribution of the paper.

Indeed, the variational formalism is a generic framework that can systematically be applied in order to derive results concerning the interpretation of the output of a trained model. For instance, we could further consider constrained distributions such as Gaussian disturbances, symmetric noise probability densities, etc. (as reviewed in [50]), and derive the associated necessary and sufficient conditions on the cost function by introducing additional constraints in the Lagrange function. Or, by using the same technique, we could search for the class of criteria that are minimized at some other summary statistics we are interested in—not only the one we study here. Another advantage of this approach is that it is constructive, that is, it provides rules for building cost functions with the desired properties.

Finally, we briefly discuss some motivations for the use of the Least Absolute Deviations (LADs) criterion minimizing the median and "**regression quantile**," especially for training radial basis function networks. In particular, and as is already well known, LAD estimation is less sensitive to the presence of outliers than the ordinary least square. On the other hand, as illustrated by Buchinsky [13], regression quantile can be used in order to provide a **confidence interval** on the y-prediction of the model. While there have been attempts to use cost functions related to the median in the neural-network community [14], [23], they are almost never considered as an alternative to the ordinary least square or the cross-entropy error function at this time.

Manuscript received April 15, 1999; revised May 30, 2000. This work was supported in part by the Project RBC-BR 216/4041 from the "Région de Bruxelles-Capitale," and funding from the SmalS-MvM. Preliminary results were published in the Proceedings of the 13th International Conference on Machine Learning, Bari, Italy, 1996.

The author is with IRIDIA Laboratory, Université Libre de Bruxelles, B-1050 Bruxelles, Belgium and also with SmalS-MvM, Research Section, B-1050 Bruxelles, Belgium (e-mail: saerens@ulb.ac.be).

Publisher Item Identifier S 1045-9227(00)09850-7.

A. Related Work

As mentioned in the introduction, it has been shown by several authors that, when training a model by minimizing a mean square error criterion—and assuming that this minimum is indeed attained after training—the output of the model provides an estimation of the conditional expectation of the desired output, given the input pattern (the explanatory variables), whatever the characteristics of the noise affecting the data (for the continuous case: [46], [60], [63]; for the binary case: [11], [20], [26], [48], and [56]; for a review, see [45]). This property of the mean square error criterion—rediscovered by the neural-network community—is in fact a result of mathematical statistics, and in particular estimation theory (see, for instance, [42, p. 175], or more generally [17], [27], [35], [36], [39], and [51]). A short overview of these related results from estimation theory is provided in [50].

Moreover, for binary desired outputs (that is, in the case of pattern classification in which the desired outputs represent the class to which the input pattern belongs to), Hampshire and Pearlmutter [22] extended this result to a larger class of cost functions. In particular, they provided conditions on the cost function used for training ensuring that the output of the model approximates the conditional probability of the desired output given the input (also called Bayesian *a posteriori* probabilities in the special case of pattern classification) when the performance criterion is minimized. Similar results were published by Savage [53] and Lindley [32] in the context of "subjective probabilities" theories (see also [16], [30]). In this framework, cost functions that minimize to the *a posteriori* probabilities are called "proper scoring rules." An extension to continuous distributions can be found in [33]. Asymmetric cost functions taking account of the fact that given positive errors may be more serious than negative errors are studied in [65].

Thereafter, Miller *et al.* [37], [38] extended these results to nonbinary desired outputs, that is, to continuous bounded outputs, for function approximation. Saerens [50], by using the calculus of variations, and Cid-Sueiro *et al.* [15] rederived these conditions for the general multioutputs case. The calculus of variations has already been proved useful in many circumstances in the context of artificial neural networks research. It has been used, for instance, for deriving learning algorithms for multilayer or recurrent neural networks [31], for regularization purposes (for a review, see [8]) or for training recurrent neural networks classifying temporal patterns [21].

In this paper, we extend these results to the conditional median (and, more generally, the q-quantile), as well as some other summary statistics.

B. Paper Outline

In the following sections, we first introduce the problem from an estimation theory perspective (Section II). In Section III, we rederive, from a calculus of variations point of view, the already known [37], [38] necessary and sufficient conditions on the cost function ensuring that the output of the trained model approximates the conditional expectation of the desired output. We also derive similar results for the conditional geometric mean and the conditional variance. Thereafter (Section IV), we extend the results to the ***q*-quantile** of the conditional distribution of the desired output, given the input. More precisely, among a class of "reasonable" performance criteria, we provide necessary and sufficient conditions on the cost function so that the optimal estimate is the q-quantile of the conditional probability density of the desired output given the input vector (the explanatory variables), whatever the noise characteristics affecting the data. In the next section (Section V), we argue that the least absolute deviations criterion (LAD) could, in some special cases, act as an alternative to the ordinary least squares criterion (OLS) for nonlinear regression. We also encourage the use of "regression quantile."

II. Statement of the Problem

Let us consider that we are given a sequence of independent m-dimensional training patterns $\mathbf{x}_k = [x_1(k), x_2(k), \ldots, x_m(k)]^T$ with $k = 1, 2, \ldots,$ as well as corresponding scalar desired outputs y_k (for the n-dimensional case, see [50]). The $\mathbf{x}_k$ and the y_k are realizations of the random variables $\mathbf{x}$ and y. Of course, we hope that the random vector $\mathbf{x}$ provides some useful information that allows to predict y with a certain accuracy on the basis of $\mathbf{x}$. The purpose is to train a model, say a neural network, in order to supply outputs, $\hat{y}_k$, that are "accurate" (in some predefined manner; see below) estimations—or predictions—of the desired outputs

$$\hat{y}_k = \mathcal{N}[\mathbf{x}_k; \mathbf{w}] \tag{1}$$

where $\mathcal{N}[.;.]$ is the function provided by the model, $\mathbf{x}_k$ the input (the explanatory variables) supplied as input to the model, and $\mathbf{w}$ is the parameter vector of the model. In order to measure how "accurate" is the estimation (1), we define a **cost function**—or loss function, penalty function, objective function, empirical risk measure, scoring rule—that provides us a measure of the discrepancy between the predicted value $\hat{y}_k$ (supplied by the model) and the desired value y_k: $\mathcal{L}[\hat{y}_k; y_k]$. The purpose of the training is, of course, to minimize this cost.

Now, since it is not generally possible to minimize the cost function for each k because of the presence of noise or disturbances [for a given value of the input $\mathbf{x}$, the desired output is distributed with a probability density function $p(y|\mathbf{x})$], the best we can do is to minimize this cost "on average." This leads to the definition of the **performance criterion** $\mathcal{C}[\hat{y}]$

$$\mathcal{C}[\hat{y}] = \iint \mathcal{L}[\hat{y}; y] p(\mathbf{x}, y)\, d\mathbf{x}\, dy = E\{\mathcal{L}[\hat{y}; y]\} \tag{2}$$

where the integral is defined on the Euclidean space $\mathbf{R}^m \times \mathbf{R}^1$. $E\{.\}$ is defined as the standard expectation. Notice that we implicitly make the assumption that the criterion is additive with respect to the cost function here. If the criterion is multiplicative (as in the case of a likelihood function), we would have to define $\log(\mathcal{C}[\hat{y}])$, which is again additive.

It is convenient to rewrite (2)

$$\mathcal{C}[\hat{y}] = \int \left\{ \int \mathcal{L}[\hat{y}; y] p(y|\mathbf{x})\, dy \right\} p(\mathbf{x})\, d\mathbf{x}. \tag{3}$$

Now, if we minimize the inner integral of (3) for every possible value of $\mathbf{x}$, then $\mathcal{C}[\hat{y}]$ will also be minimized, since $p(\mathbf{x})$ is nonnegative. We therefore select $\hat{y}$ in order to minimize

$$\mathcal{C}[\hat{y}|\mathbf{x}] = \int \mathcal{L}[\hat{y};\, y]p(y|\mathbf{x})\,dy = E\{\mathcal{L}[\hat{y};\, y]|\mathbf{x}\} \tag{4}$$

where $\mathcal{C}[\hat{y}|\mathbf{x}]$ is a function of both $\hat{y}$ and $\mathbf{x}$, and $E\{.|\mathbf{x}\}$ is the conditional expectation, given $\mathbf{x}$. This means that the minimization of (4) can be performed independently for every $\mathbf{x}$. Moreover, since $\hat{y}$ is chosen in order to minimize (4) for every value of $\mathbf{x}$, $\hat{y}$ will in general be a function of this parameter $\mathbf{x}$. The function of $\mathbf{x}$ that minimizes (4) with respect to $\hat{y}$ will be called the **best**, or **optimal**, estimator.

We assume that this optimal estimator can be approximated to any degree of accuracy by the model, $\hat{y} = \mathcal{N}[\mathbf{x};\, \mathbf{w}]$, for some optimal value of the parameters $\mathbf{w} = \mathbf{w}^*$ (perfect parameters tuning). In other words, we are making a "perfect model matching" assumption. In the Miller *et al.* terminology [37], [38], such a model is called a "sufficiently powerful model" that is able to produce the optimal estimator.

In the next sections, we derive necessary and sufficient conditions on the cost function ensuring that the output of the trained model approximates the conditional expectation, the geometric mean, the variance and the q-quantile of the desired output. Notice that we are designing cost functions for estimation of specific summary statistics without reference to 1) the statistical distribution of the noise and 2) the architecture of the model.

The reason for 1) is the following: We want to build cost functions minimizing to some summary statistics in any noise condition, that is, **whatever the distribution of the noise**. This means that the model, if perfectly trained, will approximate the summary statistics in any noise condition defined by $p(y|\mathbf{x})$ (Poisson density, Gaussian density, etc). As an example, it is well known (see, for instance, [3], [17], [42], and [57]) that if the criterion is the **mean square error**, that is, when the cost function is $\mathcal{L}[\hat{y};\, y] = (\hat{y} - y)^2$, the minimum of the criterion (4) is reached for

$$\hat{y}(\mathbf{x}) = \int yp(y|\mathbf{x})\,dy = E\{y|\mathbf{x}\}. \tag{5}$$

And this result remains true whatever the characteristics of the noise affecting the data, and represented by the probability density function $p(y|\mathbf{x})$.

The reason for (2) is the fact that neural networks are powerful nonlinear function approximators. They have universal approximation properties, and have been proved very useful in various engineering areas when carefully designed and trained. This is to be compared to, for instance, the generalized linear model approach, also very powerful and widely used in applied statistics (see, for instance, [34] or [24]). In this later methodology, the architecture is carefully chosen based on the problem at hand. First, a (possibly nonlinear) link function relating a linear transformation of the inputs to the expected output is introduced. For instance, in the case of pattern classification, the link function could be a logistic or a probit function. Then, the cost function (a likelihood or quasi-likelihood function) is designed according to the distribution of the output at hand: After having observed the output distribution and made some assumptions about the noise, the model is usually trained by maximum likelihood or quasi-likelihood. One could argue that it is useless to discuss creative cost functions that can estimate some summary statistics without discussing the relevant underlying distributions. We think that both approaches are stimulating and complementary. But, of course, if we know the real distribution of the noise, maximum likelihood estimation is usually more efficient.

III. Searching for the Class of Cost Functions $\mathcal{L}[y_1;\, y_2]$ Such That the Criterion $\mathcal{C}[\hat{\mathbf{y}}|\mathbf{x}]$ is Minimized for $\hat{\mathbf{y}} = e\{y|\mathbf{x}\}$

In this section, we derive conditions on the cost function $\mathcal{L}$ ensuring that the criterion (2)—or equivalently (4)—attains its minimum at the conditional expectation $\hat{y} = E\{y|\mathbf{x}\}$. We will consider the class of cost functions $\mathcal{L}[.;.]$ of the type

$$\mathcal{L}[y_1;\, y_2] = 0 \quad \text{if and only if } y_1 = y_2 \tag{6a}$$

$$\mathcal{L}[y_1;\, y_2] \geq 0 \tag{6b}$$

$$\mathcal{L}[y_1;\, y_2] \quad \text{is twice continuously differentiable in terms of all its arguments.} \tag{6c}$$

We also assume that $\mathcal{L}[\hat{y};\ y]$ depends on $\mathbf{x}$ only through the variable $\hat{y}$. If we want $\mathcal{C}[\hat{y}|\mathbf{x}]$ to be a minimum for $\hat{y} = E\{y|\mathbf{x}\} = \mu(\mathbf{x})$, we have

$$\mu(\mathbf{x}) = \int yp(y|\mathbf{x})\,dy = E\{y|\mathbf{x}\} \quad (\mu \text{ is the conditional expectation}) \tag{7a}$$

and the following standard optimality conditions must hold:

$$\left.\frac{\partial \mathcal{C}[\hat{y}|\mathbf{x}]}{\partial \hat{y}}\right|_{\hat{y}(\mathbf{x})=\mu(x)} = \int \left.\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}}\right|_{\hat{y}(\mathbf{x})=\mu(\mathbf{x})} p(y|\mathbf{x})\,dy = 0 \tag{7b}$$

$$\left.\frac{\partial^2 \mathcal{C}[\hat{y}|\mathbf{x}]}{\partial \hat{y}^2}\right|_{\hat{y}(\mathbf{x})=\mu(\mathbf{x})} = \int \left.\frac{\partial^2 \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}^2}\right|_{\hat{y}(\mathbf{x})=\mu(\mathbf{x})} p(y|\mathbf{x})\,dy > 0. \tag{7c}$$

These conditions must hold whatever the characteristics of the noise.

Before discussing the proof, let us provide the condition on the cost function for which the performance criterion is a minimum for $\hat{y} = \mu(\mathbf{x})$ (the conditional expectation)

$$\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}} = a[\hat{y}](y - \hat{y}) \text{ with } a[\hat{y}] < 0 \tag{8}$$

where a is an arbitrary function of $\hat{y}$ only (a does not depend on y) and is negative on the range of the possible values of $\hat{y}$.

The proof of (8) is presented in Appendix A. In this appendix, we show that the condition (8) on the cost function, and the optimality conditions (7a)–(7c) are equivalent. We therefore proceed in two steps: we first prove that (8) on the cost function implies that the criterion is a minimum at the conditional expectation [optimality conditions (7a)–(7c)] (sufficient condition; Appendix A-A). Thereafter we show that if the criterion is a minimum at the conditional expectation [optimality conditions (7a)–(7c)], the cost function verifies the condition (8) (necessary condition; Appendix A-B).

In [50], we also showed that if the cost function is a function of the difference between desired output and predicted output, the mean square error criterion is the only one that leads to the estimation of the conditional expectation of the output given the input pattern, and we extended these conditions to the more general multioutputs case.

Condition (8) allows us to verify that the optimal estimator is the conditional expectation. Some examples of cost functions verifying this condition are provided in [50]. On the other hand, (8) allows us to construct cost functions with the desired property by integration; the integration constant term is determined by imposing $\mathcal{L} = 0$ for $\hat{y} = y$ [condition (6a)].

Interestingly and surprisingly enough, there is a very close relationship between (8) and the **quasi-likelihood** functions[1] in the context of generalized linear models (see [62] or [34, p. 325]). While derived in a totally other way, log quasi-likelihood functions have exactly the same form as (8). In this framework, in (8), $V[\hat{y}] = 1/a[\hat{y}]$ is interpreted as a variance varying in function of the output $\hat{y}$. Many interesting properties of quasi-likelihood functions are reported in [34, chs. 9 and 10]. For instance, quasi-likelihood functions allow the modeling of nonconstant-variance processes, for which the variance changes with the value of the output—approximating the conditional expectation. If $V[\hat{y}] = \sigma^2 =$ constant, the variance is supposed to remain constant with respect to $\hat{y}$. This interpretation provides a motivation for using cost functions derived from (8). There are also close links between the exponential family of distributions and quasi-likelihood; the interested reader is referred to [62].

Notice that the calculus of variations approach can easily be extended to constrained problems and to other summary statistics. For instance, by introducing Lagrange parameters, we can impose that the probability density function of the desired output $p(y|\mathbf{x})$ is symmetric around the mean $\mu(\mathbf{x})$. In this case, we obtain that the cost function has to be convex. Examples of necessary and sufficient conditions for two other summary statistics, the geometric mean and the variance, are presented in Appendix B. The case of the median is somewhat more complex, and is developed in the next section.

IV. Searching for the Class of Cost Functions $\mathcal{L}[y_1; y_2]$ such that the Criterion $\rfloor[\hat{y}|\mathbf{x}]$ is Minimized for the Q-Quantile of $p(y|\mathbf{x})$

In this section, we derive conditions on the cost function $\mathcal{L}$ ensuring that the criterion (2)—or equivalently (4)—attains its minimum for $\hat{y}(\mathbf{x})$ being the q-quantile of the conditional probability density $p(y|\mathbf{x})$. The median is a special case of q-quantile where $q = 0.5$. For this problem, we will consider the class of cost functions $\mathcal{L}[.;.]$ of the type

$$\mathcal{L}[y_1; y_2] = 0 \quad \text{if and only if } y_1 = y_2 \tag{9a}$$

$$\mathcal{L}[y_1; y_2] \geq 0 \tag{9b}$$

$$\mathcal{L}[y_1; y_2] \quad \text{is twice continuously differentiable in terms of all its arguments, except at } y_1 = y_2 \text{ where it is simply continuous.} \tag{9c}$$

[1]This interesting fact was pointed out by an anonymous reviewer.

As in previous section, we assume that the cost function $\mathcal{L}$ depends on $\mathbf{x}$ only through the variable $\hat{y}$. Notice that the class of functions defined by (9a)–(9c) is more general than the one defined in Section III [(6a)–(6c)]. We do not require here that the cost function is differentiable at $y_1 = y_2$. Indeed, as will be seen later, the class of cost functions minimizing to the q-quantile will be nondifferentiable at $y_1 = y_2$. This case therefore requires a more careful mathematical treatment.

If we want $\mathcal{C}[\hat{y}|\mathbf{x}]$ to be a minimum at $\hat{y} = q\text{-quantile}[p(y|\mathbf{x})] = \upsilon(\mathbf{x})$, defined as

$$\int_{-\infty}^{\upsilon(\mathbf{x})} p(y|\mathbf{x})\, dy = q$$

and

$$\int_{\upsilon(\mathbf{x})}^{+\infty} p(y|\mathbf{x})\, dy = (1-q) \qquad (\upsilon \text{ is the } q\text{-quantile}) \tag{10a}$$

with $0 < q < 1$, the following standard optimality conditions must hold:

$$\left.\frac{\partial \mathcal{C}[\hat{y}|\mathbf{x}]}{\partial \hat{y}}\right|_{\hat{y}(\mathbf{x})=\upsilon(\mathbf{x})} = \frac{\partial}{\partial \hat{y}} \left\{ \int_{-\infty}^{\hat{y}} \mathcal{L}[\hat{y}; y] p(y|\mathbf{x})\, dy + \int_{\hat{y}}^{+\infty} \mathcal{L}[\hat{y}; y] p(y|\mathbf{x})\, dy \right\} \Bigg|_{\hat{y}(\mathbf{x})=\upsilon(\mathbf{x})} = 0 \tag{10b}$$

$$\left.\frac{\partial^2 \mathcal{C}[\hat{y}|\mathbf{x}]}{\partial \hat{y}^2}\right|_{\hat{y}(\mathbf{x})=\upsilon(\mathbf{x})} > 0. \tag{10c}$$

And these conditions (10a)–(10c) must hold whatever the characteristics of the noise.

Before discussing the proof, let us provide the conditions on the cost function for which the performance criterion is a minimum for $\hat{y} = \upsilon(\mathbf{x})$ [the q-quantile of $p(y|\mathbf{x})$]

$$\mathcal{L}[\hat{y}; y] = (1-q)[\Lambda(\hat{y}) - \Lambda(y)] \qquad \text{for } \hat{y} > y \tag{11a}$$

$$\mathcal{L}[\hat{y}; y] = q[\Lambda(y) - \Lambda(\hat{y})] \qquad \text{for } y > \hat{y} \tag{11b}$$

with $0 < q < 1$ and $\Lambda(.)$ being a strictly increasing—but otherwise arbitrary—differentiable function

$$\frac{\partial \Lambda(y)}{y} > 0. \tag{11c}$$

The proof of conditions (11a)–(11c) is presented in Appendix C. In this Appendix, we show that the conditions (11a)–(11c) on the cost function and the optimality conditions (10a)–(10c) are equivalent. We therefore proceed in two steps: we first prove that the conditions (11a)–(11c) on the cost function imply that the criterion is a minimum at the q-quantile of the conditional distribution (10b)–(10c) (sufficient conditions; Appendix C-B. Thereafter we show that if the criterion is a minimum at the q-quantile of the conditional distribution [optimality conditions (10a)–(10c)], the cost function verifies (11a)–(11c) (necessary conditions; Appendix C-A).

V. Least Absolute Deviations: An Alternative to Ordinary Least Squares?

Before using some nonusual cost function (for a review of various distance measure, see [7]), practitioners must first assess its properties in order to be able to give a clear probabilistic interpretation to the output of the model, after training. This is especially important if the model plays a crucial probabilistic role, as in the case of multilayer neural networks used in the context of hybrid models for speech recognition (hidden Markov models + neural networks) (see, for instance, [10]). Besides the probabilistic interpretation, other important issues for choosing a cost function include the robustness of the sample estimator against outliers [47], the statistical properties of the resulting finite-sample statistical estimator—in particular its variance (see for instance [57]), the shape of the disturbance's probability density [14], and the ease of computing the optimum of the criterion (is the criterion convex?). In this section, we suggest that the least absolute deviations criterion has some interesting properties that make it suitable for the training of neural networks, especially radial basis function neural networks.

We observe that the usual mean absolute deviations criterion (also called "absolute error cost function," or "L_1-norm"), for which $\mathcal{L} = |y - \hat{y}|$, verifies the conditions (11a)–(11c), and, indeed, provides the median as an estimate when minimized (see [36, p. 191] or [27, p. 343]). The finite-sample counterpart of the mean absolute deviations criterion, based on N realizations of the random variables $(\mathbf{x}, y)$, is called the LAD empirical criterion

$$\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|. \tag{12}$$

It is known for a long time that, in the case of multiple regression modeling, LAD estimation is more robust against outliers than the OLS (the finite-sample counterpart of the mean square error criterion); for a review, see [2]; for a discussion in the neural-network context, see [8]. In the univariate case, the sample median is known to be a robust estimator of the central location of a distribution [47], [25]; in contrast, the sample mean is very sensitive to outliers. In the multiple regression case, while the sample median is known to be more robust than the sample mean, there exists even more robust central location estimators which, unfortunately, are very computer-intensive to estimate [47]. In the context of using the LAD criterion for multiple regression parameters estimation, Wilson [64] and Barrodale [4] showed via Monte Carlo studies that the LAD is more efficient that the OLS in the presence of large disturbances (outliers). Blattberg and Sargent [9] showed that LAD criterion is superior to OLS in the case of fat-tailed (non-Gaussian) distributions of disturbances. Pfaffenberger and Dinkel [43] reported the same conclusions for Cauchy distributions of noise.

A drawback of the LAD comes from the fact that it is nondifferentiable at the y_i. However, it appeared that the linear regression parameters can be efficiently estimated by minimizing the LAD through linear programming [18], [58], [59]; for a review, see [2]. Optimized linear programming techniques have been especially designed for this problem, and the convergence to the global minimum is guaranteed ([5]; the Fortran code of the algorithm is published in [6]). This is also true for regression quantile [29]. An alternative to linear programming is to use iteratively reweighted least squares techniques (see [54], or for an improved algorithm, [1]), Karmakar's algorithm [49] or subgradient iterative techniques [44]. Online algorithms have also been designed [61]. In the case of nonlinear optimization, iterative techniques based on subgradient optimization can be used [40], [41].

The main reason why LAD is hardly used is the lack of large-sample properties, allowing, among others, the computation of confidence intervals for the estimated regression coefficients. This is, however, less critical in the case of connection weights estimation.

In the case of "**regression quantile**," we can similarly use the following empirical criterion:

$$\frac{1}{N}\sum_{i=1}^{N}[\delta(\hat{y}_i > y_i)(1-q)(\hat{y}_i - y_i) + \delta(y_i > \hat{y}_i)q(y_i - \hat{y}_i)] \quad (\text{with } 0 < q < 1)$$

where $\delta(\hat{y}_i > y_i) = 1$ if $\hat{y}_i > y_i$, and $= 0$ otherwise. This empirical criterion has been in use in applied statistics and econometrics for a long time (see [13] or [28] for a concrete application). Notice that regression quantile can be used in order to compute a kind of **confidence interval** on the y-forecast containing, say, 80% of the data (consisting in two curves, one at $q = 10\%$, and one at $q = 90\%$).

In consequence, LAD estimation could be a viable alternative to OLS estimation for radial basis function network parameters estimation, in the context of nonlinear regression. In this special case, the output is linear in terms of the parameters, and linear programming can be used for parameters estimation. It would be worth trying if we suspect that the noise distribution is non-Gaussian, in the presence of outliers and in the case of small training sets. As already mentioned, regression quantile can also be used in order to compute a confidence interval on the output y.

VI. Conclusion

In this paper, we used the calculus of variations in order to study the probabilistic interpretation of the output of a model trained by minimizing a cost function. In particular, we were able to provide simple conditions on the cost function ensuring that the optimal estimator is the conditional expectation of the desired output, and the q-quantile of the conditional probability density of the desired output. We argue that the variational formalism is a general framework that can systematically be applied in order to derive results concerning the interpretation of the output of a trained model. The same technique could be used in order to define the class of criteria that are minimized at some other summary statistics we are interested in (not only the conditional expectation or the q-quantile).

However, we must keep in mind that these conditions are only valid if 1) the global minimum of the criterion is indeed reached after training, and 2) the neural network is a "sufficiently powerful model" that is able to approximate the optimal estimator to any degree of accuracy (perfect model matching). Notice also

that 3) the results presented here are essentially asymptotic, and that issues regarding estimation from finite data sets are not addressed.

We must also stress that if we put some restrictions on the class of noise affecting the data (for instance, if we consider Gaussian disturbances, symmetric or convex probability distributions), a larger class of cost functions will also provide the conditional expectation as an optimal estimator (see [12], [17], and [55]; a short review is provided in [50]). In this work, we assumed that the noise is **completely arbitrary** (arbitrary conditional probability distribution).

We also argued that the least absolute deviations criterion can be used as an alternative to the ordinary least squares criterion, in the context of nonlinear regression. Moreover, regression quantile can be used in order to provide a confidence interval on the output of the model.

APPENDIX A

A. Condition (8) on the Cost Function Implies that the Performance Criterion is a Minimum for $\hat{y} = \mu(\mathbf{x})$ (The Conditional Expectation), i.e., Implies (7a)–(7c)

Let us multiply equation (8) by $p(y|\mathbf{x})$ and integrate over dy. We obtain

$$\int \frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}} p(y|\mathbf{x})\, dy = \int \{a[\hat{y}](y-\hat{y})\} p(y|\mathbf{x})\, dy = a[\hat{y}] \left\{ \int (y-\hat{y}) p(y|\mathbf{x})\, dy \right\}. \quad \text{(A1)}$$

Now, by the definition of the conditional expectation, the following identity holds:

$$\int (y - E\{y|\mathbf{x}\}) p(y|\mathbf{x})\, dy = \int (y-\mu(\mathbf{x})) p(y|\mathbf{x})\, dy = 0 \quad \text{(A2)}$$

so that, for $\hat{y} = \mu(\mathbf{x})$, the right-hand side of (A1) is zero. We therefore obtain (7b). This proves that all the cost functions verifying (8) will also verify the condition (7b).

Moreover, by differentiating (8), we evaluate (7c)

$$E\left\{ \left. \frac{\partial^2 \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}^2} \right|_{\hat{y}(\mathbf{x})=\mu(\mathbf{x})} |\mathbf{x} \right\} = E\left\{ \left. \left[\frac{\partial a[\hat{y}]}{\partial \hat{y}} (y-\hat{y}) - a[\hat{y}] \right] \right|_{\hat{y}(\mathbf{x})=\mu(\mathbf{x})} |\mathbf{x} \right\} = -a[\mu]. \quad \text{(A3)}$$

Since, from (8), $a[\hat{y}]$ is **negative** on the range of possible values of the conditional expectation μ, we obtain (7c). This completes the first part of the proof: the condition (8) implies the optimality conditions (7a)–(7c) ■

B. All the Solutions to Equations (7a)–(7c) are also Solutions to Equation (8) with $a[\hat{y}]$ Negative

Equation (7b) must hold whatever the characteristics of the noise, that is, whatever the probability density $p(y|\mathbf{x})$ of the random variable y with mean $\mu(\mathbf{x})$. This means that the integral

$$\int \left. \frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}} \right|_{\hat{y}(\mathbf{x})=\mu(\mathbf{x})} p(y|\mathbf{x})\, dy \quad \text{(A4a)}$$

remains stationary—it remains identically zero—for any variation (in the sense of the calculus of variations; see [19]) of the function $p(y|\mathbf{x})$—denoted as $\delta p(y|\mathbf{x})$, $\mathbf{x}$ being fixed—subject to the constraints

$$\int y p(y|\mathbf{x})\, dy = \mu(\mathbf{x}) \quad \text{(A4b)}$$

$$\int p(y|\mathbf{x})\, dy = 1. \quad \text{(A4c)}$$

Roughly speaking, it means that the result of the integral (A4a) is **invariant** when making a transformation $p'(y|\mathbf{x}) = p(y|\mathbf{x}) + \delta p(y|\mathbf{x})$, where $p'(y|\mathbf{x})$ and $p(y|\mathbf{x})$ are subject to the constraints (A4b)–(A4c). The constraints are important since if they were not present, we could directly deduce from (A4a) that

$$\left. \frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}} \right|_{\hat{y}(\mathbf{x})=\mu(\mathbf{x})} = 0$$

which, of course, is incorrect.

Now, this is a classical problem of the calculus of variations (often called the isoperimetric problem): the stationarity of (A4a) with respect to variations of $p(y|\mathbf{x})$, subject to the constraints (A4b) and (A4c), directly implies that the following functional (called the Lagrange function) is stationary for **any** variation of $p(y|\mathbf{x})$, without considering the constraints

$$L[p(y|\mathbf{x})] = \int \left. \frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}} \right|_{\hat{y}(\mathbf{x})=\mu(\mathbf{x})} p(y|\mathbf{x})\, dy + \lambda(\mathbf{x}) \left[\int y p(y|\mathbf{x})\, dy - \mu(\mathbf{x}) \right] + \rho(\mathbf{x}) \left[\int p(y|\mathbf{x})\, dy - 1 \right]. \quad \text{(A5)}$$

In other words, the method of Lagrange multipliers transforms the constrained problem into an unconstrained problem [19]. The λ and ρ are called Lagrange multipliers; since the stationarity property must hold for every $\mathbf{x}$, they are labeled by $\mathbf{x}$.

Let us compute the variation of L with respect to $p(y|\mathbf{x})$, and set it equal to zero

$$\delta L[p(y|\mathbf{x})] = \int \left\{ \left. \frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}} \right|_{\hat{y}(\mathbf{x})=\mu(\mathbf{x})} + \lambda(\mathbf{x}) y + \rho(\mathbf{x}) \right\} \cdot \delta p(y|\mathbf{x})\, dy = 0. \quad \text{(A6)}$$

Since the result of the integral in (A6) is zero for **any variation** $\delta p(y|\mathbf{x})$, the term into bracket must cancel:

$$\left. \frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}} \right|_{\hat{y}(\mathbf{x})=\mu(\mathbf{x})} + \lambda(\mathbf{x}) y + \rho(\mathbf{x}) = 0. \quad \text{(A7)}$$

Let us multiply equation (A7) by $p(y|\mathbf{x})$ and integrate over dy. From (7b), we obtain

$$\lambda(\mathbf{x}) \int y p(y|\mathbf{x})\, dy = -\rho(\mathbf{x}) \quad \text{(A8)}$$

so that

$$\rho(\mathbf{x}) = -\lambda(\mathbf{x})\mu(\mathbf{x}). \tag{A9}$$

And (A7) can be rewritten as

$$\left.\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}}\right|_{\hat{y}(\mathbf{x})=\mu(\mathbf{x})} = -\lambda(\mathbf{x})[y-\mu(\mathbf{x})]. \tag{A10}$$

Since we assumed that $\mathcal{L}[\hat{y};\, y]$ depends on $\mathbf{x}$ only through the variable $\hat{y}$, we obtain

$$\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}} = -\Lambda(\hat{y})[y-\hat{y}] \tag{A11}$$

which is equivalent to (8) with $a[\hat{y}] = -\Lambda(\hat{y}(\mathbf{x})) = -\lambda(\mathbf{x})$.

Moreover, we showed in previous section that (7c) is equivalent to the negativity of $a[\hat{y}]$ [see (A3)]. Therefore, the optimality conditions (7a)–(7c) imply (8) with $a[\hat{y}]$ negative. This completes the proof. ■

APPENDIX B

By following the same procedure as in Section III, we can obtain results concerning other summary statistics. In this appendix, we consider the conditional geometric mean and the conditional variance as examples.

A. The Geometric Mean

The conditional geometric mean, γ, can be defined as

$$\log(\gamma(\mathbf{x})) = \int \log(y)p(y|\mathbf{x})\,dy = E\{\log(y)|\mathbf{x}\} \tag{B1}$$

where we assume that $y > 0$. By using constraint (B1) instead of (A4b), we obtain the necessary and sufficient condition for which the performance criterion is a minimum for $\hat{y} = \gamma(\mathbf{x})$ (the conditional geometric mean)

$$\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}} = a[\hat{y}]\log\left(\frac{y}{\hat{y}}\right) \qquad \text{with } a[\hat{y}] < 0 \tag{B2}$$

where $\hat{y} > 0$, a is an arbitrary function of $\hat{y}$ only (a does not depend on y), and is negative on the range of the possible values of $\hat{y}$.

B. The Variance

The conditional variance σ is defined as

$$\sigma(\mathbf{x}) = \tfrac{1}{2}\int (y-\mu(\mathbf{x}))^2 p(y|\mathbf{x})\,dy = E\left\{\tfrac{1}{2}(y-\mu(\mathbf{x}))^2|\mathbf{x}\right\} \tag{B3}$$

where we assume that the conditional expectation $\mu(\mathbf{x})$ is known and provided by another previously trained model. By using constraint (B3) instead of (A4b), we obtain the necessary and sufficient conditions for which the performance criterion is a minimum for $\hat{y} = \sigma(\mathbf{x})$ (the conditional variance):

$$\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}} = a[\hat{y},\, \mu]\left[\frac{(y-\mu(\mathbf{x}))^2}{2} - \hat{y}\right] \qquad \text{with } a[\hat{y},\, \mu] < 0 \tag{B4}$$

where a is an arbitrary function of $\hat{y}$ and μ, and is negative on the range of the possible values of $\hat{y}$.

APPENDIX C

A. Conditions (11a)–(11c) on the Cost Function Imply that the Performance Criterion is a Minimum for $\hat{y} = \upsilon(\mathbf{x})$ (The q-Quantile), i.e., Imply (10b)–(10c)

Let us multiply $\mathcal{L}[\hat{y};\, y]$ [given by (11a)–(11c)] by $p(y|\mathbf{x})$ and integrate over dy in order to obtain the performance criterion. We obtain

$$\begin{aligned}\mathcal{C}[\hat{y}|\mathbf{x}] &= \int \mathcal{L}[\hat{y};\, y]p(y|\mathbf{x})\,dy \\ &= \int_{-\infty}^{\hat{y}} (1-q)[\Lambda(\hat{y})-\Lambda(y)]p(y|\mathbf{x})\,dy \\ &\quad - \int_{\hat{y}}^{+\infty} q[\Lambda(\hat{y})-\Lambda(y)]p(y|\mathbf{x})\,dy.\end{aligned}$$

Now, let us differentiate this expression by Leibnitz's rule. If we pose $\lambda(\hat{y}) = \partial\Lambda(\hat{y})/\partial\hat{y}$, we obtain

$$\begin{aligned}\frac{\partial \mathcal{C}[\hat{y}|\mathbf{x}]}{\partial \hat{y}} &= \int_{-\infty}^{\hat{y}} (1-q)\lambda(\hat{y})p(y|\mathbf{x})\,dy \\ &\quad \int_{\hat{y}}^{+\infty} -q\lambda(\hat{y})p(y|\mathbf{x})\,dy \\ &= \lambda(\hat{y})\left[(1-q)\int_{-\infty}^{\hat{y}} p(y|\mathbf{x})\,dy - q\int_{\hat{y}}^{+\infty} p(y|\mathbf{x})dy\right] \\ &= \lambda(\hat{y})\left[\int_{-\infty}^{\hat{y}} p(y|\mathbf{x})\,dy - q\right]\end{aligned} \tag{C1}$$

with $\lambda(\hat{y}) > 0$, from (11c). Now, for $\hat{y} = \upsilon$ (the q-quantile)

$$\int_{-\infty}^{\hat{y}} p(y|\mathbf{x})\,dy = q$$

$$\int_{\hat{y}}^{+\infty} p(y|\mathbf{x})\,dy = (1-q).$$

Which, from (C1), implies that $\mathcal{C}[\hat{y}|\mathbf{x}]$ is stationary for $\hat{y}$ equal to the q-quantile υ

$$\left.\frac{\partial \mathcal{C}[\hat{y}|\mathbf{x}]}{\partial \hat{y}}\right|_{\hat{y}(\mathbf{x})=\upsilon(\mathbf{x})} = 0.$$

This proves condition (10b).

Now, let us compute the second-order derivative of $\mathcal{C}[\hat{y}|\mathbf{x}]$

$$\frac{\partial^2 \mathcal{C}[\hat{y}|\mathbf{x}]}{\partial \hat{y}^2} = p(\hat{y}|\mathbf{x})\lambda(\hat{y}) + \left[\int_{-\infty}^{\hat{y}} p(y|\mathbf{x})\,dy - q\right]\frac{\partial \lambda(\hat{y})}{\partial \hat{y}}$$

At the q-quantile υ, we have

$$\left.\frac{\partial^2 \mathcal{C}[\hat{y}|\mathbf{x}]}{\partial \hat{y}^2}\right|_{\hat{y}(\mathbf{x})=\upsilon(\mathbf{x})} = p(\upsilon|\mathbf{x})\lambda(\upsilon) \tag{C2}$$

which is positive since $\lambda(\hat{y}) > 0$ (11c). This proves (10c). The criterion is therefore a minimum at $\hat{y} = \upsilon$ (the q-quantile). ■

B. All the Solutions to (10a)–(10c) are Also Solutions to (11a)–(11c)

Equation (10b) must hold whatever the characteristics of the noise, that is, whatever the probability density $p(y|\mathbf{x})$ of the random variable y with q-quantile $v(\mathbf{x})$. From Leibnitz's rule, this implies that the expression

$$\int_{-\infty}^{\hat{y}=v} \left.\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}}\right|_{\hat{y}(\mathbf{x})=v(\mathbf{x})} p(y|\mathbf{x})\, dy + \int_{\hat{y}=v}^{+\infty} \left.\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}}\right|_{\hat{y}(\mathbf{x})=v(\mathbf{x})} p(y|\mathbf{x})\, dy = 0 \quad \text{(C3a)}$$

remains **stationary**—it remains identically zero—for any variation (the class of admissible variations being piecewise smooth curves in the sense of the calculus of variations; see [19]) of the function $p(y|\mathbf{x})$—denoted as $\delta p(y|\mathbf{x})$, $\mathbf{x}$ being fixed—subject to the constraints

$$\int_{-\infty}^{v} p(y|\mathbf{x})\, dy = q \quad \text{(C3b)}$$

$$\int_{v}^{+\infty} p(y|\mathbf{x})\, dy = (1-q). \quad \text{(C3c)}$$

Roughly speaking, it means that the result of the integral (C3a) is **invariant** when making a transformation $p'(y|\mathbf{x}) = p(y|\mathbf{x}) + \delta p(y|\mathbf{x})$, where $p'(y|\mathbf{x})$ and $p(y|\mathbf{x})$ are subject to the constraints (C3b)–(C3c).

As in Appendix A, this is a classical problem of the calculus of variations: the stationarity of (C3a) with respect to variations of $p(y|\mathbf{x})$, and subject to the constraints (C3b) and (C3c), directly implies that the following Lagrange function is stationary for any variation of $p(y|\mathbf{x})$, without considering the constraints

$$\begin{aligned} L[p(y|\mathbf{x})] = &\int_{-\infty}^{v} \left.\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}}\right|_{\hat{y}(\mathbf{x})=v(\mathbf{x})} p(y|\mathbf{x})\, dy \\ &+ \int_{v}^{+\infty} \left.\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}}\right|_{\hat{y}(\mathbf{x})=v(\mathbf{x})} p(y|\mathbf{x})\, dy \\ &+ \lambda^{-}(\mathbf{x}) \left[\int_{-\infty}^{v} p(y|\mathbf{x})\, dy - q\right] \\ &+ \lambda^{+}(\mathbf{x}) \left[\int_{v}^{+\infty} p(y|\mathbf{x})\, dy - (1-q)\right]. \end{aligned} \quad \text{(C4)}$$

$\lambda^{-}(\mathbf{x})$ and $\lambda^{+}(\mathbf{x})$ being Lagrange multipliers. Since the stationarity property must hold for every $\mathbf{x}$, they are labeled by $\mathbf{x}$.

The stationarity of expression (C4) requires that the variations for both $y > v$ and $y < v$ cancel (see [19, pp. 61–63])

$$\left.\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}}\right|_{\hat{y}(\mathbf{x})=v(\mathbf{x})} = -\lambda^{-}(\mathbf{x}) \qquad y < v \quad \text{(C5a)}$$

$$\left.\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}}\right|_{\hat{y}(\mathbf{x})=v(\mathbf{x})} = -\lambda^{+}(\mathbf{x}) \qquad y > v \quad \text{(C5b)}$$

and the Weierstrass–Erdmann conditions at v (ε being an infinitesimal positive value)

$$\left.\frac{\partial \mathcal{L}[\hat{y};\, v-\varepsilon]}{\partial \hat{y}}\right|_{\hat{y}(\mathbf{x})=v(\mathbf{x})} + \lambda^{-} = \left.\frac{\partial \mathcal{L}[\hat{y};\, v+\varepsilon]}{\partial \hat{y}}\right|_{\hat{y}(\mathbf{x})=v(\mathbf{x})} + \lambda^{+} \quad \text{(C6)}$$

which, from (C5a) and (C5b), is automatically satisfied.

Now, from (C5a) and (C5b), (C3a) can be rewritten as

$$\int_{-\infty}^{v} \lambda^{-}(\mathbf{x}) p(y|\mathbf{x})\, dy + \int_{v}^{+\infty} \lambda^{+}(\mathbf{x}) p(y|\mathbf{x})\, dy = 0 \quad \text{(C7)}$$

so that, from (C3b) to (C3c)

$$q\lambda^{-}(\mathbf{x}) = -(1-q)\lambda^{+}(\mathbf{x}). \quad \text{(C8)}$$

Since we assumed that $\mathcal{L}[\hat{y};\, y]$ depends on $\mathbf{x}$ only through the variable $\hat{y}$, by defining $\lambda(\hat{y}(\mathbf{x})) = \lambda^{+}(\mathbf{x})/q$, we obtain from (C8) and (C5a) and (C5b)

$$\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}} = (1-q)\lambda(\hat{y}) \qquad y < \hat{y} \quad \text{(C9a)}$$

$$\frac{\partial \mathcal{L}[\hat{y};\, y]}{\partial \hat{y}} = -q\lambda(\hat{y}) \qquad y > \hat{y}. \quad \text{(C9b)}$$

By taking the primitive and defining $\Lambda(\hat{y})$ such that $\partial\Lambda(\hat{y})/\partial\hat{y} = \lambda(\hat{y})$

$$\mathcal{L}[\hat{y};\, y] = (1-q)\Lambda(\hat{y}) + \phi(y) \qquad y < \hat{y} \quad \text{(C10a)}$$

$$\mathcal{L}[\hat{y};\, y] = -q\Lambda(\hat{y}) + \theta(y) \qquad y > \hat{y} \quad \text{(C10b)}$$

where ϕ and θ are integration constants. Since we require $\mathcal{L}[y;\, y] = 0$ (9a), we finally obtain

$$\mathcal{L}[\hat{y};\, y] = (1-q)[\Lambda(\hat{y}) - \Lambda(y)] \qquad y < \hat{y} \quad \text{(C11a)}$$

$$\mathcal{L}[\hat{y};\, y] = q[\Lambda(y) - \Lambda(\hat{y})] \qquad y > \hat{y}. \quad \text{(C11b)}$$

Moreover, under conditions (C11a) and (C11b), we have [the development is the same as for (C2), and is not repeated here]

$$\left.\frac{\partial^2 \mathcal{C}[\hat{y}|\mathbf{x}]}{\partial \hat{y}^2}\right|_{\hat{y}=v} = p(v|\mathbf{x}) \left.\frac{\partial \Lambda(\hat{y})}{\partial \hat{y}}\right|_{\hat{y}=v} \quad \text{(C12)}$$

so that (10c) implies that $\lambda(v) = \partial\Lambda(\hat{y})/\partial\hat{y}|_{\hat{y}=v}$ must be positive. Thus, $\partial\Lambda(\hat{y})/\partial\hat{y}$ must be positive on the range of possible values of the q-quantile v. Therefore, the optimality conditions (10a)–(10c) imply (11a)–(11c). This completes the proof. ■

ACKNOWLEDGMENT

The author would like to thank the three anonymous reviewers for their pertinent and constructive remarks.

REFERENCES

[1] C. Adcock and N. Meade, "A comparison of two linear programming solvers and a new iteratively reweighted least squares algorithm for L_1 estimation," in *L_1-Statistical Procedures and Related Topics*, ser. Inst. Math. Statist. Lecture Notes—Monograph Series, Y. Dodge, Ed., 1997, vol. 31.

[2] T. Arthanari and Y. Dodge, *Mathematical Programming in Statistics*. New York: Wiley, 1981.

[3] Y. Bard, *Nonlinear Parameter Estimation*. New York: Academic, 1974.

[4] I. Barrodale, "L_1 approximation and the analysis of data," *Appl. Statist.*, vol. 17, pp. 51–57, 1968.

[5] I. Barrodale and F. D. Roberts, "An improved algorithm for discrete L_1 linear approximation," *SIAM J. Numer. Anal.*, vol. 10, no. 5, pp. 839–848, 1973.

[6] ——, "Solution of an overdetermined system of equations in the L_1 norm," *Commun. ACM*, vol. 17, no. 6, pp. 319–320, 1974.

[7] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, pp. 349–369, 1989.

[8] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[9] R. Blattberg and T. Sargent, "Regression with non-Gaussian stable disturbances: Some sampling results," *Econometrica*, vol. 39, no. 3, pp. 501–510, 1971.

[10] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Amsterdam, The Netherlands: Kluwer, 1994.
[11] H. Bourlard and C. Wellekens, "Links between Markov models and multilayer perceptrons," in *Advances in Neural Information Processing Systems 1*, Touretzky, Ed. San Mateo, CA: Morgan Kaufmann, 1989, pp. 502–510.
[12] J. L. Brown, "Asymmetric nonmean-square error criteria," *IEEE Trans. Automat. Contr.*, vol. AC-7, pp. 64–66, 1962.
[13] M. Buchinsky, "Changes in the U.S. Wage Structure 1963–1987: An application of quantile regression," *Econometrica*, vol. 62, pp. 405–458, 1994.
[14] P. Burrascano, "A norm selection criterion for the generalized delta rule," *IEEE Trans. Neural Networks*, vol. 2, pp. 125–130, Jan. 1991.
[15] J. Cid-Sueiro, J. Arribas, S. Urban-Munoz, and A. Figueiras-Vidal, "Cost functions to estimate *a posteriori* probabilities in multiclass problems," *IEEE Trans. Neural Networks*, vol. 10, pp. 645–656, May 1999.
[16] M. De Groot and S. Fienberg, "The comparison and evaluation of forecasters," *Statistician*, vol. 32, pp. 12–22, 1983.
[17] R. Deutsch, *Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1965.
[18] W. Fisher, "A note on curve fitting with minimum deviations by linear programming," *J Amer. Statist. Assoc.*, vol. 56, pp. 359–362, 1961.
[19] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*. Englewood Cliffs, NJ: Prentice-Hall, 1963.
[20] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 1990, pp. 1361–1364.
[21] L. Gonzales Sotelino, M. Saerens, and H. Bersini, "Classification of temporal trajectories by continuous time recurrent nets," *Neural Networks*, vol. 7, no. 5, pp. 767–776, 1994.
[22] J. B. Hampshire and B. Pearlmutter, "Equivalence proofs for multilayer perceptron classifiers and the Bayesian discriminant function," in *Proc. 1990 Connectionnist Models Summer School*, D. Touretzky, J. Elman, T. Sejnowski, and G. Hinton, Eds. San Mateo, CA: Morgan Kaufmann, 1990, pp. 159–172.
[23] S. Hanson and D. Burr, "Minkowski-r backpropagation: Learning in connectionist models with non-Euclidian error signals," in *Neural Information Processing Systems*, D. Anderson, Ed. College Park, MD: Amer. Inst. Phys., 1988, pp. 348–357.
[24] D. Hosmer and S. Lemeshow, *Applied Logistic Regression*. New York: Wiley, 1989.
[25] P. Huber, *Robust Statistics*. New York: Wiley, 1981.
[26] F. Kanaya and S. Miyake, "Bayes statistical behavior and valid generalization of pattern classifying neural networks," *IEEE Trans. Neural Networks*, vol. 2, pp. 471–475, July 1991.
[27] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[28] R. Koenker and G. Bassett, "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.
[29] R. Koenker and V. D'Orey, "Computing regression quantiles," *Appl. Statist.*, vol. 36, pp. 383–393, 1987.
[30] F. Lad, *Operational Subjective Statistical Methods*. New York: Wiley, 1996.
[31] Y. Le Cun, "A theorical framework for backpropagation," in *Proc. 1988 Connectionnist Models Summer School*, D. Touretzky, G. Hinton, and T. Sejnowski, Eds. San Mateo, CA: Morgan Kaufmann, 1989, pp. 21–28.
[32] D. Lindley, "Scoring rules and the inevitability of probability (with discussions)," *Int. Statist. Rev.*, vol. 50, pp. 1–26, 1982.
[33] J. Matheson and R. Winkler, "Scoring rules for continuous probability distributions," *Management Sci.*, vol. 22, pp. 1087–1096, 1976.
[34] P. McCullagh and J. A. Nelder, *Generalized Linear Models, 2nd ed*. London, U.K.: Chapman and Hall, 1990.
[35] J. S. Meditch, *Stochastic Optimal Linear Estimation and Control*. New York: McGraw-Hill, 1969.
[36] J. L. Melsa and D. L. Cohn, *Decision and Estimation Theory*. New York: McGraw-Hill, 1978.
[37] J. W. Miller, R. Goodman, and P. Smyth, "Objective functions for probability estimation," in *Proc. IEEE Int. Joint Conf. Neural Networks*, San Diego, CA, 1991, pp. I-881–886.
[38] ——, "On loss functions which minimize to conditional expected values and posterior probabilities," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1404–1408, 1993.
[39] N. E. Nahi, *Estimation Theory and Applications*. New York: Wiley, 1976.
[40] M. R. Osborne, *Finite Algorithms in Optimization and Data Analysis*. New York: Wiley, 1985.
[41] M. R. Osborne and G. A. Watson, "On an algorithm for discrete nonlinear L_1 approximation," *Comput. J.*, vol. 14, pp. 184–188, 1971.
[42] A. Papoulis, *Probability, Random Variables, and Stochastic Processes, 3th ed*. New York: McGraw-Hill, 1991.
[43] R. Pfaffenberger and J. Dinkel, "Absolute deviations curve fitting: An alternative to least squares," in *Contributions to Survey Sampling and Applied Statistics*, H. A. David, Ed. New York: Academic, 1978, pp. 279–294.
[44] A. Pinkus, *On L_1 Approximation*. Cambridge, U.K.: Cambridge Univ. Press, 1989.
[45] M. D. Richard and R. P. Lippmann, "Neural-network classifiers estimate Bayesian *a posteriori* probabilities," *Neural Comput.*, vol. 3, pp. 461–483, 1991.
[46] R. Rojas, "A short proof of the posterior probability property of classifier neural networks," *Neural Comput.*, vol. 8, pp. 41–43, 1996.
[47] P. Rousseuw and A. Leroy, *Robust Regression and Outlier Detection*. New York: Wiley, 1988.
[48] D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter, "The multilayer perceptron as an approximation to a Bayes optimal discriminant function," *IEEE Trans. Neural Networks*, vol. 1, pp. 296–298, 1990.
[49] S. Ruzinsky and E. Olsen, "L_1 and L_∞ minimization via a variant of Karmakar's algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 245–253, 1989.
[50] M. Saerens, "Nonmean square error criteria for the training of learning machines," in *Proc. 13th Int. Conf. Machine Learning (ICML)*, Bari, Italy, July 1996, pp. 427–434.
[51] A. P. Sage and J. L. Melsa, *Estimation Theory with Applications to Communications and Control*. New York: McGraw-Hill, 1971.
[52] S. Santini and A. Del Bimbo, "Recurrent neural networks can be trained to be maximum *a posteriori* probability classifiers," *Neural Networks*, vol. 8, no. 1, pp. 25–29, 1995.
[53] L. Savage, "Elicitation of personal probabilities and expectations," *J. Amer. Statist. Assoc.*, vol. 66, pp. 783–801, 1971.
[54] E. Schlossmacher, "An iterative technique for absolute deviations curve fitting," *J. Amer. Statist. Assoc.*, vol. 68, pp. 857–865, 1973.
[55] S. Sherman, "Nonmean-square error criteria," *IEEE Trans. Inform. Theory*, vol. IT-4, pp. 125–126, 1958.
[56] P. A. Shoemaker, "A note on least-squares learning procedures and classification by neural-network models," *IEEE Trans. Neural Networks*, vol. 2, pp. 158–160, 1991.
[57] H. W. Sorenson, *Parameter Estimation, Principles and Problems*. New York: Marcel Dekker, 1980.
[58] L. D. Taylor, "Estimation by minimizing the sum of absolute errors," in *Frontiers in Econometrics*, P. Zarembka, Ed. New York: Academic, 1974, pp. 169–190.
[59] H. M. Wagner, "Linear programming techniques for regression analysis," *J. Amer. Statist. Assoc.*, vol. 54, pp. 206–212, 1959.
[60] E. A. Wan, "Neural network classification: A Bayesian interpretation," *IEEE Trans. Neural Networks*, vol. 1, pp. 303–305, 1990.
[61] R. K. Ward, "An on-line adaptation for discrete L_1 linear estimation," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 67–71, 1984.
[62] R. Wedderburn, "Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method," *Biometrika*, vol. 61, no. 3, pp. 439–447, 1974.
[63] H. White, "Learning in artificial neural networks: A statistical perspective," *Neural Comput.*, vol. 1, pp. 425–464, 1989.
[64] H. Wilson, "Least squares versus minimum absolute deviations estimation in linear models," *Decision Sci.*, vol. 9, pp. 322–335, 1978.
[65] A. Zellner, "Bayesian estimation and prediction using asymmetric loss functions," *J. Amer. Statist. Assoc.*, vol. 81, pp. 446–451, 1986.

Marco Saerens (M'00) received the B.S. degree in physics engineering from the Faculté Polytechnique of the Université Libre de Bruxelles, and the M.S. degree in theoretical physics, also from the Université Libre de Bruxelles. He joined the IRIDIA Laboratory as a Research Assistant in 1986, where he received the Ph.D. degree equivalent in engineering in 1990, on the subject of process control by using artificial neural networks.

As a part-time Researcher at the IRIDIA Laboratory, he then joined various private companies, where he has acted as an applied statistics and artificial intelligence expert. His main research interests include artificial neural networks, pattern recognition, and speech processing.