3. Constraint-Based Modeling of Case-Based Inference

In this chapter, we adopt a constraint-based view of the CBI hypothesis, according to which the similarity of inputs imposes a constraint on the similarity of associated outcomes in the form of a lower bound. A related inference mechanism then allows for realizing CBI as a kind of constraint propagation. We also discuss representational issues and algorithms for putting the idea of *learning* within this framework into action. The chapter is organized as follows: Section 3.1 introduces the aforementioned formalization of the CBI hypothesis. A case-based inference scheme which emerges quite naturally from this formalization is proposed in Section 3.2 and further developed in Section 3.3. Case-based learning is discussed in Section 3.4. In Section 3.5, some applications of case-based inference in the context of statistics are outlined. The chapter concludes with a brief summary and some complementary remarks in Section 3.6.

3.1 Basic concepts

3.1.1 Similarity profiles and hypotheses

Proceeding from the framework introduced in Section 2.4, the system under consideration can be thought of as the triple $(\mathcal{S}, \mathcal{R}, \varphi)$.[1] The (unknown) functional relation φ completely determines the structure of this system at the *instance level*, whereas a memory of observed cases provides only partial information. In connection with CBI, we are interested in utilizing the additional information provided by a CBI setup Σ for deriving a corresponding characterization of the system at the *similarity level*. This additional information is mainly contained in the similarity measures.

Definition 3.1 (similarity profile). Consider a CBI setup Σ. The function $h_\Sigma : D_\mathcal{S} \longrightarrow [0,1]$ defined by

$$h_\Sigma(x) \overset{\mathrm{df}}{=} \inf_{s,s' \in \mathcal{S},\, \sigma_\mathcal{S}(s,s')=x} \sigma_\mathcal{R}(\varphi(s), \varphi(s'))$$

is called the similarity profile of Σ. □

[1] This is in agreement with general systems theory, where an abstract system is defined as a relation on a set [228]. It should also be mentioned that this mathematical structure, even though formally very simple, is general enough for modeling any kind of "real" system.

The similarity profile h_Σ is the "fingerprint" of the system $(\mathcal{S}, \mathcal{R}, \varphi)$ at the similarity level and (partly) defines the *similarity structure* of the setup Σ. Just like φ determines dependencies at the instance level, h_Σ depicts relations between degrees of similarity: Given the similarity of two inputs, it provides a lower bound to the similarity of the respective outcomes. It hence conveys a precise idea of the extent to which the application at hand actually meets the CBI hypothesis, i.e, it can be interpreted as a (multi-dimensional) quantification of the degree to which the CBI hypothesis holds true.[2] In fact, the stronger the similarity structure of $(\mathcal{S}, \mathcal{R}, \varphi)$ is developed, the more constraining the similarity profile will be. Note that the domain and the codomain of h_Σ are one-dimensional, whereas $\mathcal{S}$ and $\mathcal{R}$ are generally of higher dimension. Thus, a similarity profile represents knowledge about the system structure φ in a *condensed* form. (We will return to the relation between h_Σ and φ in Section 3.2.)

Needless to say, the similarity profile of a CBI setup will generally be unknown. This leads us to introduce the related concept of a *similarity hypothesis*.

Definition 3.2 (similarity hypothesis). A similarity hypothesis is identified by a function $h : [0,1] \longrightarrow [0,1]$ (and similarity measures $\sigma_\mathcal{S}, \sigma_\mathcal{R}$).[3] The intended meaning of the hypothesis h (or, more precisely, the hypothesis $(h, \sigma_\mathcal{S}, \sigma_\mathcal{R})$) is the assumption that

$$\forall\, s, s' \in \mathcal{S} \;:\; (\sigma_\mathcal{S}(s,s') = x) \Rightarrow (\sigma_\mathcal{R}(\varphi(s), \varphi(s')) \geq h(x))\,. \tag{3.1}$$

A hypothesis h is called *stronger* than a hypothesis h' if $h' \leq h$ and $h \not\leq h'$. Let Σ be a CBI setup with similarity profile h_Σ. We say that Σ *satisfies* the hypothesis h, or that h is *admissible*, if $h(x) \leq h_\Sigma(x)$ for all $x \in D_\mathcal{S}$. □

A similarity hypothesis h is thought of as an approximation of a similarity profile h_Σ. It thus defines a formal model of the CBI hypothesis for the application at hand, as represented by the setup Σ. In Section 2.4, it has already been mentioned that different types of hypotheses might be of different expressive power. This remark becomes more obvious now. Since a similarity profile h_Σ is a condensed representation of φ, a similarity hypothesis h will generally be less constraining than a hypothesis which is directly related to φ, that is, an approximation $\widehat{\varphi} : \mathcal{S} \longrightarrow \mathcal{R}$ of φ. Yet, a similarity profile has a relatively simple structure which facilitates the formulation, derivation, or adaptation of hypotheses (cf. Section 3.4).

A similarity hypothesis can originate from different sources. Firstly, it might express a purely heuristic quantification of the CBI assumption. In this case, it is often expressed as "*the more* similar two inputs are, *the more* similar the corresponding outputs are." The concept of a similarity profile, as introduced above,

[2] There are obvious ways of deriving a one-dimensional quantification, for example a (weighted) mean of the values $\{h_\Sigma(x) \,|\, x \in D_\mathcal{S}\}$.

[3] Note that is would be sufficient to define a hypothesis on $D_\mathcal{S}$. Quite often, however, it will indeed appear more convenient to let $\mathrm{dom}(h) = [0,1]$, especially if $|D_\mathcal{S}|$ is large. Otherwise, $\mathrm{dom}(h) = [0,1]$ can still be assumed without loss of generality, simply by letting $h(x) = 1$ for all $x \notin D_\mathcal{S}$.

reveals that this kind of formulation implicitly makes a stronger assumption than the simple "similar inputs imply similar outputs" hypothesis. Namely, it suggests the function h_Σ associated with a setup Σ to be increasing, or at least non-decreasing. More precisely, this formulation may be understood as "the more similar two inputs are, the larger is the lower similarity bound of the associated outcomes." Therefore, we call h a *strict hypothesis* if it is a non-decreasing function. Moreover, we say that a setup Σ satisfies the CBI hypothesis in the strict sense if h_Σ is non-decreasing.

Secondly, it is a natural idea to consider the acquisition of hypotheses as a problem of (empirical) *learning*, i.e., to learn hypotheses from observed (pairs of) cases. This way, CBI combines *instance-based learning*, which essentially corresponds to the collection of cases, and *model-based learning*, namely the learning of similarity hypotheses. The assumption that the CBI hypothesis applies in a strict sense serves an (additional) inductive bias in connection with the model-based aspect of learning. In fact, since it suffices to consider non-decreasing functions h as candidates for approximating h_Σ, the hypothesis space $\mathcal{H}$ under consideration is reduced correspondingly.

Remark 3.3. Observe that the CBI hypothesis can be enforced to hold true in the strict sense by adapting the similarity measure $\sigma_\mathcal{S}$ (and, hence, changing the CBI setup correspondingly). In fact, one can always determine a bijective mapping $f : D_\mathcal{S} \longrightarrow D_\mathcal{S}$ such that h_Σ is non-decreasing if $\sigma_\mathcal{S}$ is replaced by $\sigma'_\mathcal{S} = f \circ \sigma_\mathcal{S}$. Seen from this perspective, one may always assume that the strict CBI hypothesis is actually valid and simply explain the opposite by the inadequacy of the (originally) chosen similarity measure.[4] □

Remark 3.4. A strict similarity hypothesis h is closely related to the concept of a *gradual inference rule* in fuzzy set-based approximate reasoning. A gradual rule is a special kind of fuzzy rule of the form "the more X is in A, the more Y is in B," where A and B are fuzzy sets modeling some gradual concepts. The application of this kind of fuzzy rule in the context of CBI will be discussed in Section 6.1. □

Example 3.5. Fig. 3.1 shows the similarity profiles h_{Σ_1} and h_{Σ_2} of the CBI setups Σ_1 and Σ_2 defined by the (repetitive) ILP problems in Example 2.5.[5] As can be seen, these functions are indeed increasing. Moreover, the similarity structure of Σ_1 is developed more strongly than the structure of Σ_2. The same remarks apply to the setups Σ_1^* and Σ_2^*, the similarity profiles of which are shown in the same figure. □

[4] Though this would again degrade the CBI hypothesis to a trivial assumption (see the discussion in Section 2.2.3).

[5] We plotted the polygonal line connecting the points $\{(x, h_\Sigma(x)) \,|\, x \in D_\mathcal{S}\}$.

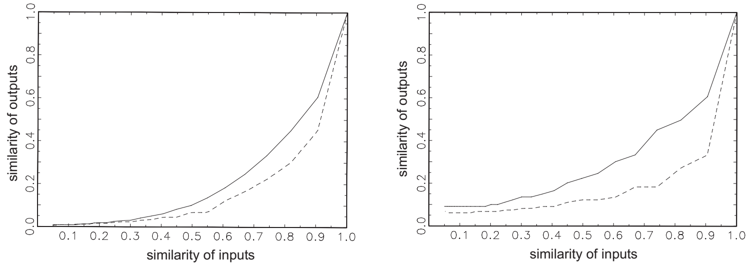

Fig. 3.1. Left: Similarity profiles h_{Σ_1} (solid line) and h_{Σ_2} of the (repetitive) ILP problems defined in Example 2.5. Right: Similarity profiles $h_{\Sigma_1^*}$ (solid line) and $h_{\Sigma_2^*}$ defined in the same example.

EXAMPLE 3.6. Let $(\mathcal{S}, \Delta_{\mathcal{S}})$ and $(\mathcal{R}, \Delta_{\mathcal{R}})$ be metric spaces and suppose $\varphi : \mathcal{S} \longrightarrow \mathcal{R}$ to be Lipschitz continuous, i.e., there is a constant $L > 0$ such that $\Delta_{\mathcal{R}}(\varphi(s), \varphi(s')) \leq L\Delta_{\mathcal{S}}(s, s')$ for all $s, s' \in \mathcal{S}$. Moreover, suppose $\sigma_{\mathcal{S}}$ to be $\Delta_{\mathcal{S}}$-related (via f) and $\sigma_{\mathcal{R}}$ to be $\Delta_{\mathcal{R}}$-related (via g). Then, $h = g \circ Lf^{-1}$ is an admissible hypothesis for the corresponding CBI setup. □

REMARK 3.7. It has already been suggested in Definition 3.2 to characterize a similarity hypothesis in a more precise way, namely as a triple $(h, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$. Indeed, the essential aspect in connection with a hypothesis h is the fact that it relates degrees x of the similarity scale $D_{\mathcal{S}}$ (resp. the unit interval) to degrees $y = h(x)$ of the scale $D_{\mathcal{R}}$ (resp. the unit interval). Thus, the meaning of a hypothesis h strongly depends on the similarity functions $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$ in the sense that changing these functions would also change the meaning of h. Particularly, two hypotheses h, h' as well as the similarity profiles associated with two systems $(\mathcal{S}, \mathcal{R}, \varphi)$ and $(\mathcal{S}, \mathcal{R}, \varphi')$ are not comparable unless the underlying similarity measures are identical. □

3.1.2 Generalized similarity profiles

There are two characteristic features of case-based reasoning which are worth mentioning in connection with the concept of a similarity profile and which suggest to generalize Definition 3.1. As will be seen, this generalization makes a similarity profile more suitable for supporting certain (case-based) problem solving strategies.

Firstly, CBI methods do usually not take the complete memory $\mathcal{M}$ of cases into account when solving a new problem. Rather, the attention is drawn to the most similar cases,[6] since less similar cases are assumed to hardly improve the solution (prediction) quality. Indeed, utilizing the complete memory may affect the system

[6] The problem of searching these cases efficiently is closely related to the topics of *case retrieval* and *case indexing* (cf. Section 2.2).

efficiency adversely, at least if the latter does not only take the quality of a solution (prediction) into consideration but also the time which has been spend on deriving it [355, 353]. Secondly, CBI problems might be solved repeatedly by using the same memory $\mathcal{M}$ of cases. One may then benefit from the fact that the memory does not change by adjusting the formalization of the similarity structure to $\mathcal{M}$.

As already announced above, we are now going to introduce some generalizations of Definition 3.1 which are motivated by the two aforementioned aspects.

Definition 3.8 (k-selection). Let $\mathcal{M} = (\langle s_1, r_1\rangle, \ldots, \langle s_n, r_n\rangle)$, $k \leq n$, and consider an input $s_0 \in \mathcal{S}$. The *extended* k-selection $\mathcal{N}_k^{ex}(\mathcal{M}, s_0)$ is defined as a subsequence of $\mathcal{M}$ such that

$$\begin{aligned} \langle s_\jmath, r_\jmath\rangle \in & \mathcal{N}_k^{ex}(\mathcal{M}, s_0) \;\Leftrightarrow \\ & \operatorname{card}\{1 \leq \imath \leq n \,|\, \sigma_{\mathcal{S}}(s_0, s_\jmath) < \sigma_{\mathcal{S}}(s_0, s_\imath)\} < k. \end{aligned}$$

The k-selection $\mathcal{N}_k(\mathcal{M}, s_0)$ is defined such that

$$\begin{aligned} \langle s_\jmath, r_\jmath\rangle \in & \mathcal{N}_k(\mathcal{M}, s_0) \;\Leftrightarrow \\ & \operatorname{card}\{1 \leq \imath < \jmath \,|\, \langle s_\imath, r_\imath\rangle \in \mathcal{N}_k^{ex}(\mathcal{M}, s_0)\} < k. \end{aligned}$$

Thus, $\mathcal{N}_k(\mathcal{M}, s_0)$ is exactly of length k, whereas $\mathcal{N}_k^{ex}(\mathcal{M}, s_0)$ might consist of more than k cases. □

Definition 3.9 ((n,k)-similarity profile). Consider a CBI setup Σ. We define the (n,k)-similarity profile

$$h_\Sigma^{(n,k)} : D_{\mathcal{S}} \longrightarrow [0,1]$$

associated with Σ as follows: For all $x \in D_{\mathcal{S}}$, the value $h_\Sigma^{(n,k)}(x)$ is given by the maximal value $y \in [0,1]$ such that

$$\begin{aligned} \forall\, \mathcal{M} \in \mathcal{M}^n \,\forall\, s_0 \in \mathcal{S} \;\forall\; \langle s, \varphi(s)\rangle \in \mathcal{N}_k(\mathcal{M}, s_0) \;: \\ \sigma_{\mathcal{S}}(s, s_0) = x \;\Rightarrow\; \sigma_{\mathcal{R}}(\varphi(s), \varphi(s_0)) \geq y, \end{aligned}$$

where $\mathcal{M}^n$ denotes the class of memories of size n. □

According to Definition 3.9, the concept of an (n,k)-similarity profile corresponds to statements of the following form: "Let $\mathcal{M}$ be an arbitrary memory of size n. If two inputs $s_0 \in \mathcal{S}$ and $s \in \mathcal{M}$ are x-similar and s is among the inputs in $\mathcal{M}$ which are most similar to s_0, then the similarity of the outcomes $\varphi(s_0)$ and $\varphi(s)$ is at least $h_\Sigma^{(n,k)}(x)$." We have $h_\Sigma \leq h_\Sigma^{(n,k)}$ for all $1 \leq k \leq n$, where $n \in \mathfrak{N}$ and $n \leq |\mathcal{S}|$ if $\mathcal{S}$ is finite. This inequality holds due to the fact that $h_\Sigma^{(n,k)}$ is less constrained than h_Σ, which can be grasped as follows: For $s, s_0 \in \mathcal{S}$ (and $\sigma_{\mathcal{S}}(s, s_0)$ small enough) it might happen that $s \in \mathcal{S}$ is *not relevant* for s_0 in the sense that

$$\forall\, \mathcal{M} \in \mathcal{M}^n \;:\; \langle s, \varphi(s)\rangle \notin \mathcal{N}_k(\mathcal{M}, s_0).$$

Now, if neither s is relevant for s_0 nor vice versa, the value $\sigma_{\mathcal{R}}(\varphi(s), \varphi(s_0))$ does no longer constrain the lower bound $h_{\Sigma}^{(n,k)}(\sigma_{\mathcal{S}}(s, s_0))$. Quite often, however, h_{Σ} and $h_{\Sigma}^{(n,k)}$ will differ but slightly, at least if $n - k$ is small in relation to the size of the set $\mathcal{S}$.

Remark 3.10. In connection with a "selective" CBI strategy it might be reasonable to require the most similar cases to be (pairwise) different. This amounts to considering only those memories induced by sequences of (pairwise) different inputs. Statistically speaking, a memory $\mathcal{M}$ is then determined by a random sample from $\mathcal{S}$ *without* replacement. Of course, Definition 3.9 can be modified accordingly. □

Definition 3.11 ($\mathcal{M}$-similarity profile). Consider a CBI setup Σ with memory $\mathcal{M}$. We define $h_{\Sigma}^{\mathcal{M}} : D_{\mathcal{S}} \longrightarrow [0,1]$ by means of

$$h_{\Sigma}^{\mathcal{M}}(x) \overset{\text{df}}{=} \inf_{s \in \mathcal{M}^{\downarrow}, s_0 \in \mathcal{S}, \sigma_{\mathcal{S}}(s, s_0) = x} \sigma_{\mathcal{R}}(\varphi(s), \varphi(s_0)).$$

This function is called the $\mathcal{M}$-similarity profile of Σ. □

Definition 3.12 (($\mathcal{M}, k$)-similarity profile). Consider a CBI setup Σ with memory $\mathcal{M}$. We define $h_{\Sigma}^{(\mathcal{M},k)} : D_{\mathcal{S}} \longrightarrow [0,1]$ as follows: For all $x \in D_{\mathcal{S}}$, the value $h_{\Sigma}^{(n,k)}(x)$ is given by the maximal value $y \in [0,1]$ such that

$$\begin{aligned} &\forall s_0 \in \mathcal{S}\; \forall \mathcal{T} \in \mathcal{N}_k(\mathcal{M}, s_0)\; \forall \langle s, r\rangle \in \mathcal{T} \;: \\ &\qquad (\sigma_{\mathcal{S}}(s, s_0) = x) \Rightarrow (\sigma_{\mathcal{R}}(r, \varphi(s_0)) \geq y) \end{aligned}$$

holds true. The function $h_{\Sigma}^{(n,k)}$ is called the $(\mathcal{M}, k)$-similarity profile of Σ. □

The above definitions reveal that a $(\cdot, k)$-profile corresponds to the idea of using only k of the stored cases for CBI. Likewise, passing from a similarity profile to an $(\mathcal{M}, \cdot)$-similarity profile is motivated by the idea of repeatedly using a fixed memory $\mathcal{M}$ of cases for solving CBI problems. A profile $h_{\Sigma}^{\mathcal{M}}$, for instance, corresponds to rules of the following form: "Given the memory $\mathcal{M}$ and two x-similar inputs $s_0 \in \mathcal{S}$ and $s \in \mathcal{M}$, the similarity of the outcomes $\varphi(s_0)$ and $\varphi(s)$ is at least $h_{\Sigma}^{\mathcal{M}}(x)$." The relations $h_{\Sigma} \leq h_{\Sigma}^{(n,k)} \leq h_{\Sigma}^{\mathcal{M},k}$ and $h_{\Sigma} \leq h_{\Sigma}^{\mathcal{M}} \leq h_{\Sigma}^{\mathcal{M},k}$ hold obviously true for all memories $\mathcal{M}$ and $k \leq n = |\mathcal{M}|$. Passing from a profile h_{Σ} to a profile $h_{\Sigma}^{\mathcal{M}}$ will generally have a considerable effect on the quantification of the similarity profile, and the smaller the memory $\mathcal{M}$ is, the stronger this effect will be. In fact, a profile h_{Σ} is determined by the similarity relations between *arbitrary* cases c and c', whereas c must be an element of $\mathcal{M}$ in connection with $h_{\Sigma}^{\mathcal{M}}$.

The generalization of Definition 3.2 in accordance with the generalization of similarity profiles is straightforward. We may then speak, e.g., of a similarity hypothesis related to an $\mathcal{M}$-similarity profile or to an (n,k)-profile. In subsequent sections of this chapter we will restrict ourselves mainly to the consideration of (ordinary) similarity profiles and related hypotheses, although a further generalization will be introduced in Section 3.3.2. Most often, it will be obvious how to transfer corresponding results.

3.2 Constraint-based inference

3.2.1 A constraint-based inference scheme

In this section, we shall introduce an inference scheme which emerges quite naturally from the constraint-based view of the CBI hypothesis as formalized in the previous section. Consider a CBI problem $\langle \Sigma, s_0 \rangle$ and suppose that Σ satisfies the hypothesis h. If the memory $\mathcal{M}$ contains the input s_0, i.e., if $\mathcal{M}$ contains a case $\langle s, r \rangle$ such that $s = s_0$, the correct outcome $r_0 = r$ can simply be retrieved from $\mathcal{M}$. Otherwise, we can derive the following restriction:

$$r_0 \in \widehat{\varphi}_{h,\mathcal{M}}(s_0) \overset{\mathrm{df}}{=} \bigcap_{\langle s,r \rangle \in \mathcal{M}} \mathcal{N}_{h(\sigma_{\mathcal{S}}(s,s_0))}(r), \tag{3.2}$$

where $\widehat{\varphi}_{h,\emptyset}(s_0) \overset{\mathrm{df}}{=} \mathcal{R}$ by convention and the *α-neighborhood* of an output $r \in \mathcal{R}$ is defined by the set of all outcomes r' which are at least α-similar to r:

$$\mathcal{N}_\alpha(r) \overset{\mathrm{df}}{=} \{ r' \in \mathcal{R} \,|\, \sigma_{\mathcal{R}}(r, r') \geq \alpha \}. \tag{3.3}$$

Thus, according to the constraint-based interpretation the task of case-based inference can be seen as one of deriving and representing the set (3.2), or an approximation thereof. This may become difficult if, for instance, the definition of the similarity $\sigma_{\mathcal{R}}$ and, hence, the derivation of a neighborhood are complicated. The sets (3.3) may also become large, in which case they cannot be represented by simply enumerating their elements.

In this connection, it should be noted that (3.2) remains correct if the intersection is taken over $k < n$ of the inputs $s \in \mathcal{M}^{\downarrow}$. Since less similar inputs will often hardly contribute to the precision of predictions, it might indeed be reasonable to proceed from k inputs maximally similar to s_0, especially if the intersection of neighborhoods (3.3) is computationally complex. Besides, it is worth mentioning that (3.2) can be approached efficiently by means of *parallel computation techniques*. In fact, the sets which have to be combined (via intersection) can be derived independently of each other. Moreover, the (associative) combination itself can be realized in an arbitrary order. Thus, a parallel implementation of

(3.2) is (more or less) straightforward and will enable the exploitation of relatively large memories.

Of course, while assuming the profile of a CBI setup to be unknown, one cannot guarantee the admissibility of a hypothesis h and, hence, the correctness of (3.2). That is, it might happen that $\varphi(s_0) \notin \widehat{\varphi}_{h,\mathcal{M}}(s_0)$. In fact, we might even have $\widehat{\varphi}_{h,\mathcal{M}}(s_0) = \emptyset$. Nevertheless, taking for granted that h is indeed a good approximation of h_Σ, it seems reasonable to derive $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ according to (3.2) as an approximation of $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s_0)$ (while keeping the hypothetical character of h in mind). This situation reflects the heuristic character of CBI as a problem solving method. Nevertheless, by quantifying the probability of obtaining correct predictions, our results in Section 3.4 will provide a sound basis of this approach.

A similarity profile as well as a similarity hypothesis relate degrees of similarity to one another: Given the similarity of two inputs, they conclude on the similarity of the related outcomes. Thus, the similarity relations between observed cases constitute the principal information from which a case-based inference scheme proceeds. This motivates the following definition.

Definition 3.13 (similarity structure). Consider a CBI setup Σ with $\mathcal{M}$ being the associated memory (2.29) of cases and let s_0 be a new input. The similarity structure of the CBI problem $\langle \Sigma, s_0 \rangle$ is defined by the similarity profile $(h_\Sigma, \sigma_\mathcal{S}, \sigma_\mathcal{R})$ of Σ resp. a corresponding hypothesis $(h, \sigma_\mathcal{S}, \sigma_\mathcal{R})$ together with the similarity structure

$$\mathsf{SST}(\mathcal{M}, s_0) \overset{\mathrm{df}}{=} \left\{ z_{\imath\jmath} = (x_{\imath\jmath}, y_{\imath\jmath}) \,|\, 1 \leq \imath < \jmath \leq n \right\} \cup \left\{ x_{0\jmath} \,|\, 1 \leq \jmath \leq n \right\}$$

of the *extended memory* $(\mathcal{M}, s_0)$. Here, the values $x_{\imath\jmath}$ and $y_{\imath\jmath}$ are defined as $x_{\imath\jmath} \overset{\mathrm{df}}{=} \sigma_\mathcal{S}(s_\imath, s_\jmath)$ and $y_{\imath\jmath} \overset{\mathrm{df}}{=} \sigma_\mathcal{R}(r_\imath, r_\jmath)$. We will generally assume the similarity profile h_Σ resp. the hypothesis h to be given and simply call $\mathsf{SST}(\mathcal{M}, s_0)$ the similarity structure of $\langle \Sigma, s_0 \rangle$. Moreover, we define the *partial* similarity structure $\mathsf{pSST}(\mathcal{M}, s_0)$ by the set $\{x_{0\jmath} \,|\, 1 \leq \jmath \leq n\}$. □

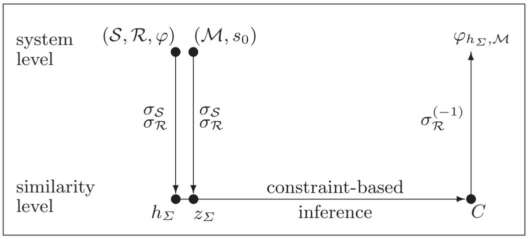
system level
$(\mathcal{S}, \mathcal{R}, \varphi)$
$(\mathcal{M}, s_0)$
$\varphi_{h_\Sigma,\mathcal{M}}$
$\sigma_\mathcal{S}$ $\sigma_\mathcal{R}$
$\sigma_\mathcal{S}$ $\sigma_\mathcal{R}$
$\sigma_\mathcal{R}^{(-1)}$
similarity level
h_Σ
z_Σ
constraint-based inference
C

Fig. 3.2. Illustration of the case-based (similarity-based) inference process.

Even though the inference scheme (3.2) is rather simple, it is worth reconsidering it from an abstract point of view. This will reveal some basic ideas of our approach to CBI, which becomes more involved within the probabilistic setting of Chapter 4. The overall CBI process as illustrated in Fig. 3.2 can be characterized as follows:

- In a first step, the problem $\langle \Sigma, s_0 \rangle$ is characterized at the *similarity level* by means of its *similarity structure*. In fact, h_Σ resp. $z_\Sigma = \mathsf{SST}(\mathcal{M}, s_0)$ can be seen as the "image" of the system $(\mathcal{S}, \mathcal{R}, \varphi)$ resp. the (extended) memory $(\mathcal{M}, s_0)$ under the transformation defined by the similarity measures $\sigma_\mathcal{S}$ and $\sigma_\mathcal{R}$. This mapping realizes a projection from an often high-dimensional (and non-numerical) *instance space* $\mathcal{S} \times \mathcal{R}$ into the two-dimensional similarity space $D_\mathcal{S} \times D_\mathcal{R}$, which is usually more accessible to analytical methods. Still, this projection is not (information-)theoretically justified like, say, dimension reduction techniques such as principal component analysis in statistics. Rather, it is guided by the heuristic assumption that the similarity structure of the problem $\langle \Sigma, s_0 \rangle$ represents useful information.
- The main step of the CBI process is then to utilize the similarity structure of the problem for constraining the unknown outcome r_0 at the similarity level. The corresponding constraints C are *implicit* in the sense that they are expressed in terms of the (bilateral) concept of similarity, i.e., they do not refer to the output itself.
- Finally, the observed outputs come into play. In conjunction with a transformation $\sigma_\mathcal{R}^{(-1)} : \mathcal{R} \times [0,1] \longrightarrow 2^\mathcal{R}$, which is inversely related to $\sigma_\mathcal{R}$ via
$$\sigma_\mathcal{R}^{(-1)}(r, \alpha) \overset{\mathrm{df}}{=} \{r' \in \mathcal{R} \,|\, \sigma_\mathcal{R}(r, r') \geq \alpha\}, \tag{3.4}$$
they are used for translating the constraints C at the similarity level into constraints on outcomes at the instance level. According to (3.2), these constraints are combined conjunctively by means of an intersection.

Two characteristics of case-based (similarity-based) inference as introduced above are worth mentioning. Firstly, CBI is *indirect* in the sense that the given information is not used for drawing inferences about the unknown output r_0 directly. Rather, it is used for deriving evidence concerning similarity degrees $\sigma_\mathcal{R}(r_0, r_k)$, which are then translated into evidence about outcomes. Secondly, CBI is *local* in the sense that the rules (3.1) associated with a hypothesis h derive evidence concerning the value r_0 from single cases. These pieces of evidence have still to be combined in order to obtain the constraint implied by the complete memory $\mathcal{M}$. Within the deterministic framework of this chapter, the combination of evidence derived from different cases is accomplished by (3.2), i.e., by means of a simple intersection of sets. As will be seen in Chapter 4, this problem becomes more complicated within a probabilistic setting.

Needless to say, the stronger the similarity structure of a setup Σ is developed, the more successful CBI will be. Within our framework, we have quantified the degree to which the CBI hypothesis holds true for the setup Σ by means of the similarity profile h_Σ. This quantification, however, may appear rather restrictive. In fact, the derivation of valid predictions according to (3.2) necessitates the use of lower similarity bounds, which leads to a kind of worst case analysis. The existence of some "exceptional" pairs of cases, for instance, might call for small values $h_\Sigma(x)$ of the similarity profile h_Σ. Consequently, the predictions (3.2) which reflect the success of the CBI process (cf. Section 3.4) might become imprecise even though the similarity structure of Σ is otherwise strongly developed. This observation serves as a main motivation for the consideration of *local* similarity profiles in Section 3.3.2 and for the probabilistic generalization of the constraint-based approach which we will turn to in Chapter 4.

From a mathematical point of view, the decisive aspect of the inference scheme in Fig. 3.2 is the fact that it is based on the analysis, not of the original data, but of *transformed data* which depicts a certain *relation* between original observations. Considering these observations in pairs, the original data (represented by the memory $\mathcal{M} \subseteq \mathcal{S} \times \mathcal{R}$) is transformed into the new set of data

$$\left\{(\sigma_{\mathcal{S}}(s,s'), \sigma_{\mathcal{R}}(r,r')) \,|\, \langle s,r\rangle, \langle s',r'\rangle \in \mathcal{M}\right\} . \tag{3.5}$$

As opposed to functional relations related to the instance level, which are mappings of the form $\mathcal{S} \longrightarrow \mathcal{R}$, the result h of the analysis of (3.5) provides information about the *relation* $\sigma_{\mathcal{R}}(\varphi(s), \varphi(s'))$ between outcomes $\varphi(s), \varphi(s')$, given the *relation* $\sigma_{\mathcal{S}}(s,s')$ between inputs s and s'. Then, given an observation $\langle s, r\rangle$ and a new input s_0 and, hence, the relation $\sigma_{\mathcal{S}}(s,s_0)$, h is used for specifying the relation $\sigma_{\mathcal{R}}(r,r_0)$ between r and $r_0 = \varphi(s_0)$. Finally, the inverse transformation $\sigma_{\mathcal{R}}^{(-1)}$ is used for translating information about r and $\sigma_{\mathcal{R}}(r,r_0)$ into information about r_0 itself. Moreover, the *combination of evidence* concerning r_0 becomes necessary if this kind of information has been derived from different observations $\langle s_1, r_1\rangle, \ldots, \langle s_n, r_n\rangle$.

In our case, the relation between observations corresponds to their similarity, the function h defines an (estimated) lower bound in the form of (an approximation of) the similarity profile, and the combination of evidence is realized by the intersection of individual predictions. This, however, is by no means compulsory. Indeed, one might think of basing inference procedures on alternative specifications, such as the differences $\sigma_{\mathcal{S}}(s,s') = s - s'$ and $\sigma_{\mathcal{R}}(r,r') = r - r'$.[7] Then, a least squares approximation h of the transformed data provides an estimation of the difference between two outcomes, given the difference between the respective inputs. Examples of this kind of inference can, e.g., be found in economic analysis where a functional relation is often assumed, not between the economic quantities themselves, but between the (temporal) *change* of these quantities. Economic time series $(x_1, \ldots, x_T)$, for instance, are often analyzed in terms of (first-order) differences $\Delta t_k = t_{k+1} - t_k$. Likewise, in preference analysis, a frequently encountered problem is to induce an absolute rating of given entities (in terms of utility

[7] In this example, $\mathcal{S}$ and $\mathcal{R}$ are assumed to be numerical, of course.

degrees) based on pairwise comparisons expressing to what extent one object is preferred to a second one.

Remark 3.14. The non-deterministic setting of Section 2.4.2 takes account of the fact that an input $s \in \mathcal{S}$ does not determine a unique outcome or that observed outputs might be imprecise. A respective generalization of the inference scheme based on (3.2) will be discussed in Section 3.2.2 below. Simple types of imprecision, however, can also be incorporated directly into (3.2). Suppose for instance, that an output cannot be observed exactly but only up to a certain (similarity) degree α of precision. That is, an observed case $\langle s, r \rangle$ does not imply $\varphi(s) = r$ but only $\varphi(s) \in \mathcal{N}_{1-\alpha}(r)$. Moreover, suppose that $\sigma_{\mathcal{R}}$ is $\top$-transitive, i.e., $\top(\sigma_{\mathcal{R}}(r, r'), \sigma_{\mathcal{R}}(r', r'')) \leq \sigma_{\mathcal{R}}(r', r'')$ for all $r, r', r'' \in \mathcal{R}$ (cf. Section 2.3). We then obtain

$$\varphi(s_0) \in \bigcap_{\langle s,r \rangle \in \mathcal{M}} \mathcal{N}_{\top(h_\Sigma(\sigma_{\mathcal{S}}(s,s_0)),1-\alpha)}(r), \tag{3.6}$$

for all $s_0 \in \mathcal{S}$ as a valid generalization of (3.2). Observe that (3.6) might be interesting in connection with non-deterministic CBI problems, namely when having to use "estimated cases" $\langle s, \widehat{\mu} \rangle$ due to the problem that the true measure μ might not be observable (cf. Section 2.4.2). In fact, this inference scheme can be applied if a minimal similarity between the true measure μ and the estimation $\widehat{\mu}$ is guaranteed. □

3.2.2 Non-deterministic problems

Within the non-deterministic setting of Section 2.4.2, a similarity profile h_Σ of a setup Σ is defined by replacing the similarity measure over outputs, $\sigma_{\mathcal{R}}$, by a similarity measure over probability distributions, $\sigma_{\mathcal{P}}$:

$$h_\Sigma : D_{\mathcal{S}} \longrightarrow [0,1]\,,\; x \mapsto \inf_{s,s' \in \mathcal{S},\, \sigma_{\mathcal{S}}(s,s')=x} \sigma_{\mathcal{P}}(\varphi(s), \varphi(s')).$$

Then, a similarity hypothesis h corresponds to the assumption that

$$\forall\, s, s' \in \mathcal{S} \;:\; \sigma_{\mathcal{S}}(s, s') = x \Rightarrow \sigma_{\mathcal{P}}(\varphi(s), \varphi(s')) \geq h(x)$$

holds true for all $x \in [0,1]$. Given a memory $\mathcal{M}$ of cases $\langle s_k, \mu_k \rangle$ $(1 \leq k \leq n)$, the inference scheme (3.2) presents itself in the form

$$\mu_0 \in \widehat{\varphi}_{h,\mathcal{M}}(s_0) \overset{\text{df}}{=} \bigcap_{\langle s,\mu \rangle \in \mathcal{M}} \mathcal{N}_{h(\sigma_{\mathcal{S}}(s,s_0))}(\mu), \tag{3.7}$$

where μ_0 is the probability measure associated with the new input s_0 and $\mathcal{N}_\alpha(\mu) \overset{\text{df}}{=} \{\mu' \in \mathcal{P}(\mathcal{R}) \,|\, \sigma_{\mathcal{P}}(\mu, \mu') \geq \alpha\}$ for $\mu \in \mathcal{P}(\mathcal{R})$ and $0 \leq \alpha \leq 1$. Thus, the set $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ now defines a class of probability measures, namely the measures which are considered as being possible in connection with the unknown measure μ_0.

Upper and lower probability bounds. For a memory $\mathcal{M}$ and a new input $s_0 \in \mathcal{S}$, the set $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ as defined in (3.7) corresponds to a *set* of probability measures. Instead of the inference result $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ itself, which might have a relatively complicated structure, one might be interested in the lower and upper probability of *individual outputs* $r \in \mathcal{R}$ according to this set, i.e.

$$\mu_0^{\downarrow}(r) = \min_{\mu \in \widehat{\varphi}_{h,\mathcal{M}}(s_0)} \mu(r) \quad \text{and} \quad \mu_0^{\uparrow}(r) = \max_{\mu \in \widehat{\varphi}_{h,\mathcal{M}}(s_0)} \mu(r). \tag{3.8}$$

Let us, therefore, consider a particular (but still reasonable) choice of the similarity $\sigma_{\mathcal{P}}$ which supports an efficient derivation of these probability bounds:

$$\sigma_{\mathcal{P}}(\mu, \mu') \overset{\text{df}}{=} 1 - f\left(\max_{r \in \mathcal{R}} |\mu(r) - \mu'(r)| \right) \tag{3.9}$$

for all $\mu, \mu' \in \mathcal{P}(\mathcal{R})$, where $f : [0,1] \longrightarrow [0,1]$ is (strictly) increasing.[8] The constraint on μ_0 induced by the k-th case $\langle s_k, \mu_k \rangle$ is now given in the form of an *interval probability* $[\mu_{0k}^l, \mu_{0k}^u]$, where

$$\mu_{0k}^l(r) = \max\left\{\mu_k(r) - f^{-1}(1 - \sigma_{\mathcal{S}}(s_0, s_k)), 0\right\}, \tag{3.10}$$

$$\mu_{0k}^u(r) = \min\left\{\mu_k(r) + f^{-1}(1 - \sigma_{\mathcal{S}}(s_0, s_k)), 1\right\}, \tag{3.11}$$

and

$$[\mu_{0k}^l, \mu_{0k}^u] \overset{\text{df}}{=} \{\mu \in \mathcal{P}(\mathcal{R}) \,|\, \forall\, r \in \mathcal{R} : \mu_{0k}^l(r) \le \mu(r) \le \mu_{0k}^u(r)\}. \tag{3.12}$$

Suppose $\widehat{\varphi}_{h,\mathcal{M}}(s_0) \neq \emptyset$ for the overall constraint (3.7). The latter is then also an interval probability:

$$\widehat{\varphi}_{h,\mathcal{M}}(s_0) = [\mu_0^l, \mu_0^u], \tag{3.13}$$

where

$$\mu_0^l(r) = \max_{1 \le k \le n} \mu_{0k}^l(r), \quad \mu_0^u(r) = \min_{1 \le k \le n} \mu_{0k}^u(r) \tag{3.14}$$

for all $r \in \mathcal{R}$. It deserves mentioning that the representation of an interval probability in the sense of (3.12) is not unique. In general, it is possible to represent a given class of probability measures μ over a set X by means of different intervals $[\mu^l, \mu^u]$ (i.e., lower and upper envelopes $\mu^l : X \longrightarrow [0,1]$ and $\mu^u : X \longrightarrow [0,1]$ such that $\mu^l \le \mu^u$). In fact, the intervals $[\mu_0^l(r_1), \mu_0^u(r_1)]$ are not necessarily minimal, i.e., the lower and upper bounds (3.14) do not necessarily correspond to the optimal bounds (3.8). That is, it might be possible that $\mu_0^l(r_1) < \mu_0^{\downarrow}(r_1)$ or $\mu_0^{\uparrow}(r_1) < \mu_0^u(r_1)$ and, hence, that one can increase $\mu_0^l(r_1)$ or reduce $\mu_0^u(r_1)$ for some $r_1 \in \mathcal{R}$ without changing the associated class (3.13) of probability measures. In other words, it might happen that $\mu_0^l(r_1)$ (resp. $\mu_0^u(r_1)$) is actually not attained by any measure $\mu \in \widehat{\varphi}_{h,\mathcal{M}}(s_0)$. In the case of finite $\mathcal{R}$, the optimal individual bounds $\mu_0^{\downarrow}(r_1)$ and $\mu_0^{\uparrow}(r_1)$ can be found by solving two simple linear programming problems:

[8] The maximum in (3.9) obviously exists.

$$\text{minimize (maximize) } \mu_0(r_1) \quad \text{s.t.} \quad \begin{cases} \mu_0^l(r) \leq \mu_0(r) \leq \mu_0^u(r) & (r \in \mathcal{R}) \\ \mu_0(r) \geq 0 \quad (r \in \mathcal{R}) \\ \sum_{r \in \mathcal{R}} \mu_0(r) = 1 \end{cases}$$

REMARK 3.15. The bounds (3.10) and (3.11) associated with a single constraint are already optimal. This can be seen as follows. Let $\alpha_0 = f^{-1}(1 - \sigma_{\mathcal{S}}(s_0, s_k))$ and $\alpha_1 = \min\{\mu_k(r_1), \alpha_0\}$ for some $r_1 \in \mathcal{R}$. That is, $\mu_{0k}^l(r_1) = \mu_k(r_1) - \alpha_1$. If $\alpha_1 = \alpha_0$ then $\mu_k(r_1) \geq \alpha_0$, i.e., there is some $r_2 \in \mathcal{R}$ such that $\mu_k(r_2) \leq 1 - \alpha_0$. The probability measure μ defined by

$$\mu(r) = \begin{cases} \mu_k(r) - \alpha_0 & \text{if} \quad r = r_1 \\ \mu_k(r) + \alpha_0 & \text{if} \quad r = r_2 \\ \mu_k(r) & \text{if} \quad r_1 \neq r \neq r_2 \end{cases}$$

is then an element of $[\mu_{0k}^l, \mu_{0k}^u]$, i.e., the lower bound $\mu_{0k}^l(r_1) = \mu_k(r_1) - \alpha_0$ is indeed attained. Now, suppose $\alpha_1 < \alpha_0$ which means that $\mu_{0k}^l(r_1) = 0$. Since $\mu_k(r_1) = 1 - \sum_{r_1 \neq r \in \mathcal{R}} \mu_k(r)$ and $\mu_k(r_1) < \alpha_0$ it is obviously possible to distribute the probability mass $\alpha_1 = \mu_k(r_1)$ over the elements $r \neq r_1$ such that the measure μ defined by

$$\mu(r) = \begin{cases} 0 & \text{if} \quad r = r_1 \\ \mu_k(r) + \alpha(r) & \text{if} \quad r \neq r_1 \end{cases}$$

for all $r \in \mathcal{R}$ is an element of the class $[\mu_{0k}^l, \mu_{0k}^u]$, where $\alpha(r) \geq 0$ and $\sum_{r_1 \neq r \in \mathcal{R}} \alpha(r) = \alpha_1$. Thus, the lower bound $\mu_{k0}^l(r_1) = 0$ is again attained. Analogously it is shown that the upper bound $\mu_{k0}^u(r_1)$ is always attained. □

A Maximum Likelihood approach. In Section 2.4.2, we have pointed out that it might not be possible to observe the probability measure μ associated with an input s. Rather, a case is often given in the form of a tuple $\langle s, x \rangle$, where x has been chosen at random according to μ. We shall now consider a framework which allows for deriving estimated cases $\langle s, \widehat{\mu} \rangle$ by means of a MAXIMUM LIKELIHOOD (ML) approach.

Let $\mathcal{P}(\mathcal{R})$ consist of a class of parameterized probability measures μ_θ $(\theta \in \Theta)$ and suppose that $\sigma_{\mathcal{P}} : \mathcal{P}(\mathcal{R}) \times \mathcal{P}(\mathcal{R}) \longrightarrow [0, 1]$ can be expressed as a function of parameter vectors, i.e., the similarity $\sigma_{\mathcal{P}}(\mu_\theta, \mu_{\theta'})$ can be written in terms of the parameter vectors θ and θ' for all $\theta, \theta' \in \Theta$. Thus, we can associate a parameter θ with each input s. By thinking of the parameter as an output, we can also identify Θ by the set of outputs, $\mathcal{R}$, and write $\sigma_{\mathcal{P}}(\mu_\theta, \mu_{\theta'}) = \sigma_{\mathcal{R}}(\theta, \theta')$.[9]

Now, consider a non-deterministic CBI problem. Suppose that n cases $\langle s_k, x_k \rangle$ $(1 \leq k \leq n)$ have been observed. A reasonable approach to estimating the probability measures μ_k associated with the inputs s_k is to maximize the likelihood function

[9] Observe that this assumption does not exclude (3.9).

$$\lambda : (\theta_1, \ldots, \theta_n) \mapsto \prod_{1 \leq k \leq n} \mu_{\theta_k}(x_k)$$

subject to the constraints

$$\forall\, 1 \leq \imath, \jmath \leq n \, : \, \sigma_{\mathcal{R}}(\theta_\imath, \theta_\jmath) \geq \sigma_{\mathcal{S}}(s_\imath, s_\jmath).$$

That is, we let $\widehat{\mu}_k = \mu_{\widehat{\theta}_k}$, where the parameter vectors $\widehat{\theta}_1, \ldots, \widehat{\theta}_n$ denote the (constrained) ML estimations. The measure μ_0 associated with a new input s_0 is then estimated according to

$$\mu_0 \in \bigcap_{1 \leq k \leq n} \mathcal{N}_{h(\sigma_{\mathcal{S}}(s_0, s_k))}(\widehat{\mu}_k).$$

3.3 Case-based approximation

Suppose a hypothesis h (with associated similarity functions $\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}$) and a memory $\mathcal{M}$ to be given. By applying (3.2) to all $s \in \mathcal{S}$ (not only to one input $s_0 \in \mathcal{S}$), we obtain a set-valued mapping $\widehat{\varphi}_{h,\mathcal{M}} : \mathcal{S} \longrightarrow 2^{\mathcal{R}}$:[10]

$$\widehat{\varphi}_{h,\mathcal{M}} : s \mapsto \bigcap_{\langle s', r' \rangle \in \mathcal{M}} \mathcal{N}_{h(\sigma_{\mathcal{S}}(s, s'))}(r'). \tag{3.15}$$

It is readily shown that $\widehat{\varphi}_{h,\mathcal{M}}$ defines an outer approximation of φ in the sense that $\varphi(s) \in \widehat{\varphi}_{h,\mathcal{M}}(s)$ for all $s \in \mathcal{S}$ if the hypothesis h is admissible. The mapping $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}$, induced by the similarity structure of a CBI setup, can be seen as a simplified but imprecise representation of the system structure φ. We call $\widehat{\varphi}_{h,\mathcal{M}}$ a *case-based approximation* (CBA) of φ. Clearly, the stronger the (admissible) hypothesis h is, the more precise the approximation $\widehat{\varphi}_{h,\mathcal{M}}$ becomes. The CBA obtained for the similarity profile h_Σ, $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}$, is the smallest outer approximation of φ in the sense that $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s) \subseteq \widehat{\varphi}_{h,\mathcal{M}}(s)$ holds true for all $s \in \mathcal{S}$ and admissible hypotheses h.

REMARK 3.16. Definition (3.15) is not exactly in agreement with our CBI approach in the sense that we may have $\widehat{\varphi}_{h,\mathcal{M}}(s) \neq \{r\}$ for some case $\langle s, r \rangle \in \mathcal{M}$. That is, the prediction $\widehat{\varphi}_{h,\mathcal{M}}(s)$ might contain additional outcomes even though the output r could be retrieved from the memory. It can easily be verified, however, that $\widehat{\varphi}_{h,\mathcal{M}}(s) = \{r\}$ is guaranteed if both measures $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$ are separating. Clearly, a further way of ensuring $\widehat{\varphi}_{h,\mathcal{M}}(s) = \{r\}$ is to modify the definition of a case-based approximation as follows: $\widehat{\varphi}_{h,\mathcal{M}}(s)$ is determined according to (3.15) only if $s \notin \mathcal{M}^{\downarrow}$, otherwise the output is retrieved from $\mathcal{M}$ and is hence given by $\{\varphi(s)\}$. □

[10] This mapping corresponds in some way to what is called the *extensional concept description* in instance-based learning [11].

Let us again mention that (3.2) resp. (3.15) are easily generalized such that only $k < n$ of the most similar cases (represented by a sub-memory $\mathcal{M}' \subseteq \mathcal{M}$) are used for constraining the outcome. Then, we can define an approximation $\widehat{\varphi}_{h,\mathcal{M},k} : \mathcal{S} \longrightarrow 2^{\mathcal{R}}$ by means of

$$\widehat{\varphi}_{h,\mathcal{M},k} : s \mapsto \bigcap_{\langle s',r' \rangle \in \mathcal{T}(s)} \mathcal{N}_{h(\sigma_{\mathcal{S}}(s,s'))}(r'), \tag{3.16}$$

where $\mathcal{T}(s) \overset{\mathrm{df}}{=} \mathcal{N}_k(\mathcal{M}, s)$ or $\mathcal{T}(s) \overset{\mathrm{df}}{=} \mathcal{N}_k^{ex}(\mathcal{M}, s)$.

3.3.1 Properties of case-based approximation

It deserves mentioning that the similarity measures principally play the role of *ordinal* concepts within our approach.[11] According to (3.2), the set $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ depends only on the relative order of similarity degrees, as specified by the hypothesis h (cf. Remark 3.7). In other words, the sets $D_{\mathcal{S}}$ and $D_{\mathcal{R}}$ can be interpreted as linearly ordered scales of similarity for which only the ordering of the grades of similarity is important. In fact, the numerical encoding is just a matter of convenience and the interval $[0, 1]$ could be replaced by any other linearly ordered scale. In fact, the inference scheme (3.2) can even be generalized in a straightforward way to similarity measures which are defined on a (complete) lattice structure [56, 283].

In order to make the ordinal character of similarity more explicit let us call two similarity measures σ and σ' (defined over a set A) *coherent* if

$$\sigma(a, b) \leq \sigma(c, d) \Leftrightarrow \sigma'(a, b) \leq \sigma'(c, d) \tag{3.17}$$

holds true for all $a, b, c, d \in A$. This definition is in accordance with the relational approach to similarity discussed in Section 2.3 (coherent similarity measures induce the same relation R).

Lemma 3.17. Let $\sigma : A \times A \longrightarrow [0, 1]$ and $\sigma' : A \times A \longrightarrow [0, 1]$ be coherent similarity measures and let $X = \{\sigma(a, b) \,|\, a, b \in A\}$. Then, a strictly increasing function $f : X \longrightarrow [0, 1]$ exists such that $\sigma' = f \circ \sigma$. □

Proof. For $a, b \in A$, let $x = \sigma(a, b)$ and define $f(x) = \sigma'(a, b)$. Obviously, f is well-defined, since the coherency of σ and σ' implies

$$\sigma(a, b) = \sigma(c, d) \Leftrightarrow \sigma'(a, b) = \sigma'(c, d) \tag{3.18}$$

for all $a, b, c, d \in A$. Moreover, f is strictly increasing, since (3.18) remains valid when replacing the equality relation by the <-relation. □

[11] This should be regarded as a reasonable property. Indeed, considering similarity as a cardinal concept complicates its formalization and raises some difficult semantical questions.

Proposition 3.18. Consider a system $(\mathcal{S}, \mathcal{R}, \varphi)$ and a memory $\mathcal{M}$ of cases and let $\sigma_{\mathcal{S}}$ and $\sigma'_{\mathcal{S}}$ resp. $\sigma_{\mathcal{R}}$ and $\sigma'_{\mathcal{R}}$ be coherent similarity measures. Moreover, denote by h_Σ resp. h'_Σ the similarity profiles induced by these measures and let $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}$ resp. $\widehat{\varphi}_{h'_\Sigma,\mathcal{M}}$ be the case-based approximations defined by $(h_\Sigma, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$ resp. $(h'_\Sigma, \sigma'_{\mathcal{S}}, \sigma'_{\mathcal{R}})$ via (3.15). We then have $\widehat{\varphi}_{h_\Sigma,\mathcal{M}} = \widehat{\varphi}_{h'_\Sigma,\mathcal{M}}$. □

Proof. According to Lemma 3.17 there are strictly increasing functions f and g such that $\sigma'_{\mathcal{S}} = f \circ \sigma_{\mathcal{S}}$ and $\sigma'_{\mathcal{R}} = g \circ \sigma_{\mathcal{R}}$. From (3.18) and $f(\sigma_{\mathcal{R}}(r,r')) \leq g(\sigma_{\mathcal{S}}(s,s')) \Leftrightarrow \sigma'_{\mathcal{R}}(r,r') \leq \sigma'_{\mathcal{S}}(s,s')$ for all $s,s' \in \mathcal{S}$ and $r,r' \in \mathcal{R}$ then follows that $h'_\Sigma \circ f = g \circ h_\Sigma$. Now, consider $s,s' \in \mathcal{S}$, $r,r' \in \mathcal{R}$ and suppose that $(h_\Sigma \circ \sigma_{\mathcal{S}})(s,s') \leq \sigma_{\mathcal{R}}(r,r')$. It follows that

$$\begin{aligned}
\sigma'_{\mathcal{R}}(r,r') &= (g \circ \sigma_{\mathcal{R}})(r,r') \\
&\geq (g \circ h_\Sigma \circ \sigma_{\mathcal{S}})(s,s') \\
&= (h'_\Sigma \circ f \circ \sigma_{\mathcal{S}})(s,s') \\
&= (h'_\Sigma \circ \sigma'_{\mathcal{S}})(s,s').
\end{aligned}$$

In the same way it is shown that $(h'_\Sigma \circ \sigma'_{\mathcal{S}})(s,s') \leq \sigma'_{\mathcal{R}}(r,r')$ implies $(h_\Sigma \circ \sigma_{\mathcal{S}})(s,s') \leq \sigma_{\mathcal{R}}(r,r')$. Consequently, we have $\mathcal{N}_{h_\Sigma(\sigma_{\mathcal{S}}(s,s'))}(r) = \mathcal{N}_{h'_\Sigma(\sigma'_{\mathcal{S}}(s,s'))}(r)$ for all $s,s' \in \mathcal{S}$, $r \in \mathcal{R}$ and, hence, $\widehat{\varphi}_{h_\Sigma,\mathcal{M}} = \widehat{\varphi}_{h'_\Sigma,\mathcal{M}}$. □

In Section 2.4, it was already pointed out that similarity measures might be more or less "discriminating." We are now in the position to put this into more precise terms. Let us call a similarity measure σ a *refinement* of a measure σ' if $\sigma' = f \circ \sigma$, where f is non-decreasing (i.e., order-preserving) but not (strictly) increasing. Loosely speaking, the measure σ uses a richer similarity scale which includes more degrees of similarity, that is $\mathrm{rg}(\sigma') \subsetneq \mathrm{rg}(\sigma)$.

Proposition 3.19. Consider a system $(\mathcal{S}, \mathcal{R}, \varphi)$ and a memory $\mathcal{M}$ of cases. Let $\sigma_{\mathcal{S}}$ be a refinement of $\sigma'_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$ a refinement of $\sigma'_{\mathcal{R}}$. Moreover, denote by h_Σ resp. h'_Σ the similarity profiles induced by these measures and let $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}$ resp. $\widehat{\varphi}_{h'_\Sigma,\mathcal{M}}$ be the case-based approximations defined by $(h_\Sigma, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$ resp. $(h'_\Sigma, \sigma'_{\mathcal{S}}, \sigma'_{\mathcal{R}})$ via (3.15). Then, $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s) \subseteq \widehat{\varphi}_{h'_\Sigma,\mathcal{M}}(s)$ for all $s \in \mathcal{S}$. □

Proof. Consider values $s,s' \in \mathcal{S}$ and $r,r' \in \mathcal{R}$. Suppose that $r' \in \mathcal{N}_{h_\Sigma(\sigma_{\mathcal{S}}(s,s'))}(r)$, i.e., $\sigma_{\mathcal{R}}(r,r') \geq h_\Sigma(\sigma_{\mathcal{S}}(s,s'))$. Thus, we find $t,t' \in \mathcal{S}$ such that $\sigma_{\mathcal{R}}(r,r') \geq \sigma_{\mathcal{R}}(\varphi(t),\varphi(t'))$ and $\sigma_{\mathcal{S}}(s,s') = \sigma_{\mathcal{S}}(t,t')$. Since $\sigma'_{\mathcal{S}} = f \circ \sigma_{\mathcal{S}}$ and $\sigma'_{\mathcal{R}} = g \circ \sigma_{\mathcal{R}}$ for non-decreasing functions f,g, we have $\sigma'_{\mathcal{S}}(s,s') = \sigma'_{\mathcal{S}}(t,t')$ and $\sigma'_{\mathcal{R}}(r,r') \geq \sigma'_{\mathcal{R}}(\varphi(t),\varphi(t'))$. Therefore,

$$h'_\Sigma(\sigma'_{\mathcal{S}}(s,s')) \leq \sigma'_{\mathcal{R}}(\varphi(t),\varphi(t')) \leq \sigma'_{\mathcal{R}}(r,r')$$

and, hence, $r' \in \mathcal{N}_{h'_\Sigma(\sigma_{\mathcal{S}}(s,s'))}(r)$. □

Of course, generally we will not only have $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s) \subseteq \widehat{\varphi}_{h'_\Sigma,\mathcal{M}}(s)$, as guaranteed by Proposition 3.19, but also $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s) \neq \widehat{\varphi}_{h'_\Sigma,\mathcal{M}}(s)$ for some $s \in \mathcal{S}$. As an obvious example consider the "least discriminating" case where $g \equiv 1$ on $D_\mathcal{R}$ and, hence, $\sigma'_\mathcal{R} \equiv 1$ on $\mathcal{R} \times \mathcal{R}$, which leads to the trivial prediction $\widehat{\varphi}_{h'_\Sigma,\mathcal{M}} \equiv \mathcal{R}$ on $\mathcal{S}$.

For us to be able to study the approximation capability of (3.15) more thoroughly the system $(\mathcal{S}, \mathcal{R}, \varphi)$ must have a structure which allows us to quantify the quality of a case-based approximation. To this end, let us endow $\mathcal{S}$ and $\mathcal{R}$ with a metric, i.e., let $(\mathcal{S}, \Delta_\mathcal{S})$ and $(\mathcal{R}, \Delta_\mathcal{R})$ be metric spaces. Clearly, a good approximation of φ can only be expected if the similarity measures $\sigma_\mathcal{S}$ and $\sigma_\mathcal{R}$ are related to the distance measures $\Delta_\mathcal{S}$ and $\Delta_\mathcal{R}$. We can prove the following result.

Proposition 3.20. Suppose that $\sigma_\mathcal{S} = f \circ \Delta_\mathcal{S}$ and $\sigma_\mathcal{R} = g \circ \Delta_\mathcal{R}$ with strictly decreasing functions f and g, and

$$\exists\, \varepsilon > 0 \,\exists\, \mathcal{S}' \subseteq \mathcal{S} \,:\, \operatorname{card}(\mathcal{S}') < \infty \,\wedge\, \mathcal{S} = \bigcup_{s \in \mathcal{S}'} \bar{\mathfrak{B}}_\varepsilon(s), \tag{3.19}$$

where $\bar{\mathfrak{B}}_\varepsilon(s) \overset{\mathrm{df}}{=} \{s' \in \mathcal{S} \,|\, \Delta_\mathcal{S}(s, s') \leq \varepsilon\}$. Moreover, assume the Lipschitz condition

$$\exists\, L > 0 \,\forall\, s, s' \in \mathcal{S} \,:\, \Delta_\mathcal{R}(\varphi(s), \varphi(s')) \leq L\, \Delta_\mathcal{S}(s, s') \tag{3.20}$$

to hold. Then, a finite memory $\mathcal{M}$ exists such that

$$\operatorname{diam}(\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s)) \overset{\mathrm{df}}{=} \max\{\Delta_\mathcal{R}(r, r') \,|\, r, r' \in \widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s)\} \leq 2\, L\, \varepsilon$$

for all $s \in \mathcal{S}$. □

Proof. Let $\varepsilon > 0$ and $\mathcal{S}' \subseteq \mathcal{S}$ satisfy (3.19) and define $\mathcal{M} = \bigcup_{s' \in \mathcal{S}'} \langle s', \varphi(s') \rangle$. For $s, s' \in \mathcal{S}$ such that $\sigma_\mathcal{S}(s, s') = x \in D_\mathcal{S}$ we have $\Delta_\mathcal{S}(s, s') = f^{-1}(x)$. Thus, according to (3.20), $\sigma_\mathcal{R}(\varphi(s), \varphi(s')) \geq g(Lf^{-1}(x))$, which means $h_\Sigma(x) \geq g(Lf^{-1}(x))$ for all $x \in D_\mathcal{S}$. Now, consider some $s \in \mathcal{S}$. According to (3.19), the memory $\mathcal{M}$ contains a case $\langle s_0, r_0 \rangle$ such that $\Delta_\mathcal{S}(s, s_0) \leq \varepsilon$. Hence, $h_\Sigma(\sigma_\mathcal{S}(s, s_0)) \geq g(Lf^{-1}(\sigma_\mathcal{S}(s, s_0))) \geq g(L\varepsilon)$, which means that $\Delta_\mathcal{R}(r_0, r') \leq L\,\varepsilon$ for all $r' \in \mathcal{N}_{h_\Sigma(\sigma_\mathcal{S}(s,s_0))}(r_0)$. The result then follows from $\Delta_\mathcal{R}(r, r') \leq \Delta_\mathcal{R}(r, r_0) + \Delta_\mathcal{R}(r_0, r')$ for all $r, r' \in \mathcal{N}_{h_\Sigma(\sigma_\mathcal{S}(s,s_0))}(r_0)$ and $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s) \subseteq \mathcal{N}_{h_\Sigma(\sigma_\mathcal{S}(s,s_0))}(r_0)$. □

Since $\varphi(s) \in \widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s)$ for all $s \in \mathcal{S}$, Proposition 3.20 guarantees the existence of a case-based approximation of φ which determines all outcomes up to a precision of $\delta = 2\, L\, \varepsilon$. The following corollaries follow immediately.

Corollary 3.21. Suppose the assumptions of Proposition 3.20 to hold true with "$\exists\, \varepsilon > 0$" in (3.19) replaced by "$\forall\, \varepsilon > 0$." Then, the mapping φ can be approximated to any degree of accuracy $\delta > 0$ via (3.15) with a finite memory $\mathcal{M}$. □

Corollary 3.22. Let $\mathcal{S} \subseteq \mathfrak{Q}^p$ be bounded, $\mathcal{R} \subseteq \mathfrak{Q}^q$, and $\varDelta_{\mathcal{S}}$ and $\varDelta_{\mathcal{R}}$ be defined by the corresponding Euclidean distances. Moreover, suppose that φ satisfies (3.20) and that $\sigma_{\mathcal{S}} = f \circ \varDelta_{\mathcal{S}}$ and $\sigma_{\mathcal{R}} = g \circ \varDelta_{\mathcal{R}}$ with f, g strictly decreasing. Then, φ can be approximated to any degree of accuracy $\delta > 0$ via (3.15) with a finite memory $\mathcal{M}$. □

Assumption (3.19), which requires the existence of a finite cover of $\mathcal{S}$, cannot be dropped, as can easily be seen by constructing a counter-example with $\varDelta_{\mathcal{S}}$ defined by $\varDelta_{\mathcal{S}}(s, s') = 0$ for $s = s'$ and $\varDelta_{\mathcal{S}}(s, s') = 1$ for $s \neq s'$ (and $\mathrm{card}(\mathcal{S}) = \aleph_0$). Likewise, (3.20) is necessary, as an example with $\varphi(s)$ defined on $[0, 1] \cap \mathfrak{Q}$ by $\varphi(s) = 1$ for $s = 0$ and $\varphi(s) = 0$ for $s > 0$ (and $\varDelta_{\mathcal{R}}$ the standard metric) shows.

The discussion so far has shown that the inference scheme presented in Section 3.2 can basically be seen as a *set-valued* approximation method. The essential part of this inference procedure is realized in what we have called the *similarity space*, not in the instance space itself (cf. Fig. 3.2). That is, CBI is not directly based on the information provided at the system level. Rather, the concept of similarity, quantified in terms of similarity functions $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$, is exploited in order to transform this information into information which is represented at the similarity level. An approximation at the instance level is then derived within a two-stage process from inferences about the *similarity* of an unknown outcome to already observed ones.

It is this indirect derivation of approximations that constitutes the main difference between CBA and other approximation techniques. In fact, an implicit notion of similarity is also present in other methods, since the (local) transfer of observed outputs is generally based on the concept of *distance*. Typically, a (scalar) estimation of an unknown value $f(x)$ of a function f is derived in the form of a weighted combination of training examples $f(x_1), \ldots, f(x_n)$, where the weight of an example $f(x_k)$ decreases with the distance of the associated point in the input space, x_k, to the query point x.[12] Consider an approximation of the form

$$\widehat{f}(x) = \frac{\sum_{k=1}^{n} K(x_k - x) \cdot f(x_k)}{\sum_{k=1}^{n} K(x_k - x)},$$

where $K(\cdot)$ is a *kernel function* (centered at 0), as an example.

In some approximation methods the observed outcomes $f(x_k)$ appear only implicitly, in the sense that they determine parameters of an approximating function. In a special version of locally weighted regression, for instance, the parameters of a linear function $\widehat{f}(\cdot)$ are determined such that

$$\sum_{k=1}^{n} (f(x_k) - \widehat{f}(x_k))^2 K(d(x, x_k))$$

[12] The input space must hence be endowed with a distance measure.

is minimized, where $d(\cdot)$ is a distance measure, and $K(\cdot)$ is a kernel function. The value of $f(\cdot)$ for the query point x is then estimated by $\widehat{f}(x)$.

As a further example consider again the k-Nearest Neighbor (kNN) method (cf. Section 2.2) from which several instance-based learning algorithms have emerged. It derives predictions according to

$$\widehat{f}(x) = F(f(x_1), \ldots, f(x_k)),$$

where $f(x_1), \ldots, f(x_k)$ are the training examples associated with the k points which are most similar to (or have the smallest distance from) the query point x. If $f(\cdot)$ is a numerical function, $F(\cdot)$ is often defined as a weighted average, i.e.

$$\widehat{f}(x) = \sum_{\jmath=1}^{k} \left(1 - \frac{|x - x_\jmath|}{\sum_{\imath=1}^{k} |x - x_\imath|} \right) \cdot f(x_\jmath).$$

If $\mathrm{rg}(f)$ is discrete, $F(\cdot)$ generally returns the value which is most frequent among $f(x_1), \ldots, f(x_k)$.

As can be seen, the approximation methods outlined above are based on the same data as CBA, namely a set of observed values of a function (= outcomes) and some kind of similarity or distance relation between points (= inputs) in the input space. This data can be defined as an extension of the similarity structure (cf. Definition 3.13 and Fig. 3.3).

Definition 3.23 (outcome structure). Let Σ be a CBI setup, s_0 a new input, and $\mathcal{M}$ the memory (2.29) associated with Σ. The set of values

$$\mathsf{OST}(\mathcal{M}, s_0) \overset{\mathrm{df}}{=} \mathsf{SST}(\mathcal{M}, s_0) \cup \{ r_\jmath \,|\, 1 \leq \jmath \leq n \}$$

(together with $(h_\Sigma, \sigma_\mathcal{S}, \sigma_\mathcal{R})$) defines the outcome structure of the CBI problem $\langle \Sigma, s_0 \rangle$. □

Usual approximation methods employ the outcome structure directly within one inference step. As opposed to this, the first step of the CBA scheme uses only the similarity structure, and the observed outcomes r_k are called in for the second inference step.

The aforementioned difference becomes obvious, e.g., when comparing CBA to the kNN algorithm. Firstly, this algorithm applies the similarity measures directly at the instance level in order to find the most similar cases, whereas in CBA these measures are used for defining the similarity structure h_Σ. Secondly, the kNN method does also perform the inference step at the instance level, in the sense that predictions are derived directly from the observed outcomes. As opposed to this, CBA uses the given information for drawing inferences, not about outputs, but about similarities. It makes use of observed outcomes by more indirect means,

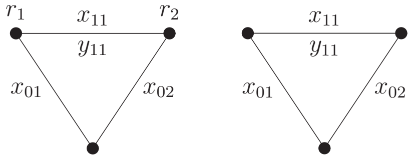

Fig. 3.3. The outcome (left) and similarity structure (right) of a CBI problem can be illustrated as a graph, where the nodes are associated with (information about) cases and the edges are labeled with information concerning the (similarity) relation between cases. This figure shows the graphs for a memory with two cases.

in the sense that each output defines an instantiation of a similarity constraint at the system level.

In connection with the kNN method it should also be observed that CBA (especially (3.16)) can be seen as an interesting *set-valued* version of this algorithm. As an advantage of CBA let us mention that it also takes the quality of the similarity structure into account when predicting an outcome. In fact, (3.15) will not be very constraining if this structure is poorly developed, thus indicating that the application of the NEAREST NEIGHBOR principle (and, hence, the original kNN method) does not seem advisable. We shall come back to this point in Chapter 4.

The following points deserve mentioning when comparing case-based to other local approximation methods. On the one hand, CBA is less demanding in the sense that it requires the specification of a similarity hypothesis, i.e., a relatively simple one-dimensional function, whereas other methods derive approximating functions with $\mathrm{dom}(f) = \mathcal{S}$ and $\mathrm{codom}(f) = \mathcal{R}$. Moreover, CBA still works if $\mathcal{S}$ and $\mathcal{R}$ are not as well-structured as certain number spaces, a situation regularly encountered within the context of CBR. In fact, the assignment of similarity degrees can then be seen as a reasonable quantification of the approximation problem. This kind of quantification will often be more obvious than a quantification of $\mathcal{S}$ and $\mathcal{R}$ which allows for deriving a good approximation $\widehat{f} : \mathcal{S} \longrightarrow \mathcal{R}$.

On the other hand, the transformation from a high-dimensional (instance) space into a low-dimensional (similarity) space is usually afflicted with a loss of information. This becomes especially apparent in connection with the (pseudo-)inverse of the similarity measure $\sigma_{\mathcal{R}}$. In fact, this transformation will generally be a *set-valued* mapping.

In any case, a comparison between (indirect) case-based and direct approximation methods remains a difficult (if not meaningless) task. Firstly, the success of any approximation method largely depends on the application and properties of the data.[13] Thus, it will generally not be possible to qualify one approach as being superior in comparison to other methods. Secondly, a case-based approximation is not scalar-valued but derives set-valued approximations which either cover the unknown outcome or, as will be seen in Section 3.4, define some kind of confidence region. Thus, the usefulness of an approximation method will also depend on whether the problem at hand requires an estimation in the form of a scalar value $\widehat{r}_0$ or whether it is important to have information about r_0 in the form of outer bounds. As can be seen, the aforementioned differences between case-based and direct approximation suggest to combine (rather than to compare) these approaches.

3.3.2 Local similarity profiles

In Section 3.2.1, it has already been pointed out that CBA is *local* in the sense that the information provided by different cases is processed and combined independently.[14] It is, however, *global* in the sense that the similarity profile represents information which holds true for the complete similarity space. In fact, the constraint $\mathcal{N}_{h_\Sigma(\sigma_\mathcal{S}(s,s_0))}(r)$ provided by a case $\langle s, r\rangle$ for the prediction of an unknown outcome $\varphi(s_0)$ contains a local component, namely the case $\langle s, r\rangle$ itself, as well as a global component, namely the similarity hypothesis h. CBA can thus be characterized as a local processing of global information.

Often, the CBI assumption is not satisfied equally well for all parts of the instance space $\mathcal{S} \times \mathcal{R}$.[15] The global validity of the similarity profile might then prevent one from defining tight bounds for those regions where the CBI hypothesis actually applies rather well. In fact, a globally admissible similarity hypothesis might lead to (local) predictions which are unnecessarily imprecise. This is illustrated by the following simple example.

Example 3.24. Let $\mathcal{S} = \mathcal{R} = [-1, 1] \setminus \{0\}$,[16] $\varphi(s) = -1$ if $-1 \leq s < 0$, and $\varphi(s) = 1$ if $0 < s \leq 1$. Moreover, let $\sigma_\mathcal{S}(u, v) = \sigma_\mathcal{R}(u, v) = 1 - |u - v|/2$. Obviously, for all $1 \neq x \in D_\mathcal{S}$ there are $s, s' \in \mathcal{S}$ such that $\sigma_\mathcal{S}(s, s') = x$ and $\sigma_\mathcal{R}(\varphi(s), \varphi(s')) = 0$. We hence have $h_\Sigma(x) = 0$ for all $x \in D_\mathcal{S} \setminus \{1\}$, which means that $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s_0) = [-1, 1]$ if $\langle s_0, \varphi(s_0)\rangle \notin \mathcal{M}$. □

Loosely speaking, a CBI strategy is not applicable in Example 3.24 because the CBI hypothesis is not globally valid. Still, it seems desirable to make use of the observation that this assumption is satisfied at least *locally*. One possibility of doing this is to partition the set $\mathcal{S}$ of inputs and to derive respective local

[13] Recall the selective superiority problem mentioned in footnote 9.

[14] CBA is also local in the sense that it is a local approximation method. These two meanings of locality should not be confused.

[15] In a game playing context, for instance, the CBI principle hardly applies to certain “tactical” situations [310].

[16] More specifically, to comply with our formal framework, we should set $\mathcal{S} = \mathcal{R} = ([-1, 1] \cap \mathfrak{Q}) \setminus \{0\}$.

approximations.[17] In Example 3.24, it suggests itself to partition $\mathcal{S}$ into $[-1, 0)$ and $(0, 1]$. However, since φ is generally unknown, the definition of such a partition will not always be obvious, all the more if $\mathcal{S}$ is non-numerical. Here, we consider a second possibility, namely that of associating an individual similarity profile with each input of the memory. This approach is somehow comparable to the use of local kernels in kernel-based density estimation [385], and to the use of *local metrics* in kNN algorithms and instance-based learning (e.g., metrics which allow feature weights to vary as a function of the instance [342, 157, 9, 311]). It leads us to introduce the concept of a *local similarity profile*.

Definition 3.25 (local similarity profile). Consider a CBI setup Σ and let $s \in \mathcal{S}$. We define $h^s_\Sigma : D_\mathcal{S} \longrightarrow [0, 1]$ by the mapping

$$x \mapsto \inf_{s' \in \mathcal{S},\, \sigma_\mathcal{S}(s,s')=x} \sigma_\mathcal{R}(\varphi(s), \varphi(s')).$$

This function is called the local similarity profile associated with s, or the s-similarity profile of Σ. A collection $h^\mathcal{M}_\Sigma = \{h^s_\Sigma \,|\, s \in \mathcal{M}^\downarrow\}$ of local profiles is referred to as the local $\mathcal{M}$-similarity profile. □

The following relations hold between the different types of similarity profiles:

$$h_\Sigma = \bigwedge_{s \in \mathcal{S}} h^s_\Sigma, \quad h^\mathcal{M}_\Sigma = \bigwedge_{s \in \mathcal{M}^\downarrow} h^s_\Sigma.$$

That is, the similarity profile h_Σ and $\mathcal{M}$-similarity profile $h^\mathcal{M}_\Sigma$ are lower envelopes of the class of local profiles associated with inputs in $\mathcal{S}$ and $\mathcal{M}^\downarrow$, respectively. Consequently, $h_\Sigma \leq h^\mathcal{M}_\Sigma \leq h^s_\Sigma$ for all memories $\mathcal{M}$ and inputs $s \in \mathcal{M}^\downarrow$.

As can be seen, a local similarity profile is closely related to the idea of an $\mathcal{M}$-similarity profile. In fact, an s-profile corresponds to the $\mathcal{M}$-profile with $\mathcal{M}^\downarrow = (s)$. Besides, a class of local profiles will generally be specified – by means of respective learning methods (cf. Section 3.4) – for a memory which does not change frequently. In connection with approximation methods, the inputs which constitute the memory and for which local profiles are defined play a role somewhat similar to the so-called *knots* in, say, approximation with spline functions, and the local profiles correspond to basis functions.

Given a hypothesis $h^\mathcal{M} = \{h^s \,|\, s \in \mathcal{M}^\downarrow\}$ related to a local $\mathcal{M}$-similarity profile and a new input $s_0 \in \mathcal{S}$, the inference scheme (3.2) is replaced by

$$\varphi(s_0) \in \widehat{\varphi}_{h^\mathcal{M},\mathcal{M}}(s_0) \overset{\mathrm{df}}{=} \bigcap_{\langle s,r\rangle \in \mathcal{M}} \mathcal{N}_{h^s(\sigma_\mathcal{S}(s,s_0))}(r). \tag{3.21}$$

The respective case-based approximation, i.e., the local counterpart to (3.15), is called a *local case-based approximation*:

[17] This idea is related to that of *feature space partitioning* in classification [77]. See also [261] for a related idea in connection with memory-based learning.

$$\widehat{\varphi}_{h^{\mathcal{M}},\mathcal{M}} : s \mapsto \bigcap_{\langle s',r'\rangle \in \mathcal{M}} \mathcal{N}_{h^{s'}(\sigma_{\mathcal{S}}(s,s'))}(r').$$

Example 3.26. Consider again Example 3.24 and suppose that the memory $\mathcal{M}$ contains the cases $\langle -1,-1\rangle$ and $\langle 1,1\rangle$. The respective local profiles are given by

$$x \mapsto \begin{cases} 1 & \text{if} \quad 1/2 \leq x \leq 1 \\ 0 & \text{if} \quad 0 \leq x < 1/2 \end{cases} .$$

These two profiles can already guarantee an exact representation of φ. That is, $\widehat{\varphi}_{h^{\mathcal{M}}_{\Sigma}}(s) = \{\varphi(s)\}$ for all $s \in \mathcal{S}$ with $\mathcal{M} = (\langle -1,-1\rangle, \langle 1,1\rangle)$. □

Note that a local profile indicates the validity of the CBI hypothesis for *individual* cases. That is, the local profile associated with an input $s \in \mathcal{S}$ can be utilized for rating the *quality* of the case $\langle s, \varphi(s)\rangle$.[18] An input with a strongly developed local profile (i.e., its outcome is locally representative) will generally support precise predictions, whereas an input with a poorly developed profile will hardly be useful from the viewpoint of CBA. Local profiles might hence serve as a (complementary) criterion for selecting "competent" cases to be stored in (or removed from) the memory [357]. It should be noted, however, that the similarity profile can only be taken as an indication of the precision of predictions. In fact, the predictions also depends on the neighborhood structure of $\mathcal{R}$. For instance, it is quite possible that $\text{card}(\mathcal{N}_{\alpha}(r)) < \text{card}(\mathcal{N}_{\beta}(r'))$ for two outcomes $r \neq r'$, even though $\beta < \alpha$.

3.4 Learning similarity hypotheses

3.4.1 The learning task

The inference scheme (3.2) reveals that CBI can essentially be seen as an *instance-based* approach. Still, it also contains a *model-based* component, namely the similarity hypothesis h. Consequently, *learning* can be realized in (at least) two ways in CBI: By storing new cases in the memory and by estimating the similarity profile. Here, we concentrate on the latter (model-based) aspect.

Definition 3.27 (CBL). Consider a CBI setup Σ with a memory

$$\mathcal{M} \subseteq \mathcal{D} = \mathcal{D}_N = (c_1, \ldots, c_N),$$

where $\mathcal{D}$ denotes the sequence of cases which have been encountered so far (these are the first N cases, given the assumption that cases arrive successively). Moreover, let $\mathcal{H}$ be a *hypothesis space* of functions $h : [0,1] \longrightarrow [0,1]$. The task of *case-based learning* (CBL) is understood as deriving an optimal hypothesis $h_* \in \mathcal{H}$ from the data given. □

[18] See Section 4.6 for a more detailed discussion of the assessment of cases.

Observe that different similarity measures define different similarity structures of the system under consideration and that the measures originally chosen might not be optimal in the sense that similarity structures induced by alternative measures are, in a certain sense, more suitable for CBI. Suppose, for instance, that we have measures $(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$ and $(\sigma'_{\mathcal{S}}, \sigma'_{\mathcal{R}})$ and let $\widehat{\varphi}_{h,\mathcal{M}}$ resp. $\widehat{\varphi}'_{h,\mathcal{M}}$ denote the case-based approximations induced by these measures via (3.15) with $h = h_{\Sigma}$. If $\widehat{\varphi}_{h,\mathcal{M}}(s) \subseteq \widehat{\varphi}'_{h,\mathcal{M}}(s)$ for all $s \in \mathcal{S}$, then $(\sigma'_{\mathcal{S}}, \sigma'_{\mathcal{R}})$ should not (at least not strictly) be preferred to $(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$. This gives rise to defining a partial order relation on a class of measures. Therefore, it might also be reasonable to allow for the adaptation of similarity measures. The problem of CBL can thus be extended as follows.

Definition 3.28 (extended CBL **problem).** Let a set $\mathcal{S}$ of inputs, a set $\mathcal{R}$ of outputs, and a memory $\mathcal{M} \subseteq \mathcal{D} = (c_1, \ldots, c_N)$ be given, where $\mathcal{D}$ denotes the sequence of cases which have been encountered so far. Moreover, let $\mathcal{H}$ be a class of functions $h : [0,1] \longrightarrow [0,1]$ and $\mathcal{H}_{\mathcal{S}}$, $\mathcal{H}_{\mathcal{R}}$ classes of similarity measures over $\mathcal{S}$ and $\mathcal{R}$, respectively. The task of (extended) CBL is defined as searching the hypothesis space $\mathcal{H} \times \mathcal{H}_{\mathcal{S}} \times \mathcal{H}_{\mathcal{R}}$ for an optimal hypothesis $h_* = (h, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$. □

Remark 3.29. Relating the interpretation of a similarity hypothesis h (resp. a similarity profile h_{Σ}) to the idea of modifying the measure $\sigma_{\mathcal{S}}$ has already been suggested in Remark 3.4. If h is strict, such a modification corresponds to a "stretching" and "squeezing" of the similarity scale underlying $\sigma_{\mathcal{S}}$. Moreover, the modification is restricted in the sense that the original measure $\sigma_{\mathcal{S}}$ and its modified version $\sigma'_{\mathcal{S}}$ are coherent in the sense of (3.17). As opposed to this, a non-monotone hypothesis additionally puts the similarity degrees $x \in D_{\mathcal{S}}$ in a different order, which corresponds to a re-arranging of the (ordinal) similarity scale $D_{\mathcal{S}}$. Then, (3.17) holds true only with $\leq$ replaced by the equality relation. In other words, two inputs s_1, s_2 which are more similar than the inputs s_3, s_4 according to $\sigma_{\mathcal{S}}$ might be seen as being less similar according to $\sigma'_{\mathcal{S}}$. Now, one possibility to approach the extended CBL problem is to allow for a re-arranging of the similarity scale underlying $\sigma_{\mathcal{R}}$ as well, i.e., to allow for replacing $\sigma_{\mathcal{R}}$ by $\sigma'_{\mathcal{R}} = m \circ \sigma_{\mathcal{R}}$ for some $m : [0,1] \longrightarrow [0,1]$. A similarity hypothesis h is then related to $(\sigma_{\mathcal{S}}, \sigma'_{\mathcal{R}})$ instead of $(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$. In connection with the extended CBL problem, this amounts to defining $\mathcal{H}_{\mathcal{R}}$ as the class of all measures which can be written in the form $m \circ \sigma_{\mathcal{R}}$. □

Definition 3.27 has not commented on the criteria which decide on the optimality of hypotheses. In order to derive such criteria we fall back on two principles. The first one is the obvious demand that an optimal hypothesis h_* should be consistent with observed data in the sense that (3.1) is satisfied at least for elements of $\mathcal{D}$, i.e.

$$(\sigma_{\mathcal{S}}(s, s') = x) \;\Rightarrow\; (\sigma_{\mathcal{R}}(r, r') \geq h_*(x)) \tag{3.22}$$

should hold true for all $\langle s, r \rangle, \langle s', r' \rangle \in \mathcal{D}$. This *consistency principle* is closely related to the *inductive learning hypothesis* in machine learning. Namely, we suspect a hypothesis h, which is consistent with a large number of observations, also to be consistent with the overall similarity structure of the system (in the sense that it is admissible). Observe that (3.22) implies $\varphi(s) \in \widehat{\varphi}_{h,\mathcal{M}}(s)$ for all s with $\langle s, \varphi(s)\rangle \in \mathcal{D}$ and $\mathcal{M} \subseteq \mathcal{D}$. Again, we may assume that a mapping which defines an outer approximation of $\varphi|\mathcal{S}'$ for a (large) subset $\mathcal{S}' \subseteq \mathcal{S}$ also defines an outer approximation of the complete mapping $\varphi = \varphi|\mathcal{S}$. We denote by $\mathcal{H}_{\mathcal{D}} \subseteq \mathcal{H}$ the class of hypotheses which are consistent with a set $\mathcal{D}$ of cases in the sense of (3.22).

As will be seen in Section 3.4.3, it may become necessary to weaken the aforementioned consistency principle. In fact, testing consistency of a hypothesis according to (3.22) requires the consideration of all pairs $(c, c') \in \mathcal{D} \times \mathcal{D}$ of cases. However, as suggested by Definition 3.27, the memory $\mathcal{M}$ of stored cases will generally be a (proper) subset of the set $\mathcal{D}$ of *successively* encountered cases. It is hence not possible to take the tuple, say, (c_1, c_N) into consideration if c_1 was not stored long enough and has been removed before the arrival of c_N. Thus, a weaker version of the consistency principle should require (3.22) to hold true for all

$$(c, c') \in \mathcal{C} = \mathcal{C}_N \overset{\mathrm{df}}{=} \bigcup_{1 \leq n \leq N-1} \mathcal{M}_n \times (c_{n+1}),$$

where $\mathcal{M}_n$ denotes the memory after the observation of the n-th case c_n. We denote by $\mathcal{H}_{\mathcal{C}}$ the class of hypotheses which are consistent with $\mathcal{D}$ in this weaker sense. Thus, we generally have $\mathcal{H}_{\mathcal{D}} \subseteq \mathcal{H}_{\mathcal{C}} \subseteq \mathcal{H}_{\mathcal{M}}$, where $\mathcal{H}_{\mathcal{M}}$ is defined in a canonical way.

In order to motivate the second principle recall that the case-based approximation (3.15), which is induced by a hypothesis $(h, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$ and a memory $\mathcal{M}$, can be seen as a simplified representation of the system structure φ. Indeed, $\widehat{\varphi}_{h,\mathcal{M}}$ is represented by card($\mathcal{M}$) cases and the hypothesis $(h, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$, whereas the representation of φ – if it cannot be expressed in closed form – requires the enumeration of the complete set

$$\mathcal{D}^* \overset{\mathrm{df}}{=} \{\langle s, \varphi(s)\rangle \,|\, s \in \mathcal{S}\}$$

of cases. Of course, in passing from φ to $\widehat{\varphi}_{h,\mathcal{M}}$ it is usually unavoidable to loose some information. The corresponding increase in uncertainty is reflected by the fact that $\widehat{\varphi}_{h,\mathcal{M}}$ is a *set-valued* mapping and that we will generally have $\{\varphi(s)\} \subsetneq \widehat{\varphi}_{h,\mathcal{M}}(s)$ for at least some inputs $s \in \mathcal{S}$. According to the *principle of minimum uncertainty*, which is one of the general principles of systems theory, one should, among a set of candidates, accept only those simplifications of a system for which the increase in uncertainty is minimal [231]. Thus, let U be some measure which quantifies the uncertainty associated with $\widehat{\varphi}_{h,\mathcal{M}}$.[19] A hypothesis h_* is then *optimal* if $h_* \in \mathcal{H}_{\mathcal{C}}$ and $U\left(\widehat{\varphi}_{h_*,\mathcal{M}}\right) \leq U\left(\widehat{\varphi}_{h,\mathcal{M}}\right)$ holds true for all $h \in \mathcal{H}_{\mathcal{C}}$. We denote by $\mathcal{H}_* \subseteq \mathcal{H}_{\mathcal{C}}$ the class of all optimal hypotheses. Of course, this definition does neither guarantee the existence nor the uniqueness of an optimal hypothesis.

[19] Various proposals for such uncertainty measures can be found in systems science literature.

In connection with the learning of hypotheses it makes sense to consider *admissibility* as a further property which is more restricting than consistency. We denote by $\mathcal{H}^*$ the class of optimal admissible hypotheses. Thus, $\mathcal{H}^*$ consists of those uncertainty minimizing hypotheses h^* which are consistent with $\mathcal{D}^*$.[20]

Let us now consider the CBL problem in its basic form. Of course, deriving the uncertainty $U(\widehat{\varphi}_{h,\mathcal{M}})$ associated with a hypothesis h is intractable if it requires the computation of the complete mapping $\widehat{\varphi}_{h,\mathcal{M}}$. Observe, however, that any reasonable measure U should satisfy $U(\widehat{\varphi}_{h,\mathcal{M}}) \leq U(\widehat{\varphi}_{h',\mathcal{M}})$ if $\widehat{\varphi}_{h,\mathcal{M}}(s) \subseteq \widehat{\varphi}_{h',\mathcal{M}}(s)$ for all $s \in \mathcal{S}$. Since the latter holds true if $h' \leq h$, U should be consistent with the partial order defined by $\leq$ over $\mathcal{H}$.

Observation 3.30. Suppose the hypothesis space $\mathcal{H}$ to satisfy $h \equiv 0 \in \mathcal{H}$ and $(h, h' \in \mathcal{H}) \Rightarrow (h \vee h' \in \mathcal{H})$, where $h \vee h'$ is defined by the mapping $x \mapsto \max\{h(x), h'(x)\}$. Moreover, suppose the measure U to satisfy

$$(h' \leq h) \;\Rightarrow\; (U(\widehat{\varphi}_{h,\mathcal{M}}) \leq U(\widehat{\varphi}_{h',\mathcal{M}}))$$

for all $h, h' \in \mathcal{H}$ and memories $\mathcal{M}$. Then, a unique optimal hypothesis $h_* \in \mathcal{H}$ exists, and $\mathcal{H}_{\mathcal{C}} = \{h \in \mathcal{H} \,|\, h \leq h_*\}$. □

Given the assumptions of Observation 3.30, CBL can be realized as a *candidate-elimination* algorithm [269], where h_* is a compact representation of the *version space*, i.e., the subset $\mathcal{H}_{\mathcal{C}}$ of hypotheses from $\mathcal{H}$ which are consistent with the training examples.

Note that (3.22) guarantees consistency in the "empirical" sense that $r \in \widehat{\varphi}_{h,\mathcal{M}}(s)$ for all observed cases $\langle s, r\rangle \in \mathcal{D}$. Still, one might think of demanding furthermore a kind of "logical" consistency, namely $\widehat{\varphi}_{h,\mathcal{M}}(s') \neq \emptyset$ for the set of all possible inputs $s' \in \mathcal{S}$. Of course, this additional demand would greatly increase the complexity of testing consistency. Moreover, the assumptions of Observation 3.30 would no longer guarantee the existence of a unique optimal hypothesis.

Since two hypotheses h and h' are only comparable for the same underlying similarity measures (cf. Remark 3.7), the above remarks do not apply to the extended CBL problem. Thus, considering the maps $\widehat{\varphi}_{h,\mathcal{M}}$ themselves cannot be avoided in this case. Nevertheless, one can think of efficient (heuristic) approaches for realizing corresponding learning procedures. A value $U(\widehat{\varphi}_{h,\mathcal{M}})$ might be approximated, for instance, by some value $\widehat{U}(\{\widehat{\varphi}_{h,\mathcal{M}}(s) \,|\, s \in \mathcal{S}'\})$ derived from a sample $\mathcal{S}' \subseteq \mathcal{S}$. The usefulness of different (generalized) learning procedures will, however, highly depend on characteristics of the similarity measures and the way in which these measures can be adapted, i.e., on the classes $\mathcal{H}_{\mathcal{S}}$ and $\mathcal{H}_{\mathcal{R}}$. In this section, we shall restrict ourselves to the basic version of the CBL problem.

[20] Observe that $\mathcal{H}^* \subseteq \mathcal{H}_*$ does generally not hold.

3.4.2 A learning algorithm

Let hypotheses be represented by step functions

$$h : x \mapsto \sum_{k=1}^{m} \beta_k \cdot \mathbb{I}_{A_k}(x), \tag{3.23}$$

where $A_k = [\alpha_{k-1}, \alpha_k)$ for $1 \leq k \leq m-1$, $A_m = [\alpha_{m-1}, \alpha_m]$ and $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m = 1$ defines a partition of $[0,1]$.[21] The hypothesis h can then be associated with a set of rules (implications) of the form

$$(\sigma_{\mathcal{S}}(s,s') \in A_k) \;\Rightarrow\; (\sigma_{\mathcal{R}}(\varphi(s), \varphi(s')) \geq \beta_k). \tag{3.24}$$

Observe that by simply defining one interval for each element $x \in D_{\mathcal{S}}$, h_{Σ} itself can be seen as a step function if $\mathcal{S}$ is finite. A combination (3.24) of such similarity degrees seems still reasonable if $\mathcal{S}$ is not finite (or even if $\operatorname{card}(\mathcal{S})$ is large).

The class $\mathcal{H}_{step}$ of functions (3.23), defined for a fixed partition, does obviously satisfy the assumptions of Observation 3.30. The optimal hypothesis h_* is defined by the values

$$\beta_k \overset{\text{df}}{=} \min_{(s,s') \in \mathcal{C}^{\downarrow}, \sigma_{\mathcal{S}}(s,s') \in A_k} \sigma_{\mathcal{R}}(\varphi(s), \varphi(s')) \tag{3.25}$$

for $1 \leq k \leq m$, where $\min \emptyset \overset{\text{df}}{=} 1$ by convention; see Fig. 3.4 for an illustration. Since this hypothesis is directly derived from the case base $\mathcal{M}$, we also call it the *empirical similarity profile*.

Now, suppose that $\mathcal{M}$ is the current memory and that a new case $c_0 = \langle s_0, r_0 \rangle$ has been observed. Updating h_* can then be accomplished by passing the iteration

$$\beta_{\kappa(s_0,s_\jmath)} = \min\{\beta_{\kappa(s_0,s_\jmath)}, \sigma_{\mathcal{R}}(r_0, r_\jmath)\} \tag{3.26}$$

for $1 \leq \jmath \leq \operatorname{card}(\mathcal{M})$; the index $1 \leq \kappa(s,s') \leq m$ is defined for inputs $s, s' \in \mathcal{S}$ by $\kappa(s,s') = k \overset{\text{df}}{\Leftrightarrow} \sigma_{\mathcal{S}}(s,s') \in A_k$. As (3.26) shows, the representation (3.23) is computationally efficient. In fact, the time complexity of updating a hypothesis is linear in the size of the memory.[22] In other words, the model-based part of learning in CBI is not critical from a computational point of view. We refer to the algorithm defined by (3.26) as CBLA and denote by CBLA($\mathcal{C}$) the hypothesis (3.25).

For obvious reason we call $h^* \in \mathcal{H}_{step}$ defined by

$$\beta_k^* \overset{\text{df}}{=} \inf_{x \in D_{\mathcal{S}} \cap A_k} h_{\Sigma}(x) \tag{3.27}$$

($1 \leq k \leq m$) the *optimal admissible* hypothesis. Since admissibility (in the sense of Definition 3.2) implies consistency, we have $h^* \leq h_*$.

[21] In Section 3.3.1 we have hinted at the *ordinal* character of the similarity measures $\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}$. In connection with the representation of hypotheses according to (3.23) it should, therefore, be noticed that a scaling of $\sigma_{\mathcal{S}}$ might influence the optimal similarity hypothesis if the underlying partition is assumed to be *fixed*.

[22] We assume that κ is computed in constant time.

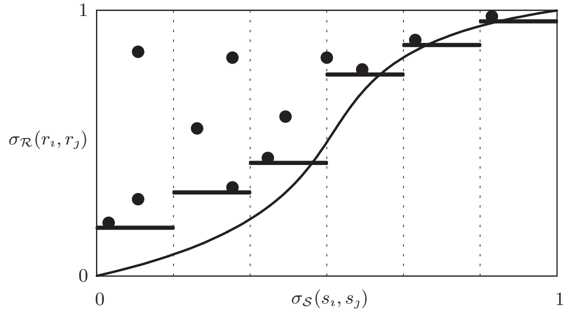

Fig. 3.4. Each pair of observed cases $\langle s_\imath, r_\imath\rangle$ and $\langle s_\jmath, r_\jmath\rangle$ contributes a point (x, y) in the "similarity space", where $y = \sigma_{\mathcal{S}}(s_\imath, s_\jmath)$ and $r = \sigma_{\mathcal{R}}(r_\imath, r_\jmath)$. By definition, these points are located above the similarity profile, which is here shown by the solid curve. The optimal similarity hypothesis h_* is given by the step function indicated by the solid horizontal lines.

Remark 3.31. Assuming the CBI hypothesis to hold true in the strict sense restricts the class $\mathcal{H}_{step}$ to the class $\mathcal{H}^{\uparrow}_{step}$ of non-decreasing step functions, which is also closed under $\vee$. Consider a hypothesis $h_* \in \mathcal{H}_{step}$ represented by values $(\beta_1, \ldots, \beta_m)$. Moreover, denote by $h^{\uparrow}_* \in \mathcal{H}^{\uparrow}_{step}$ the corresponding strict hypothesis represented by values $(\beta^{\uparrow}_1, \ldots, \beta^{\uparrow}_m)$. The relation between h_* and $h^{\uparrow}_*$ is obviously given by $\beta^{\uparrow}_k = \min\{\beta_\jmath \,|\, k \leq \jmath \leq m\}$ for all $1 \leq k \leq m$. Thus, an optimal strict hypothesis can always be derived easily from h_*. □

Remark 3.32. If a similarity hypothesis h is defined by a step function, the same is actually true for a case-based approximation $\widehat{\varphi}_{h,\mathcal{M}}$ itself. Namely, for $s, s' \in \mathcal{S}$ we have $\widehat{\varphi}_{h,\mathcal{M}}(s) = \widehat{\varphi}_{h,\mathcal{M}}(s')$ if $\kappa(s, s_\imath) = \kappa(s', s_\imath)$ for all $1 \leq \imath \leq n$. A corresponding equivalence relation on $\mathcal{S} \times \mathcal{S}$, where each equivalence class is identified by some vector $(k_1, \ldots, k_n)$ of indices $k_\jmath = \kappa(s, s_\jmath) \in \{1, \ldots, m\}$, offers some interesting possibilities of representing the mapping $\widehat{\varphi}_{h,\mathcal{M}}$ and deriving values thereof. For instance, since $\widehat{\varphi}_{h,\mathcal{M}}(s) = \widehat{\varphi}_{h,\mathcal{M}}(s')$ whenever s and s' are elements of the same equivalence relation, the values associated with the equivalence classes might be computed in advance and stored by means of an adequate data structure. The derivation of a value $\widehat{\varphi}_{h,\mathcal{M}}(s)$ then reduces to a (simple) "look-up" procedure. Admittedly, the number m^n of (potential) classes is generally extremely large, even though most of them will be empty. □

3.4.3 Properties of case-based learning

We shall now consider an iterative scheme which is in accordance with the idea of CBI as a repeated process of problem solving and learning. This case-based learning process, called CBLP and outlined in Algorithm 1, is based on a random sequence $(S_N)_{N \geq 1}$ of inputs $S_N \in \mathcal{S}$ which are independent and identically distributed according to $\mu_{\mathcal{S}}$, and a sequence $p = (p_N)_{N \geq 1} \in [0,1]^\infty$.

Algorithm 1 CBLP

```
Input: a sequence of query inputs
Output: a sequence of estimation for outputs
1: $\mathcal{M}_0 = \emptyset$, $h_0 \equiv 1$
2: $N = 0$
3: repeat
4:     compute $\widehat{r}_{N+1} = \widehat{\varphi}_{h_N, \mathcal{M}_N}(s_{N+1})$
5:     solve-problem($s_{N+1}, \widehat{r}_{N+1}$)
6:     $h_{N+1} =$ update($h_N, c_{N+1}, \mathcal{M}_N$)
7:     $\mathcal{M}_{N+1} = \begin{cases} \mathcal{M}_N \cup (c_{N+1}) & \text{with probability } p_{N+1} \\ \mathcal{M}_N & \text{with probability } 1 - p_{N+1} \end{cases}$
8:     $N = N + 1$
9: until no more queries exist
```

Here, `solve-problem` is a procedure in which the prediction $\widehat{r}_{N+1}$ is used for supporting the derivation of the true outcome $\varphi(s_{N+1})$. Moreover, the procedure `update`$(h_N, c_{N+1}, \mathcal{M}_N)$ returns the hypothesis obtained from h_N by passing the iteration (3.26) for $\mathcal{M}_N$ and the case $c_{N+1} = \langle s_{N+1}, \varphi(s_{N+1})\rangle$. Observe that CBLP guarantees $h_N = \text{CBLA}(\mathcal{C}_N)$ but that we generally have $h_N \neq \text{CBLA}(\mathcal{D}_N \times \mathcal{D}_N)$. The probabilistic extension of the memory in CBLP takes into account that adding all observations to $\mathcal{M}$, i.e., taking $p \equiv 1$, might not be advisable [353]. Of course, efficient problem solving will generally assume a more sophisticated strategy for the instance-based aspect of learning, i.e., for maintaining the memory of cases. It might be reasonable, e.g., to take the "quality" of individual cases into account and to allow for removing already stored cases from the memory [355, 286]. Nevertheless, the probabilistic extension in CBLP allows for gaining insight into theoretical properties of the learning scheme. Observe that $p_N = 0$ for $N \geq N_0$ (with N_0 being a constant number) comes down to using a fixed memory $\mathcal{M}$.

Given a CBI setup and the sequence $(p_N)_{N \geq 1}$, the hypotheses h_N induced by CBLP are random functions with well-defined (even though tremendously complicated) distributions. We are now going to derive some important properties of the sequence $(h_N)_{N \geq 1}$. It goes without saying that one of the first questions arising in connection with our learning scheme concerns the relation between $(h_N)_{N \geq 1}$ and the optimal admissible hypothesis h^*.

Proposition 3.33. Suppose $p \geq \delta > 0$, i.e., $p_N \geq \delta$ for all $N \in \mathfrak{N}$, and let $(h_N)_{N \geq 1}$ be the sequence of hypotheses induced by CBLP. Then, $h_N \searrow h^*$ stochastically as $N \to \infty$. That is, $h_N \geq h^*$ for all $N \in \mathfrak{N}$ and

$$\mathbb{P}(\|h_N - h^*\|_\infty \geq \varepsilon) \to 0$$

for all $\varepsilon > 0$. □

Proof. From the definition of h^* and the updating scheme (3.26) it becomes obvious that $h^* \leq h_N$ for all $N \geq 1$ and that the sequence of functions $(h_N)_{N \geq 0}$ is decreasing. Let $\varepsilon > 0$ and consider some $1 \leq k \leq m$. According to (3.27), there is some $x \in A_k$ such that $|h_\Sigma(x) - \beta_k^*| < \varepsilon/2$. Since we have $h_\Sigma(x) = \inf\left\{\sigma_\mathcal{R}(\varphi(s), \varphi(s')) \,|\, s, s' \in \mathcal{S}, \sigma_\mathcal{S}(s, s') = x\right\}$, there are also values $s_{k_1}, s_{k_2} \in \mathcal{S}$ such that $\sigma_\mathcal{S}(s_{k_1}, s_{k_2}) = x$ and $|\sigma_\mathcal{R}(\varphi(s_{k_1}), \varphi(s_{k_2})) - h_\Sigma(x)| < \varepsilon/2$. Hence, $|\sigma_\mathcal{R}(\varphi(s_{k_1}), \varphi(s_{k_2})) - \beta_k^*| < \varepsilon$. This implies $|h_{\mathcal{M}_N}(x) - \beta_k^*| < \varepsilon$ as soon as the memory $\mathcal{M}_N$ contains the inputs s_{k_1} and s_{k_2}, where $h_{\mathcal{M}_N} = \text{CBLA}(\mathcal{M}_N)$. Since this argumentation applies to all $1 \leq k \leq m$ and since $h^* \leq h_N \leq h_{\mathcal{M}_N}$, we obtain

$$\|h_N - h^*\|_\infty \leq \|h_{\mathcal{M}_N} - h^*\|_\infty = \max_{0 \leq x \leq 1} |h_{\mathcal{M}_N}(x) - h^*(x)| < \varepsilon$$

if $\mathcal{M}_N$ contains the (at most $2\,m$) inputs s_{k_1}, s_{k_2} $(1 \leq k \leq m)$. Since $\mu_\mathcal{S}(s_{k_1}) > 0$ and $\mu_\mathcal{S}(s_{k_2}) > 0$ for all $1 \leq k \leq m$ and $p_N \geq \delta > 0$ for all $N \in \mathfrak{N}$, the probability for this tends toward 1 for $N \to \infty$. □

Observe that the stochastic convergence (from above) of the hypotheses $(h_N)_{N \geq 0}$ toward $h^* \in \mathcal{H}_{step}$, which is guaranteed by Proposition 3.33, does not imply that $h_N(x) \to h_\Sigma(x)$ for all $x \in D_\mathcal{S}$. In fact, it might happen that $h^*|D_\mathcal{S}$ is already a poor approximation of h_Σ (at least in the strong sense of the $\|\cdot\|_\infty$ metric) regardless of the (finite) partition underlying the definition of the hypothesis space $\mathcal{H}_{step}$. The following example shows that this cannot be avoided even if the system $(\mathcal{S}, \mathcal{R}, \varphi)$ satisfies strong structural assumptions:

Example 3.34. Let $\mathcal{S} = \{s_k = k - (1/2)^k \,|\, k \in \mathfrak{N}_0\}$, $\mathcal{R} = \{0, 1\}$, and

$$\varphi(s_k) = \begin{cases} 0 & \text{if} \quad \lfloor k/2 \rfloor \text{ is odd} \\ 1 & \text{if} \quad \lfloor k/2 \rfloor \text{ is even} \end{cases} .$$

Moreover, let $\sigma_\mathcal{S}(s, s') = |s - s'|^{-1}$ and $\sigma_\mathcal{R}(r, r') = 1 - |r - r'|$ (and note that $\varphi : (\mathcal{S}, |\cdot|) \longrightarrow (\mathcal{R}, |\cdot|)$ does even satisfy a Lipschitz condition). Now, for $\alpha_k = 2^k/(2^k + 1)$ $(k \in \mathfrak{N})$ there are exactly two inputs $s, s' \in \mathcal{S}$ such that $\sigma_\mathcal{S}(s, s') = \alpha_k$, namely $s = s_{k-1}$ and $s' = s_k$ (or vice versa). Thus, we have

$$h_\Sigma(\alpha_k) = \sigma_\mathcal{R}(\varphi(s_{k-1}), \varphi(s_k)) = \begin{cases} 1 & \text{if} \quad \text{k is odd} \\ 0 & \text{if} \quad \text{k is even} \end{cases} .$$

Obviously, each finite partition of $[0, 1]$ contains an interval A such that $\alpha_k, \alpha_{k+1} \in A$ for some $k \geq 1$. Consequently, $h^*|A \equiv 0$ and, hence, $\|h^*|D_\mathcal{S} - h_\Sigma\|_\infty = 1$. □

The convergence from above established by Proposition 3.33 already suggests that we will generally have $h_N(x) > h_\Sigma(x)$ for some $x \in D_\mathcal{S}$ in the course of a CBL process. Thus, we might work with inadmissible hypotheses (see also Fig. 3.4, where $h_* \leq h_\Sigma$ does not hold). This, of course, seems to conflict with the objective of providing an outer approximation of φ. Indeed, it can easily be shown that h_Σ is the largest function h (defined on $D_{\mathcal{S}}$) such that $\varphi(s) \in \widehat{\varphi}_{h,\mathcal{M}}(s)$ for all $s \in \mathcal{S}$ is guaranteed regardless of the memory $\mathcal{M}$. In other words, for each function h with $h(x) > h_\Sigma(x)$ for at least one $x \in D_{\mathcal{S}}$, a memory $\mathcal{M}$ can be found such that $\varphi(s) \not\in \widehat{\varphi}_{h,\mathcal{M}}(s)$ for at least one $s \in \mathcal{S}$. Observe, however, that the approximation $\widehat{\varphi}_{h_N,\mathcal{M}_N}$ is derived from the *specific* memory $\mathcal{M}_N$. Thus, the fact that $h_N(x) > h_\Sigma(x)$ for some $x \in D_{\mathcal{S}}$ does by no means rule out the possibility of $\widehat{\varphi}_{h_N,\mathcal{M}_N}$ being an outer approximation of φ. In connection with CBLP one might therefore be interested in the probabilities

$$q_{N+1} = \mathbb{P}\left(\varphi(S_{N+1}) \not\in \widehat{\varphi}_{h_N,\mathcal{M}_N}(S_{N+1})\right) \tag{3.28}$$

of incorrect predictions.

Consider a memory $\mathcal{M}$, a hypothesis h, and an input $s_0 \in \mathcal{S}$. We call s_0 *extremal*[23] (with respect to $\mathcal{M}$ and h) if $h \neq \texttt{update}(h, s_0, \mathcal{M})$, i.e., if there is some $1 \leq k \leq m$ and a case $\langle s, r\rangle \in \mathcal{M}$ such that $\sigma_{\mathcal{S}}(s, s_0) \in A_k$ and

$$\forall \, \langle s', r'\rangle \in \mathcal{M} \, : \, (\sigma_{\mathcal{S}}(s, s') \in A_k) \Rightarrow (\sigma_{\mathcal{R}}(r, r_0) < \sigma_{\mathcal{R}}(r, r')) \, .$$

Lemma 3.35. For a memory $\mathcal{M}$, a hypothesis $h \leq \mathrm{CBLA}(\mathcal{M})$, and an input $s_0 \in \mathcal{S}$ suppose that $\varphi(s_0) \not\in \widehat{\varphi}_{h,\mathcal{M}}(s_0)$. Then, s_0 is extremal. □

Proof. Suppose $r_0 \not\in \widehat{\varphi}_{h,\mathcal{M}}(s_0)$. Then, we find a case $\langle s, r\rangle \in \mathcal{M}$ such that $r_0 \not\in \mathcal{N}_{h(\sigma_{\mathcal{S}}(s,s_0))}(r)$. This means that $\sigma_{\mathcal{R}}(r, r_0) < h(\sigma_{\mathcal{S}}(s, s_0))$ and, since $h \leq \mathrm{CBLA}(\mathcal{M})$, $\sigma_{\mathcal{R}}(r, r_0) < \sigma_{\mathcal{R}}(r, r')$ for all cases $\langle s', r'\rangle \in \mathcal{M}$ satisfying $\sigma_{\mathcal{S}}(s, s') \in A_{\kappa(s,s_0)}$. Hence, s_0 is extremal. □

Proposition 3.36. The following estimation holds true for the probability (3.28):

$$\begin{aligned} q_{N+1} \; &\leq \; \sum_{n=0}^{N} \frac{2m}{n+1} \cdot \mathbb{P}(\mathrm{card}(\mathcal{M}_N) = n) && (3.29) \\ &\leq \; \frac{2m}{1 + \mathbb{E}(\mathrm{card}(\mathcal{M}_N))} = \frac{2m}{1 + \sum_{k=1}^{N} p_k}, && (3.30) \end{aligned}$$

where m is the size of the partition underlying $\mathcal{H}_{step}$ and $\mathbb{E}$ denotes the expected value operator. □

Proof. Suppose M_N to consist of $n \leq N$ cases, i.e., $\mathcal{M}_N$ is defined by some random (sub-)sequence $(S_{\pi(1)}, \ldots, S_{\pi(n)})$ of inputs, where $1 \leq \pi(1) < \pi(2) < \ldots < \pi(n) \leq N$. Moreover, consider a new input $S_0 = S_{N+1}$ and observe that

$$\mathbb{P}(\varphi(S_0) \not\in \widehat{\varphi}_{h_N,\mathcal{M}_N}(S_0)) \leq \mathbb{P}(\varphi(S_0) \not\in \widehat{\varphi}_{h_{\mathcal{M}_N},\mathcal{M}_N}(S_0)),$$

[23] This definition of being extremal is to some extent related to the concept of "strangeness" of an observation in the context of so-called confidence machines [162, 301].

where $h_{\mathcal{M}_N} = \text{CBLA}(\mathcal{M}_N)$. From the random sequence $(S_{\pi(1)}, \ldots, S_{\pi(n)}, S_0)$ of inputs we can choose a set $\mathcal{M}'$ of (at most) $2m$ inputs resp. associated cases such that $\text{CBLA}(\mathcal{M}_N \cup \{\langle S_0, \varphi(S_0)\rangle\}) = \text{CBLA}(\mathcal{M}')$. Obviously, $\langle S_0, \varphi(S_0)\rangle \notin \mathcal{M}'$ implies that S_0 is not extremal. Now, recall that inputs are independent and identically distributed according to $\mu_{\mathcal{S}}$. Thus, the value $2m/(n+1)$ defines an (upper) bound to the probability that $\langle S_0, \varphi(S_0)\rangle \in \mathcal{M}'$ due to reasons of symmetry. We hence obtain

$$\begin{aligned} &\mathbb{P}(\varphi(S_0) \notin \widehat{\varphi}_{h_N,\mathcal{M}_N}(S_0) \mid \text{card}(\mathcal{M}_N) = n) \leq \\ &\qquad \mathbb{P}(\varphi(S_0) \notin \widehat{\varphi}_{h_{\mathcal{M}_N},\mathcal{M}_N}(S_0) \mid \text{card}(\mathcal{M}_N) = n) \leq 2m/(n+1) \end{aligned}$$

from Lemma 3.35. Then, (3.29) and (3.30) follow from the theorem of total probability and Jensen's inequality, respectively. □

Corollary 3.37. Suppose $p \geq \delta > 0$. Then, $q_{N+1} \leq 2m/(\delta N + 1)$. Particularly, $q_{N+1} \leq 2m/(N+1)$ if $p \equiv 1$. □

According to the above results, the probability of an incorrect prediction becomes small for large memories, even though the hypotheses h_N might be inadmissible. Under the assumptions of Corollary 3.37, this probability tends toward 0 with a convergence rate of order $O(1/N)$.

Corollary 3.38. Suppose $p \geq \delta > 0$. Then, the expected proportion of incorrect predictions in connection with CBLP converges toward 0. □

Proof. Define the random variable V_n $(n \geq 1)$ by means of $V_n = 1$ if the n-th prediction is incorrect, i.e., if $\varphi(S_n) \notin \widehat{\varphi}_{h_{n-1},\mathcal{M}_{n-1}}(S_n)$, and $V_n = 0$ otherwise. Then, $\mathbb{E}(V_n) = q_n$, where $\mathbb{E}(V_n)$ denotes the expected value of V_n, and

$$\begin{aligned} \mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N} V_n\right) &= \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}(V_n) \\ &\overset{\text{(Cor. 3.37)}}{\leq} \frac{1}{N}\sum_{n=1}^{N} 2m/(\delta n) \\ &\leq \frac{2m(1+\ln(N))}{\delta N} \to 0 \end{aligned}$$

as $N \to \infty$. □

Example 3.39. Fig. 3.5 shows the optimal hypothesis h^* for the setup Σ_1 defined in Example 2.5 (cf. Section 2.4.1) and the hypothesis $h_{\mathcal{M}}$ for a typical memory $\mathcal{M}$ of size 250, generated by a sequence of inputs chosen at random. The underlying partition has been defined by the values $\alpha_k = k/10$ $(k = 0, \ldots, 10)$. The same figure shows a characterization of the evolution of the approximation quality in the form of the values $\|h^* - h_{\mathcal{M}_n}\|_2$ and $\|h^* - h_{\mathcal{M}_n}\|_\infty$ $(n = 1, \ldots, 300)$, where $\|\cdot\|_p$ denotes the corresponding $\mathcal{L}^p$-norm. □

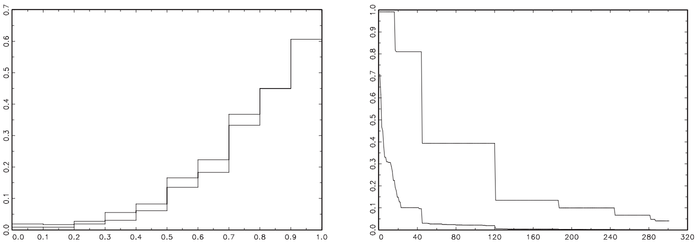

Fig. 3.5. Left: Optimal hypothesis h^* for the setup Σ_1 in Example 2.5 and the hypothesis $h_{\mathcal{M}}$ for a memory $\mathcal{M}$ of size 250. Right: Evolution of approximation quality $\|h^* - h_{\mathcal{M}_n}\|_2$ and $\|h^* - h_{\mathcal{M}_n}\|_\infty$ (cf. Example 3.39).

The upper bound established in Proposition 3.36 might suggest to reduce the probability of an incorrect prediction by reducing the size m of the partition underlying $\mathcal{H}_{step}$. Observe, however, that this will also lead to a less precise approximation of h_Σ and, hence, to less precise predictions of outcomes. “Merging” two neighbored intervals A_k and A_{k+1}, for instance, means to define a new hypothesis h with $h|(A_k \cup A_{k+1}) \equiv \min\{\beta_k, \beta_{k+1}\}$. In fact, the probability of an incorrect prediction can be made arbitrarily small by increasing the size of the memory. The precision of the predictions, however, is limited by the precision to which h_Σ can be approximated by h^* and, hence, by the granularity of the partition underlying the definition of the hypothesis space $\mathcal{H}_{step}$. Of course, nothing prevents us from extending our approach to CBL such that it allows for the adaptation of the partition. A refinement of the latter will make sense, e.g., if the size of the memory becomes large.

Let us now consider the *fixed memory-model*, i.e., the case where CBI is based on a fixed memory $\mathcal{M} = (c_1, \ldots, c_n)$ of size $n \geq 1$. The objective of CBL is then to find an approximation of the $\mathcal{M}$-similarity profile $h_\Sigma^{\mathcal{M}}$. Thus, the consistency principle (3.22) should hold true for $\mathcal{C} = \mathcal{M} \times \mathcal{D}$. Again, the class $\mathcal{H}_*$ consists of the uncertainty minimizing hypotheses in $\mathcal{H}_{\mathcal{C}}$. Likewise, $\mathcal{H}^*$ is made of those uncertainty minimizing hypotheses that satisfy (3.22) for $\mathcal{M} \times \mathcal{D}^*$. Observation 3.30 does obviously remain correct. The hypothesis $h_* = \mathrm{CBLA}_{\mathcal{M}}(\mathcal{D})$ is now defined by the values

$$\beta_k = \min \left\{ \sigma_{\mathcal{R}}(r, r') \,|\, \langle s, r\rangle \in \mathcal{M}, \langle s', r'\rangle \in \mathcal{D}, \sigma_{\mathcal{S}}(s, s') \in A_k \right\}.$$

Thus, given a new observation, the update of the current hypothesis is realized by passing the iteration (3.26) for the n cases in $\mathcal{M}$. The fixed-memory version of CBLP, denoted $\mathrm{CBLP}_{\mathcal{M}}$, is outlined in Algorithm 2.

For the hypotheses h_N induced by $\mathrm{CBLP}_{\mathcal{M}}$ we do not only obtain an upper approximation but even $h_N = \mathrm{CBLA}_{\mathcal{M}}(\mathcal{D}_N)$.

Algorithm 2 CBLP$_{\mathcal{M}}$

```
Input: a sequence of query inputs
Output: a sequence of estimation for outputs
$h_0 = \mathrm{CBLA}(\mathcal{M})$
$N = 0$
repeat
    compute $\widehat{r}_{N+1} = \widehat{\varphi}_{h_N,\mathcal{M}}(s_{N+1})$
    solve-problem($s_{N+1}, \widehat{r}_{N+1}$)
    $h_{N+1} =$ update($h_N, c_{N+1}, \mathcal{M}$)
    $N = N + 1$
until no more queries exist
```

Proposition 3.40. For the sequence $(h_N)_{N \geq 1}$ induced by CBLP$_{\mathcal{M}}$ it holds true that $h_N \searrow h^*$ stochastically as $N \to \infty$, where h^* is defined by the values $\beta_k^* = \inf\{h_{\Sigma}^{\mathcal{M}}(x) \,|\, x \in D_{\mathcal{S}} \cap A_k\}$ $(1 \leq k \leq m)$. □

Proposition 3.41. In connection with the fixed memory-model we obtain the estimation $q_{N+1} \leq 2m/(N+1)$ for the probability (3.28), where m is the size of the partition underlying $\mathcal{H}_{step}$. □

Proof. Consider the random sequence $(S_1, \ldots, S_N, S_0)$ of $N+1$ inputs. From this sequence we can choose a set $\mathcal{D}$ of (at most) $2m$ inputs resp. associated cases such that $\mathrm{CBLA}_{\mathcal{M}}(\mathcal{D}_N \cup \{\langle S_0, \varphi(S_0)\rangle\}) = \mathrm{CBLA}_{\mathcal{M}}(\mathcal{D})$. Now, recall that $\langle S_0, \varphi(S_0)\rangle \notin \mathcal{D}$ implies that S_0 is not extremal with respect to h_N and $\mathcal{M}$ and that inputs are independent and identically distributed according to $\mu_{\mathcal{S}}$. Thus, the value $2m/(N+1)$ defines an (upper) bound to the probability that $\langle S_0, \varphi(S_0)\rangle \in \mathcal{D}$ due to reasons of symmetry. The rest follows from Lemma 3.35. □

Corollary 3.42. The expected proportion of incorrect predictions in connection with CBLP$_{\mathcal{M}}$ converges toward 0. □

It should be noticed that CBLP$_{\mathcal{M}}$ is closely related to CBLP in the case where some $N_0 \in \mathfrak{N}$ exists such that $p_N = 0$ for all $N \geq N_0$. Suppose for instance, that $p_N = 1$ for $1 \leq N < N_0$ and $p_N = 0$ for $N \geq N_0$. Then, Proposition 3.41 remains correct with CBLP$_{\mathcal{M}}$ replaced by CBLP. Proposition 3.40 remains correct if, moreover, $h_{\Sigma}^{\mathcal{M}}$ is replaced by $h_{\Sigma}^{\mathcal{M}_{N_0-1}}$. The result of Proposition 3.41 can also be used for deriving the following generalizations of Proposition 3.36 and Corollary 3.38.

Proposition 3.43. Let $N_0 \in \mathfrak{N}$ and suppose $p_N = 1$ for $N \geq N_0$. We then obtain the estimation

$$q_{N+1} \leq 2m \left(1 + \max\{0, N - N_0\} + \sum_{k=1}^{N_0-1} p_k \right)^{-1},$$

where m is the size of the partition underlying $\mathcal{H}_{step}$. □

Corollary 3.44. Let $N_0 \in \mathfrak{N}$ and suppose $p_N = 1$ for $N \geq N_0$. Then, the expected proportion of incorrect predictions in connection with CBLP converges toward 0. □

Summing up, the results of this section throw light on some interesting properties of our approach to case-based learning. In fact, the combination of case-based inference and case-based learning, i.e., the application of the prediction scheme of Section 3.2.1 with a hypothesis derived by means of CBLA, allows for deriving a set-valued prediction $\widehat{\varphi}(s_0) = \widehat{\varphi}_{h,\mathcal{M}}(s_0)$ which covers the true outcome with a high probability. In a statistical sense, $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ can thus be seen as a kind of confidence region or *credible output set*, a justification for designating the above inference scheme as *credible case-based inference*.

Remark 3.45. In many applications one is interested in both, a credible output set and a "point-estimation" of the output r_0, i.e., a distinguished element $\hat{r}_0 \in \mathcal{R}$ that can be considered as representative. The latter can be derived from the credible output set $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ as a *generalized median*:

$$\hat{r}_0 \overset{\text{df}}{=} \arg \max_{r \in \widehat{\varphi}_{h,\mathcal{M}}(s_0)} \sum_{r' \in \widehat{\varphi}_{h,\mathcal{M}}(s_0)} \sigma_{\mathcal{R}}(r, r') \tag{3.31}$$

As can be seen, the generalized median is a kind of center-point, namely the element of the credible output set which is maximally similar to all other elements. □

Note that the concrete probability of a correct prediction depends on the number of observed cases and can thus be estimated in advance. Moreover, it can be made arbitrarily large by extending the size of the memory. CBLP, the combination of CBI and CBL, can thus be seen as an interesting method of statistical inference. Principally, it defines a generalized instance-based learning algorithm which takes uncertainty in connection with the prediction of outcomes into account. This aspect will be discussed in more detail in Section 3.5 below.

Let us finally mention that results similar to the ones derived in this section can also be obtained in connection with other types of similarity profiles. Recall, for instance, the concept of a *local* similarity profile: Let $\mathcal{M}$ be a memory of cases, namely a subset $\mathcal{M} \subseteq \mathcal{D}$ of the cases $\langle s_n, r_n \rangle$ $(1 \leq n \leq N)$ which have been encountered so far. For $\langle s, r \rangle \in \mathcal{M}$ we define the *local hypothesis* h^s by the values

$$\beta_k \overset{\text{df}}{=} \min_{1 \leq n \leq N : \sigma_{\mathcal{S}}(s, s_n) \in A_k} \sigma_{\mathcal{R}}(\varphi(s), \varphi(s_n)). \tag{3.32}$$

The *local $\mathcal{M}$-hypothesis* is given by $h^{\mathcal{M}} \overset{\text{df}}{=} \{h^s \,|\, s \in \mathcal{M}^{\downarrow}\}$. We can then prove a result similar to Proposition 3.36:

Proposition 3.46. Suppose that N (independent and identically distributed) cases have been encountered so far. For a subset $\mathcal{M}$ containing $|\mathcal{M}|$ cases let a local $\mathcal{M}$-hypothesis be defined according to (3.32). Moreover, let $s_0 \in \mathcal{S}$ be a new problem (chosen at random from $\mathcal{S}$). The probability that the true outcome $r_0 = \varphi(s_0)$ is not covered by

$$\widehat{\varphi}_{h^{\mathcal{M}},\mathcal{M}}(s_0) = \bigcap_{\langle s,r\rangle \in \mathcal{M}} \mathcal{N}_{h^s(\sigma_{\mathcal{R}}(s,s_0))}(r) \tag{3.33}$$

is bounded from above by $|\mathcal{M}|m/(N+1)$. □

A prediction (3.33) based on a local $\mathcal{M}$-hypothesis is generally more precise than a prediction (3.2). At the same time, however, the associated confidence level is smaller. Still, Proposition 3.46 shows that this level can be made arbitrarily large by increasing the number of observed cases.

Note that it might not be possible to compute the hypothesis (3.32) exactly if only some of the encountered cases $\langle s_n, r_n\rangle \in \mathcal{D}$ are added to $\mathcal{M}$. However, Proposition 3.46 remains valid (up to some minor modifications) if the minimum in (3.32) is not taken over all (pairs) of cases.

3.4.4 Experimental results

The basic learning scheme presented in Section 3.4.2 offers a convenient framework which enables the realization of methods for predicting unknown outcomes based on a sequence of observed cases. The results of Section 3.4.3 show that corresponding predictions take the form of confidence regions which cover the unknown output with a certain probability. In this section, we shall present some small examples in order to convey how this approach works in practice. These examples are not meant as an empirical evaluation of our CBI method, they are only intended to provide an illustration of the theoretical results derived above.

We have organized two experimental studies as follows: First of all, a target function φ with domain $\mathcal{S}$ and range $\mathcal{D}$ is specified. A single run of a simulation corresponds to the CBLP scheme presented in Section 3.4.3, where $p \equiv 1$, a new input is chosen according to the uniform distribution, and the length of the generated random sequence of inputs is 1000. The size of the partition underlying the learned similarity hypothesis is $m = 20$. Given a new input S_{N+1}, a prediction $\widehat{\varphi}_{h_N,\mathcal{M}_N}(S_{N+1})$ is derived from the hypothesis h_N and the memory $\mathcal{M}_N$ according to (3.15) or (3.16). Two characteristic quantities are recorded for this estimation. Firstly, the *correctness* is captured by means of $V_N \in \{0,1\}$, where $V_N = 1$ iff $(\varphi(S_{N+1}) \in \widehat{\varphi}_{h_N,\mathcal{M}_N}(S_{N+1}))$. Secondly, the *precision* is specified by $P_N \overset{\mathrm{df}}{=} \operatorname{diam}(\widehat{\varphi}_{h_N,\mathcal{M}_N}(S_{N+1}))$. The behavior of the prediction method can then be characterized by means of the expected values $\mathbb{E}(V_N)$ and $\mathbb{E}(P_N)$ associated with the sequences $(V_1,\ldots,V_{1000})$ and $(P_1,\ldots,P_{1000})$, respectively. Approximations of these expected values have been obtained by deriving mean values $\overline{V}_N$ and $\overline{P}_N$ from a large number of simulation runs. The respective sequences $(\overline{V}_1, \ldots, \overline{V}_{1000})$ and $(\overline{P}_1, \ldots, \overline{P}_{1000})$ constitute the results which are finally presented in Appendix D. Note that $1-\overline{V}_N$ is an estimation of the probability q_N specified in (3.28).

For the first example, we have chosen the relatively simple function

$$\varphi : s \mapsto \sin(s+1) \cdot \cos^2(s),$$

where $\mathcal{S} = [0, \pi/2] \cap \mathfrak{Q}$ (and $\mathcal{R} = \varphi(\mathcal{S}) \subseteq [0, 1.2]$). The results are shown in Fig. D.1–Fig. D.3. As it was to be expected from the theoretical results of Section 3.4.3, the probability of an incorrect prediction soon becomes very small. Of course, the more cases are used for constraining the outcome, the more precise the predictions become. At the same time, however, this also increases the probability of an incorrect prediction. The approximation (3.16), using a constant number of $k = 10$ cases, shows that the expected precision of a prediction is not necessarily a monotone function of the size of the memory (cf. Fig. D.2). This effect is not restricted to (3.16) but can also occur in connection with (3.15), i.e., if all cases are used. It is caused by two opposite effects related to the extension of a memory. On the one hand,

$$\mathcal{M}' \subseteq \mathcal{M} \;\Rightarrow\; \widehat{\varphi}_{h,\mathcal{M}}(s) \subseteq \widehat{\varphi}_{h,\mathcal{M}'}(s)$$

for all hypotheses h, memories $\mathcal{M}, \mathcal{M}'$, and $s \in \mathcal{S}$. That is, the larger a memory is, the more precise the approximation becomes. On the other hand,

$$h \leq h' \;\Rightarrow\; \widehat{\varphi}_{h',\mathcal{M}}(s) \subseteq \widehat{\varphi}_{h,\mathcal{M}}(s)$$

for all hypotheses h, h', i.e., the less strong a hypothesis is, the less precise the approximation becomes. The aforementioned effect is then explained by the fact that a case-based approximation is derived from a memory $\mathcal{M}$ and the associated hypothesis $h_{\mathcal{M}}$ and that $\mathcal{M}' \subseteq \mathcal{M}$ implies $h_{\mathcal{M}} \leq h_{\mathcal{M}'}$.

The simulation results might give the impression that the expected precision of predictions converges toward some value which is larger than 0. Even though this might happen in certain cases, it is actually not true for our example. In fact, this example reflects a typical situation where the expected precision indeed converges toward 0, but where the improvement due to additional observations decreases with the size of the memory. In other words, the convergence rate might be rather low. This can also be illustrated by means of the simple example $\varphi : s \mapsto s^2$, $s \in \mathcal{S} = [0, 1]$.[24] For the CBI setup using $\sigma_{\mathcal{S}} : (s, s') \mapsto 1 - |s - s'|$ and $\sigma_{\mathcal{R}} : (r, r') \mapsto 1 - |r - r'|$ we obtain $h_{\Sigma}(x) = x^2$. Moreover, it can be shown that (3.15) leads to $\widehat{\varphi}_{h_{\Sigma},\mathcal{M}}(0) = [0, 2\min\{s_1, \ldots, s_n\}]$, where $s_1, \ldots, s_n$ denote the inputs which have already been observed, i.e., which define the memory $\mathcal{M}$. That is, the expected precision of the prediction of $\varphi(0)$, i.e., the length of the above interval, is given by the random variable $X \overset{\mathrm{df}}{=} 2\min\{S_1, \ldots, S_n\}$,

[24] For the sake of simplicity, we put up with the fact that $\mathcal{S}$ violates our assumption of countability.

where $S_1, \ldots, S_n$ are independent random variables distributed according to $\mu_{\mathcal{S}}$. If the latter is taken as the uniform measure over $[0,1]$, it is not difficult to show that $\mathbb{E}(X) = 2/(n+1)$. Thus, the expected precision converges toward 0 with a convergence rate of $O(1/n)$.

The second experimental study uses the CBI setup Σ_1 which has been introduced in Example 2.5, i.e., a value $\varphi(s)$ is defined as the cost of the optimal solution associated with the combinatorial optimization problem encoded by s. The results of this study, shown in Fig. D.4–Fig. D.7, are qualitatively similar to those of the first experiment. As can be seen in Fig. D.4, the non-monotone behavior of the expected precision of predictions now also occurs in connection with the case-based approximation (3.15). It should be remarked that the results are quite satisfactory in the sense that a rather small fraction of the $\text{card}(\mathcal{S}) = 7^5$ cases suffices for deriving relatively precise predictions of cost values (which are between 0 and 48). A memory of size 1000, for instance, corresponds to a fraction of approximately 6/100, i.e., a prediction based on the 10 most similar cases uses only slightly more than 0.06% of the cases.

Let us finally consider a "real-world" application. In connection with the Housing Database,[25] we have used CBI for predicting prices of houses which are characterized by 13 attributes. Similarity was defined as an affine-linear function of the distance between (real-valued) attribute values. For randomly chosen memories of size 30 we have used 450 cases as training examples in order to learn the respective local $\mathcal{M}$-profiles. Based on (local) hypotheses thus obtained, CBI allowed for predicting prices of the remaining 56 cases with a precision of approximately 10,000 dollars and a confidence level around 0.85. Taking the generalized median (3.31) as a point-estimation, which here simply corresponds to the center of the interval, one thus obtains predictions of the form $x \pm 5,000$ dollars. As can be seen, these estimations are quite reliable but not extremely precise (the average price of a house is approximately 22,500 dollars). In fact, this example clearly points out the limits of an inference scheme built upon the CBI hypothesis. Our approach takes these limits into account and makes them explicit: A case-based prediction of prices cannot be confident and extremely precise at the same time, simply because the housing data meets the CBI hypothesis but moderately. Needless to say, problems of such type are of a general nature and by no means specific to case-based inference. Linear regression, for example, assumes a linear relationship between the dependent and independent variables. It yields poor predictions and imprecise confidence intervals if this assumption is not satisfied (which is often the case in practice).

[25] Available at `http://www.ics.uci.edu/~mlearn`.

3.5 Application to statistical inference

It has already been mentioned that our approach to case-based learning (Section 3.4) gives rise to an extension of the inference scheme of Section 3.2 which provides us with an interesting statistical inference mechanism. In fact, it is just the attached level of confidence which makes a (set-valued) prediction (3.2) attractive from a statistical perspective. In order to emphasize this point, we have already used the term *credible* CBI, referring to the combination of the inference scheme (3.2) and the case-based learning algorithm of Section 3.4: Given a randomly chosen memory $\mathcal{M}$ of cases and a new input s_0, CBI derives a hypothesis $h = \mathrm{CBLA}(\mathcal{M})$ and delivers a prediction

$$(\widehat{\varphi}_{h,\mathcal{M}}(s_0), \alpha)$$

such that

$$\mathbb{P}\left(\varphi(s_0) \in \widehat{\varphi}_{h,\mathcal{M}}(s_0)\right) \geq 1 - \alpha.$$

This section is meant to outline briefly two applications which show that credible CBI can complement existing statistical methods in a reasonable way.

3.5.1 Case-based parameter estimation

In order to show how credible CBI might support classical approaches to statistical inference let us consider the idea of *case-based parameter estimation.* Thus, the task is to estimate an unknown parameter $\vartheta \in \Theta$, where Θ denotes an underlying class of parameters. Quite often, the estimation of ϑ according to, say, the Maximum Likelihood (ML) principle, is a computationally complex problem involving numerical optimization methods. The computation of an ML estimation (MLE) is hence impossible if such estimations have to be made available frequently, perhaps even under strict time constraints. As an example one might think of a control problem where data is obtained from monitoring a technical system and where the MLE serves as a control parameter [219]. Likewise, online data analysis and estimation problems arise in mining so-called *data streams* [92, 161].

If the (repeated) derivation of an MLE is computationally too complex, credible CBI might be used for estimating it. More specifically, we can derive a confidence region for the MLE based on a set of data–MLE tuples and a new set of data. Using our terminology, the data plays the role of an input and the MLE corresponds to the output. The data–MLE tuples which constitute the memory may originate from other estimations or may have been derived during a less time-critical preprocessing phase.

The CBI hypothesis now means that similar data leads to similar ML estimations, an assumption which appears reasonable for many applications. Still, the choice of an adequate measure for determining the similarity between two sets of data will generally not be obvious. Since the adequacy of a measure depends on the respective application, we will not go into detail here. Let us only mention that it will often be possible to simplify the problem by passing from the data itself to *sufficient statistics* thereof, i.e., to consider sufficient statistics as inputs which determine the output in the form of an MLE.

In general, one will be interested in a confidence region not for the MLE ϑ_{ML} but for the *true* parameter ϑ of an underlying stochastic model. Suppose that a confidence region for ϑ takes the form $\vartheta_{ML} \oplus C_{ML}$, where $C_{ML} \subseteq \mathfrak{R}^n$ can be constructed from the data and does not depend on ϑ. A simple example is the estimation of the mean μ of a normal distribution with standard deviation σ. In this case, the $(1-\alpha)$-confidence region C_{ML} corresponds to an interval $[-t_\alpha \cdot \sigma/\sqrt{n}, t_\alpha \cdot \sigma/\sqrt{n}]$,[26] i.e., C_{ML} depends only on the number of observations. Now, let $(\widehat{\varphi}_{h,\mathcal{M}}, \beta)$ be a CBI prediction of ϑ_{ML}. Since

$$(\vartheta_{ML} \in \widehat{\varphi}_{h,\mathcal{M}}) \wedge (\vartheta \in \vartheta_{ML} \oplus C_{ML}) \Rightarrow (\vartheta \in \widehat{\varphi}_{h,\mathcal{M}} \oplus C_{ML}),$$

we obtain

$$\mathbb{P}(\vartheta \in \widehat{\varphi}_{h,\mathcal{M}} \oplus C_{ML}) \geq (1-\alpha)(1-\beta).$$

That is, the set $\widehat{\varphi}_{h,\mathcal{M}} \oplus C_{ML}$ defines a $(1-\alpha)(1-\beta)$-confidence region for ϑ. This way, a confidence region for the true parameter ϑ can be derived by means of purely *case-based* reasoning, i.e., without any reference to a likelihood function and corresponding maximization problems.

3.5.2 Case-based prior elicitation

The determination of prior probability distributions is a main burden of Bayesian analysis, and it has become a focus of criticism of the Bayesian approach. As a second application let us therefore consider the possibility of exploiting (credible) CBI in order to support the elicitation of such priors, i.e., the determination of prior distributions from previous cases. The idea is thus to treat a CBI prediction $(\widehat{\varphi}_{h,\mathcal{M}}, \alpha)$ of an MLE ϑ_{ML} as prior information about the unknown parameter ϑ.

In general, there will exist several possibilities of utilizing a CBI prediction. A relatively straightforward choice of a prior based on a prediction $(\widehat{\varphi}_{h,\mathcal{M}}, \alpha)$ is defined by the associated probability density function

$$f : \vartheta \mapsto \left\{ \begin{array}{ll} (1-\alpha)(\int_{\widehat{\varphi}_{h,\mathcal{M}}} dt)^{-1} & \text{if} \quad \vartheta \in \widehat{\varphi}_{h,\mathcal{M}} \\ \alpha(\int_{\Theta \setminus \widehat{\varphi}_{h,\mathcal{M}}} dt)^{-1} & \text{if} \quad \vartheta \not\in \widehat{\varphi}_{h,\mathcal{M}} \end{array} \right. ,$$

where we assume $(\int_\Theta dt) < \infty$.[27] For very small α one might even completely concentrate on the predicted region and define a corresponding uniform prior only over $\widehat{\varphi}_{h,\mathcal{M}}$:

[26] The value t_α is defined through the equality $\int_{-t_\alpha}^{t_\alpha} \phi(t)\, dt = 1-\alpha$, where ϕ denotes the probability density function of the standard normal distribution.

[27] Otherwise it might still be possible to work with improper priors.

$$f : \vartheta \mapsto \begin{cases} (\int_{\widehat{\varphi}_{h,\mathcal{M}}} dt)^{-1} & \text{if } \vartheta \in \widehat{\varphi}_{h,\mathcal{M}} \\ 0 & \text{if } \vartheta \not\in \widehat{\varphi}_{h,\mathcal{M}} \end{cases} .$$

The prior distribution is often assumed to belong to a certain parameterized class $\mathcal{C} = \{f_\gamma \,|\, \gamma \in \Gamma\}$ of distributions, where $\mathcal{C}$ is chosen in such a way that the prior is *conjugate* to the likelihood function. This guarantees that the posterior distribution belongs to the same class. A CBI prediction can then be utilized for constraining (or even determining) the parameters of a prior distribution, the so-called *hyper-parameters*. More precisely, a prediction $(\widehat{\varphi}_{h,\mathcal{M}}, \alpha)$ serves as a constraint in the sense that the parameter γ has to satisfy $\int_{\widehat{\varphi}_{h,\mathcal{M}}} f_\gamma(t)\, dt = \alpha$. For example, if the prior is normal with mean μ and standard deviation σ, the CBI prediction $([\beta^-, \beta^+], \alpha)$ entails $\int_{\beta^-}^{\beta^+} \phi_{\mu,\sigma}(t)\, dt = \alpha$, which in turn suggests

$$\mu = \frac{\beta^- + \beta^+}{2}, \qquad \sigma = \frac{\beta^+ - \beta^-}{2t_\alpha}.$$

3.6 Summary and remarks

Summary

- We have adopted a *constraint-based* view of the CBI hypothesis, according to which the similarity of inputs imposes a constraint on the similarity of associated outcomes in the form of a lower bound. This interpretation allows for exploiting the reasoning principle underlying CBI within a formal inference process.
- The concept of a *similarity profile* has been introduced. It establishes a connection between the system level and the similarity level and represents the similarity structure of a CBI setup. Several generalizations of this concept have been proposed in order to take special characteristics of CBI into consideration and to improve case-based inference.
- A *similarity hypothesis* is thought of as an approximation of a similarity profile. It thus defines a formal model of the CBI hypothesis for the system under consideration.
- CBI has been realized as a process of constraint propagation which allows for predicting an unknown output $r_0 \in \mathcal{R}$ by means of a set $\widehat{\varphi}_{h,\mathcal{M}}(s_0) \subseteq \mathcal{R}$ of possible outcomes. This set is derived from an underlying hypothesis h and a memory $\mathcal{M}$ of cases. It is guaranteed to cover r_0 if h is admissible. An efficient implementation of this inference scheme can be realized by means of parallel computation techniques.
- We have studied some properties of *case-based approximations*, i.e., set-valued mappings $\widehat{\varphi}_{h,\mathcal{M}} : \mathcal{S} \longrightarrow 2^{\mathcal{R}}$ derived from a hypothesis h and a memory of cases $\mathcal{M}$.

- The idea of *case-based learning* can be realized in different ways within our framework. Here, we have concentrated on the learning of a suitable similarity hypothesis from a sequence of observations.
- Utilizing the hypothesis space $\mathcal{H}_{step}$, which consists of a class of step functions on $[0, 1]$, allows for realizing CBL by means of an efficient candidate-elimination algorithm, CBLA. Particularly, the time complexity of updating a hypothesis $h_{\mathcal{M}}$ is linear in the size of the memory $\mathcal{M}$.
- A sequence of hypotheses derived by CBLA from a random sequence of cases converges stochastically toward the optimal admissible hypothesis $h^* \in \mathcal{H}$. Even though these hypotheses may be inadmissible, they allow for deriving predictions which define outer bounds with high probability. We thus obtain a method of *credible case-based inference* that produces predictions in the form of *credible output sets* which cover the true output with high probability. In fact, our CBI method can be seen as a non-parametric approach to estimating confidence regions.
- In Section 3.5, it has been argued that credible CBI is also interesting in the context of classical statistical inference. More specifically, we have outlined the ideas of case-based parameter estimation and case-based prior elicitation in Bayesian analysis.

Remarks

- Within our framework, the concept of *similarity* should be seen as an essential but at the same time *auxiliary* concept. Indeed, the inference procedure outlined in this chapter principally works with *any* pair of similarity functions $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$, each of which defines a certain similarity structure. Of course, the more suitably these functions are chosen, the more precise the inference results will be. However, since our inference scheme takes into account the degree to which the CBI hypothesis applies these results remain valid even if similarity is not quantified in a meaningful way. The interpretation as an auxiliary concept contrasts with other formalizations of CBI [99, 141, 296], in which inference becomes more or less meaningless without a reasonable measure of similarity.
- It has already been remarked that the CBI scheme in Section 3.2.1 is based on the transformation of original data, i.e., instances in the space $\mathcal{S} \times \mathcal{R}$, into points of the similarity space $D_{\mathcal{S}} \times D_{\mathcal{R}}$. In this connection, it is interesting to note that the transformation of data from a high-dimensional into a low-dimensional space is also used by several other methods, e.g., in statistical data analysis or self-organizing neural networks. Of course, the underlying objective which is common to these methods is to capture essential properties of a system structure by means of a simplified representation.
- In [221], an instance-based prediction method has been advocated as an alternative to linear regression techniques. By deriving set-valued instead of point

estimations, credible CBI somehow combines advantages from both methods: It requires less structural assumptions than (parametric) statistical methods as does the instance-based approach. Still, it allows for quantifying the uncertainty related to predictions by means of confidence regions. We shall return to this point in the following chapter.

- We have argued that our approach to CBI combines model-based and instance-based learning (cf. Section 3.1). Let us mention, therefore, another idea of establishing a relationship between model-based and instance-based reasoning. According to the point of view adopted in [236], an instance-based prediction is obtained within a Bayesian framework by marginalizing over all possible model families and all (parameterized) individual models within those families. The basic idea can be expressed by writing (in a somewhat sloppy notation)

$$\mathbb{P}(x \,|\, X) = \int_{\mathcal{M}} \mathbb{P}(x \,|\, M)\, \mathbb{P}(M \,|\, X)\, dM, \tag{3.34}$$

  where X and x denote, respectively, the observed data and a new vector x, and $\mathcal{M}$ is a class of models. Equation (3.34) suggests that the prediction does not depend on a model, only on the data X. However, apart from some technical difficulties, this approach is not very convincing. In fact, (3.34) is nothing else than the standard approach to higher-level Bayesian analysis (Bayesian averaging): A prediction is derived by taking the average of the predictions made by each possible model, weighted by the plausibilities of these models. Thus, (3.34) corresponds to a weighted average of models of a certain class (sometimes called the ensemble average).[28] It is by no means "model-free" since the bias of the model class is actually not "integrated out" by (3.34). Besides, it deserves mentioning that our approach to combining model-based and instance-based inference is very different. This becomes especially obvious by realizing that we do not consider a model of any underlying data-generating process, but rather of the CBI principle itself.

- The construction of confidence regions[29] in the context of CBLP is in line with classical (Neyman-Pearson) statistical inference. Particularly, the inference procedure does not condition on the (structure of the) actually observed data (as likelihood methods do). Rather, the claim that the n-th outcome is covered with probability $1-\alpha_n$ by the confidence region C_n derived from the first $n-1$ cases should be interpreted in a "frequentistic" way: Let an experiment consist of drawing a random sample of n cases, constructing a confidence region from the first $n-1$ cases, and noting a success if the outcome of the n-th case is covered by that region. By repeating this type of experiment over and over again, the relative frequency of successes will converge toward $1-\alpha_n$. In other words, the probability α_n is a property which has to be ascribed to the *inference procedure*, not to the result.

[28] Taking all model families (whatever this means) into account is impossible anyway. In practice, one only considers one class, e.g., a certain type of neural networks.

[29] Note that these regions are random variables.

- In Section 3.1, we have hinted at limitations of a similarity-based analysis which can occur due to the low dimensionality of the similarity space. In order to overcome such limits one might think of using a more general, multi-dimensional formalization of the concept of similarity. Such representations have indeed been advocated in literature (e.g. [283]).
- In [247], the authors consider the problem to quantify the extent to which the CBR hypothesis holds for a particular application at hand. To this end, they propose a measure of the *problem–solution regularity*. In contrast to our concept of a similarity profile, however, this is a one-dimensional measure. Besides, it is not used for the purpose of prediction but rather as a kind of trigger for the maintenance of the CBR system.