Analytical Results on Style-Constrained Bayesian Classification of Pattern Fields

Sriharsha Veeramachaneni and
George Nagy, *Fellow*, *IEEE*

Abstract—We formalize the notion of *style* context, which accounts for the increased accuracy of the field classifiers reported in this journal recently. We argue that style context forms the basis of all order-independent field classification schemes. We distinguish between intraclass style, which underlies most adaptive classifiers, and interclass style, which is a manifestation of interpattern dependence between the features of the patterns of a field. We show how style-constrained classifiers can be optimized either for field error (useful for short fields like zip codes) or for singlet error (for long fields, like business letters). We derive bounds on the reduction of error rate with field length and show that the error rate of the optimal style-constrained field classifier converges asymptotically to the error rate of a style-aware Bayesian singlet classifier.

Index Terms—Style context, field classification, adaptive classification, Bayesian classification.

1 Introduction and Motivation

In statistical pattern recognition, it is often assumed that the test patterns are independently and identically distributed (i.i.d.), therefore they are classified one at a time. One consequence of the i.i.d. assumption is that the labels assigned to the singlets in a test set are independent of the order in which the test singlets are presented to the classifier. However, classifying groups (or *fields*) of patterns is often more accurate than classifying single patterns (or *singlets*) because of interpattern statistical dependence or context. We defined in [1] and [2] a particular kind of interpattern dependence, called *style context*, which we exploited in field classification. We show below that such dependence also results in identical classification of singlets in the field independent of their order (unlike, for example, morphological and lexical context in Optical Character Recognition (OCR) and Automated Speech Recognition (ASR)).

The interpattern dependence among the patterns in the test field, induced by the fact that they belong to the same mixture component, is called *style context*. In many applications style context is the consequence of each group (field) of patterns to be recognized having been generated by one of several sources, as exemplified in Fig. 1.

Style consistency forms the basis for all *adaptive* classification, i.e., order-independent classifiers that modify their decision regions by exploiting the statistics of the test set. For instance, clustering using the expectation-maximization or K-means algorithms exploits some assumed similarity of patterns from the same class. The necessary conditions occur frequently in OCR and ASR, but, so far, adaptive methods have exploited only the consistency of patterns of the same class generated by a given source. In terms of style, we can define adaptation succinctly as *style-constrained field classification where the field encompasses the entire test set.*

Nagy suggested exploiting "spatial context" in [3] without any specific notion how this was to be done. The word "style" came to be used with similar meaning four or five years later. We call algorithms developed specifically under the three assumptions listed in the next section style-constrained or style-consistent classification. Discrete-style field classifiers were demonstrated using templates in [4], on simulated Gaussian distributions in [5], and on printed digits in [6]. We reported application of a continuous-style quadratic classifier to printed and hand-printed digits in [7], [8]. Sarkar and Veeramachaneni also derived fast, suboptimal approximations to the optimal style classifiers. An "adaptive" (within-class style) classifier was presented in [9]. We proposed in [10] some conjectures on the nature of class and style distributions in high-dimensional feature space, with supportive evidence on hand-printed characters. Most of this material, with detailed derivation of the classifier formulas, was assembled in three journal papers [11], [1], [2] which also include additional experimental results. In [12], we differentiated, by means of Bayesian networks (directed graphical models), style context from several other kinds of context that occur in OCR. Nagy reviewed progress in adaptive character recognition in [13], almost 40 years after his first attempts in that direction [14]. Adaptation in a commercial OCR engine was reported in [15], but no commercial exploitation of interclass style is known to us.

Our objective here is the formalization of the conditions where style context is beneficial and of the equations that govern its exploitation. We investigate four salient aspects of style-constrained classification:

1. order independence,
2. intraclass versus interclass style,
3. singlet versus field error optimization, and
4. dependence of error rate on field length.

Order independence distinguishes style context from linguistic context. The definitions of intraclass and interclass style clarify the difference between conventional adaptive classifiers and the broader notion of style-constrained classifiers. Since the traditional assumption (in Bayesian classification) of class-conditional independence between the features of different patterns is dispensed with, we can build classifiers optimized for either the field error rate or the singlet error rate. In addition to showing that the error rate of a (Bayesian) style-constrained classifier asymptotically converges to that of the optimal style-aware classifier, our bound provides insight into the properties of the classification problem that influence the reduction in error rate with style-constrained classification.

2 Notation and Assumptions

For simplicity we restrict our notation and discussion to two-class problems. We consider the problem of classifying a field-feature vector $\boldsymbol{y} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_L)$ (each $\boldsymbol{x}_i$ represents d feature measurements for one of L patterns in the field) produced in one style $s \in \mathcal{S}$. The field feature vector is an instance of the random vector $\mathbf{y} = (\mathbf{x}_1, \dots, \mathbf{x}_L)$. Let $\mathcal{C} = \{A, B\}$ be the set of singlet-class labels. Let $\mathbf{c}^i$ represent the class of the ith pattern of the field.[1] We make the following assumptions on the style, class, and feature distributions.

1. $p(\mathbf{c}^1, \mathbf{c}^2, \dots, \mathbf{c}^L) = p(\mathbf{c}^1)p(\mathbf{c}^2)\dots p(\mathbf{c}^L)$. That is, there is no higher order linguistic dependence[2] than the prior class probabilities $p(A)$ and $p(B) = 1 - p(A)$.

- S. Veeramachaneni is with the Automated Reasoning Systems Division (SRA), IRST-Istituto per la Ricerca Scientifica e Tecnologica, Via Sommarive 18, Povo, 38050, Trento, Italy. E-mail: hveera@gmail.com.
- G. Nagy is with the Department of Electrical, Computer, and Systems Engineering, 6020 Johnsson Engineering Center, Rensselaer Polytechnic Institute, Troy, NY 12180-3590. E-mail: nagy@ecse.rpi.edu.

Manuscript received 7 Feb. 2006; revised 1 Aug. 2006; accepted 23 Oct. 2006; published online 18 Jan. 2007.
Recommended for acceptance by H. Wechsler.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0134-0206. Digital Object Identifier no. 10.1109/TPAMI.2007.1030.

[1] If the fifth pattern of the field is a B, then it is denoted $\mathbf{c}^5 = B$.

[2] This assumption is desirable for exploring style context independently from linguistic context which is already widely used in classification.

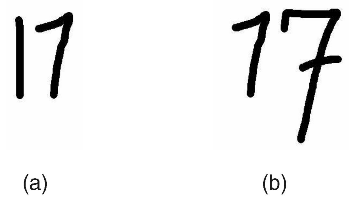


Fig. 1. A hypothetical example with style induced dependence—"17" written by two writers, 1 and 2. We can achieve higher accuracy by classifying a group of patterns from the same writer simultaneously than classifying the singlets in the group independently. Note that the label of the writer of the test field is not needed to improve the accuracy.

2. $p(A|s) = p(A)\,\forall\, s \in \mathcal{S}$. The prior class probabilities are style-independent. For multiwriter word recognition, this assumption states that the handwriting style of a writer does not influence his or her vocabulary.
3. $p(\mathbf{y}|\mathbf{c}^1, \mathbf{c}^2, \ldots, \mathbf{c}^L, s) = \prod_{i=1}^{L} p(\mathbf{x}_i|\mathbf{c}^i, s)\,\forall\, s \in \mathcal{S}$. The features of each pattern in the field are class-conditionally independent of the features of every other pattern in the same field. For multifont word recognition, this assumption states that, for the word $ABBA$ in a particular font, the noise in the first A is independent of the noise in the second one as well as of the noise in the Bs.

3 Order Independence

A consequence of our assumptions is order independence, which is central to the idea of exchangeability in modern Bayesian statistics. An infinite sequence of random variables is finitely exchangeable if the joint distribution of any finite subset of them is equal to that of any permutation of the subset. The theorem of De Finetti states that the probability distribution of a finitely exchangeable sequence must be a (possibly uncountably infinite) mixture of probability distributions of i.i.d. sequences [16]. In other words the sequence is conditionally i.i.d. We render this latent conditioning variable explicit and call it *style*.

Result 1. *Under our assumptions, for any permutation* $(i_1, \ldots, i_L)$ *of* $(1, \ldots, L)$

$$p(\mathbf{x}_1 = \boldsymbol{x}_1, \ldots, \mathbf{x}_l = \boldsymbol{x}_L \mid \mathbf{c}^1 = c_1, \ldots, \mathbf{c}^L = c_L)$$
$$= p(\mathbf{x}_1 = \boldsymbol{x}_{i_1}, \ldots, \mathbf{x}_L = \boldsymbol{x}_{i_L} \mid \mathbf{c}^1 = c_{i_1}, \ldots, \mathbf{c}^L = c_{i_L}).$$

Proof Outline. The left-hand side can be written as

$$p(\mathbf{x}_1, \ldots, \boldsymbol{x}_L \mid \mathbf{c}^1 = c_1, \ldots, \mathbf{c}^L = c_L)$$
$$= \sum_{s \in \mathcal{S}} p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, s \mid \mathbf{c}^1 = c_1, \ldots, \mathbf{c}^L = c_L).$$

The result follows straightforwardly from Assumptions 2 and 3 above. □

This result implies that the probability of a pattern field given a field class is equal to the probability of any permutation of the pattern field given the field class which is the same permutation of the original field class. Note that such order independence does not hold when there is class-label dependence due to linguistic context or sequence-induced interpattern feature dependence due to ligatures or coarticulation. Conversely, the assumption of Markov dependence is not appropriate to model the feature dependence arising due to a common source. We shall make further use of order independence for deriving a bound for the error rate as a function of field length.

4 Intraclass and Interclass Style Context

As reported in [1], style constrained classifiers trade-off loss of accuracy on same-class fields against gain on mixed-class fields. We, therefore, define two kinds of style context: *intraclass* style and *interclass* style. Intraclass style is present when there is statistical dependence between patterns of the same class in a field, i.e., there is at least one class $c \in \mathcal{C}$ such that

$$p(\mathbf{x}_1, \mathbf{x}_2|c, c) \neq p(\mathbf{x}_1|c)p(\mathbf{x}_2|c). \tag{1}$$

Interclass style is defined as the statistical dependence between patterns of different classes in the same field. That is, there exist two different classes $c_i, c_j \in \mathcal{C}$ such that

$$p(\mathbf{x}_1, \mathbf{x}_2|c_i, c_j) \neq p(\mathbf{x}_1|c_i)p(\mathbf{x}_2|c_j). \tag{2}$$

The above definitions are only for $L = 2$, but they generalize to longer fields. Class-conditional statistical dependence between triples of different-class patterns may arise even if all pairs are class-conditionally independent. This cannot happen with same-class patterns because dependence between any three patterns from the same class in a field implies dependence between all pairs. For simplicity, we shall avoid considering any egregious higher-order dependence without lower-order dependence.

Result 2. *The existence of interclass style context implies intraclass style context.*

Proof Outline. If there is no intraclass style context, then for every class, the class-conditional distributions are identical for all styles. This implies that inter-class style is absent as well. □

As mentioned, adaptive classifiers reported to date exploit only intraclass style context. They use the patterns in the test set to refine the estimates of the underlying distributions via clustering, expectation maximization, or decision-directed estimation, and the patterns in the training set to assign labels to these distributions. The essential idea is crystallized in [17], [18].

Adaptation can compensate both for an insufficient number of training samples and for nonrepresentative training sets, but classifiers that use only intraclass style context are clearly suboptimal for data that also exhibit interstyle context. As shown experimentally in [11], exploiting interclass style can increase the accuracy over using only intraclass style, especially on short fields. We now examine the essential characteristics of field classifiers that exploit the existing interclass style context.

5 Style-Constrained Classifiers

Most of the experiments we reported earlier were based on Gaussian class-and-style-conditional feature distributions. Here, we take a broader view to examine issues common to all style-constrained classifiers. We formulate criteria for singlet-error optimized and field-error optimized classification, propose an efficient approximation, and define, for analytical purposes, an abstract classifier.

As is customary, we assume that the cost of all interclass confusions is the same and, therefore, we wish to minimize the error rate. When classifying fields, however, we can minimize either the number of misclassified fields (a field is considered misclassified if any pattern in the field is misclassified), or the total number of misclassified patterns, regardless of how the errors are distributed among fields. In operational applications with short fields (bank check amounts and ZIP codes) field error rate is paramount, because regardless of the number of errors, after proofreading it is more convenient to reenter the whole field. However, in longer fields, like a business letter, the number of singlet errors, which will be individually corrected, must be minimized. We note also that the field error rate increases with field length, which must be taken into account in experimental comparisons with different field lengths [2].

When we classify field-patterns into field-classes, the Bayes decision rule using the zero-one loss function minimizes the field error rate: the classification decision that minimizes field error rate for the test field $\boldsymbol{y} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)$ is, as in [1]

$$\hat{\boldsymbol{c}} = \operatorname*{argmax}_{\boldsymbol{c}\in\mathcal{C}^L} p(\boldsymbol{c}|\boldsymbol{y}) = \operatorname*{argmax}_{\boldsymbol{c}\in\mathcal{C}^L} \sum_{s\in\mathcal{S}} p(\boldsymbol{y}|\boldsymbol{c}, s)p(\boldsymbol{c})p(s). \tag{3}$$

We call this field classifier *FOPT* (for Field error OPTimized).

To minimize the singlet error rate, we construct the so-called *SOPT* classifier (for Singlet error OPTimized), which is a Bayes decision rule with the Hamming distance loss function between the true and assigned field class labels. The *SOPT* classifier assigns the label $\hat{c}_l$ to $\boldsymbol{x}_l$, the lth pattern in the field $\boldsymbol{y} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)$, where

$$\begin{aligned} \hat{c}_l &= \operatorname*{argmax}_{c\in\mathcal{C}} p(\mathbf{c}^l = c|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) \\ &= \operatorname*{argmax}_{c\in\mathcal{C}} \sum_s p(c|\boldsymbol{x}_l, s)p(s|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L), \end{aligned} \tag{4}$$

where

$$p(s|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) = \frac{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L|s)p(s)}{\sum_s p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L|s)p(s)} = \frac{p(s)\prod_{l=1}^L p(\boldsymbol{x}_l|s)}{\sum_s p(s)\prod_{l=1}^L p(\boldsymbol{x}_l|s)}. \tag{5}$$

These optimal style classifiers can be computationally demanding. The *FOPT* classifier requires the computation of posterior probabilities for all field classes, the number of which increases exponentially with field length. Both the *FOPT* and *SOPT* classifiers require the averaging of posterior class probabilities over all styles.

The performance of the optimal field classifiers can be approximated by that of a *style-first* (*SF*) classifier[3] which first recognizes the style of the test field and then uses the appropriate style-conditional maximum a posteriori classifier. The *style-first* classifier assigns the label $\hat{c}_l$ to $\boldsymbol{x}_l$, the lth pattern in the field according to

$$\hat{c}_l = \operatorname*{argmax}_{c\in\mathcal{C}} p(c|\mathbf{x}_l, \hat{s}), \tag{6}$$

where

$$\begin{aligned} \text{where } \hat{s} &= \operatorname*{argmax}_s p(s|\boldsymbol{y}) = \operatorname*{argmax}_s p(s|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) \\ &= \operatorname*{argmax}_s \{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L|s)p(s)\}. \end{aligned} \tag{7}$$

To help us obtain bounds on the error rate, we now define a new style-constrained classifier called the *GIBBS* classifier.[4] For the test field $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L)$, the *GIBBS* classifier chooses a style s randomly according to the posterior distribution of the styles given the test field $p(s|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L)$ and then classifies each singlet $\boldsymbol{x}_l$ in the field according to the style-conditional distributions of style s. The *GIBBS* classifier assigns the label $\hat{c}_l$ to $\boldsymbol{x}_l$, the lth pattern in the field according to

$$\hat{c}_l = \operatorname*{argmax}_{c\in\mathcal{C}} p(c|\boldsymbol{x}_l, \hat{s}) \tag{8}$$

$$\text{where } \hat{s} \sim p(s|\boldsymbol{y}) = p(s|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) \tag{9}$$

From the above classification functions and Result 1, it is clear that all of the above classifiers classify a test field order-independently, i.e., every singlet in the field is classified indepedently of its position in the field. Note that the difference between the *SOPT*, *SF*, and *GIBBS* classifiers stems from their different usage of the posterior style distribution. By construction, the singlet error rate of the *SOPT* classifier is lower than that of the *SF* or the *GIBBS* classifier.

Although we have so far implied that there is a finite number of discrete styles, it may sometimes be advantageous to consider a continuous distribution of styles. In handwriting, a training set may contain a large number of writers, but it is unlikely that any writers in the test set will duplicate the patterns of some writer in the training set. For example, in [2], we assumed that the class-conditional distributions of every writer are identical, but their means vary continuously. We showed that, if both distributions are Gaussian, then all the patterns of a class will be distributed according to a single Gaussian distribution obtained by convolving the distribution of the style means with the class-and-style-conditional distribution.

6 Error Rate versus Field Length

We now present some results on the error rate of the *SOPT* style-constrained classifier. Since, due to order independence, the singlet error rate is independent of the position of the singlet in the field, the probability of singlet error of the *SOPT* classifier (see (4)) acting on fields of length L is

$$\begin{aligned} p_L(e) &= \int_{\boldsymbol{x}_1\ldots\boldsymbol{x}_L} \min_{c\in\mathcal{C}}\{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, \mathbf{c}^1 = c)\} \\ &= \int_{\boldsymbol{x}_1\ldots\boldsymbol{x}_L} \min_{c\in\mathcal{C}}\left\{\sum_s p(c|\boldsymbol{x}_1, s)p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L|s)p(s)\right\}. \end{aligned} \tag{10}$$

Result 3. *$p_L(e)$ is a monotonically nonincreasing function of L.*

Proof Outline. It follows from (10) and the inequality

$$\int_x \min\{a(x), b(x)\} \le \min\left\{\int_x a(x), \int_x b(x)\right\}.$$

□

In particular, the error rate of an optimal style-constrained field classifier for any field length is no greater than the error rate of a singlet classifier. We show later that except under pathological conditions, it is less.

We can calculate the probability of singlet error of a classifier realizable only if the styles are known (called *style-aware* classifier). It is given by

$$\begin{aligned} p^\star(e) &= \sum_s p(s)\int_{\boldsymbol{x}} \min_{c\in\mathcal{C}}\{p(\boldsymbol{x}, \mathbf{c} = c|s)\} \\ &= \sum_s p(s)\int_{\boldsymbol{x}} \min_{c\in\mathcal{C}}\{p(\mathbf{c} = c|\boldsymbol{x}, s)\}p(\boldsymbol{x}), \end{aligned} \tag{11}$$

$$\triangleq \sum_s p(s)\int_{\boldsymbol{x}} p^\star(e|\boldsymbol{x}, s)p(\boldsymbol{x}). \tag{12}$$

Result 4. $p_L(e) \ge p^\star(e)\,\forall\, L.$

Proof Outline. It follows from (10), (11), and the inequality

$$\sum_x \min\{a(x), b(x)\} \le \min\left\{\sum_x a(x), \sum_x b(x)\right\}.$$

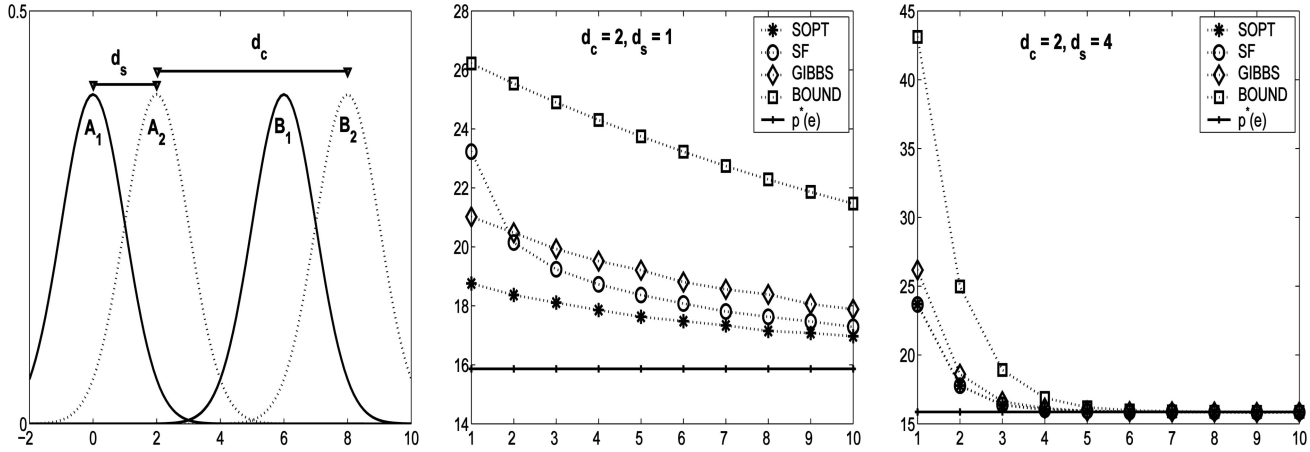
□

That is, the probability of error for the style-constrained field classifier is never lower than that of the *style-aware* classifier. Also, we note that $\lim_{L\to\infty} p_L(e) \triangleq p_\infty(e)$ is not always equal to $p^\star(e)$.

Result 5. *$p_\infty(e) = p^\star(e)$, the style-aware error rate, if the styles are distinguishable.*

[3] A different suboptimal approximation is derived in [5].

[4] Our *GIBBS* classifier is, in principle, similar to its namesake in [19], but we use it in a different context. Here, the singlet being classified also plays a part in altering the posterior distribution according to which the style is sampled.



(a) (b) (c)

Fig. 2. (a) Shows the style-and-class-conditional densities. (b) and (c) Show the plots of error rate versus the field length. In both cases, $p^\star(e) = 15.87\%$.

Proof Outline. Let the test field be $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L)$. From (4), the label $c_l^\star$ assigned by the *SOPT* classifier to $\boldsymbol{x}_l$ is

$$\hat{c}_l = \operatorname*{argmax}_{c \in \mathcal{C}} \sum_s p(\mathbf{c}^l = c | \boldsymbol{x}_l, s) p(s | \boldsymbol{x}_1, \ldots, \mathrm{x}_L). \tag{13}$$

If the styles are statistically distinguishable, i.e., for two distinct styles s_1 and s_2, $p(\boldsymbol{x}|s_1)$ is "different" from $p(\boldsymbol{x}|s_2)$, then we have

$$\lim_{L \to \infty} p(s | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) = \delta(s, s^\star), \tag{14}$$

where $s^\star$ is the identity of the style that generated the field and $\delta(.)$ is the Kronecker delta function. Thus, we have from (13) and (14) that

$$\hat{c}_l = \operatorname*{argmax}_{c \in \mathcal{C}} \sum_s p(\mathbf{c}^l = c | \boldsymbol{x}_l, s) \delta(s, s^\star) = \operatorname*{argmax}_{c \in \mathcal{C}} p(\mathbf{c}^l = c | \boldsymbol{x}_l, s^\star).$$

Thus, when the styles are distinguishable, a style-constrained field classifier optimized for character error rate converges to the *style-aware* classifier asymptotically with the length of the test field. Consequently, the error rate of the *SOPT* classifier converges to $p^\star(e)$. □

In the Appendix we derive an upper bound on the error rate of the *SOPT* classifier by bounding the error rate of the *GIBBS* classifier. The upper bound is related to the error rate of the *style-aware* classifier (i.e., the lowest achievable error rate), the pairwise difference of style-specific classifiers and the pairwise difference in style-conditional feature distributions. Let the *style-aware* classifier for style $s_i \in \mathcal{S}$ be denoted $\phi_i(.)$.

Result 6.

$$p_L(e) \leq p_L^{GIBBS}(e) \leq p^\star(e) + 2 \sum_{i \neq j} \mu(s_i, s_j) \nu(s_i, s_j)^{(L-1)}, \tag{15}$$

where $p^\star(e)$ is the error rate when the style of the field is known and

$$\begin{aligned} &\mu(s_i, s_j) \\ &\quad = \int_{\phi_j(\boldsymbol{x}) \neq \phi_i(\boldsymbol{x})} (1 - p^\star(e|\boldsymbol{x}, s_i) - p^\star(e|\boldsymbol{x}, s_j)) \ldots \\ &\quad\quad \ldots \sqrt{p(\boldsymbol{x}, s_i) p(\boldsymbol{x}, s_j)} d\boldsymbol{x} \end{aligned}$$

depends on the difference in the style-conditional classification boundaries of styles s_i and s_j in the singlet feature space, and

$$\nu(s_i, s_j) = \int_{\boldsymbol{x}} \sqrt{p(\boldsymbol{x}|s_i) p(\boldsymbol{x}|s_j)}$$

is a measure of the difference between the style-conditional singlet feature distributions of s_i and s_j. In addition, $\nu(s_i, s_j) < 1$ when the style-conditional distributions are different, implying that the bound approaches $p^\star(e)$ asymptotically with L.

In Fig. 2, we show the singlet error rates obtained by simulations for various classifiers along with the bound as a function of the field length L. We use synthetic data to illustrate our analytical results for *Bayesian* classification. We cannot compute the bound exactly for real data because we can only estimate the true distributions. The simulations were conducted for Gaussian style-and-class-conditional feature distributions given by (see Fig. 2)

$$\begin{aligned} &(\mathbf{x}|A, s_1) \sim N(0, 1),\ (\mathbf{x}|A, s_2) \sim N(d_s, 1), \\ &(\mathbf{x}|B, s_1) \sim N(d_c, 1),\ \text{and}\ (\mathbf{x}|B, s_2) \sim N(d_c + d_s, 1). \end{aligned}$$

Figs. 2b and 2c show the error rates for the *SOPT*, *SF*, and the *GIBBS* classifiers, along with the bound and the error rate of the *style-aware* classifier for two choices of d_c and d_s as a function of the field length. Note that, because in Fig.2c, the styles are more distinguishable than in Fig. 2a, the error rates of the classifiers (and the bound) approach $p^\star(e)$ more rapidly with increasing field length.

The presence of style dependence does not necessarily decrease the Bayes Risk achievable by field classification. Equation (15) helps us to intuit situations where there is style dependence which does not translate to a decrease in error rate. When does this happen? In situations where:

1. the singlet error rate or the error rate in classifying the style of a singlet is zero, or
2. the style-specific classifiers are identical across styles, or
3. the error rate of the *style-aware* classifier is 50 percent, or
4. the style-conditional singlet feature distributions are identical.

In the first three situations, the error rate of the singlet classifier already matches the accuracy of the *style-aware* classifier, and in the last case, although the error rate of the *style-aware* classifier may be lower, it cannot be achieved by field classification because the styles are not distinguishable. These situations are not representative of most application domains.

7 Summary

The decrease in error rate due to style-constrained classification has already been amply demonstrated experimentally. We believe that the analytical findings communicated above provide guidance for further development of field classifiers based on less restrictive assumptions. We explored the connection between style-constrained classification and exchangeability. We defined intraclass and interclass style and showed how the commonly accepted notion of adaptive classification fits into the style framework. Even though it is difficult to find situations where intraclass style occurs without interclass style, we drew a distinction between them because many adaptive classification algorithms exploit only intraclass style. Only recently has attention been focused on interclass style. We gave a general formulation for style classification and showed that minimizing the singlet error rate and field error rate require different algorithms and have different applications. We defined an abstract Gibbs classifier, by means of which we investigated the decrease in error rate with field length. We proved that when the styles are distinguishable, the error rate of the optimal style-constrained classifier converges asymptotically to that of the style-aware Bayes singlet classifier.

Appendix I

Proof Outline for Result 6

Proof Outline. Let us consider the singlet error rate of the *GIBBS* classifier, say on $\boldsymbol{x}_1$, from the test field $\boldsymbol{y} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L)$ generated in style s_i. The probability of error is the sum, over all j, of the probability of choosing a style s_j times the probability of misclassifying $\boldsymbol{x}_1$ by style s_j, given that it was generated in s_i. Recall that the label assigned to $\boldsymbol{x}_1$ using the classification boundaries for style $s_i \in \mathcal{S}$ by $\phi_i(\boldsymbol{x}_1)$.

$$p_L^{GIBBS}(e|\boldsymbol{y}, s_i) \;=\; \sum_j p(s_j|\boldsymbol{y}) p(\overline{\phi_j(\boldsymbol{x}_1)}|\boldsymbol{x}_1, s_i),$$

where $\overline{\phi_j(\boldsymbol{x}_1)}$ denotes that the style-conditional classifier of s_j makes an error on $\boldsymbol{x}_1$.

Now (for a two class problem), if $\phi_j(\boldsymbol{x}_1) = \phi_i(\boldsymbol{x}_1)$, i.e., if both styles s_i and s_j classify $\boldsymbol{x}_1$ identically, $p(\overline{\phi_j(\boldsymbol{x}_1)}|\boldsymbol{x}_1, s_i) = p^\star(e|\boldsymbol{x}_1, s_i)$ and if $\phi_j(\boldsymbol{x}_1) \neq \phi_i(\boldsymbol{x}_1)$, then $p(\overline{\phi_j(\boldsymbol{x}_1)}|\boldsymbol{x}_1, s_i) = 1 - p^\star(e|\boldsymbol{x}_1, s_i)$. Therefore,

$$\begin{aligned}
&p_L^{GIBBS}(e|\boldsymbol{y}, s_i)\\
&\quad= \sum_j p(s_j|\boldsymbol{y}) \ldots\\
&\qquad \{p^\star(e|\boldsymbol{x}_1, s_i) + I(\phi_j(\boldsymbol{x}_1) \neq \phi_i(\boldsymbol{x}_1))(1 - 2p^\star(e|\boldsymbol{x}_1, s_i))\},
\end{aligned}$$

where $I(.)$ is the indicator function, which is 1 when its argument is true, and 0 otherwise.

Removing the conditioning on the field and the style, the error rate of the *GIBBS* classifier is

$$\begin{aligned}
&p_L^{GIBBS}(e)\\
&\quad= \int_{\boldsymbol{y}} \sum_i \sum_j p(s_i|\boldsymbol{y}) p(s_j|\boldsymbol{y}) \ldots\\
&\qquad \ldots \{p^\star(e|\boldsymbol{x}_1, s_i) + I(\phi_j(\boldsymbol{x}_1) \neq \phi_i(\boldsymbol{x}_1)) \ldots\\
&\qquad \ldots (1 - 2p^\star(e|\boldsymbol{x}_1, s_i))\} p(\boldsymbol{y})\\
&\quad= p^\star(e) + \int_{\boldsymbol{y}} \sum_i \sum_j p(s_i|\boldsymbol{y}) p(s_j|\boldsymbol{y}) \ldots\\
&\qquad \ldots I(\phi_j(\boldsymbol{x}_1) \neq \phi_i(\boldsymbol{x}_1))(1 - 2p^\star(e|\boldsymbol{x}_1, s_i)) p(\boldsymbol{y}).
\end{aligned}$$

Now, since $p(s_i|\boldsymbol{y}) p(s_j|\boldsymbol{y}) \leq \min\{p(s_i|\boldsymbol{y}), p(s_j|\boldsymbol{y})\}$, we have

$$\begin{aligned}
&p_L^{GIBBS}(e) - p^\star(e)\\
&\quad\leq \sum_i \sum_j \int_{\boldsymbol{x}_1} I(\phi_j(\boldsymbol{x}_1) \neq \phi_i(\boldsymbol{x}_1))(1 - 2p^\star(e|\boldsymbol{x}_1, s_i)) \ldots\\
&\qquad \ldots \int_{\boldsymbol{x}_2 \ldots \boldsymbol{x}_L} \min\{p(s_i, \boldsymbol{y}), p(s_j, \boldsymbol{y})\}\\
&\quad= \sum_i \sum_j \int_{\boldsymbol{x}_1} I(\phi_j(\boldsymbol{x}_1) \neq \phi_i(\boldsymbol{x}_1))(1 - 2p^\star(e|\boldsymbol{x}_1, s_i)) \ldots\\
&\qquad \ldots \int_{\boldsymbol{x}_2 \ldots \boldsymbol{x}_L} \min\{p(s_i, \boldsymbol{x}_1) p(\boldsymbol{x}_2 \ldots, x_L|s_i), \ldots\\
&\qquad \ldots p(s_j, \boldsymbol{x}_1) p(\boldsymbol{x}_2 \ldots, \boldsymbol{x}_L|s_j)\}\\
&\quad= \sum_i \sum_j \int_{\boldsymbol{x}_1} I(\phi_j(\boldsymbol{x}_1) \neq \phi_i(\boldsymbol{x}_1))(1 - 2p^\star(e|\boldsymbol{x}_1, s_i)) \ldots\\
&\qquad \ldots \int_{\boldsymbol{x}_2 \ldots \boldsymbol{x}_L} \min\{p(s_i, \boldsymbol{x}_1) p(\boldsymbol{x}_2|s_i) \ldots p(\boldsymbol{x}_L|s_i), \ldots\\
&\qquad \ldots p(s_j, \boldsymbol{x}_1) p(\boldsymbol{x}_2|s_j) \ldots p(\boldsymbol{x}_L|s_j)\}\\
&\quad\leq \sum_i \sum_j \int_{\boldsymbol{x}_1} I(\phi_j(\boldsymbol{x}_1) \neq \phi_i(\boldsymbol{x}_1))(1 - 2p^\star(e|\boldsymbol{x}_1, s_i)) \ldots\\
&\qquad \sqrt{p(\boldsymbol{x}_1, s_i) p(\boldsymbol{x}_1, s_j)} \left\{ \int_{\boldsymbol{x}} \sqrt{p(\boldsymbol{x}|s_i) p(\boldsymbol{x}|s_j)} \right\}^{L-1}\\
&\quad= 2 \sum_{i \neq j} \int_{\boldsymbol{x}} I(\phi_j(\boldsymbol{x}_1) \neq \phi_i(\boldsymbol{x}_1)) \ldots\\
&\qquad \ldots (1 - p^\star(e|\boldsymbol{x}, s_i) - p^\star(e|\boldsymbol{x}, s_j)) \sqrt{p(\boldsymbol{x}, s_i) p(\boldsymbol{x}, s_j)} \ldots\\
&\qquad \ldots \left\{ \int_{\boldsymbol{x}} \sqrt{p(\boldsymbol{x}|s_i) p(\boldsymbol{x}|s_j)} \right\}^{L-1}\\
&\quad= 2 \sum_{i \neq j} \mu(s_i, s_j) \nu(s_i, s_j)^{(L-1)}.
\end{aligned}$$

The fact that $\nu(s_i, s_j) < 1$, if the styles are distinguishable, follows from the Cauchy-Schwarz inequality. □

Acknowledgments

George Nagy gratefully acknowledges the support of IRST, Trento, and of Hitachi CRL, Tokyo. The authors would like to thank Dr. Prateek Sarkar for suggesting the relationship with exchangeability and several clarifications and Dr. Diego Sona for discussions on the error rate bounds. The authors thank the anonymous reviewers for their suggestions to improve the presentation.

References

[1] P. Sarkar and G. Nagy, "Style Consistent Classification of Isogenous Patterns," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 1, pp. 14-22, Jan. 2005.

[2] S. Veeramachaneni and G. Nagy, "Style Context with Second-Order Statistics," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 1, pp. 88-98, Jan. 2005.

[3] G. Nagy, "Teaching a Computer to Read," *Proc. 11th Int'l Conf. Pattern Recognition,* vol. 2, pp. 225-229, 1992.

[4] P. Sarkar and G. Nagy, "Heeding More than the Top Template," *Proc. Fifth Int'l Conf. Document Analysis and Recognition,* pp. 382-385, 1999.

[5] P. Sarkar and G. Nagy, "Classification of Style-Constrained Pattern-Fields," *Proc. 15th Int'l Conf. Pattern Recogntion,* pp. 859-862, 2000.

[6] P. Sarkar and G. Nagy, "Style Consistency in Isogenous Patterns," *Proc. Sixth Int'l Conf. Document Analysis and Recognition,* pp. 1169-1174, 2001.

[7] S. Veeramachaneni, H. Fujisawa, C.-L. Liu, and G. Nagy, "Classifying Isogenous Fields," *Proc. Eighth Int'l Workshop Frontiers in Handwriting Recognition,* pp. 41-46, 2002.

[8] S. Veeramachaneni, H. Fujisawa, C.-L. Liu, and G. Nagy, "Style-Conscious Quadratic Field Classifier," *Proc. 16th Int'l Conf. Pattern Recogntion,* vol. II, pp. 72-75, 2002.

[9] S. Veeramachaneni and G. Nagy, "Classifier Adaptation with Non-Representative Training Data," *Proc. Fifth Int'l Workshop Document Analysis Systems,* pp. 123-133, 2002.

[10] S. Veeramachaneni and G. Nagy, "Towards a Ptolemaic Model for OCR," *Proc. Seventh Int'l Conf. Document Analysis and Recognition,* pp. 1060-1064, 2003.

[11] S. Veeramachaneni and G. Nagy, "Adaptive Classifiers for Multisource OCR," *Int'l J. Document Analysis and Recognition,* vol. 6, no. 3, pp. 154-166, 2003.

[12] S. Veeramachaneni, P. Sarkar, and G. Nagy, "Modeling Context as Statistical Dependence," *Proc. Fifth Int'l and Interdisciplinary Conf. Modeling and Using Context,* pp. 515-528, 2005.

[13] G. Nagy, "Classifiers that Improve with Use," *Proc. Int'l Conf. Pattern Recognition and Multimedia,* pp. 79-86, 2004.

[14] G. Nagy and G.L. Shelton Jr., "Self-Corrective Character Recognition System," *IEEE Trans. Information Theory,* vol. 12, no. 2, pp. 215-222, 1966.

[15] I. Marosi and L. Toth, "OCR Voting Methods for Recognizing Low Contrast Printed Documents," *Proc. Second Int'l Conf. Document Image Analysis for Libraries,* pp. 108-115, 2006.

[16] D. Heath and W. Sudderth, "De Finetti's Theorem on Exchangeable Variables," *Am. Statistician,* vol. 30, no. 4, pp. 188-189, 1976.

[17] V. Castelli and T.M. Cover, "On the Exponential Value of Labeled Samples," *Pattern Recognition Letters,* vol. 16, no. 1, pp. 105-111, 1995.

[18] V. Castelli and T.M. Cover, "The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition with an Unknown Mixing Parameter," *IEEE Trans. Information Theory,* vol. 42, no. 6, pp. 2102-2117, 1996.

[19] D. Haussler, M. Kearns, and R. Schapire, "Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension," *Proc. Fourth Ann. Workshop Computational Learning Theory,* pp. 61-74, 1991.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.