J Glob Optim (2010) 46:307–315
DOI 10.1007/s10898-009-9417-z

Convexification for data fitting

James Ting-Ho Lo

Received: 18 February 2009 / Accepted: 17 March 2009 / Published online: 31 March 2009
© Springer Science+Business Media, LLC. 2009

Abstract The main results reported in this paper are two theorems concerning the use of a newtype of risk-averting error criterion for data fitting. The first states that the convexity region of the risk-averting error criterion expands monotonically as its risk-sensitivity index increases. The risk-averting error criterion is easily seen to converge to the mean squared error criterion as its risk-sensitivity index goes to zero. Therefore, the risk-averting error criterion can be used to convexify the mean squared error criterion to avoid local minima. The second main theorem shows that as the risk-sensitivity index increases to infinity, the risk-averting error criterion approaches the minimax error criterion, which is widely used for robustifying system controllers and filters.

Keywords Convexification · Global optimization · Local minima · Data fitting · Neural network · Nonlinear regression · Minimax · Robustifying error crition · Degree of robustness

Mathematics Subject Classification (2000) 90C30 · 90C31 · 62M45 · 62G08

1 Introduction

A method of training neural networks into robust approximators of functions and robust identifiers of dynamical systems was reported in two papers [7,8] presented at the 2001 International Joint Conference on Neural Networks. The method was called the adaptive

This material is based upon work supported in part by the National Science Foundation and Army Research Office, but does not necessarily reflect the position or policy of the Government.

J. T.-H. Lo (✉)
Department of Mathematics and Statistics, University of Maryland Baltimore County,
Baltimore, MD 21250, USA
e-mail: jameslo@umbc.edu

risk-averting training method. It uses a new type of risk-averting error criterion (4), which is a modified version of a criterion with the same name employed for deriving robust controllers and filters [5,11,12]. The use of the risk-averting error criterion (the modified version) was motivated by its emphasizing large individual deviations in approximating functions and identifying dynamical systems in an exponential manner, thereby avoiding such large individual deviations and achieving robust performances.

Close examination of the numerical results in [7,8] reveals that the neural networks resulting from training with the risk-averting error criterion do actually have smaller mean squared errors than the neural networks trained with the mean squared error criterion. This observation confirms the well-known local-minimum difficulty that there are local minima in the mean squares error criterion for training neural networks that are difficult, if not impossible, to escape. In addition, the observation shows that the risk-averting error criterion enables us to escape these local-minima with ease.

This capability of the risk-averting error criterion (4) motivated an investigation and led to the discovery of the first main result reported here: the process of increasing the value of the risk-sensitivity index λ gradually in the adaptive risk-averting training method [7,8] expands the convexity region of the risk-averting error criterion, creating tunnels (or worm holes) to allow escape from poor local minima.

More specifically, the first theorem states that when λ in the criterion (4) increases to infinity, the region in the weight or parameter space of the neural network or regression model on which the risk-averting error criterion is convex expands monotonically to the entire space except the intersection of a finite number of lower dimensional sets, the number of sets increasing rapidly as the number K of exemplary input/output pairs in the training data increases. Roughly speaking, λ and K control the size of the convexity region of (4): the greater λ or K, the larger the convexity region of the risk-averting error criterion.

As mentioned, the purpose of employing the risk-averting error criterion in [7,8] is to induce robust performance in the function approximators and system identifiers. A natural question is how the robustness induced by the risk-averting error criterion is related to the robustness induced by the minimax error criterion used in the robust control theory [1,3]. An answer to this question is reported as the second main theorem: as $\lambda \to \infty$, the risk-averting error criterion approaches the minimax error criterion.

The idea of convexifying a nonconvex function for global optimization is not new. Two well-known methods are the graduated nonconvexity method [2] and the Liu–Floudas convexification method [6,13]. In theory, the Liu–Floudas convexification method can be applied to data fitting where the error criterion is twice continuously differentiable. However, if the number of weights or parameters in the error criterion is very large, as is usually the case with training neural networks, determining the weight α of the added quadratic function for convexifying the error criterion involves much computation.

2 The risk-averting error criterion

If a set of exemplary input/output pairs, $\{(x_k, y_k), k = 1, \ldots, K\}$ is to be fitted to by a feedforward neural network or a nonlinear functional regression model, $y = f(x, w)$, with a weight or parameter vector w, a standard mean squared error criterion is

$$\sum_{k=1}^{K} \|y_k - f(x_k, w)\|_Q^2 \tag{1}$$

where Q is a positive definite matrix and the symbol $\|\cdot\|_Q^2$ denotes the quadratic form $(\cdot)^T Q(\cdot)$.

For system identification, filter/controller synthesis, or parametric or nonparametric dynamical regression model estimation, a set of exemplary time sequences of input/output pairs, $\{(x_t(s), y_t(s)), t = -B, \ldots, T, s = 1, \ldots, S\}$ is to be fitted to by a recurrent neural network or a dynamical regression model with a weight or parameter vector w. Denoting the output of the recurrent neural network or dynamical regression model at time t, after its initial is set properly for time $-B$ and it has processed $x_\tau(s)$, $\tau = -B, \ldots, t$, one at a time in the given order, a standard mean squared error criterion is

$$\sum_{s=1}^{S} \sum_{t=1}^{T} \|y_t(s) - f(x_t(s), w)\|_Q^2 \tag{2}$$

where $\|\cdot\|_Q^2$ is defined as before.

For notational simplicity, (1) and (2) are written as

$$J_0(w) = \sum_{k=1}^{K} e_k^T(w) Q e_k(w) \tag{3}$$

where $e_k(w) := y_k - \hat{y}_k(w)$ and $\hat{y}_k(w)$ denotes either $f(x_k, w)$ or $f(x_t(s), w)$ depending on whether (1) or (2) is concerned. The dimensionalities of $e_k(w)$, Q and w are denoted by m, $m \times m$ and N, respectively.

The risk-averting error criterion $J_\lambda(w)$ corresponding to the standard mean squared error criterion (3) is

$$J_\lambda(w) = \sum_{k=1}^{K} \exp\left[\lambda e_k^T(w) Q e_k(w)\right] \tag{4}$$

which is a modified version of the risk-averting error criterion used in robust control [5,11]. Here, λ is a positive number called the risk-sensitivity index.

3 The Hessian matrix

Assume that the the neural network or nonlinear regression model $\hat{y}_k(w)$ is twice continuously differentiable with respect to the vector w. Then,

$$\frac{\partial J_\lambda(w)}{\partial w_j} = -2\lambda \sum_{k=1}^{K} \alpha_k(w) e_k^T(w) Q \frac{\partial \hat{y}_k(w)}{\partial w_j}$$

where

$$\alpha_k(w) := \exp\left[\lambda e_k^T(w) Q e_k(w)\right]$$

Denoting a matrix whose $(i \times j)$th entry is a_{ij} by $[a_{ij}]$, the $N \times N$ Hessian matrix $H_\lambda(w) := [\partial^2 J_\lambda(w)/\partial w_i \partial w_j]$ of $J_\lambda(w)$ is

$$H_\lambda(w) = 2\lambda \sum_{k=1}^{K} \alpha_k(w)\{2\lambda A_k(w) + B_k(w) - C_k(w)\} \tag{5}$$

where

$$\begin{aligned} A_k(w) &= \left[e_k^T(w)Q(\partial \hat{y}_k(w)/\partial w_i)(\partial \hat{y}_k^T(w)/\partial w_j)Qe_k(w)\right] \\ B_k(w) &= \left[(\partial \hat{y}_k(w)/\partial w_i)Q(\partial \hat{y}_k^T(w)/\partial w_j)\right] \\ C_k(w) &= \left[e_k^T(w)Q(\partial^2 \hat{y}_k(w)/\partial w_i \partial w_j)\right] \end{aligned}$$

are all $N \times N$ matrices.

4 Convexity region

In this section, we will examine the convexity region of the criterion $J_\lambda(w)$, namely the region on which $J_\lambda(w)$ is convex. We note first that

$$\begin{aligned} B_k(w) &= F_k(w)QF_k^T(w) \\ F_k^T(w) &:= [\partial \hat{y}_k(w)/\partial w_1 \quad \cdots \quad \partial \hat{y}_k(w)/\partial w_N] \end{aligned}$$

and

$$\begin{aligned} A_k(w) &= D_k(w)D_k^T(w) \\ D_k^T(w) &:= [(\partial \hat{y}_k^T(w)/\partial w_1)Qe_k(w), \ldots, (\partial \hat{y}_k^T(w)/\partial w_N)Qe_k(w)] \end{aligned}$$

Note that $B_k(w)$ and $A_k(w)$ are positive semi-definite for all w, but $C_k(w)$ may be indefinite.

Straightforward matrix calculation yields

$$\begin{aligned} \sum_{k=1}^{K} \alpha_k B_k(w) &= \mathcal{F}_K(w)\mathcal{B}_K(w)\mathcal{F}_K^T(w) \\ \mathcal{F}_K(w) &:= [F_1(w) \quad \cdots \quad F_K(w)] \\ \mathcal{A}_K(w) &:= diag[\alpha_1(w)Q \quad \cdots \quad \alpha_K(w)Q] \end{aligned} \tag{6}$$

and

$$\begin{aligned} \sum_{k=1}^{K} \alpha_k A_k(w) &= \mathcal{D}_K(w)\mathcal{A}_K(w)\mathcal{D}_K^T(w) \\ \mathcal{D}_K(w) &:= [D_1(w) \quad \cdots \quad D_K(w)] \\ \mathcal{A}_K(w) &:= diag[\alpha_1(w) \quad \cdots \quad \alpha_K(w)] \end{aligned} \tag{7}$$

Using these newly established notations, the Hessian matrix can be written as

$$H_\lambda(w) = 2\lambda \left\{ 2\lambda \mathcal{D}_K(w)\mathcal{A}_K(w)\mathcal{D}_K^T(w) + \mathcal{F}_K(w)\mathcal{B}_K(w)\mathcal{F}_K^T(w) - \sum_{k=1}^{K} \alpha_k(w)C_k(w) \right\} \tag{8}$$

Since $\mathcal{A}_K(w)$ and $\mathcal{B}_K(w)$ are positive definite matrices, both $\mathcal{D}_K(w)\mathcal{A}_K(w)\mathcal{D}_K^T(w)$ and $\mathcal{F}_K(w)\mathcal{B}_K(w)\mathcal{F}_K^T(w)$ are positive-semidefinite. Note that while the first and third terms within the curly brackets in (8) are quadratic and linear functions of the deviations $e_k(w)$, the second term $\mathcal{F}_K(w)\mathcal{B}_K(w)\mathcal{F}_K^T(w)$ is independent of $e_k(w)$. If the deviations $e_k(w)$ approach zero toward the end of applying the risk-averting error criterion $J_\lambda(w)$ for data fitting, the second term becomes dominant in the Hessian matrix $H_\lambda(w)$ in (8). The positive-semidefiniteness of the second term (positive-definiteness if $\mathcal{F}_K(w)$ is of full rank) makes it less necessary to increase λ much toward the end of using $J_\lambda(w)$ for data fitting.

If the $(N \times K)$-matrix $\mathcal{D}_K(w)$ is of full rank (i.e., rank $\mathcal{D}_K(w) = N$), then $\mathcal{D}_K(w)\mathcal{A}_K(w)$ $\mathcal{D}_K^T(w)$ is positive definite, and the Hessian matrix $H_\lambda(w)$ is strictly monotone increasing in the risk-sensitivity index λ. In other words, $H_{\lambda_2}(w) > H_{\lambda_1}(w)$ for $\lambda_2 > \lambda_1$. It follows that the sequence of sets $P_\lambda := \{w \in R^N : H_\lambda(w) > 0\}$ is strictly monotone increasing in the sense that $P_{\lambda_1} \subset P_{\lambda_2}$ for $\lambda_2 > \lambda_1$.

Notice that within the pair of curly brackets on the right of (8), the second and third terms are independent of λ, the second being positive semi-definite. If λ is sufficiently large, the first term dominates the third with respect to the matrix inequality. Hence, if the matrix $\mathcal{D}_K(w)$ is of full rank, there is a positive number $\Lambda(w)$ such that $H_\lambda(w) > 0$ for $\lambda \geq \Lambda(w)$. Therefore, $\{w \in R^N : \text{rank}\, \mathcal{D}_K(w) = N\} \subset \cup_{\lambda>0} P_\lambda$ and $\{w \in R^N : \text{rank}\, \mathcal{D}_K(w) < N\} \supset (\cup_{\lambda>0} P_\lambda)^c$, the superscript c denoting the set complement.

The condition rank $\mathcal{D}_K(w) < N$ means that the determinant of every $N \times N$ submatrix of $\mathcal{D}_K(w)$ is zero. The number of such submatrices in $\mathcal{D}_K(w)$ is the number $C(K, N)$ of combinations of K columns of $\mathcal{D}_K(w)$ taken N at a time. Since the determinant of such a submatrix being equal to zero is an equation in w, there are $C(K, N)$ equations. The intersection of these $C(K, N)$ solution sets is the set $\{w \in R^N : \text{rank}\, \mathcal{D}_K(w) < N\}$, which contains $(\cup_{\lambda>0} P_\lambda)^c$.

Summarizing the above discussion, we have the following theorem:

Theorem 1 *Assume that the risk-averting error criterion $J_\lambda(w)$ in (4) is twice continuously differentiable. The sequence of sets $P_\lambda := \{w \in R^N : H_\lambda(w) > 0\}$ is monotone increasing as λ increases. The set $M := \{w \in R^N : \text{rank}\, \mathcal{D}_K(w) < N\}$, which is the intersection of the solution sets of $C(K, N)$ algebraic equations defined by setting the $C(K, N)$ submatrices of $\mathcal{D}_K(w)$ equal to zero, contains the complement of the set $\cup_{\lambda>0} P_\lambda$. In other words, as λ increases to ∞, the set P_λ expands monotonically to the entire weight or parameter space except the set $(\cup_{\lambda>0} P_\lambda)^c$, which is contained in the intersection M.*

Remark As the number K of input/output pairs in the training data increases, the number $C\,(K, N)$ of solution sets increases rapidly, and the intersection M of these solution sets shrinks monotonically.

5 A range of robustness

In the risk-averting error criterion $J_\lambda(w)$, the greater the risk-sensitivity index λ is, the more emphasis is place on large individual deviations $e_k(w)$. As λ ranges from 0 to ∞ (excluding 0 and ∞, of course), $J_\lambda(w)$ induces a range of robustness. To obtain some intuitive understanding of this range, we will show, in the following, that

1. $J_\lambda(w)$ acts like the mean squared error criterion (3) as $\lambda \to 0$; and
2. $J_\lambda(w)$ acts like the minimax error criterion, $\inf_w \max_k \|e_k(w)$, as $\lambda \to \infty$.

The meaning of the word, “acts,” is specifically defined in the following.

With λ as a parameter, $\{J_\lambda(w)|\lambda > 0\}$ is a parametrized collection of criteria. Observing that

$$\frac{1}{K}J_\lambda(w) = \frac{1}{K}\sum_{k=1}^{K}\left[1 + \lambda e_k^T(w)Qe_k(w) + O(\lambda^2)\right]$$
$$= 1 + \frac{1}{K}\sum_{k=1}^{K}\lambda e_k^T(w)Qe_k(w) + O(\lambda^2)$$

Recalling the power expansion formula,

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \cdots, \quad \text{for } -1 < x < 1$$

we have, for λ sufficiently small,

$$\frac{1}{\lambda}\ln\left[\frac{1}{K}J_\lambda(w)\right] = \frac{1}{K}\sum_{k=1}^{K}e_k^T(w)Qe_k(w) + O(\lambda)$$

It follows that

$$\lim_{\lambda\to 0}\frac{1}{\lambda}\ln\left[\frac{1}{K}J_\lambda(w)\right] = \frac{1}{K}\sum_{k=1}^{K}e_k^T(w)Qe_k(w) \tag{9}$$

This completes the proof of item 1 above. Note that $\frac{1}{\lambda}\ln\left[\frac{1}{K}(\cdot)\right]$ is a strictly monotone increasing function, and hence $\frac{1}{\lambda}\ln\left[\frac{1}{K}J_\lambda(w)\right]$ and $J_\lambda(w)$ share the same local and global minimizers.

Let us now consider an error criterion $J_{\lambda,p}(w)$ more general than $J_\lambda(w)$ in proving a slightly more general version of item 2 above. Denote the L_p norm by $\|\cdot\|_p$, i.e., $\|a\|_p = \left(\sum_{i=1}^{m}|a_i|^p\right)^{1/p}$, and let $J_{\lambda,p}(w) := \sum_{k=1}^{K}\exp\left[\lambda\|y_k - \hat{y}_k(w)\|_p^p\right]$.

Theorem 2 *Let* $\{u_\lambda \in R^N, \lambda > 0\}$ *be a sequence of weight vectors such that*

$$\lim_{\lambda\to\infty}\max_k\|e_k(u_\lambda)\|_p = \inf_w\max_k\|e_k(w)\|_p \tag{10}$$

If a sequence $\{w_\lambda \in R^N, \lambda > 0\}$ *satisfies* $J_{\lambda,p}(w_\lambda) \le J_{\lambda,p}(u_\lambda)$ *for all* $\lambda \ge \Lambda$ *for some* $\Lambda > 0$*, then*

$$\lim_{\lambda\to\infty}\max_k\|e_k(w_\lambda)\|_p = \inf_w\max_k\|e_k(w)\|_p \tag{11}$$

where $\max_k$ *means the maximum over* $k \in \{1,\ldots K\}$ *and* $\inf_w$ *means the infimum over* $w \in R^N$.

Proof Note first that the existence of $\{u_\lambda \in R^N, \lambda > 0\}$ is implied by the definition of $\inf_w\max_k\|e_k(w)\|_p$. Define the notations, $\Phi(w) := \arg\max_k\|e_k(w)\|_p^p$, $b := \inf_w\max_k \|e_k(w)\|_p$ and $b_\lambda := \inf_w J_{\lambda,p}(w)$. Note that $\Phi(w)$ may be a set.

Rewirte the two sides of $J_{\lambda,p}(w_\lambda) \le J_{\lambda,p}(u_\lambda)$ as $\exp\left[\lambda\left\|e_{\phi(w_\lambda)}(w_\lambda)\right\|_p^p\right] + \sum_{k\neq\phi(w_\lambda)}\exp\left[\lambda\|e_k(w_\lambda)\|_p^p\right] \le \exp\left[\lambda\left\|e_{\phi(u_\lambda)}(u_\lambda)\right\|_p^p\right] + \sum_{k\neq\phi(u_\lambda)}\exp\left[\lambda\|e_k(u_\lambda)\|_p^p\right]$, where $\phi(u_\lambda) \in \Phi(u_\lambda)$, $\phi(w_\lambda) \in \Phi(w_\lambda)$, the summations $\sum$ are taken over $k = 1,\ldots,K$ with the exception indicated below the summation signs. Because the first term on the right side is a dominant term, it follows from this inequality that $\exp\left[\lambda\left\|e_{\phi(w_\lambda)}(w_\lambda)\right\|_p^p\right] < (K+1)\exp\left[\lambda\left\|e_{\phi(u_\lambda)}(u_\lambda)\right\|_p^p\right]$. By (10), the sequence $d_\lambda := \max_k\|e_k(u_\lambda)\|_p^p - b^p$ converges to 0 as λ approaches ∞. Rewrite this inequality as

$$\exp\left[\lambda\left\|e_{\phi(w_\lambda)}(w_\lambda)\right\|_p^p\right] < (K+1)\exp\left[\lambda(b^p + d_\lambda)\right] \tag{12}$$

Note that this inequality holds for all $\lambda \geq \Lambda$.

Assume that (11) fails to hold for this sequence $\{w_\lambda, \lambda > 0\}$. Then $\lim_{\lambda\to\infty} \max_k \|e_k(w_\lambda)\|_p > b$ or $\lim_{\lambda\to\infty} \max_k \|e_k(w_\lambda)\|_p$ does not exist. In either case, there is $c > 0$ such that for every $L > 0$, there exists $\lambda > L$ such that $\|e_{\phi(w_\lambda)}(w_\lambda)\|_p^p \geq b^p + c$. Since $\lim_{\lambda\to\infty} d_\lambda = 0$, there is $L_1 > 0$ such that $d_\lambda < c/2$ for all $\lambda \geq L_1$. It follows from the above inequality (12) that for all $\lambda \geq \max\{\Lambda, L_1\}$,

$$\exp\left[\lambda\|e_{\phi(w_\lambda)}(w_\lambda)\|_p^p\right] < (K+1)\exp\left[\lambda(b^p + c/2)\right] \tag{13}$$

Then for $L_2 = \max\{\Lambda, L_1, 2\ln(K+2)/c\}$, there exists a $\lambda > L_2$ such that $\|e_{\phi(w_\lambda)}(w_\lambda)\|_p^p \geq b^p + c$, which implies that for this λ, $\exp\left[\lambda\|e_{\phi(w_\lambda)}(w_\lambda)\|_p^p\right] \geq \exp[\lambda b^p + \lambda c] > \exp[\lambda b^p + \lambda c/2 + \ln(K+2)] > (K+2)\exp[\lambda(b^p + c/2)]$, which contradicts (12). Therefore, the above assumption is false, and $\{w_\lambda, \lambda > 0\}$ satisfies (11), completing the proof of the theorem. □

The following corollary is an immediate consequence of Theorem 2.

Corollary 3 *If* $w_\lambda \in \arg\min_w J_{\lambda,p}(w)$ *for all* $\lambda > \Lambda$ *for some* $\Lambda > 0$, *then*

$$\lim_{\lambda\to\infty} \max_k \|e_k(w_\lambda)\|_p = \inf_w \max_k \|e_k(w)\|_p$$

Note that Theorem 2 and its proof with the pth power $\|\cdot\|_p^p$ of the L_p norm $\|\cdot\|_p$ replaced with any monotone increasing function or any norm are still valid. An example is obtained by replacing $\|\cdot\|_p^p$ with $\|\cdot\|_Q^2$ as used in (4). Another example is obtained by replacing $\|\cdot\|_p^p$ with the supremum norm $\|\cdot\|_\infty$, which is defined for a m-vector y by $\|y\|_\infty := \max_i\{y_i : i = 1, \ldots m\}$.

The two items above have now been proved. They show that $J_\lambda(w)$ induces a range of robustness from the minimum mean squared error to the minimax error as the risk-sensitivity index λ ranges from 0 to ∞. The value of λ can be used as an index or degree of robustness. Some numerical experiments on applying $J_\lambda(w)$ to approximating functions and dynamical systems in noisy data with various values of λ and the resulting robustification effects are reported in [9,10].

We remark that the robustness addressed in this paper is the robustness usually found in the engineering literature [1,3,5,11,12]. It is the opposite of the robustness usually found in the robust statistics literature [4], where robustness means de-emphasizing large deviations or outliers. In fact, $J_\lambda(w)$ with a negative risk-sensitivity index λ is well suited to inducing robustness for de-emphasizing large deviations. Obviously, the more negative the risk-sensitivity index λ is, the more de-emphasized large deviations $e_k(w)$ are. $J_\lambda(w)$ with negative risk-sensitivity index λ is called a risk-seeking error criterion. It is proven that as the risk-sensitivity index λ decreases to negative infinity, $J_\lambda(w)$ acts like the maximin error criterion. The risk-seeking error criteria $J_\lambda(w)$ with λ ranging from 0 to $-\infty$ induce another range of robustness for studying robust statistics in the sense of [4]. Work on risk-seeking error criteria in the context of robust statistics [4] will be reported elsewhere.

6 Centering and bounding for computation

The reader is referred to [7,8] for a method of training neural networks using $J_\lambda(w)$. Numerical examples in those papers show the effectiveness of the convexification method even if the training or fitting data are noisy.

Theorem 2 states that the greater λ is, the larger P_λ is. However, in applying the convexification method, the magnitude of λ has to be restricted to avoid overflow or underflow of the registers in the computer. In the following, we determine the maximum λ value that can be properly handled by a given computer. Recall that a typical term in $J_\lambda(w)$ is $\alpha_k(w) := \exp\left[\lambda e_k^T(w)Qe_k(w)\right]$. When $\lambda e_k^T(w)Qe_k(w)$ is too large or too small, $\alpha_k(w)$ may go beyond the maximum and minimum positive numbers, denoted by $\exp\zeta_2$ and $\exp\zeta_1$, respectively, that the computer can properly do arithmetic with. For instance, ζ_1 and ζ_2 for a Pentium PC are -13 and 13, respectively. The interval $[\exp\zeta_1, \exp\zeta_2]$ may also be the range that we choose to have the arithmetic done within. To best use the range $[\exp\zeta_1, \exp\zeta_2]$, we treat $\alpha_k(w)$ as $\exp(\lambda b)\cdot\exp\left[\lambda(e_k^T(w)Qe_k(w)-b)\right]$ numerically, and determine b and the largest value of λ allowed for the computer as follows.

Obviously, to keep $\exp\left[\lambda\left(e_k^T(w)Qe_k(w)-b\right)\right]$ within the range $[\exp\zeta_1, \exp\zeta_2]$, we need $\lambda\left(e_k^T(w)Qe_k(w)-b\right)\le\zeta_2$ and $\lambda(e_k^T(w)Qe_k(w)-b)\ge\zeta_1$, or equivalently, $q_{\max}-\zeta_2/\lambda\le b\le q_{\min}-\zeta_1/\lambda$, where $q_{\max}:=\max_k\left\{e_k^T(w)Qe_k(w)\right\}$ and $q_{\min}:=\min_k\left\{e_k^T(w)Qe_k(w)\right\}$. From $q_{\max}-\zeta_2/\lambda\le q_{\min}-\zeta_1/\lambda$, it follows that $\lambda<(\zeta_2-\zeta_1)/(q_{\max}-q_{\min})$, which is the maximum value of λ allowed for the computer. Notice that the range of b for a selected λ is $[q_{\max}-\zeta_2/\lambda, q_{\min}-\zeta_1/\lambda]$. A reasonable choice for b is the middle point $(q_{\max}-\zeta_2/\lambda+q_{\min}-\zeta_1/\lambda)/2$ of this range, which places $\lambda(q_{\min}-b)$ and $\lambda(q_{\max}-b)$ equi-distant from $(\zeta_2+\zeta_1)/2$ and makes a good use of the range $[\exp\zeta_1, \exp\zeta_2]$.

In the process of training a neural network or estimating a regression model with a Pentium PC or any computer with $\zeta_2+\zeta_1=0$, we may want to set $b=(q_{\max}+q_{\min})/2$ and $\lambda=0.9\times 26/(q_{\max}-q_{\min})$ and fix them for a certain number of iterations in an iterative algorithm, and repeat. As $q_{\max}-q_{\min}$ decreases in the training process, λ increases.

7 Conclusion

It is proven that the greater the risk-sensitivity index λ, the greater the region on which the risk-averting error criterion is convex. This explains the ability of the adaptive risk-averting training method in [7,8], which increases λ gradually, to avoid poor local minima. Intuitively, increasing λ creates tunnels (or worm holes) for a local-search minimization procedure (e.g., quasi-Newton and conjugate gradient methods) to travel through to a good local minimum.

Nevertheless, it is still not clear under what condition or for what value of λ , a global minimum can be reached. However, is a global minimum always desirable? Perhaps not. A global minimum at the bottom of a narrow notch with a small opening on the “landscape” of the error criterion may not be as desirable as a local minimum at the bottom of a slightly shallower but much wider trough. The latter may represent a neural network with a better generalization capability of a regression model less sensitive to sampling bias or errors.

Intuitively, increasing λ creates tunnels leading to a wider trough before leading to a narrower notch. Some of the most narrow notches may not be opened up by increasing λ before a satisfactory wide trough is obtained. This may be a blessing. However, more study is needed to understand the effects and noneffects of increasing λ and to understand the relationship between the width of the notch and the generalization capability of the neural network or the regression model at the bottom of the notch.

The minimax error criterion comes from the game theory and is known to be too pessimistic as a robustifying error criterion. Moreover, it is difficulty to use. The risk-averting error criterion with a very large risk-sensitivity index acts like the minimax error criterion, and is easier to use. More important perhaps, the risk-averting error criterion induces a continuous range of robustness indexed by λ ranging from 0 to ∞. As λ goes to zero at one end, the risk-averting error criterion approaches the mean squared error criterion. As λ goes to infinity at the other end, the risk-averting error criterion approaches the minimax error criterion. The risk-averting error criterion provides different degrees of robustness for a robust system designer to choose from. Developing a probabilistic decision theory to determine the optimal value of λ for a give application is an open research topic.

References

1. Basar, T., Bernhard, P.: H-infinity Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach. 2nd edn. Birkhauser, Boston (1995)
2. Blake, A., Zisserman, A.: Visual Reconstruction. The MIT Press, Cambridge (1987)
3. Glover, K., Doyle, J.C.: State-space formulae for all stabilizing controllers that satisfy an H-infinity norm bound and relations to risk-sensitivity. Syst. Control Lett. **11**, 167–172 (1988)
4. Huber, P.: Robust Statistics. Wiley, New York (1982)
5. Jacobson, D.H.: Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic games. IEEE Trans. Automat. Contr. **AC-18**(2), 124–131 (1973)
6. Liu, W.B., Floudas, C.A.: A remark on the GOP algorithm for global optimization. J. Glob. Optim. **3**, 519–531 (1993)
7. Lo, J.T.-H., Bassu, D.: An adaptive method of training multilayer perceptrons. In: Proceedings of the 2001 International Joint Conference on Neural Networks, vol. 3, pp. 2013–2018. IEEE Xplore, The IEEE Press, Piscataway (2001)
8. Lo, J.T.-H., Bassu, D.: Robust identification of dynamic systems by neurocomputing. In: Proceedings of the 2001 International Joint Conference on Neural Networks, vol. 2, pp. 1285–1290. IEEE Xplore, The IEEE Press, Piscataway (2001)
9. Lo, J.T.-H., Bassu, D.: Robust approximation of uncertain functions where adaptation is impossible. In: Proceedings of the 2002 International Joint Conference on Neural Networks, vol. 2, pp. 1956–1961. IEEE Xplore, The IEEE Press, Piscataway (2002)
10. Lo, J.T.-H., Bassu, D.: Robust identification of uncertain dynamical systems where adaptation is impossible. In: Proceedings of the 2002 International Joint Conference on Neural Networks, vol. 2, pp. 1558–1563. IEEE Xplore, The IEEE Press, Piscataway (2002)
11. Speyer, J., Deyst, J., Jacobson, D.H.: Optimization of stochastic linear systems with additive measurement and process noise using exponential performance criteria. IEEE Trans. Automat. Contr. **AC-19**, 358–366 (1974)
12. Whittle, P.: Risk Sensitive Optimal Control. Wiley, New York (1990)
13. Zlobec, S.: On the Liu–Floudas convexification of smooth programs. J. Glob. Optim. **32**, 401–407 (2005)