As this is valid for any Borel-measurable set C, the distribution of (X, Y) is determined by (μ, η). The function η is sometimes called the *a posteriori probability*.

Any function $g : \mathcal{R}^d \to \{0, 1\}$ defines a *classifier* or a *decision function*. The error probability of g is $L(g) = \mathbf{P}\{g(X) \neq Y\}$. Of particular interest is the Bayes decision function

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

This decision function minimizes the error probability.

Theorem 2.1. *For any decision function $g : \mathcal{R}^d \to \{0, 1\}$,*

$$\mathbf{P}\{g^*(X) \neq Y\} \leq \mathbf{P}\{g(X) \neq Y\},$$

that is, g^* is the optimal decision.

PROOF. Given $X = x$, the conditional error probability of any decision g may be expressed as

$$\begin{aligned} &\mathbf{P}\{g(X) \neq Y | X = x\} \\ &\quad = 1 - \mathbf{P}\{Y = g(X) | X = x\} \\ &\quad = 1 - (\mathbf{P}\{Y = 1, g(X) = 1 | X = x\} + \mathbf{P}\{Y = 0, g(X) = 0 | X = x\}) \\ &\quad = 1 - \left(I_{\{g(x)=1\}} \mathbf{P}\{Y = 1 | X = x\} + I_{\{g(x)=0\}} \mathbf{P}\{Y = 0 | X = x\}\right) \\ &\quad = 1 - \left(I_{\{g(x)=1\}} \eta(x) + I_{\{g(x)=0\}} (1 - \eta(x))\right), \end{aligned}$$

where I_A denotes the indicator of the set A. Thus, for every $x \in \mathcal{R}^d$,

$$\begin{aligned} &\mathbf{P}\{g(X) \neq Y | X = x\} - \mathbf{P}\{g^*(X) \neq Y | X = x\} \\ &\quad = \eta(x) \left(I_{\{g^*(x)=1\}} - I_{\{g(x)=1\}}\right) + (1 - \eta(x)) \left(I_{\{g^*(x)=0\}} - I_{\{g(x)=0\}}\right) \\ &\quad = (2\eta(x) - 1) \left(I_{\{g^*(x)=1\}} - I_{\{g(x)=1\}}\right) \\ &\quad \geq 0 \end{aligned}$$

by the definition of g^*. The statement now follows by integrating both sides with respect to $\mu(dx)$. □

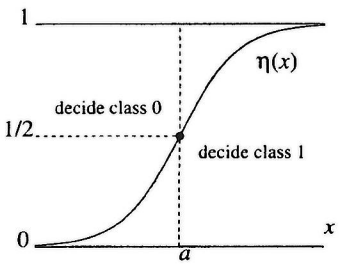

FIGURE 2.1. *The Bayes decision in the example on the left is 1 if $x > a$, and 0 otherwise.*

REMARK. g^* is called the Bay
the Bayes probability of erro
reveals that

$$L(g) = 1 - \mathbf{E}$$

and in particular,

$$L^* = 1 - \mathbf{E}\{I_{\{\eta(}$$

We observe that the a post

$$\eta(x) =$$

minimizes the squared error
$f : \mathcal{R}^d \to \mathcal{R}$:

$$\mathbf{E}\{(\eta(X$$

To see why the above inequa

$$\begin{aligned} &\mathbf{E}\{(f(X) - Y)^2 | \\ &\quad = \mathbf{E}\{(f(x) - \\ &\quad = (f(x) - \eta(\\ &\qquad + \mathbf{E}\{(\eta(X) \\ &\quad = (f(x) - \eta(\end{aligned}$$

The conditional median, i.e.,
$Y|\}$ is even more closely rela

2.2 A Simple Exan

Let us consider the predictio
when given a number of imp
$Y = 0$ stand for failure. The
per week. This, in itself, is n
because for that we would ne
mind, health, and social habit
probably monotonically incr
$c > 0$, say, our problem woul

$$g^*(x) = \{$$