

Robust Classification under Uncertainty in the Prior Probabilities

Carmit Keren, Miriam Zacksenhouse and Yakov Ben-Haim

Faculty of Mechanical Engineering
Technion - Israel Institute of Technology
Haifa 32000 Israel

Abstract

Optimal classification methods and most notably Bayesian classifiers are sensitive to uncertainty in the statistics of the patterns to be classified. Errors in the associated probabilities and distributions would degrade the performance of optimal classifiers. We present here a robust-satisficing classifier whose robustness to uncertainty in the priors is maximized given a performance demand. We apply the method to a two-class medical classification problem. We show that the robust-satisficing classifier is more robust to uncertainty in the priors than the optimal Bayesian classifier at sub-optimal performance levels. We present 4 propositions which characterize the robust-satisficing classifier.

Keywords. Bayesian classification; Robust classification; Info-gap; Priors; Decision making; Uncertainties

1 Introduction

Consider the basic model of a two-class classification problem, based on a single measured parameter x . A classification algorithm determines the threshold θ for classifying a new measurement to one of the two classes, ω_1 and ω_2 (Duda, Hart, & Stork 2001):

$$\text{Decision algorithm: } \begin{cases} \omega_1 & x < \theta \\ \omega_2 & x > \theta \end{cases}$$

This decision model can be applied in many disciplines, where in some applications the cost of making the wrong decision can be critical. For example detection of heart failure conditions can be considered as a classification problem, where relevant indicators of the heart-condition should be classified as either indicating Normal or Failure conditions.

Given the two class-conditional probability distribution functions (pdfs), $p_1(x|\omega_1)$ and $p_2(x|\omega_2)$, two types of errors may occur: ε_1 for deciding class ω_2 while the true state of nature is ω_1 , and ε_2 for deciding class ω_1 while the true state is ω_2 . An example of this two category decision is depicted in Fig.1 for Normal pdfs. The Bayesian classification goal is to minimize the combined probability of error, given the prior probabilities of the two classes.

In most practical applications, however, the estimated prior probabilities are uncertain, due, for example, to inherent differences between the testing population, from which the priors were estimated, and the target population used for the decision algorithm. Moreover, the estimated pdfs of measured quantities in such applications are rarely available, and their estimate from history records is poor. In medicine, erroneous fault detection due to improper estimation can result in serious adverse outcome on the one hand and in consuming substantial health care resources on the other. While both sources of uncertainties are important, this paper deals only with uncertainties in the prior probabilities, assuming that the class-conditioned pdfs are known.

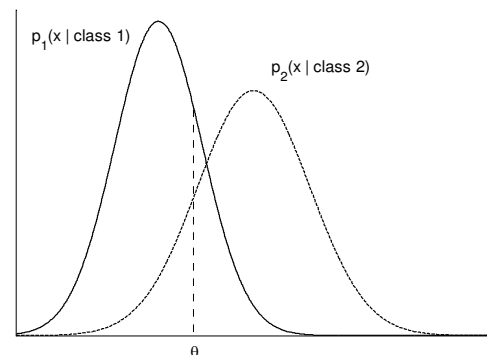


Figure 1: Decision boundary for two-class pdfs

Information-gap decision theory (Ben-Haim, 2006) provides a quantitative tool for decision-making under severe uncertainty and is applied here to assess the robustness of the Bayesian classification algorithm. In addition we provide a robust-satisficing algorithm based on maximization of the level of uncertainties at which satisfactory performance can be guaranteed.

2 Info-Gap model

Given a two-class classification problem, ω_1 and ω_2 , the best estimated (but highly uncertain) prior probability of state ω_1 is \tilde{P}_1 . To represent the uncertainty in the estimated prior probability we use an info-gap model denoted as $U(\alpha, \tilde{P}_1)$. A common info-gap model is the fractional-error model, given by:

$$U(\alpha, \tilde{P}_1) = \{ P_1 : 0 \leq P_1 \leq 1, \\ |P_1 - \tilde{P}_1| \leq \alpha \tilde{P}_1 \}, \quad \alpha \geq 0 \quad (1)$$

The set $U(\alpha, \tilde{P}_1)$ contains all the priors probabilities allowed at horizon of uncertainty α . These uncertainty-sets become more inclusive as α increases. The horizon of uncertainty, α , is unknown and the info-gap model is a family of nested sets of possible priors.

Since the prior probabilities, P_1 and P_2 , of the two-categories are summed to one, only one probability has to be determined. Note however that there are alternative info-gap models that should be evaluated as well, including: (i) fractional uncertainty in the minimum between the two priors, and (ii) constant uncertainty.

Given the two class-conditioned pdfs, $p_1(x|\omega_1)$ and $p_2(x|\omega_2)$, the expected weighted classification error is:

$$\varepsilon(\theta, P_1) = P_1 \varepsilon_1(\theta) + P_2 \varepsilon_2(\theta), \quad P_1 + P_2 = 1 \quad (2)$$

where $\varepsilon_1(\theta)$ and $\varepsilon_2(\theta)$ are the two types of errors multiplied by non-negative weights λ_1 and λ_2 respectively, which define their relative importance ($\lambda_1 + \lambda_2 = 1$):

$$\varepsilon_1(\theta) = \lambda_1 \int_{\theta}^{\infty} p_1(x|\omega_1) dx; \quad \varepsilon_2(\theta) = \lambda_2 \int_{-\infty}^{\theta} p_2(x|\omega_2) dx$$

The optimal Bayesian method aims to choose the threshold θ so that $\varepsilon(\theta, P_1)$ is minimized. However, only the highly uncertain estimation \tilde{P}_1 is available. Thus, using the info-gap approach we seek to choose θ so that $\varepsilon(\theta, P_1)$ is limited to a desired performance under the greatest level of uncertainty in P_1 .

Given a performance demand ε_c , we define the robustness to uncertainty in the estimated prior as the greatest horizon of uncertainty α up to which all realizations of the prior lead to adequate performance:

$$\hat{\alpha}(\varepsilon_c, \theta) = \max \left\{ \alpha : \left(\max_{P_1 \in U(\alpha, \tilde{P}_1)} \varepsilon(\theta, P_1) \right) \leq \varepsilon_c \right\} \quad (3)$$

Equation (3) determines the robustness of a classifier with threshold θ given a performance criterion ε_c , while establishing a performance/robustness trade-off. Greater robustness can be achieved only by relinquishing performance, i.e. when degraded performance (larger ε_c) is satisfactory. In contrast, stringent performance demands (small ε_c) possess low robustness. In particular, the optimal performance obtained by the Bayesian algorithm has zero robustness, as shown in Section 3.

3 Robustness to uncertainty in the priors

The robustness of a classifier with threshold θ to uncertainty in the priors can be assessed by substituting (2) into (3) using the uncertainty model (1).

Def. 1 Given the estimation \tilde{P}_1

$$\tilde{\varepsilon}(\theta) = \varepsilon(\theta, \tilde{P}_1) = \tilde{P}_1 \varepsilon_1(\theta) + (1 - \tilde{P}_1) \varepsilon_2(\theta) \quad (4)$$

is the nominal classification error of a classifier with threshold θ .

Prop. 1 Given a classification threshold θ , its robustness to fractional uncertainties in the prior-probabilities around nominal priors (Eq. 1), increases linearly with the accepted classification error ε_c over the range $\tilde{\varepsilon}(\theta) \leq \varepsilon_c \leq \max[\varepsilon_1(\theta), \varepsilon_2(\theta)]$ and is given by (Appendix A):

$$\hat{\alpha}(\varepsilon_c, \theta) = \begin{cases} 0 & \varepsilon_c < \tilde{\varepsilon}(\theta) \\ \frac{\varepsilon_c - \tilde{\varepsilon}(\theta)}{\tilde{P}_1 |\varepsilon_1(\theta) - \varepsilon_2(\theta)|} & \tilde{\varepsilon}(\theta) \leq \varepsilon_c \leq \max[\varepsilon_1(\theta), \varepsilon_2(\theta)] \\ \infty & \varepsilon_c > \max[\varepsilon_1(\theta), \varepsilon_2(\theta)] \end{cases} \quad (5)$$

Equation (5) implies that poor performance ($\varepsilon_c > \max[\varepsilon_1(\theta), \varepsilon_2(\theta)]$) can be guaranteed due to its infinite robustness. In contrast, any performance demands in

excess of the nominal demands ($\varepsilon_c < \tilde{\varepsilon}(\theta)$) possess zero robustness.

In order to demonstrate the results obtained in (5) we used a well-studied medical classification problem, for the identification of patients with Normal and Failure heart condition [4]. We used the Thoracic Fluid Content (TFC) as a classification indicator assuming normal distributions for the class-conditioned pdfs with parameters taking from [4]. The two types of errors ($\varepsilon_1(\theta)$ and $\varepsilon_2(\theta)$) corresponds, respectively, to false-alarm (FA) for failure decision while the heart condition is normal, and miss-detection (MD) for normal decision while it is prone to failure. MD which may result in unidentified heart failure is considered more important than FA, which may result in unnecessary checkup. Assuming that MD is nine times more important, we set the weights to: $\lambda_1 = 0.9$ and $\lambda_2 = 0.1$. The effect of choosing different weights is explored in Figure 4.

Figure 2 presents two robustness curves as a function of the performance criterion ε_c . Note (1) the linear increase in robustness (larger $\hat{\alpha}$) with decrease in performance (larger ε_c) and (2) the zero robustness for small ε_c values. In addition we observe curve crossing. Let ε_x denote the value of ε_c at which the curves cross. If superior performance is desired (i.e. $\varepsilon_c < \varepsilon_x$), then threshold θ_2 is preferred over θ_1 . On the other hand, if ε_c in excess of ε_x is acceptable, than θ_1 is preferred.

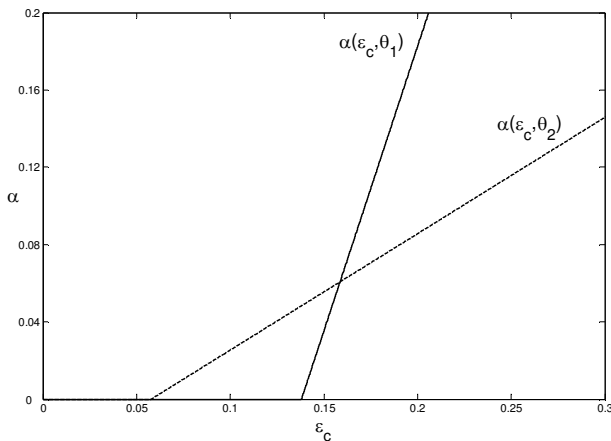


Figure 2: Robustness vs. performance criteria

The next proposition defines the conditions for crossing of the robustness curves.

Let θ_b denote the Bayesian threshold for which the weighted classification error is minimized, i.e.

$$\theta_b = \arg \min_{\theta} \varepsilon(\theta, \tilde{P}_1),$$

and let θ_l denote the limiting threshold satisfying:

$$\varepsilon_1(\theta_l) = \varepsilon_2(\theta_l)$$

Prop. 2 The linear regions of the robustness curves in Eq. (5) cross each other for any two thresholds within the range (Appendix B):

$$\min[\theta_l, \theta_b] \leq \theta \leq \max[\theta_l, \theta_b] \quad (6)$$

Figure 3 presents the robustness curves for different threshold values over the range $\theta_l < \theta \leq \theta_b$. Notice that within that range, an upper envelop is formed by the crossing curves. This envelop determines a decision rule for choosing the threshold θ which maximizes the robustness for a given performance demand.

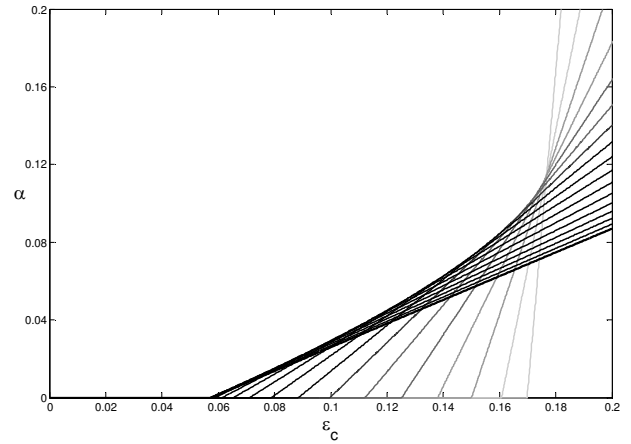


Figure 3: Robustness vs. performance criteria – crossing region

Def. 2 Given a performance criteria ε_c

$$\theta_{rs} = \arg \max_{\theta} \hat{\alpha}(\varepsilon_c, \theta) \quad (7)$$

is the robust-satisficing (RS) threshold which maximizes the robustness to uncertainty in the priors for specified performance.

Let $\tilde{\varepsilon}(\theta_b)$ denote the nominal classification error of the Bayesian classifier (Eq. 4) and $\tilde{\varepsilon}(\theta_l)$ - the nominal error of the limiting threshold classifier. The following proposition asserts the existence of a RS threshold.

Prop. 3 Given a performance requirement over the range $\tilde{\varepsilon}(\theta_b) < \varepsilon_c < \tilde{\varepsilon}(\theta_l)$, there is a RS threshold, θ_{rs} , within the range defined in (6), that maximizes the robustness and is given implicitly by (Appendix C):

$$\lambda_2 p_2(\theta_{rs})(\varepsilon_1(\theta_{rs}) - \varepsilon_c) + \lambda_1 p_1(\theta_{rs})(\varepsilon_2(\theta_{rs}) - \varepsilon_c) = 0 \quad (8)$$

where $p_1(\theta_{rs})$ and $p_2(\theta_{rs})$ are the pdfs of ω_1 and ω_2 respectively, for the RS threshold.

The next proposition shows that for any positive robustness the Bayesian and the RS thresholds differ.

Prop. 4 Let $\hat{\alpha}_{rs}(\varepsilon_c) = \hat{\alpha}(\varepsilon_c, \theta_{rs}(\varepsilon_c))$ for values of ε_c for which $0 \leq \hat{\alpha}_{rs} \leq 1$. The RS threshold $\theta_{rs}(\varepsilon_c)$ is the Bayesian optimal threshold for the prior probability given by: $P_1 = (1 - \hat{\alpha}_{rs})\tilde{P}_1$.

4 Threshold Curves

Figure 4 presents thresholds curves, specifying the RS threshold as a function of the performance criteria ε_c for different λ_2 values with $\lambda_1 = 1 - \lambda_2$. Specifying the relative importance of the two types of errors and the total acceptable error, determines the decision threshold level such that the robustness to uncertainty in the priors is maximized.

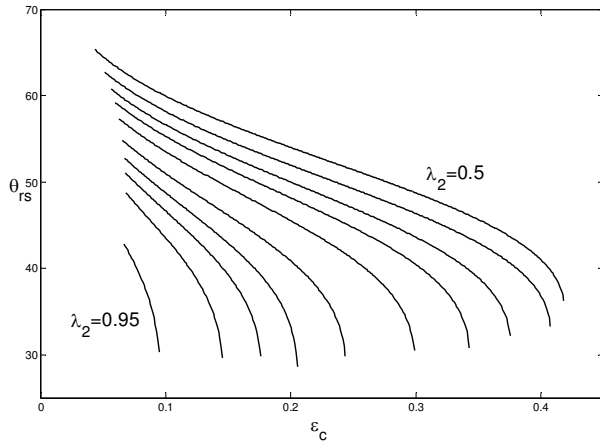


Figure 4: Robust-satisficing threshold vs. performance criteria – operating curve

5 Summary and Conclusions

In this paper we propose a robust approach for the standard two-class classification problem under severe uncertainty in the prior probabilities. Information-gap decision theory was applied to compare the robustness of the standard Bayesian classifier with other classifiers and determine the robust-satisficing classifier that maximizes the robustness.

The issue of uncertainty in the prior probability has been addressed by Augustin (2001), by using interval-valued

probabilities. Given the limits of the interval, the max-min approach was developed to determine the classifier that maximizes the minimum performance. The info-gap model presented here extends the interval-valued probabilities by considering an unbounded set of nested intervals. Thus, while the max-min approach requires knowledge of the interval-values probabilities, the info-gap approach requires only nominal values.

The info-gap model facilitates the development of a robust-satisficing classifier that maximizes the uncertainty for which a desired performance can be guaranteed. In situations where such a critical demand exists, the robust-satisficing classifier is expected to have higher chance of success in achieving that demand (exact proof of this claim is under work).

Future extensions of this work would include: 1) Extension to multi-category case. 2) Uncertainties in the probability distribution functions.

Acknowledgements

The authors gratefully acknowledge the support of the Abramson Center for the Future of Health, Houston, and the Funds for the promotion of research at the Technion.

References

- [1] Duda R. O. Hart, P. E. and Stork, D. G., *Pattern Classification*. New York: Wiley-Interscience 2001.
- [2] Ben-Haim Y, Fault detection with uncertain priors, 47th Structures, Structural Dynamics and Materials (SDM) conference, AIAA, Newport Rhode Island, 1-4 May 2006.
- [3] Ben-Haim Y, *Info-gap decision theory: decisions under severe uncertainty*, 2nd edition, Academic Press, London, 2006.
- [4] Packer M, et al, Utility of Impedance Cardiography for the Identification of Short-Term Risk of Clinical Decompensation in Stable Patients With Chronic Heart Failure, *Journal of the american college of cardiology* Vol. 47, No.11, 2006.
- [5] Augustin T, On decision making under ambiguous prior and sampling information, 2nd Int. Symp. Imprecise Probabilities and Their Applications, 9-15, 2001.

Appendix A – Proof of Proposition 1

Here we derive the robustness (3) of classifiers with threshold θ to uncertainty in the priors. The probability of error classification (2) is used as a cost function with the goal of limiting it to a desired performance criteria ε_c . The uncertainty in the priors is given by (1).

The inner maximization in (3) depends on whether $\varepsilon_1(\theta)$ is larger or smaller than $\varepsilon_2(\theta)$. When $\varepsilon_1(\theta) < \varepsilon_2(\theta)$, maximization of $\varepsilon(\theta, P_1)$ is obtained by setting $P_1 = \max[0, (1-\alpha)\tilde{P}_1]$, while when $\varepsilon_1(\theta) > \varepsilon_2(\theta)$, the maximization is obtained by setting $P_1 = \min[(1+\alpha)\tilde{P}_1, 1]$.

Given $\varepsilon_1(\theta) < \varepsilon_2(\theta)$, if $\alpha < 1$:

$$\begin{aligned} \max_{P_1 \in U(\alpha, \tilde{P}_1)} \varepsilon(\theta, P_1) = \\ (1-\alpha)\tilde{P}_1\varepsilon_1(\theta) + [1-(1-\alpha)\tilde{P}_1]\varepsilon_2(\theta) = \\ \tilde{\varepsilon}(\theta) + \alpha\tilde{P}_1(\varepsilon_2(\theta) - \varepsilon_1(\theta)) \leq \varepsilon_c \end{aligned}$$

and the robustness is:

$$\hat{\alpha}(\varepsilon_c, \theta) = \frac{\varepsilon_c - \tilde{\varepsilon}(\theta)}{\tilde{P}_1(\varepsilon_2(\theta) - \varepsilon_1(\theta))}, \quad \varepsilon_c > \tilde{\varepsilon}(\theta)$$

If $\varepsilon_c \leq \tilde{\varepsilon}(\theta)$, the performance demands are not satisfied by the nominal condition and the robustness is set to zero.

If $\alpha > 1$:

$$\max_{P_1 \in U(\alpha, \tilde{P}_1)} \varepsilon(\theta, P_1) = \varepsilon_2(\theta) \leq \varepsilon_c$$

for which the robustness is infinite. By the same method we derive the robustness given that $\varepsilon_1(\theta) > \varepsilon_2(\theta)$. Combining this, we derive the robustness function given in (5).

Appendix B – Proof of Proposition 2

Here we prove crossing of the robustness curves and provide the conditions for crossing. The following analysis refers to simply connected decision regions, i.e. one decision boundary, and can be extended to several decision regions (see the remark in the end of Appendix C).

Part 1

Let θ_l denote the limiting threshold for which $\varepsilon_1(\theta_l) = \varepsilon_2(\theta_l)$. Given the Bayesian threshold satisfies $\varepsilon_1(\theta_b) < \varepsilon_2(\theta_b)$, then $\theta_l < \theta_b$ and for each θ within the range $\theta_l < \theta < \theta_b$: $|\varepsilon_2(\theta) - \varepsilon_1(\theta)| < |\varepsilon_2(\theta_b) - \varepsilon_1(\theta_b)|$. Moreover, the nominal error of any threshold θ satisfies: $\tilde{\varepsilon}(\theta) \geq \tilde{\varepsilon}(\theta_b)$. Hence the robustness curve of a threshold $\theta_l < \theta < \theta_b$ has a steeper slope than the robustness curve of the Bayesian threshold while intersecting the ε_c axis at a larger point. Consequently, it crosses the robustness curve of the Bayesian threshold at some point.

The above analysis can be extended to any two thresholds θ_1 and θ_2 satisfying $\theta_l < \theta_1 < \theta_2 \leq \theta_b$ to obtain the following inequalities:

$$\begin{aligned} 1. |\varepsilon_1(\theta_1) - \varepsilon_2(\theta_1)| < |\varepsilon_1(\theta_2) - \varepsilon_2(\theta_2)| \\ 2. \tilde{\varepsilon}(\theta_1) > \tilde{\varepsilon}(\theta_2) \end{aligned} \quad (10)$$

Inequalities (10) assure curve crossing of the robustness curves of any two thresholds within the range $\theta_l < \theta \leq \theta_b$.

For $\theta > \theta_b$: $|\varepsilon_1(\theta) - \varepsilon_2(\theta)| > |\varepsilon_1(\theta_b) - \varepsilon_2(\theta_b)|$ while $\tilde{\varepsilon}(\theta) > \tilde{\varepsilon}(\theta_b)$ and together with (5) curve crossing is eliminated.

Given θ_b satisfying $\varepsilon_1(\theta_b) > \varepsilon_2(\theta_b)$, then $\theta_b < \theta_l$ and by the same method it can be shown that curve crossing is guaranteed for any two thresholds within the range $\theta_b \leq \theta < \theta_l$.

Part 2

Here we show that crossing of the robustness curves for any two thresholds within the range defined in (6) occurs at the linear region in (5).

Given the Bayesian threshold satisfies $\varepsilon_1(\theta_b) < \varepsilon_2(\theta_b)$ and two thresholds $\theta_l < \theta_1 < \theta_2 \leq \theta_b$, the robustness curves of the two threshold cross (part 1). The robustness value at which the curves cross can be extracted from:

$$\varepsilon_c(\hat{\alpha}_c, \theta_1) = \varepsilon_c(\hat{\alpha}_c, \theta_2) \quad (11)$$

to obtain:

$$\hat{\alpha}_c = \frac{\tilde{\varepsilon}(\theta_2) - \tilde{\varepsilon}(\theta_1)}{\tilde{\varepsilon}(\theta_2) - \tilde{\varepsilon}(\theta_1) + \varepsilon_2(\theta_1) - \varepsilon_2(\theta_2)} \quad (12)$$

Using the second inequality in (10) and the fact that $\varepsilon_2(\theta_1) < \varepsilon_2(\theta_2)$, the robustness crossing value satisfies

$\hat{\alpha}_c < 1$, which assure crossing in the linear region (Appendix A).

Using the same considerations it can be shown that for θ_b satisfying $\varepsilon_1(\theta_b) > \varepsilon_2(\theta_b)$ and two thresholds within the range $\theta_b \leq \theta_l < \theta_l$, the robustness crossing value is within the linear range.

Appendix C – Proof of Proposition 3

Here we prove the existence of a robust-satisficing threshold θ_{rs} which maximizes the robustness for a given performance criteria ε_c . As in appendix B, the analysis refers to simply connected decision regions.

Given θ_b satisfying $\varepsilon_1(\theta_b) < \varepsilon_2(\theta_b)$, the robust-satisficing threshold can be found from:

$$\left. \frac{\partial \hat{\alpha}(\varepsilon_c, \theta)}{\partial \theta} \right|_{\theta_{rs}} = \left. \frac{f_1(\theta)}{f_2(\theta)} \right|_{\theta_{rs}} = 0 \quad (13)$$

where

$$f_1(\theta) = \lambda_2 p_2(\theta) (\varepsilon_1(\theta) - \varepsilon_c) \tilde{P}_1 + \lambda_1 p_1(\theta) (\varepsilon_2(\theta) - \varepsilon_c) \tilde{P}_1 \quad (14)$$

and $f_2(\theta) = \tilde{P}_1^2 (\varepsilon_2(\theta) - \varepsilon_1(\theta))^2 > 0 \quad \forall \theta \neq \theta_l$.

This sets the following condition:

$$f_1(\theta_{rs}) = \lambda_2 p_2(\theta_{rs}) (\varepsilon_1(\theta_{rs}) - \varepsilon_c) \tilde{P}_1 + \lambda_1 p_1(\theta_{rs}) (\varepsilon_2(\theta_{rs}) - \varepsilon_c) \tilde{P}_1 = 0 \quad (15)$$

Claim: Equation (15) has a solution given that $\tilde{\varepsilon}(\theta_b) < \varepsilon_c < \tilde{\varepsilon}(\theta_l)$.

Proof: Substituting $\theta = \theta_l$ in (14) gives:

$$f_1(\theta_l) = \tilde{P}_1 (\lambda_2 p_2(\theta_l) + \lambda_1 p_1(\theta_l)) (\tilde{\varepsilon}(\theta_l) - \varepsilon_c) > 0 \quad (16) \\ \forall \varepsilon_c < \tilde{\varepsilon}(\theta_l)$$

The bayesian threshold is known to satisfy the following relation:

$$p_1(\theta_b) = p_2(\theta_b) \frac{\lambda_2 (1 - \tilde{P}_1)}{\lambda_1 \tilde{P}_1} \quad (17)$$

Substituting $\theta = \theta_b$ in (14) gives:

$$f_1(\theta_b) = \lambda_2 p_2(\theta_b) (\tilde{\varepsilon}(\theta_b) - \varepsilon_c) < 0 \\ \forall \varepsilon_c > \tilde{\varepsilon}(\theta_b) \quad (18)$$

Thus given $p_1(\theta)$ and $p_2(\theta)$ continues and ε_c within the range $\tilde{\varepsilon}(\theta_b) < \varepsilon_c < \tilde{\varepsilon}(\theta_l)$, there is a threshold $\theta_l < \theta < \theta_b$ which satisfies (15) and sets the robustness first derivative (with respect to θ) to zero.

In addition, given that $f_1(\theta_l) > 0$ and $f_1(\theta_b) < 0$, where $\theta_l < \theta_b$ (since $\varepsilon_1(\theta_b) < \varepsilon_2(\theta_b)$), then at least one of thresholds satisfying (15), defines a maximum. The threshold at which the robustness is maximal determines the robust-satisficing threshold, θ_{rs} .

By the same method it can be shown that given θ_b satisfying $\varepsilon_1(\theta_b) > \varepsilon_2(\theta_b)$, there is a robust-satisficing threshold, $\theta_b \leq \theta_{rs} < \theta_l$, which maximizes the robustness for each ε_c within the range $\tilde{\varepsilon}(\theta_b) < \varepsilon_c < \tilde{\varepsilon}(\theta_l)$.

Remark: As mentioned, the analysis presented in Appendices B & C refers to simply connected decision regions, i.e. classification based on one decision boundary. When there are multiple decision boundaries, the decision space can be separated into several sub-spaces, each containing one decision boundary, where the analysis holds. Thus, the analysis can be extended to multiple decision boundaries by treating each decision boundary separately