

Nearly Optimal Statistical Recognition and Learning

M.I.Schlesinger , E.V.Vodolazskiy

International Research and Training Center of Informational Technologies and Systems, Ukrainian Academy of Sciences, Kiev

schles@irtc.org.ua, waterlaz@gmail.com

Abstract. *We formulate problems of generative learning and recognition without learning in a common framework of complex hypothesis testing. Based on arguments from multi-criteria optimization, we identify strategies that are improper for solving these problems and derive a common form of the remaining strategies. We show that some widely used approaches to recognition and learning are improper in this sense. We then propose a generalized formulation of the recognition and learning problem which embraces the whole range of sizes of the learning sample, including the zero size. Learning becomes a special case of recognition without learning. We define the concept of nearly optimal Bayesian strategy, being a solution to the formulated problem. On several illustrative cases, the strategy is shown to be superior to the widely used learning methods based on maximal likelihood estimation.*

Keywords

complex object recognition, learning, multi-criteria optimization, Bayesian strategy, small sample problem

1 Complex object recognition

Definition 1. The tuple $\langle X, Y, \Theta, p : X \times Y \times \Theta \rightarrow \mathbb{R} \rangle$ is called a complex object where X is the finite set of observable signals, Y is the finite set of hidden states, Θ is the finite set of models, $p(x, y; \theta)$ is the probability of the pair (signal, state) depending on the model. \square

Recognition means making a reasonable decision about the hidden state based on the observed signal without knowing the model. It is formalised in the following way.

Definition 2. The function $q : Y \times X \rightarrow \mathbb{R}$ is called a recognition strategy and defines a conditional probability $q(y|x)$ of making a decision that the object is in state y when observing signal x . \square

The loss of making one decision or another is defined by a loss function $w : Y \times Y \rightarrow \mathbb{R}$ and defines the loss $w(y, y')$ of making the decision that the object state is y' while the true object state is y . Let Q be the set of all recognition strategies. The mathematical expectation of the loss for some strategy $q \in Q$ on the model $\theta \in \Theta$ is called the risk $R(q, \theta)$,

$$R(q, \theta) = \sum_{x \in X} \sum_{y' \in Y} \sum_{y \in Y} q(y' | x) p(y, x; \theta) w(y, y'). \quad (1)$$

Though the risk is represented by $|\Theta|$ numbers, not by one number, some strategies are obviously worse than any other.

Definition 3. A strategy q^0 is called improper if a strategy q' exists such that the inequality $R(q^0, \theta) > R(q', \theta)$ holds for every model $\theta \in \Theta$. \square

We want to exclude all improper strategies from consideration and derive the common form of all other strategies. Such a dichotomy is possible due to the known result in multi-criteria optimization [3, Theorem 3.5]. Let $\tau : \Theta \rightarrow \mathbb{R}$ be

a function, called the weight function, satisfying $\tau(\theta) \geq 0$ for every $\theta \in \Theta$ and $\sum_{\theta \in \Theta} \tau(\theta) = 1$. Let the set of such functions be denoted by T .

Definition 4. The strategy

$$q(\tau) = \operatorname{argmin}_{q \in Q} \sum_{\theta \in \Theta} \tau(\theta) R(q, \theta)$$

is called Bayesian with respect to the weight function τ . A strategy q is called Bayesian if a weight function $\tau \in T$ exists with respect to which the strategy q is Bayesian. \square

The concept of Bayesian strategies allows to partition the set of all strategies into the set of improper strategies and the set of all other ones. This dichotomy is given by the following theorem, which states that every strategy is either Bayesian or improper.

Theorem 1. For every strategy $q^0 \in Q$, either a weight function $\tau^* \in T$ exists such that

$$q^0 = \operatorname{argmin}_{q \in Q} \sum_{\theta \in \Theta} \tau^*(\theta) R(q, \theta) \quad (2)$$

or a strategy q^* exists such that the inequality

$$R(q^*, \theta) < R(q^0, \theta) \quad (3)$$

holds for every $\theta \in \Theta$. These two properties of strategies are incompatible. \square

Now it is obvious that any reasonable strategy for complex object recognition has to minimize the weighted sum of risks

$$q^* = \operatorname{argmin}_{q \in Q} \sum_{\theta \in \Theta} \tau(\theta) R(q, \theta)$$

for some weights $\tau(\theta)$. Therefore, it has to make the decision of the type

$$y^* = \operatorname{argmin}_{y' \in Y} \sum_{y \in Y} \left[\sum_{\theta \in \Theta} \tau(\theta) p(x, y; \theta) \right] w(y, y') \quad (4)$$

for observation x .

2 Minimax and nearly optimal strategies

We excluded from consideration all obviously bad strategies, which left us only with Bayesian strategies. Deciding what particular strategy to choose from the set of all Bayesian strategies depends on the requirements on the strategy. Perhaps, the most popular requirement is the so-called minimax requirement [4]. For given sets X, Y, Θ and probabilities $p(x, y; \theta)$, $x \in X, y \in Y, \theta \in \Theta$, the strategy $q: Y \times X \rightarrow \mathbb{R}$ has to be found that minimizes c subject to the conditions

$$R(q, \theta) \leq c, \quad \theta \in \Theta. \quad (5)$$

In other words, the strategy

$$q^* = \operatorname{argmin}_{q \in Q} \max_{\theta \in \Theta} R(q, \theta) \quad (6)$$

is called a minimax strategy. This strategy is clearly Bayesian and looks reasonable. Minimax strategies are fruitfully used in the theory and practice of recognition without learning [2, 8]. However, the minimax requirement is not the only possible meaningful one. In multi-criteria decision making [9] other reasonable requirement is used that differs from the minimax one.

It is obvious that for any model θ no strategy can have the risk lower than $\min_{q \in Q} R(q, \theta)$. The strategy q^* with the risks $R(q^*, \theta) = \min_{q \in Q} R(q, \theta)$, $\theta \in \Theta$, will be called an optimal strategy. Let us define the strategy that aims to be as close to the optimal strategy as possible in the following way. For given sets X, Y, Θ and probabilities $p(x, y; \theta)$, $x \in X, y \in Y, \theta \in \Theta$, the strategy $q: Y \times X \rightarrow \mathbb{R}$ has to be found that minimizes c subject to the conditions

$$R(q, \theta) - \min_{q \in Q} R(q, \theta) \leq c, \quad \theta \in \Theta. \quad (7)$$

Definition 5. The strategy

$$q^* = \operatorname{argmin}_{q \in Q} \left[\max_{\theta \in \Theta} (R(q, \theta) - \min_{q \in Q} R(q, \theta)) \right] \quad (8)$$

is called a nearly optimal Bayesian strategy or simply a nearly optimal strategy. \square

The minimax approach (5) and the nearly optimal one (7) result in different strategies. If our task is only recognition without learning, we have no means to decide which of these two approaches is better. However, the situation becomes different when the approaches are applied to learning. Then it turns out that minimax strategies are not suitable for recognition learning at all. They simply ignore the learning sample. This defect is considered in details in the next section and we show that nearly optimal strategies are free of this defect.

3 Generative learning as a special case of complex object recognition

As in the previous sections, we consider complex objects, defined by sets X, Y, Θ and probabilities $p(x, y; \theta)$, $x \in X$, $y \in Y$, $\theta \in \Theta$. Recognition learning arises if an additional information z , called learning information, is available, taking values from a given set Z . The learning information z is random and depends on the model $\theta \in \Theta$, so that the probability $p(z; \theta)$ is defined for every $z \in Z$ and $\theta \in \Theta$. It is crucial that for a fixed model θ , the learning information z depends neither on the current state y nor on the current signal x of the object under recognition, so that $p(z, x, y; \theta) = p(z; \theta) p(x, y; \theta)$. The learning procedure observes the learning information z and constructs a recognition strategy.

In case of supervised learning, the learning information z may be a sequence $(x_1, y_1; x_2, y_2; \dots x_n, y_n)$, $Z = (X \times Y)^n$, and

$$p(z; \theta) = p(x_1, y_1, x_2, y_2, \dots, x_n, y_n; \theta) = \prod_{i=1}^n p(x_i, y_i; \theta).$$

Learning information may be a sequence $(x_1, x_2, \dots x_n)$, $Z = X^n$ in case of unsupervised learning and

$$p(z; \theta) = p(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n p(x_i; \theta).$$

Definition 6. A learning procedure $g: Z \rightarrow Q$ is a function that for any learning information $z \in Z$ defines a recognition strategy $g(z) \in Q$. \square

This means that after learning with a random information z and observing a signal x , the decision "the object is in state y " is made with probability $g(z)(y' | x)$. Recognition learning is defined as finding a learning procedure $g: Z \rightarrow Q$ from given sets X, Y, Θ, Z and probabilities $p(x, y; \theta)$ and $p(z; \theta)$. Recall that complex object recognition without learning is defined as finding a recognition strategy $q: Y \times X \rightarrow \mathbb{R}$, $q \in Q$, from given sets X, Y, Θ and probabilities $p(x, y; \theta)$. When treating the problems this way, one can see that recognition learning does not differ from recognition with no learning. Indeed, *finding the learning procedure $g: Z \rightarrow Q$ for given sets X, Y, Θ, Z and probabilities $p(x, y; \theta)$ and $p(z; \theta)$ does not differ from finding the recognition strategy $q': Y \times X' \rightarrow \mathbb{R}$ for given sets X', Y, Θ and probabilities $p'(x', y; \theta)$* . The former is reduced to the latter by defining

$$\begin{aligned} X' &= Z \times X, \\ p'(x', y; \theta) &= p(z; \theta) p(x, y; \theta), \\ q'(y' | x') &= q'(y' | z, x) = g(z)(y' | x). \end{aligned}$$

Thus, the recognition learning is a special case of complex object recognition without learning. In this special case, the data available for recognition consist of two parts. The first part (observed signal) depends both on the hidden object state and on the unknown model. The second part (learning data) directly depends on the model and does not depend on the state when the model is fixed.

Theorem 1, proved in the general form, remains valid for learning procedures. Then it has the following form. As before, $R(q, \theta)$ is the risk of recognition strategy $q \in Q$ with respect to model $\theta \in \Theta$. The number $R(g(z), \theta)$ is the risk of the strategy $g(z)$, obtained by applying learning procedure $g: Z \rightarrow Q$ to learning information $z \in Z$. The risk $R(g(z), \theta)$ is random because it depends on the random information z . Let $R_G(g, \theta)$ denote the expectation of the risk over the set of all learning informations,

$$R_G(g, \theta) = \sum_{z \in Z} p(z; \theta) R(g(z), \theta). \quad (9)$$

Definition 7. A learning procedure g^* is called Bayesian if a weight function $\tau: \Theta \rightarrow \mathbb{R}$ exists such that

$$g^* = \operatorname{argmin}_{g \in G} \sum_{\theta \in \Theta} \tau(\theta) R_G(g, \theta).$$

□

Now the restriction of Theorem 1 to learning procedures reads as follows.

Theorem 2. For every learning procedure $g^0 \in G$, either a weight function $\tau^* \in T$ exists such that

$$g^0 = \operatorname{argmin}_{g \in G} \sum_{\theta \in \Theta} \tau^*(\theta) R_G(g, \theta)$$

or a learning procedure g^* exists such that for every model $\theta \in \Theta$ the inequality

$$R_G(g^*, \theta) < R(g^0, \theta)$$

holds. These two properties of learning procedures are incompatible. □

The theorem says that every learning procedure is either Bayesian or improper. In the special case when a sample $(x_1, y_1; x_2, y_2; \dots; x_n, y_n)$ is a learning information and a signal x_0 is observed, the decision y^* about the current object state has to be

$$y^* = \operatorname{argmin}_{y' \in Y} \sum_{y_0 \in Y} \sum_{\theta \in \Theta} \tau(\theta) \cdot \prod_{i=0}^n p(x_i, y_i; \theta) \cdot w(y_0, y') \quad (10)$$

for some weights $\tau(\theta)$, $\theta \in \Theta$. Strategies that are commonly used in pattern recognition are of other form. Most widely known is the method based on maximum likelihood estimation [2, 8]. According to this method, the model

$$\theta^{\text{ML}} = \operatorname{argmax}_{\theta \in \Theta} \prod_{i=1}^n p(x_i, y_i; \theta) \quad (11)$$

is found and then for an observed signal x the decision

$$y^{\text{ML}} = \operatorname{argmin}_{y' \in Y} \sum_{y \in Y} p(x, y; \theta^{\text{ML}}) \cdot w(y, y') \quad (12)$$

is made, as if the maximum likelihood model was identical to the true one. In general, *the strategy (11, 12) differs from (10) and thus it is improper in the exactly defined meaning of the word.*

We restrict our further considerations to Bayesian learning procedures, focusing in particular on the minimax procedure

$$g^{\text{minimax}} = \operatorname{argmin}_{g \in G} \max_{\theta \in \Theta} R_G(g, \theta) \quad (13)$$

and the nearly optimal procedure

$$g^{\text{nearopt}} = \operatorname{argmin}_{g \in G} \max_{\theta \in \Theta} [R_G(g, \theta) - \min_{q \in Q} R(q, \theta)]. \quad (14)$$

Both strategies belong to the Bayesian class and there is no reason to exclude them from recognition without learning. However, minimax strategies have a fundamental drawback that they sometimes do not make use of a learning sample, no matter how large. For a rather wide class of complex objects, the minimax learning procedure simply ignores the learning sample.

Theorem 3. Let for sets X, Y, Θ and probabilities $p(x, y; \theta)$, $x \in X, y \in Y, \theta \in \Theta$, a model θ^* exist such that the strategy

$$q^* = \operatorname{argmin}_{q \in Q} R(q, \theta^*) \quad (15)$$

satisfies the inequality system

$$R(q^*, \theta^*) \geq R(q^*, \theta), \quad \theta \in \Theta. \quad (16)$$

Then any set Z , any probabilities $p(z; \theta)$, $z \in Z, \theta \in \Theta$, and any learning procedure $g: Z \rightarrow Q$ satisfy the inequality

$$\max_{\theta \in \Theta} R_G(g, \theta) \geq \max_{\theta \in \Theta} R(q^*, \theta). \quad (17)$$

□

The theorem shows that there are complex objects for which the minimax approach is particularly inappropriate. Inequality (17) states that any learning procedure, however sophisticated, is useless from the minimax point of view. They cannot yield a recognition strategy that would be better than some strategy that does not use the learning sample at all.

Theorem 3 is a strong negative result and makes minimax learning useless in most cases. As shown by the following theorem similar to Theorem 3, nearly optimal learning procedures are also useless for a certain class of complex objects.

Theorem 4. Let for sets X, Y, Θ and probabilities $p(x, y; \theta)$, $x \in X, y \in Y, \theta \in \Theta$ a model θ^* exists such that the strategy

$$q^* = \operatorname{argmin}_{q \in Q} [R(q, \theta^*) - \min_{q \in Q} R(q, \theta^*)] \quad (18)$$

satisfies the inequality system

$$R(q^*, \theta^*) - \min_{q \in Q} R(q, \theta^*) \geq R(q^*, \theta) - \min_{q \in Q} R(q, \theta), \quad \theta \in \Theta. \quad (19)$$

Then any set Z , any probabilities $p(z; \theta)$, $z \in Z, \theta \in \Theta$, and any learning procedure $g: Z \rightarrow Q$ satisfy the inequality

$$\max_{\theta \in \Theta} [R_G(g, \theta) - \min_{q \in Q} R(q, \theta)] \geq \max_{\theta \in \Theta} [R(q^*, \theta) - \min_{q \in Q} R(q, \theta)].$$

□

However, the consequences of this theorem for nearly optimal learning are not so destructive as those of Theorem 3 for minimax learning. In fact, conditions (18) and (19) imply that an optimal Bayesian strategy exists for the recognized object, namely the strategy q^* . In this case, not only nearly optimal or minimax learning is useless. Any other learning approach is also useless because no recognition strategy can be better than q^* .

Conditions (18, 19) fully characterize the set of models for which nearly optimal learning is useless. For any complex model, either an optimal Bayesian strategy exists or a strategy obtained with nearly optimal learning converges to optimal Bayesian strategy in the sense of the following theorem.

Theorem 5. Let the learning procedure $g_n^*: (X \times Y)^n \rightarrow Q$ be defined by

$$g_n^* = \operatorname{argmin}_{g \in G} \max_{\theta \in \Theta} [R_G(g, \theta) - \min_{q \in Q} R(q, \theta)].$$

Then

$$\lim_{n \rightarrow \infty} \max_{\theta \in \Theta} [R_G(g_n^*, \theta) - \min_{q \in Q} R(q, \theta)] = 0. \quad (20)$$

□

In such way the concept of nearly optimal learning formalizes our requirements on learning procedures. It embraces the whole range of sample sizes, including the zero size. If the learning sample is long enough, the risks $R_G(g, \theta)$ approach the best possible risks $\min_{q \in Q} R(q, \theta)$. This property is shared by other commonly used learning procedures, such as those based on maximal likelihood estimates. But besides that, we require that the difference between $R_G(g, \theta)$ and $\min_{q \in Q} R(q, \theta)$ is as small as possible for every, even a very small, size of the learning sample.

4 Examples

We compare the nearly optimal learning procedure g^0 with the maximum likelihood learning procedure g^1 and with another learning procedure g^2 (to be defined later) on several simplest examples. Nevertheless, according to the formal definition they are already complex but the nearly optimal Bayesian learning procedures for them can be implemented almost exactly. Even for such simple complex objects, the minimax learning procedures $g^{\min\max} = \operatorname{argmin}_g \max_{\theta} R_G(g, \theta)$ are inappropriate because they ignore the learning sample.

To use the technique with a continuous model set Θ , the set was discretized. In all the examples, we have $X = \mathbb{R}$, $Y = \{1, 2\}$, and the probability density distribution $p(x, y; \theta)$ has the form

$$p(x, y; \theta) = p_y \cdot (\sqrt{2\pi})^{-1} \cdot e^{-\frac{1}{2}(x - \mu_y)^2}$$

with some unknown parameters specified later.

Example 1. [6] Let $\mu_1 = 1$, $\mu_2 = (-1)$, and only the a priori probabilities p_y , $y \in \{1, 2\}$, be unknown. Thus, the set of models is $\Theta = \{\theta \mid 0 \leq \theta \leq 1\}$ and the a priori probabilities p_y of the states are $p_1 = \theta$, $p_2 = 1 - \theta$. This is the example used by H.Robbins [5] to explain his idea of empirical Bayesian approach. We simplify the example even further by assuming that the learning information is not a signal sample (x_1, x_2, \dots, x_n) but a state sample (y_1, y_2, \dots, y_n) .

Figure 1 shows how the risks $R_G(g^1, \theta)$ and $R_G(g^0, \theta)$ depend on the model θ for the sample sizes $n = 1, 2, 4$.

It is striking how high the risk $R_G(g^1, \theta)$ is for some models. Of course, one cannot expect the risk to be too low because the learning samples are very small and thus they bring little information about the true model. Nevertheless, however small the amount of information, it is non-zero and it should be used to improve subsequent recognition, at least to some extent. But the example shows that it is much better to ignore this information than to use it with maximum likelihood learning. The example illustrates quite transparently the main drawback of maximum likelihood learning. \square

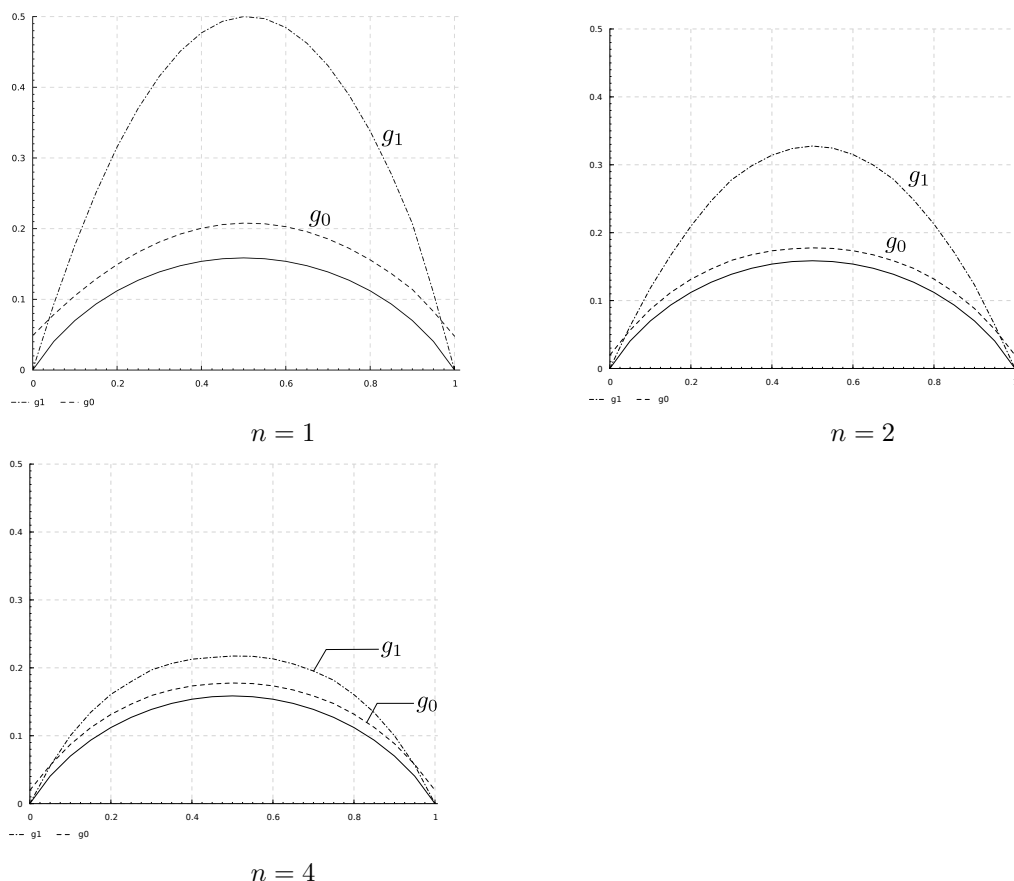


Fig. 1. EXAMPLE 1. The dependency of $R_G(g^0, \theta)$ and $R_G(g^1, \theta)$ on θ for sample sizes $n = 1, 2, 4$. The solid curve shows risk $\min_{q \in Q} R(q, \theta)$.

Example 2. For the same complex object as in Example 1, let the learning information be a sequence (x_1, x_2, \dots, x_n) rather than (y_1, y_2, \dots, y_n) . This is exactly the case considered by H.Robbins [5]. The risk $R_G(g^0, \theta)$ of nearly optimal learning is compared with the risk $R_G(g^1, \theta)$ of maximum likelihood learning. Even for this simple complex object, it is not easy to find the maximum likelihood model

$$\theta_1 = \operatorname{argmax}_{\theta \in \Theta} \sum_{i=1}^n \log \left[\theta \cdot e^{-\frac{1}{2}(x_i-1)^2} + (1-\theta) \cdot e^{-\frac{1}{2}(x_i+1)^2} \right].$$

Therefore, we also consider the heuristic procedure by H.Robbins, which is denoted by g^2 . The procedure is not based on the maximum likelihood estimate θ_1 but on the consistent estimate $\theta_2 = \frac{1}{2n} \sum_{i=1}^n x_i + \frac{1}{2}$, which can be easily calculated.

Figure 2 shows how the risks $R_G(g^0, \theta)$, $R_G(g^1, \theta)$ and $R_G(g^2, \theta)$ depend on the model θ for the sample sizes $n = 1, 2, 4$. \square

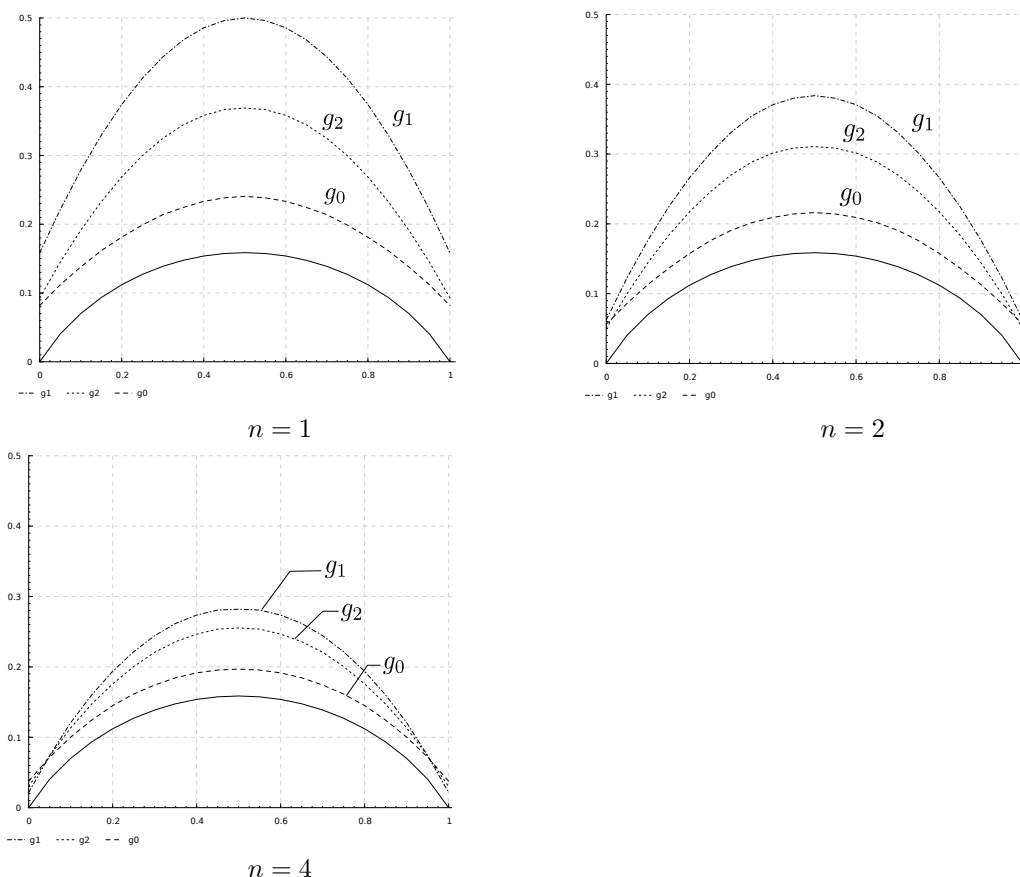


Fig. 2. EXAMPLE 2. The dependency of $R_G(g^0, \theta)$, $R_G(g^1, \theta)$, $R_G(g^2, \theta)$ on θ for sample sizes $n = 1, 2, 4$. The solid curve shows the risk $\min_{q \in Q} R(q, \theta)$.

Example 3. Here we have the same object as in the previous example but the learning information is more complicated. It is a sequence (x_1, x_2, \dots, x_n) of signals generated by the object with unknown states. This is the case of unsupervised learning [7]. Normally the so-called EM algorithms are used [1]. In this case, the maximum likelihood estimate

$$\theta_1 = \operatorname{argmax}_{\theta \in \Theta} \sum_{i=1}^n \log \left[e^{-\frac{1}{2}x_i^2} + e^{-\frac{1}{2}(x_i - \theta)^2} \right]$$

is not easy to find. Therefore, besides the maximum likelihood estimate we also consider the consistent estimate $\theta_2 = \frac{2}{n} \sum_{i=1}^n x_i$ along with the corresponding learning procedure g^2 . Figure 3 shows how the risks $R_G(g^0, \theta)$, $R_G(g^1, \theta)$, $R_G(g^2, \theta)$ depend on the model θ for the sample sizes $n = 1, 2, 4$. One can see that neither of the maximal likelihood procedure g^1 and the heuristic procedure g^2 dominates the other. However, the nearly optimal procedure g^0 dominates both g^1 and g^2 , especially for small sample sizes. We have not observed any significant difference between the compared procedures for the learning sample sizes larger than 10.

A case of an empty learning sample ($n = 0$) is of a particular interest. In this case the maximum-likelihood learning problem cannot be even formulated because any model estimate is meaningless when empirical data are absent. The minimax requirement to recognition strategy may be formally stated but it may result in a strategy that makes wrong decision with probability 0.5 for each true model. As for nearly optimal requirement, it can be formulated for this case too and results in a quite reasonable strategy. Figure 3 shows how the risk $R_G(g^0, \theta)$ depends on θ when the learning sample is absent at all.

\square

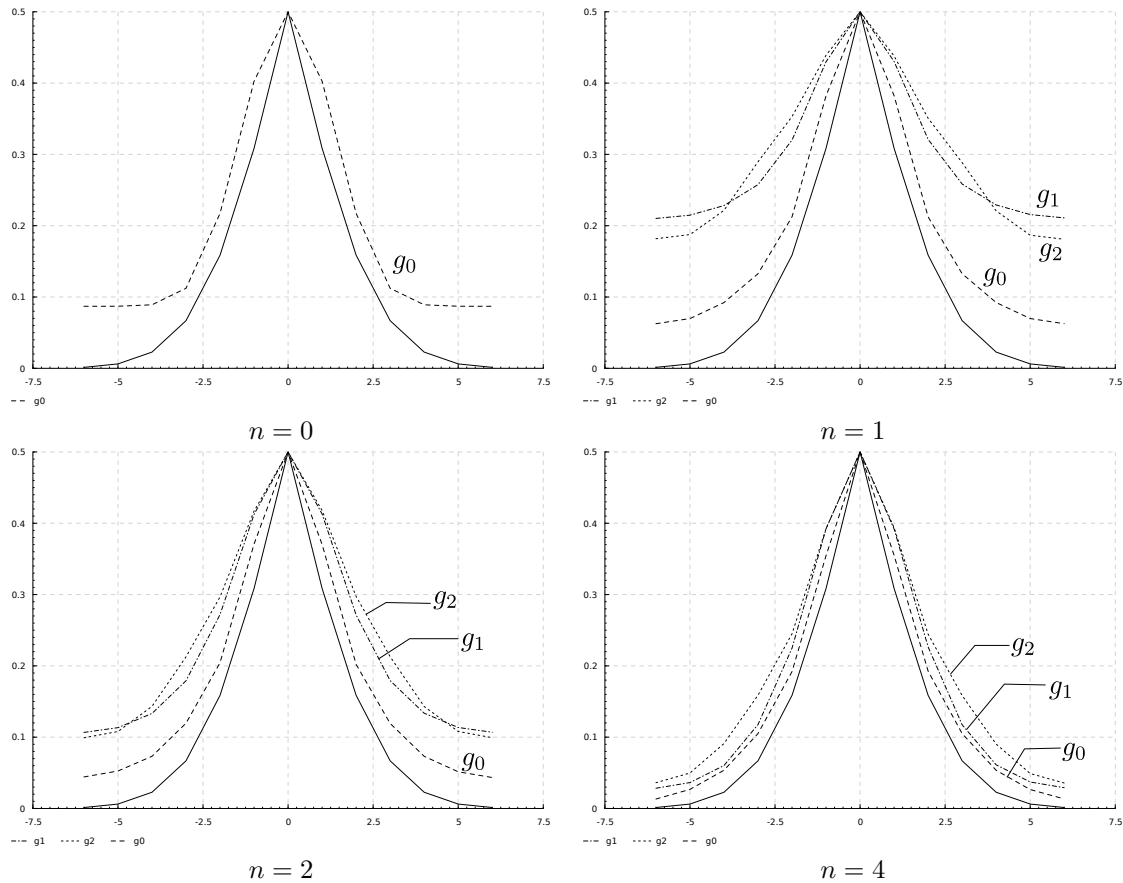


Fig. 3. EXAMPLE 3. Dependency of the risks $R_G(g^0, \theta)$, $R_G(g^1, \theta)$, $R_G(g^2, \theta)$ on the model θ for sample sizes $n = 0, 1, 2, 4$. The solid curve shows the risk $\min_{q \in Q} R(q, \theta)$.

References

- [1] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
- [2] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, 2000.
- [3] M. Ehrgott. *Multicriteria Optimization*. Lecture notes in economics and mathematical systems. Springer, 2005.
- [4] E.L.A. Lehmann. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer Verlag, 1986.
- [5] Herbert Robbins. Asymptotically Subminimax Solutions of Compound Statistical Decision Problems. In Jerzy Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 131–148. University of California Press, 1951.
- [6] M.I. Schlesinger and A. V. Bondarenko. On pattern recognition learning problem formulation. (in russian). *Control Systems and Computers*, 2:4–19, 2009.
- [7] M.I. Shlezinger. The interaction of learning and self-organization in pattern recognition. *Kibernetika*, 2:81–88, 1968.
- [8] Andrew R. Webb. *Statistical Pattern Recognition*. Wiley, 2002.
- [9] M. Zeleny. *Multiple criteria decision making*. McGraw-Hill series in quantitative methods for management. McGraw-Hill, 1982.