1.3.2 Orthogonality Principle

There are two important conclusions that follow from the proof of Thm. 1.3.1, namely, the orthogonality properties (1.3.5) and (1.3.6). The first one states that the difference

$$\boldsymbol{x} - \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$$

is orthogonal to any function $g(\cdot)$ of $\boldsymbol{y}$. Now since we already know that the conditional expectation, $\mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$, is the optimal least-mean-squares estimator of $\boldsymbol{x}$, we can re-state this result by saying that the estimation error $\tilde{\boldsymbol{x}}$ is orthogonal to any function of $\boldsymbol{y}$,

$$\boxed{\mathsf{E}\,\tilde{\boldsymbol{x}}\, g(\boldsymbol{y}) = 0} \tag{1.3.12}$$

We shall sometimes use a geometric notation to refer to this result and write instead

$$\boxed{\tilde{\boldsymbol{x}} \;\perp\; g(\boldsymbol{y})} \tag{1.3.13}$$

where the symbol $\perp$ is used to signify that the two random variables are orthogonal; a schematic representation of this orthogonality property is shown in Fig. 1.6.

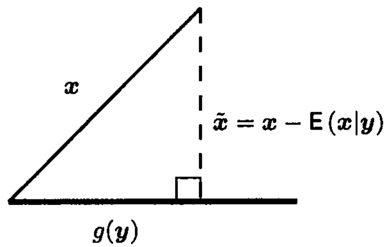

Figure 1.6. The orthogonality condition: $\tilde{\boldsymbol{x}} \perp g(\boldsymbol{y})$.

Relation (1.3.13) admits the following interpretation. It states that the optimal estimator $\hat{\boldsymbol{x}} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$ is such that the resulting error, $\tilde{\boldsymbol{x}}$, is orthogonal to (and, in fact, also uncorrelated with) any transformation of the data $\boldsymbol{y}$. In other words, the optimal estimator is such that no matter how we modify the data $\boldsymbol{y}$, there is no way we can extract additional information in order to reduce the variance of $\tilde{\boldsymbol{x}}$ any further other than the information that has already been extracted by $\hat{\boldsymbol{x}}$. This is because any additional processing of $\boldsymbol{y}$ will remain uncorrelated with $\tilde{\boldsymbol{x}}$.

The second orthogonality property (1.3.6) is a special case of (1.3.13). It states that

$$\boxed{\tilde{\boldsymbol{x}} \perp \hat{\boldsymbol{x}}}$$

That is, the estimation error is orthogonal to (or uncorrelated with) the estimator itself. This is a special case of (1.3.13) since $\hat{\boldsymbol{x}}$ is a function of $\boldsymbol{y}$ by virtue of the result $\hat{\boldsymbol{x}} = \mathsf{E}(\boldsymbol{x}|\boldsymbol{y})$.

In summary, the optimal least-mean-squares estimator is such that the estimation error is orthogonal to the estimator and, more generally, to any function of the observation. It turns out that the converse statement is also true so that the orthogonality condition (1.3.13) is in fact a *defining* property of optimality in the least-mean-squares sense.

Theorem 1.3.2 (Orthogonality condition) Given two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, an estimator $\hat{\boldsymbol{x}} = h(\boldsymbol{y})$ is optimal in the least-mean-squares sense (1.3.3) if, and only if, $\hat{\boldsymbol{x}}$ is unbiased (i.e., $\mathsf{E}\,\hat{\boldsymbol{x}} = \bar{x}$) and $\boldsymbol{x} - \hat{\boldsymbol{x}} \perp g(\boldsymbol{y})$ for any function $g(\cdot)$.

Proof: One direction has already been proven prior to the statement of the theorem, namely, if $\hat{\boldsymbol{x}}$ is the optimal estimator and hence, $\hat{\boldsymbol{x}} = \mathsf{E}(\boldsymbol{x}|\boldsymbol{y})$, then we already know from (1.3.13) that $\tilde{\boldsymbol{x}} \perp g(\boldsymbol{y})$, for any $g(\cdot)$. Moreover, we know from Thm. 1.3.1 that this estimator is unbiased.

Conversely, assume $\hat{\boldsymbol{x}}$ is an unbiased estimator for $\boldsymbol{x}$ and that it satisfies $\boldsymbol{x} - \hat{\boldsymbol{x}} \perp g(\boldsymbol{y})$, for any $g(\cdot)$. Define the random variable $\boldsymbol{z} = \hat{\boldsymbol{x}} - \mathsf{E}(\boldsymbol{x}|\boldsymbol{y})$ and let us show that it is the zero variable with probability one. For this purpose, we note first that $\boldsymbol{z}$ is zero mean since

$$\mathsf{E}\,\boldsymbol{z} = \mathsf{E}\,\hat{\boldsymbol{x}} \;-\; \mathsf{E}\,(\mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})) \;=\; \bar{x} - \bar{x} \;=\; 0$$

Moreover, from (1.3.5) we have $\boldsymbol{x} - \mathsf{E}(\boldsymbol{x}|\boldsymbol{y}) \perp g(\boldsymbol{y})$ and, by assumption, we have $\boldsymbol{x} - \hat{\boldsymbol{x}} \perp g(\boldsymbol{y})$ for any $g(\cdot)$. Subtracting these two conditions we conclude that

$$\boldsymbol{z} \;\perp g(\boldsymbol{y})$$

which is the same as $\mathsf{E}\,\boldsymbol{z}g(\boldsymbol{y}) = 0$. Since the variable $\boldsymbol{z}$ itself is a function of $\boldsymbol{y}$, we choose $g(\boldsymbol{y}) = \boldsymbol{z}$ to get $\mathsf{E}\,\boldsymbol{z}^2 = 0$. We thus find that $\boldsymbol{z}$ is zero mean and has zero variance, so that, from Remark 1 at the end of Section 1.1, we conclude that $\boldsymbol{z} = 0$, or equivalently, $\hat{\boldsymbol{x}} = \mathsf{E}(\boldsymbol{x}|\boldsymbol{y})$, with probability one.

$\diamond$