Does the Cost Function Matter in Bayes Decision Rule?

Ralf Schlüter, Member, IEEE, Markus Nussbaum-Thom, Student Member, IEEE, and Hermann Ney, Fellow, IEEE

Abstract—In many tasks in pattern recognition, such as automatic speech recognition (ASR), optical character recognition (OCR), part-of-speech (POS) tagging, and other string recognition tasks, we are faced with a well-known inconsistency: The Bayes decision rule is usually used to minimize string (symbol sequence) error, whereas, in practice, we want to minimize symbol (word, character, etc.) error. When comparing different recognition systems, we do indeed use symbol error rate as an evaluation measure. The topic of this work is to analyze the relation between string (i.e., 0-1) and symbol error (i.e., metric, integer valued) cost functions in the Bayes decision rule, for which fundamental analytic results are derived. Simple conditions are derived for which the Bayes decision rule with integer-valued metric cost function and with 0-1 cost gives the same decisions or leads to classes with limited cost. The corresponding conditions can be tested with complexity linear in the number of classes. The results obtained do not make any assumption w.r.t. the structure of the underlying distributions or the classification problem. Nevertheless, the general analytic results are analyzed via simulations of string recognition problems with Levenshtein (edit) distance cost function. The results support earlier findings that considerable improvements are to be expected when initial error rates are high.

Index Terms—Statistical pattern recognition, classifier design and evaluation, Bayes decision rule, cost/loss function.

1 INTRODUCTION

In its general formulation, the Bayes decision rule requires the definition of a cost function suitable for the intended pattern classification task. In many pattern classification tasks such as automatic speech recognition (ASR), optical character recognition (OCR), part-of-speech (POS) tagging, etc., system performance is measured on the symbol level, i.e., by counting word, character, tagging, etc., errors. Nevertheless, the standard Bayes decision rule used for such tasks optimizes string (symbol sequence) error. This well-known inconsistency [7, pp. 4-5] is supported by empirical evidence: Experiments in areas such as ASR, machine translation (MT), or POS tagging [1], [8], [10], [11], [14] indicate that improvements to be obtained from using task-related cost functions for the Bayes decision rule seem limited, at least at low levels of error rate. In this work, partial evidence for this observation will be given, showing a clear analytic relation between the Bayes decision rule with general metric, integer-valued cost functions and the standard approach using 0-1 cost. After a general discussion of the Bayes decision rule in Section 2, in Section 3, we present several conditions for which the Bayes decision rule with general metric cost functions and with 0-1 cost function lead to the same decisions. Also, conditions to efficiently reduce the search space for the Bayes decision rule with integer-valued metric cost functions to the vicinity of a reference class are derived without lack of optimality. The analytic results are valid for arbitrary class definitions, especially including string classes, as they are used in ASR, MT, or POS-tagging. In Section 4, the analytic results are analyzed by way of simulations for the case of fixed length string classes with the Levenshtein (edit) cost function.

2 BAYES DECISION RULE

Consider an observation vector \( x \), class indices \( c, c' \in C \), and a cost function \( L(c', c) \) which measures the cost for misclassifying a class \( c' \) as \( c \). Also define the posterior probability \( p(c'|x) \). Then, the posterior risk \( R(c|x) \) with respect to cost function \( L \) for a class \( c \) is defined as

\[
R(c|x) = \sum_{c'} p(c'|x) \cdot L(c', c),
\]

which defines the average cost for classifying observation \( x \) as class \( c \). The minimum of the posterior risk is called Bayes risk:

\[
R(c_{\text{Bayes}}(x)|x) = \min_c R(c|x).
\]

Its minimization leads to the Bayes decision rule w.r.t. cost function \( L \):

\[
c_{\text{Bayes}}(x) = \arg\min_c R(c|x).
\]

Provided the cost function \( L \) represents the error measure for the intended classification task, the Bayes risk represents the expected error for an observation \( x \). The optimal decision depends on both the posterior distribution as well as the cost (or loss) function for the problem. In most statistical pattern recognition tasks, the most basic 0-1 cost function \( L_{0-1} \) is used for the Bayes decision
rule, giving zero cost for equal classes and unit cost for different classes

\[ L_{0-1}(c', c) = 1 - \delta_{c, c'} = \begin{cases} 0 & \text{for } c' = c, \\ 1 & \text{for } c' \neq c. \end{cases} \]

The Bayes risk in this case reduces to a simple function of the posterior probability:

\[ R_{0-1}(c|x) = \sum_{c' \in \mathcal{C}} p(c'|x) \cdot L_{0-1}(c', c) = 1 - p(c|x). \]  

(2)

For the case of a 0-1 cost function, the Bayes decision rule reduces to finding the maximum posterior class

\[ c_{\text{MAP}}(x) := c_{\text{Bayes}}(x) = \arg \min_c R_{0-1}(c|x) = \arg \max_c p(c|x). \]

The corresponding decision rule will be called the MAP rule in the following, and the maximizing class will be called the MAP class.

For example, even in automatic speech recognition and machine translation with their combinatorial complex string classes, the 0-1 cost-based Bayes decision rule is still the standard case. In speech recognition and machine translation, the classes \( c \) become word sequences \( W = w_1, \ldots, w_N \). Therefore, in these cases, the MAP rule corresponds to a minimization of the expected sentence (or, more precisely, segment [4]) error rate. This is partly justified by the fact that, especially for short word sequences, there is a correlation between sentence and word errors. Nevertheless, for longer sentences, the correlation necessarily decreases. Therefore, theoretically, the choice of a 0-1 (or sentence error) cost function is expected to be suboptimal for such tasks. In automatic speech recognition, the evaluation measure is the (empirical) word error rate (WER). Therefore, the number of word errors, i.e., the Levenshtein or edit distance, would be expected to be the ideal cost function for an automatic speech recognition decision rule. Due to the complexity of the Levenshtein alignment needed to compute the number of word errors, it was prohibitive to use the number of word errors as a cost function for the Bayes decision rule for a long time. It still remains unclear if the posterior risk using Levenshtein cost analytically can be simplified as easily as for 0-1 cost, cf. (2), where the expectation sum can be performed analytically. Nevertheless, a number of approximation approaches exist which were shown to be efficient to evaluate [2], [3], [9], [15], [16]. Most importantly, it can be observed experimentally that for a large percentage of test samples, the decisions using the simple 0-1 cost function are the same as those using more sophisticated decision rules based on Levenshtein, Hamming [5], or similar cost functions, as used in ASR, MT, or POS tagging [1], [6], [8], [10], [11].

### 3 Metric Cost Functions

A fundamental relation between the Bayes decision rule using 0-1 cost function, i.e., the MAP rule, and using any integer-valued metric cost function will be derived in the following. A metric loss function is positive, symmetric, fulfills the triangle inequality, and it is zero if and only if both arguments are equal. Metric loss functions include, but are not limited to, word/character/tagging error counts as used in ASR, OCR, POS tagging, or position independent word error counts as partly used in machine translation, to name but a few. Nevertheless, it should be noted that the following derivation is not limited to string classes. It is valid for any definition of classes together with any integer valued, metric cost function.

#### 3.1 The Low-Cost Principle

The title of this section relates to a theorem which will show that candidates for the Bayes class can be reduced to subsets of classes with low cost w.r.t. some “seed” class, provided their cumulative probability is large enough.

Assume classes \( c, c' \in \mathcal{C} \), and a metric integer-valued cost function \( \mathcal{L} : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{N}^+ \). Also assume an observation \( x \), and a posterior distribution \( p(c|x) \). Further, assume a partitioning of all classes \( c \) into subspaces \( \mathcal{C}_i \) depending on their distance \( i = \mathcal{L}(c, c_i) \) to a “seed” class \( c_i \):

\[ \mathcal{C}_i = \{ c \in \mathcal{C} | \mathcal{L}(c, c_i) = i \}. \]  

(3)

Furthermore, define \( p(\mathcal{C}_i|x) \) to be the overall probability mass covered by subspace \( \mathcal{C}_i \):

\[ p(\mathcal{C}_i|x) = \sum_{c \in \mathcal{C}_i} p(c|x). \]  

(4)

Also define the maximum probability \( q(c_i|x) \) of all classes contained in subspace \( \mathcal{C}_i \):

\[ q(c_i|x) = \max_{c \in \mathcal{C}_i} p(c|x). \]  

(5)

**Theorem 1.** Assume there exists an integer \( j > 0 \) which fulfills

\[ 2 \sum_{i=0}^{j-1} (j - i) \cdot p(\mathcal{C}_i|x) + q(c_j|x) - q(c_j|x) - j \geq 0. \]  

(6)

Then, all classes \( c \) with cost \( \mathcal{L}(c, c_j) \geq j \) have a larger posterior risk than the seed class \( c_j \):

\[ R(c_i|x) \leq R(c_j|x) \quad \forall c \in \mathcal{C} \text{ with } \mathcal{L}(c, c_j) \geq j. \]

In other words, if (6) is fulfilled, only those classes with distance lower than \( j \) are possible candidates for the Bayes class:

\[ \mathcal{L}(c_{\text{Bayes}}(x), c_i) < j. \]  

(7)

**Proof.** Using the partitioning of the set of classes defined in (3) and using the corresponding posterior distribution from (4), the posterior risk for the seed class can be rewritten as

\[ R(c_j|x) = \sum_{i \geq j} i \cdot p(\mathcal{C}_i|x). \]  

(8)

For a metric cost function \( \mathcal{L} \) the application of the triangle inequality results in

\[ \mathcal{L}(c_i, c_j) \begin{cases} \geq |j - i| & \forall i \neq j \land c_i \in \mathcal{C}_i \land c_j \in \mathcal{C}_j, \\ \geq 1 & \forall i = j \land c_i, c_j \in \mathcal{C}_i \land c_i \neq c_j, \\ = 0 & \forall i = j \land c_i, c_j \in \mathcal{C}_i \land c_i = c_j. \end{cases} \]  

(9)

1. Provided the position independent error measure is defined on a set of words instead of a word sequence to ensure uniqueness.
Now, the posterior risk for a class \( c_j \in C_j \), i.e., for a class with cost \( j \) compared to the seed class \( c_s \), gives the following inequality:

\[
\begin{align*}
R(c_j|x) &= \sum_{c} \gamma(c|x) \cdot \mathcal{L}(c, c_j) \\
&= \sum_{i=0}^{j} \sum_{c \in C_i} \gamma(c|x) \cdot \mathcal{L}(c, c_j) \\
&= \sum_{i=0}^{j} \sum_{c \in C_i} \gamma(c|x) \cdot \mathcal{L}(c, c_j) \\
&\geq \sum_{i=0}^{j} |j-i| \cdot p(C_i|x) + p(C_j|x) - q(C_j|x) \quad \forall \ c_j \in C_j.
\end{align*}
\]

Using (8) and (10), a lower bound for the difference between the posterior risk for a class \( c_j \in C_j \) with cost \( j \) compared to the seed class, and the posterior risk for the seed class itself can be established:

\[
\begin{align*}
R(c_j|x) - R(c_s|x) \\
&\geq \sum_{i=0}^{j} |j-i| \cdot p(C_i|x) + p(C_j|x) - q(C_j|x) - \sum_{i=0}^{j} i \cdot p(C_i|x) \\
&= \sum_{i=0}^{j-1} (j-i) \cdot p(C_i|x) + \sum_{i=0}^{j-1} (i-j) \cdot p(C_i|x) \\
&\quad - \sum_{i=0}^{j-1} i \cdot p(C_i|x) + p(C_j|x) - q(C_j|x) \\
&= \sum_{i=0}^{j-1} (j-2i) \cdot p(C_i|x) - j \cdot \sum_{i=0}^{j-1} p(C_i|x) - j \cdot p(C_j|x) \\
&\quad + p(C_j|x) - q(C_j|x) \\
&= 2 \sum_{i=0}^{j-1} (j-i) \cdot p(C_i|x) + p(C_j|x) - q(C_j|x) - j.
\end{align*}
\]

Therefore, if there exists an integer \( j > 0 \) fulfilling (6), then (6) by induction is also fulfilled for any integer \( j' \geq j \). Hence, substituting (6) into (11) and (12) finally proves Theorem 1.

Now, define the minimal \( j = j_{\min}(x) \) that fulfills (6):

\[
j_{\min}(x) = \min \left\{ j \left| \sum_{i=0}^{j-1} (j-i) \cdot p(C_i|x) + p(C_j|x) - q(C_j|x) - j \geq 0 \right. \right\}.
\]

Then, (11) shows that the posterior risk of all hypotheses \( c \) with cost \( \mathcal{L}(c, c_s) \geq j_{\min}(x) \) is larger than or equal to the posterior risk of the seed class \( c_s \). Therefore, these classes can be discarded from the search for the class \( c_{\text{Bayes}}(x) \) minimizing the posterior risk.

In the following, two corollaries of Theorem 1 are presented. They show that parts of the analytic results in [12] and [14] present special cases of Theorem 1.

### 3.2 Cost One Dominance

**Corollary 1.** Assume (6) is fulfilled for \( j = 1 \), i.e., assume

\[
2p(C_0|x) + p(C_1|x) - q(C_1|x) - 1 \geq 0.
\]

Then, the seed class minimizes the posterior risk, and the seed class is equal to the posterior maximizing class, i.e., \( c_s = c_{\text{MAP}}(x) = c_{\text{Bayes}}(x) \).

**Proof.** Using Theorem 1, and using that cost function \( \mathcal{L} \) is both metric and integer valued, (13) leads to the following specific form of (7):

\[
\mathcal{L}(c_{\text{Bayes}}(x), c_s) < 1
\]

\[
\iff \mathcal{L}(c_{\text{Bayes}}(x), c_s) = 0
\]

\[
\iff c_{\text{Bayes}}(x) = c_s.
\]
Also, we have
\[
p(c_*|x) = p(C_0|x) \\
\geq 1 - p(C_0|x) - p(C_1|x) + q(C_1|x) \quad \text{(cf. Ineq. (13))}
\]
\[
= 1 - \sum_{i=0} p(C_i|x) + q(C_1|x)
\]
\[
= 1 - \sum_{i=0} p(C_i|x) + \max_{i \geq 1} p(C_i|x) + q(C_1|x)
\]
\[
= \max_{c \neq c_*} p(c|x) \quad \text{(cf. (5))},
\]
and, therefore, \( c_{MAP}(x) = c_* = c_{Bayes}(x) \).

**3.3 MAP Dominance**

**Corollary 2.** Assume
\[
p(c_{MAP}(x)|x) \geq \frac{1}{2}.
\]

Then, the MAP class minimizes the posterior risk.

**Proof.** With the choice \( c_* = c_{MAP}(x) \), from (4) we obtain
\[
p(C_0|x) = p(c_{MAP}(x)|x).
\]

Thus, substituting (14) into the left side of (13) leads to
\[
2p(C_0|x) + p(C_1|x) - q(C_1|x) - 1 \geq p(C_1|x) - q(C_1|x) \geq 0
\]
since, by definition, we have \( p(C_1|x) \geq q(C_1|x) \). With Corollary 1, it then follows that \( c_{Bayes}(x) = c_{MAP}(x) \). □

It should be noted that Corollary 2 can also be proven without the assumption of an integer-valued cost function. In [12], a proof for the MAP dominance is given which only assumes a metric cost function.

**3.4 Low Risk Dominance**

In case of a risk lower than or equal to 1/2, a special case of the MAP dominance presented in Corollary 2 can be observed.

**Corollary 3.** Assume a posterior risk based on a metric integer-valued cost function \( \mathcal{L} \), and assume a class \( c \) with risk less or equal to 1/2:
\[
\mathcal{R}(c|x) \leq \frac{1}{2}.
\]

Then, \( c = c_{MAP}(x) = c_{Bayes}(x) \) follows and \( p(c_{MAP}(x)|x) \geq \frac{1}{2} \).

**Proof.** Substitute the metric and integer-valued cost function in the definition of the posterior risk with a lower estimate:
\[
\frac{1}{2} \geq \mathcal{R}(c|x)
\]
\[
= \sum_{e'} p(e'|x) \mathcal{L}(e', c)
\]
\[
\geq 1 - p(c|x)
\]
\[
\Rightarrow p(c|x) \geq \frac{1}{2}.
\]

Due to normalization and Corollary 2, \( c = c_{MAP}(x) = c_{Bayes}(x) \) follows. Without loss of correctness of the proof, in case of \( p(c_{MAP}(x)|x) = 1/2 \) the maximizing class \( c_{MAP}(x) \) might not be unique. □

**3.5 Discussion**

The low-cost principle derived in Section 3.1 shows a close relation between the standard MAP rule and the Bayes decision rule with any metric and integer-valued cost function, in the following called the general Bayes rule. For sufficient probabilistic evidence of a specific (“seed”) class and those classes with limited cost compared to it, the decision rules are shown to either coincide or the remaining candidates that might optimize the general Bayes rule are limited to being within a certain distance of the above specific class. This result leads to a number of relatively simple criteria (cost one/low risk/MAP dominance) which can be checked efficiently and indicate coincidence of the standard MAP and the general Bayes decision rule. Empirically, it can be observed that the decision of the standard MAP and the general Bayes rule often coincide. This behavior can now at least partly be attributed to the analytic results presented in this section, and will be further analyzed by way of simulations in the following section.

**4 Simulations**

In this section, the relevance of the analytic relation between the MAP rule, i.e., the Bayes decision rule with 0-1 cost, and the Bayes decision rule with an integer-valued metric cost function will be analyzed by way of simulations for the case of string classes and the widely used word error measure based on edit distance/Levenshtein alignment cost. String classes are relevant for a number of pattern recognition tasks, including automatic speech recognition, handwriting recognition, optical character recognition, and machine translation. For these tasks, the standard decision rule is based on the Bayes decision rule with 0-1 cost and therefore minimizes the string error rate. Nevertheless, the evaluation criterion for string classes often is word error rate. Hence, to minimize word error rate, the Bayes decision rule with the Levenshtein cost function needs to be used, which in the following will be called the minimum word error rule or MWE rule for short. For the simulations, the word error rate of a string is measured as the posterior risk using the Levenshtein cost function. For the MWE rule, the posterior risk certainly reduces to the Bayes risk or Bayes error with the Levenshtein cost function.

The aim of this section is to investigate the potential of the MWE rule without approximation and to analyze the effect of the analytic results presented in Section 3 for the case of string classes. In addition, the effect of the structure of the string class distribution is investigated. Due to the complexity of the MWE rule, its application to medium and large vocabulary string recognition tasks with exhaustive search and without approximations is prohibitive. Hence, we did simulations with limited vocabulary and string length which are introduced and analyzed in the following. It should be noted, for example, that in ASR, in practice, vocabulary sizes of more than 100,000 and string lengths of 100 or more words do occur. On the other hand, speech
utterances can usually be segmented to give lower string lengths. Also, effectively, the number of words with significant probabilities within a specific position usually is limited, as can be observed in ASR search. In our simulations, a number of small but increasing values of vocabulary size and string length are chosen to give at least a qualitative idea of the dependence of the Bayes decision rule and the criteria presented in Section 3 on these parameters.

4.1 Modeling

For the specific case of string classes \( w_N = w_1, w_2, \ldots, w_N \) of fixed length \( N \) with words \( w_n \in \mathcal{V} \), for each “utterance,” represented by a set of observations \( x_N^v \), a distribution over all word strings has to be simulated.

We consider simulations of two types of distributions:

- **Global context dependence**, i.e., full dependence of all string positions. In this case, the distribution \( p(w_N^N|x_N^N) \) is simulated directly by a distribution \( p(w_N^N) \), ignoring the specific class structure of word strings. The observation sequence therefore does not have to be simulated explicitly.

- **Local context dependence using emission probabilities and a bigram prior**, i.e.,

\[
p(w_N^N|x_N^N) \propto \prod_{n=1}^{N} p_n(w_n) \cdot p(w_n|x_{n-1}),
\]

with

\[
p_n(w) := \frac{p(x_n|w)}{\sum_v p(x_n|v)}.
\]

This case serves as a simple, nontrivial example for local context dependence as observed and utilized in string recognition tasks. For each simulation of an utterance represented by an observation sequence \( x_N^N \), distributions \( p(w_N^N) \) (global context dependence) or \( p_n(w) \) for all \( n = 1, \ldots, N \) and \( p(w|x) \) (local context dependence) are generated, i.e., the observation sequence \( x_N^N \) also does not have to become explicit in these simulations.

The simulation approach chosen here is summarized in the following. For more information on this approach, the reader is referred to an internal report describing the simulation and corresponding algorithms in detail in [13].

In case of local context dependence, a specific bigram prior was simulated to match a given perplexity; a corresponding algorithm is described in [13].

The probability distribution of word sequences observed during testing and the string prior distribution should ideally coincide. Also, the emission probabilities usually are significant for those words that are observed in the corresponding positions in the test data (cf. task: word sequences spoken, written, etc.). To take this into account, for each simulation, the word sequence \( \tilde{w}_N^N \) whose words \( \tilde{w}_n \) got assigned the maximum string emission probability \( p_n(\tilde{w}_n) = p_{n,\text{max}} \) in each respective word position \( n = 1, \ldots, N \) were drawn from the simulated bigram prior distribution conditioned on the predecessor word, as shown in [13]. The maximum probability \( p_{n,\text{max}} \) of the local emission probability distribution then was drawn from identical Beta distributions for each position \( n = 1, \ldots, N \). The parameters of the Beta distributions control both the average and the variance of the maximum emission probability \( p_{n,\text{max}} \) in each word position. The average will be connected to the average error rate, whereas the variance models the variability of the emission probability maxima encountered (due to, e.g., acoustic conditions in automatic speech recognition, or legibility in handwriting recognition).

Subsequently, for each word position \( n = 1, \ldots, N \), the distribution \( p_n(w) \) over the remaining vocabulary \( w \in \mathcal{V}_n = \mathcal{V} \setminus \{\tilde{w}_n\} \) is generated iteratively from a uniform distribution, conserving both the already drawn maximum \( p_{n,\text{max}} = p_n(\tilde{w}_n) = \max_w p_n(w) \) and the normalization over all words, for further details, cf. [13].

In case of global context dependence, for each distribution simulated, an arbitrary string is chosen to which the maximum string probability is assigned. The maximum string probability is drawn from a product of \( N \) identical Beta distributions to enable direct comparison to the case of local emission probabilities. The probabilities of all remaining string classes were drawn iteratively from a uniform distribution, conserving the already drawn maximum and normalization over all strings, for details cf. [13].

For each vocabulary size \( V = |\mathcal{V}| \), string length \( N \), and choice of Beta distribution parameters for the maxima, 10,000 simulations were performed. In case of a bigram prior, 100 different bigram distributions with the same perplexity were simulated and, for each bigram simulation, 100 different sets of local emission distributions for each position in the string were computed.

The simulations were done for vocabulary sizes \( V \in \{2, 3, \ldots, 25\} \) and string lengths \( N = 2, 3, \ldots, 10 \), with an additional limit to the number of (string) classes \( V^N \leq 65,536 \). Bigram priors were generated with perplexities \( PP = 1.1 \), and \( PP = 1 + \gamma \cdot (V - 1) \) with \( \gamma \in \{0.2, 0.4, 0.6, 0.8, 1.0\} \). Unless stated otherwise, the Beta distribution used for the simulation of the maxima of the local emission distributions was parameterized to obtain mean 0.8 and variance 0.2. The rationale for simulating the maxima of the local emission distributions around a fixed mean is the assumption of having a fixed average evidence of the maximum word in each position, as could be expected in corresponding string recognition tasks.

4.2 Evaluation

In this section, string recognition simulations are presented to analyze both the effect of the low-cost principle, as derived in Section 3, as well as the general effect of using word error cost (MWE rule) instead of standard 0-1 cost (MAP rule) in recognition. In Section 4.2.1, first of all, the amount of coincidence between the MWE rule and the MAP rule as covered by the low-cost principle and its corollaries is analyzed for the different cases of context dependency. Furthermore, the effect of the low-cost principle and its corollaries on the complexity of the MWE rule is investigated in Section 4.2.2. Finally, the simulations are used to further analyze the potential of the MWE rule and its effect on word error rate compared to the MAP rule in Section 4.2.3.
with vocabulary size conditions presented in Section 3, we consider simulations accumulated for cost one dominance. Abilities and, therefore, a decrease of the probability mass of per-string probability with string length. This leads to a general decrease of the individual string's emission probability distributions influences the above analysis of the MAP and MWE rules cannot be explained by the criteria derived in Section 3.

In addition, how the structure of the underlying probability distributions influences the above analysis of the MWE and MAP rule behavior is analyzed. In Fig. 1, the relative frequency of cases in which the cost one dominance, as introduced in Section 3.2, applies is shown, for which the MAP and MWE rules give identical decisions, i.e., where the MAP (string) class, or MAP string short, minimizes the posterior risk with the Levenshtein cost function. As Fig. 1 shows, the relevance of cost one dominance increases with decreasing bigram perplexity. It is lowest for global context dependence.

Also, for perplexities considerably larger than one, the relevance of cost one dominance decreases with string length. The latter can be explained by way of the decrease of per-string probability with string length. This leads to a general decrease of the individual string's emission probabilities and, therefore, a decrease of the probability mass accumulated for cost one dominance.

To further analyze the contribution of the different conditions presented in Section 3, we consider simulations with vocabulary size \( V = 8 \) at string lengths \( N \in \{2, 3, 4, 5\} \) to analyze the effect of each of the conditions presented in Section 3. For the case of local context dependence using a bigram prior, Fig. 2a shows the contribution of simulations covered by the criteria from Section 3.

The results are relative frequencies plotted as cumulative bar graphs. Since low-cost dominance is covered by the MAP dominance, which in turn is covered by cost one dominance, their bars are plotted covering each other. In addition, the plots include those cases which are not covered by the criteria, separately for the MAP and MWE decision rules, giving equal and unequal results. In Fig. 2b, similar results are shown for the case of global context dependent simulations.

The decrease in coverage by cost one dominance with increasing string length \( N \) can clearly be seen. The decrease is more pronounced for the case of global context dependence in Fig. 2b than it is for local context dependence using a bigram prior in Fig. 2a. It is interesting to notice that in the case of the bigram prior in Fig. 2a, the ratio of simulations for which the decision rules result in the same decision does not even vary strongly with string length \( N \), although the coverage by cost one dominance decreases with \( N \). It might be expected that under consideration of the distributions' structure, further properties of the Bayes decision rule might be revealed.

4.2.2 Complexity Gains by Low-Cost Principle

In the case of the Levenshtein cost function, an analytic simplification of the MWE rule is not known without further approximations, cf., e.g., [2], [3], [9], [15], [16]. In this section, the MWE rule without approximations is analyzed, whose exhaustive complexity is squared in the number of classes, \( (V^N)^2 \), cf. the definition of the posterior risk in (1),
which is to be evaluated with Levenshtein cost to apply the MWE rule. Here, the effect of each of the criteria derived in Section 3 on the complexity of the MWE rule is analyzed by way of simulations.

The exhaustive complexity of the MWE rule can be reduced considerably. The low-cost principle can be used to constrain the set of candidate (string) classes to those with a certain maximum Levenshtein distance to the MAP (string) class. In the extreme case, the low-cost principle reduces to its special case cost one dominance, where the only remaining candidate string for the MWE rule is the MAP string. The complexity for computing the statistics necessary to evaluate the low-cost principle (and cost one dominance) is proportional to the number of classes, i.e., here the number of strings, \( V^N \). Therefore, in case of cost one dominance, the complexity of the MAP rule is retained. In addition to using the low-cost principle, the search complexity is further reduced by early pruning. Early pruning is applied to those strings not fulfilling cost one dominance. The posterior risk calculation has to be done for each string by carrying out the summation over all strings in (1). In early pruning, this summation can be aborted once the accumulated posterior risk exceeds the posterior risk of the current best string hypothesis since the summands are all positive.

After considering low-cost principle and early pruning, we measured the remaining search complexity defined as the number of summands considered in the posterior risk summation from (1) over all hypothesized classes for which the posterior risk is computed. In our analysis, we then considered the normalized complexity, defined as the search complexity after considering low-cost principle and early pruning, divided by the exhaustive complexity \( V^{2N} \) for applying the MWE rule.

The previously simulations with vocabulary size \( V = 8 \) at string lengths \( N \in \{2,3,4,5\} \), which were already discussed above, now are further analyzed for the effect of each of the conditions from Section 3 on the search complexity. For the case of local context dependence using a bigram prior, Fig. 3a shows the contribution to the normalized search complexity of simulations covered by the criteria from Section 3. Since low-cost dominance is covered by MAP dominance, which in turn is covered by cost one dominance, their bars are plotted covering each other. In addition, the plots include those cases which are not covered by the criteria, separately for the MAP and MWE decision rules, giving equal and unequal results. In Fig. 3b, similar results are shown for the case of global context dependent simulations.

Clearly, the reduction in complexity is much more pronounced for the case of local context dependence using a bigram prior, as shown in Fig. 3a, than it is for global context dependence, as shown in Fig. 3b. Due to the large difference in complexities, the bar graphs for local and global context dependence are scaled differently. Fig. 3a only showing the codomain up to 2 percent and Fig. 3b showing the codomain up to 60 percent. Recall that the (unnormalized) exhaustive complexity of the MWE rule increases exponentially with the string length \( N \). The results show that with local context dependence, the normalized complexity even decreases with the string length. The low-cost principle apparently takes advantage of distributions with local context dependency, which introduces correlation to strings with the low Levenshtein distance. For longer strings this effect is intensified, partly because cost one dominance, provided it is fulfilled, reduces the search complexity of the MWE rule from \( V^{2N} \) to \( V^N \), i.e., to the complexity of the MAP rule, leading to a normalized complexity of \( V^{-N} \). A similar, though lesser, effect can be observed for those simulations for which cost one dominance does not apply, but for which the low-cost principle can be used to constrain the search. In contrast to this, for global context dependence, the normalized complexity even increases with increasing the string length. This can be attributed to the lack of correlation in between different strings, irrespective of their Levenshtein distance. Therefore, in case of global context dependence, the low-cost principle cannot take as much advantage of strings with the low Levenshtein distance from the MAP string.

In Figs. 4 and 5, the effect of the low-cost principle and early pruning on the normalized complexity is further analyzed as a function of the bigram perplexity for strings
of length $N = 2$ and $N = 5$ and varying vocabulary sizes $V$, and results are compared to the case of global context dependence.

For local context dependence using a bigram prior, the low-cost principle and early pruning lead to an approximate linear decrease of the normalized complexity as a function of vocabulary size $V$ when shown in a log-log plot. This means the normalized complexity is exponential as a function of vocabulary. In the extreme of very low bigram perplexity, mostly cost one dominance is observed, reducing the normalized complexity to $V^{-N}$. For higher perplexities, in case of string length $N = 5$, the normalized complexity seems to converge to being proportional to $V^{-A}$ with a lower exponent $A < N$, whereas the prefactor increases with perplexity (parallel lines in the log-log plot in Fig. 5). In case of global context dependence and low string length of $N = 2$, in Fig. 4, less decrease in the complexity of the MWE rule as function of the vocabulary is observed since the amount of context dependence at string length $N = 2$ with bigram is not much lower than for global context dependence simulating full strings. Using the low-cost principle and early pruning, for a larger string length $N = 5$, the normalized complexity in case of global context dependence even stays constant at high level, cf. Fig. 5. Here, the difference to local context dependence using a bigram prior is much more pronounced.

4.2.3 Overall Effect of Word Error Cost

Considering the high complexity of the MWE rule compared to the MAP rule, or the corresponding necessity to apply approximations to the MWE rule to reduce complexity, in this section, we analyze the potential of the MWE rule w.r.t. to the word error rate gains possible, compared to the standard MAP rule.

Fig. 6 shows the effect of low-cost principle and early pruning on the search complexity of the MWE rule as a function of the expected word error rate (or Bayes risk) for the case of local context dependence with a bigram prior with a vocabulary size $V = 5$ and a string length $N = 5$.

The different error rates in Figs. 6 and 7 are obtained by simulating the local emission probabilities, starting with pronounced high maxima in each position (leading to low word error rate), which are gradually decreased down to a uniform distribution (leading to high error rate). The level of the maxima is controlled statistically by drawing the

---

**Fig. 4.** Relative complexity remaining for local context dependence with different bigram perplexities as well as for global context dependence, plotted against vocabulary size, for string length $N = 2$. For each point in the plot, 10,000 simulations were performed.

**Fig. 5.** Relative complexity remaining for local context dependence with different bigram perplexities as well as for global context dependence, plotted against vocabulary size, for string length $N = 5$. For each point in the plot, 10,000 simulations were performed.

**Fig. 6.** Number of simulations for which Bayes decision rules with string error cost (MAP) and with word error cost (MWE) give different decisions, plotted against vocabulary size $V = 5$, string length $N = 5$, and the bigram perplexities ($PP$) given in the legend. For each point in the plot, 10,000 simulations were performed.

**Fig. 7.** Improvement in WER obtained by the Bayes decision rule with word error cost (MWE) plotted against WER using the MAP rule for simulations with vocabulary size $V = 5$, string length $N = 5$, and the bigram perplexities ($PP$) given in the legend. For each point in the plot, 10,000 simulations were performed.
corresponding maximum values from Beta distributions with corresponding mean and variance, see Section 4.1. For each bigram with fixed perplexity, the highest word error rate is obtained with uniform local emission distributions. It should be noted that this maximum in word error rate decreases with bigram perplexity. Therefore, the plots in Figs. 6 and 7 cover the complete range of obtainable word error rates for each bigram perplexity chosen.

Certainly, the choice of bigram perplexity leads to different results. Nevertheless, the general trend of the normalized complexity as function of word error rate in Fig. 6 is comparable in each case. Clearly, the normalized complexity increases with increasing error rate. Higher error rates generally come with distributions of higher entropy, which leads to less probability mass available for strings with the low Levenshtein distance to the MAP string, and therefore less effect of the low-cost principle on the complexity.

In contrast to this, as shown in Fig. 7, the possible relative improvement in word error rate for the MWE rule compared to the word error rate when applying the MAP rule more strongly depends on the bigram perplexity, in addition to the dependence on the initial MAP word error rate.

In general, the relative improvements that can be obtained clearly decrease with decreasing word error rate, due to the fact that at low word error rate the decisions of the MWE and the MAP rules do coincide ever more often. Similar observations were made in speech recognition using approximated versions of the MWE rule, cf. [8, Tables 5 and 7].

The relative improvement from the MWE rule also increases with decreasing bigram perplexity unless per-plexities are relatively low. Nevertheless, in the latter case, i.e., for perplexities $PP \in \{1.4, 1.8\}$, it is interesting to notice that with increasing MAP word error rate the obtainable relative improvement for the MWE rule decreases again after reaching a maximum. As pointed out above, for fixed bigram perplexity, the highest word error rates (the rightmost points for each perplexity in Figs. 6 and 7) are obtained when the local emission distributions approach a uniform distribution. In the extreme case of uniform emission distributions in each word position, the string probabilities from (15) are reduced to a product of bigrams only, with no explicit position dependence remaining. Although the overall word error rate still increases, when the local emission distributions approach a uniform distribution, the probability mass will concentrate more clearly on that string that maximizes the prior probability and, due to the local context dependency introduced by the prior, on those strings with the low Levenshtein distance from the string maximizing the prior probability. It is interesting to notice that the fraction of simulations with equal decisions either becomes very large again (99 percent for the highest attainable word error rate at $PP = 1.4$, of which 96 percent are explained by cost one dominance), or at least is leveled to some degree for higher perplexities, cf. Fig. 7. For perplexity $PP = 1.8$, no intermediate maximum in the number of equal decisions is observed in Fig. 6, but still 70 percent of the decisions from the MAP and MWE rule coincide in case of uniform emission distributions (rightmost point for $PP = 1.8$), although in this case only 3 percent can be explained by cost one dominance. As motivated above, this behavior might be attributed to the pure bigram structure of the string probabilities and might be the starting point for further investigations that take into account the specific structure of the underlying distributions.

Finally, it should be noted that part of the analytic results presented also were analyzed in automatic speech recognition experiments in [12] and [14].

5 Conclusions and Outlook

In this work, fundamental properties of the Bayes decision rule with integer-valued metric cost functions were presented. Analytical results show that the posterior-maximizing class can dominate the Bayes decision rule, which leads to equality of the Bayes decision rule with any integer-valued metric cost function to the maximum-a-posteriori (MAP) decision rule resulting from a 0-1 cost function. These general analytic findings support empirical evidence in areas such as automatic speech recognition, machine translation, or part-of-speech tagging [1], [6], [8], [10], [11], which indicate that the potential of considering task-related cost functions for the Bayes decision rule seems limited, at least at low levels of error rate.

For the analytic results presented, no assumption whatsoever has been made concerning the structure of the underlying probability distributions. Nevertheless, probability distributions for string classification tasks like, e.g., automatic speech recognition or handwriting recognition, show a large amount of local context dependency, and simulations for this case show a higher ratio of identical decisions for both decision rules than predicted by the criteria presented in this work. This might lead to further, more specific properties of the Bayes decision rule with word error-based cost functions and distributions showing local context dependency, which could be the starting point for further investigations.

References


Ralf Schützer received the Dipl degree in computer science in 2008 from RWTH Aachen University. He studied computer science at RWTH Aachen University, Germany. He is currently working on automatic speech recognition as a doctoral student researcher. He is a member of the IEEE.

Markus Nussbaum-Thom received the Diploma degree in computer science in 2008 from RWTH Aachen University. He studied computer science at RWTH Aachen University, Germany. He is currently working on automatic speech recognition as a doctoral student researcher. He is a member of the IEEE.

Hermann Ney is a full professor of computer science at RWTH Aachen University, Germany. Previously, he headed the Speech Recognition Group at Philips Research. His main research interests include the area of statistical methods for pattern recognition and human language technology and their specific applications to speech recognition, machine translation, and image object recognition. In particular, he has worked on dynamic programming for continuous speech recognition, language modeling, and phrase-based approaches to machine translation. He has authored and coauthored more than 500 papers in journals, books, conferences, and workshops. In 2006, he was the recipient of the Technical Achievement Award of the IEEE Signal Processing Society. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society from 1997 to 2000. He is a fellow of the IEEE.