
How many faces can be recognized? Performance extrapolation for multi-class classification

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 The difficulty of multi-class classification generally increases with the number of
2 classes. Using data from a subset of the classes, can we predict how well a classifier
3 will scale with an increased number of classes? Under the assumption that the
4 classes are sampled exchangeably, and under the assumption that the classifier is
5 generative (e.g. QDA or Naive Bayes), we show that the expected accuracy when
6 the classifier is trained on k classes is the $k - 1$ st moment of a *conditional accuracy*
7 *distribution*, which can be estimated from data. This provides the theoretical
8 foundation for performance extrapolation based on pseudolikelihood, unbiased
9 estimation, and high-dimensional asymptotics. We find empirically that some of
10 the methods work well even for non-generative classifiers.

11 1 Introduction

12 In multi-class classification, one observes pairs (z, y) where $y \in \mathcal{Y} \subset \mathbb{R}^p$ are feature vectors, and z
13 are unknown labels, which lie in a countable label set \mathcal{Z} . The goal is to construct a classification rule
14 for predicting the label of a new data point; generally, the classification rule $h : \mathcal{Y} \rightarrow \mathcal{Z}$ is learned
15 from previously observed data points. In many applications of multi-class classification, such as face
16 recognition or image recognition, the space of potential labels is practically infinite. In such a setting,
17 one might consider a sequence of classification problems on finite label subsets $\mathcal{Z}_1 \subset \dots \subset \mathcal{Z}_K$,
18 where in the i th problem, one constructs the classification rule $h^{(i)} : \mathcal{Y} \rightarrow \mathcal{Z}_i$. Supposing that (Z, Y)
19 have a joint distribution, define the accuracy for the i th problem as

$$\text{acc}^{(i)} = \Pr[h^{(i)}(Y) = Z | Z \in \mathcal{Z}_i].$$

20 Using data from only \mathcal{Z}_k , can one predict the accuracy achieved on the larger label set \mathcal{Z}_K , with
21 $K > k$? This is the problem of *performance extrapolation*.

22 A practical instance of performance extrapolation occurs in neuroimaging studies, where the number
23 of classes k is limited by experimental considerations. Kay et al. (2008) obtained fMRI brain
24 scans which record how a single subject's visual cortex responds to natural images. The label set \mathcal{Z}
25 corresponds to the space of all grayscale photographs of natural images, and the set \mathcal{Z}_1 is a subset
26 of 1750 photographs used in the experiment. They construct a classifier which achieves over 0.75
27 accuracy for classifying the 1750 photographs; based on exponential extrapolation, they estimate that
28 it would take on the order of $10^{9.5}$ photographs before the accuracy of the model drops below 0.10!
29 Directly validating this estimate would take immense resources, so it would be useful to develop the
30 theory needed to understand how to compute such extrapolations in a principled way.

31 However, in the fully general setting, it is impossible to construct non-trivial bounds on the accuracy
32 achieved on the new classes $\mathcal{Z}_K \setminus \mathcal{Z}_k$ based only on knowledge of \mathcal{Z}_k : after all, \mathcal{Z}_k could consist
33 entirely of well-separated classes while the new classes $\mathcal{Z}_K \setminus \mathcal{Z}_k$ consist entirely of highly inseparable

34 classes, or vice-versa. Thus, the most important assumption for our theory is that of *exchangeable*
 35 *sampling*. The labels in \mathcal{Z}_i are assumed to be an exchangeable sample from \mathcal{Z} . The condition of
 36 exchangeability ensures that the separability of random subsets of \mathcal{Z} can be inferred by looking at the
 37 empirical distributions in \mathcal{Z}_k , and therefore that some estimate of the achievable accuracy on \mathcal{Z}_K can
 38 be obtained.

39 In addition to the assumption of exchangeability, we restrict the set of classifiers considered. We focus
 40 on *generative classifiers*, which are classifiers that work by training a model separately on each class.
 41 This convenient property allows us to characterize the accuracy of the classifier by selectively condi-
 42 tioning on one class at a time: in section 3, we use this technique to reveal an equivalence between
 43 the expected accuracies of \mathcal{Z}_k to moments of a common distribution. This moment equivalence result
 44 allows standard approaches in statistics, such as U-statistics and nonparametric pseudolikelihood,
 45 to be directly applied to the extrapolation problem, as we discuss in section 4. In non-generative
 46 classifiers, the classification rule has a joint dependence on the entire set of classes, and cannot be
 47 analyzed by conditioning on individual classes. Nevertheless, in Section 5, we see that our methods
 48 achieve similarly accurate extrapolation for both generative and non-generative classifiers in real data
 49 examples.

50 2 Setting

51 Having motivated the problem of performance extrapolation, we now reformulate the problem for
 52 notational and theoretical convenience. Instead of requiring \mathcal{Z}_k to be a random subset of \mathcal{Z} as we
 53 did in section 1, take $\mathcal{Z} = \mathbb{N}$ and $\mathcal{Z}_k = \{1, \dots, k\}$. We fix the size of \mathcal{Z}_k without losing generality,
 54 since any monotonic sequence of finite subsets can be embedded in a sequence with $|\mathcal{Z}_k| = k$. In
 55 addition, rather than randomizing the labels, we will randomize the marginal distribution of each
 56 label; Towards that end, let $\mathcal{Y} \subset \mathbb{R}^p$ be a space of feature vectors, and let $\mathcal{P}(\mathcal{Y})$ be a measurable
 57 space of probability distributions on \mathcal{Y} . Let \mathcal{F} be a probability measure on \mathcal{P} , and let F_1, F_2, \dots
 58 be an infinite sequence of i.i.d. draws from \mathbb{F} . We refer to \mathbb{F} , a probability measure on probability
 59 measures, as a *meta-distribution*. The distributions F_1, \dots, F_k are the marginal distributions of the
 60 first k classes. We therefore rewrite the accuracy as

$$\text{acc}^{(i)} = \frac{1}{t} \sum_{i=1}^t \Pr_{F_i} [h^{(t)}(Y) = i].$$

61 where the probabilities are taken over $Y \sim F_i$.

62 In order to construct the classification rule $h^{(t)}$, we need data from the classes F_1, \dots, F_t . In most
 63 instances of multi-class classification, one observes independent observations from each F_i which
 64 are used to construct the classifier. Since the order of the observations does not generally matter, a
 65 sufficient statistic for the training data for the t th classification problem is the collection of empirical
 66 distributions $\hat{F}_1^{(t)}, \dots, \hat{F}_t^{(t)}$ for each class. Henceforth, we make the simplifying assumption that
 67 the training data for the i th class remains fixed from $t = i, i + 1, \dots$, so we drop the superscript on
 68 $\hat{F}_i^{(t)}$. Write $\hat{\mathbb{F}}(F)$ for the conditional distribution of \hat{F}_i given $F_i = F$; also write $\hat{\mathbb{F}}$ for the marginal
 69 distribution of \hat{F} when $F \sim \mathbb{F}$. As an example, suppose every class has the number of training
 70 examples $r \in \mathbb{N}$; then \hat{F} is the empirical distribution of r i.i.d. observations from F , and $\hat{\mathbb{F}}(F)$ is the
 71 *empirical meta-distribution* of \hat{F} . Meanwhile, $\hat{\mathbb{F}}$ is the meta-distribution of the empirical distribution
 72 of r i.i.d. draws from a random $F \sim \mathbb{F}$.

73 2.1 Multiclass classification

74 Extending the formalism of Tewari and Bartlett (2007)¹, we define a classifier as a collection of
 75 mappings $\mathcal{M}_i : \mathcal{P}(\mathcal{Y})^k \times \mathcal{Y} \rightarrow \mathbb{R}$ called *margin functions*. Intuitively speaking, each margin function
 76 *learns a model* from the first k arguments, which are the empirical marginals of the k classes, which
 77 it uses to assign a *margin* or *score* to the *query point* $y \in \mathcal{Y}$. A higher score $\mathcal{M}_i(\hat{F}_1, \dots, \hat{F}_k, y)$

¹We borrow their terminology of *margin functions*, but introduce the notion of a classifier as a multiple-argument functional on empirical distributions. The functional formulation of a classifier echoes the functional formulation of estimators common in the statistical literature.

indicates a higher estimated probability that y belongs to the k th class. Therefore, the classification rule corresponding to a classifier \mathcal{M}_i assigns a class with maximum margin to y :

$$h(y) = \operatorname{argmax}_{i \in \{1, \dots, k\}} \mathcal{M}_i(y).$$

For some classifiers, the margin function \mathcal{M}_i is especially simple in that \mathcal{M}_i is only a function of \hat{F}_i and y . Furthermore, due to symmetry, in such cases one can write

$$\mathcal{M}_i(\hat{F}_1, \dots, \hat{F}_k, y) = \mathcal{Q}(\hat{F}_i, y)$$

where \mathcal{Q} is called a *single-class margin* (or simply *margin*.) Quadratic discriminant analysis and Naive Bayes are two examples of generative classifiers².

For notational convenience, we assume that ties occur with probability zero: that is, $\hat{\mathbb{F}}$ and \mathcal{Q} jointly satisfy the *tie-breaking* property:

$$\Pr[\mathcal{Q}(\hat{F}, y) = \mathcal{Q}(\hat{F}', y)] = 0 \quad (1)$$

for all $y \in \mathcal{Y}$, where $\mathbb{F}, \mathbb{F}' \stackrel{iid}{\sim} \hat{\mathbb{F}}$. Note that the tie-breaking property implies that \mathbb{F} contains no atoms. The *generative* property allows us to prove strong results about the accuracy of the classifier under the exchangeable sampling assumption, as we see in Section 3.

3 Performance extrapolation for generative classifiers

Let us specialize to the case of a generative classifier, with scoring rule \mathcal{Q} . Consider estimating the expected accuracy at time t ,

$$p_t \stackrel{def}{=} \mathbf{E}[\operatorname{acc}(t)].$$

Define the *conditional accuracy* function $u(\hat{F}, y)$ which maps a distribution \hat{F} on \mathcal{Y} and a *test* observation y to a real number in $[0, 1]$. The conditional accuracy gives the probability that for independently drawn \hat{F}' from $\hat{\mathbb{F}}$, that $\mathcal{Q}(\hat{F}, y)$ will be greater than $\mathcal{Q}(\hat{F}', y)$:

$$u(\hat{F}, y) = \Pr_{\hat{F}' \sim \hat{\mathbb{F}}} [\mathcal{Q}(\hat{F}, y) > \mathcal{Q}(\hat{F}', y)].$$

Define the *conditional accuracy* distribution ν as the law of $u(\hat{F}, Y)$ where \hat{F} and Y are generated as follows: (i) a true distribution F is drawn from \mathbb{F} ; (ii) the query Y is drawn from F , and (iii) the empirical distribution \hat{F} is drawn from $\hat{\mathbb{F}}(F)$ (e.g., the distribution of the empirical distribution of r i.i.d. observations drawn from F), with Y independent of \hat{F} . The significance of the conditional accuracy distribution is that the expected generalization error p_t can be written in terms of its moments.

Theorem 3.1. *Let \mathcal{Q} be a single-distribution margin, and let $\mathbb{F}, \hat{\mathbb{F}}(F)$ be a distribution on $\mathcal{P}(\mathcal{Y})$. Let U be defined as the random variable*

$$U = u(\hat{F}, Y)$$

for $F \sim \mathbb{F}$, $Y \sim F$, and $\hat{F} \sim \hat{\mathbb{F}}(F)$ with $Y \perp \hat{F}$. Recall the definition

$$p_k = \mathbf{E}[\operatorname{acc}^{(k)}] = \mathbf{E} \left[\frac{1}{k} \sum_{i=1}^k \Pr_{Y \sim F_i} [\mathcal{Q}(\hat{F}_i, Y) > \max_{j \neq i} \mathcal{Q}(\hat{F}_j, Y)] \right].$$

²For QDA, the margin is given by

$$\mathcal{Q}_{QDA}(\hat{F}, y) = -(y - \mu(\hat{F}))^T \Sigma(\hat{F})^{-1} (y - \mu(\hat{F})) - \log \det(\Sigma(\hat{F}))$$

where $\mu(F) = \int y dF(y)$ and $\Sigma(F) = \int (y - \mu(F))(y - \mu(F))^T dF(y)$. In Naive Bayes, the margin is

$$\mathcal{Q}_{NB}(\hat{F}, y) = \sum_{i=1}^n \log \hat{f}_i(y_i)$$

where \hat{f}_i is a density estimate for the i th component of \hat{F} .

104 Then

$$p_k = \mathbf{E}[U^{k-1}].$$

105

106 **Proof.** Write $q^{(i)}(y) = \mathcal{Q}(\hat{F}_i, y)$. By using conditioning and conditional independence, p_k can be
 107 written

$$\begin{aligned} p_k &= \mathbf{E} \left[\frac{1}{k} \sum_{i=1}^k \Pr[q^{(i)}(Y) > \max_{j \neq i} q^{(j)}(Y)] \right] \\ &= \mathbf{E} \left[\Pr[q^{(1)}(Y) > \max_{j \neq 1} q^{(j)}(Y)] \right] \\ &= \mathbf{E}_{F_1} [\Pr[q^{(1)}(Y) > \max_{j \neq 1} q^{(j)}(Y) | \hat{F}_1, Y]] \\ &= \mathbf{E}_{F_1} [\Pr[\cap_{j>1} q^{(1)}(Y) > q^{(j)}(Y) | \hat{F}_1, Y]] \\ &= \mathbf{E}_{F_1} [\prod_{j>1} \Pr[q^{(1)}(Y) > q^{(j)}(Y) | \hat{F}_1, Y]] \\ &= \mathbf{E}_{F_1} [\Pr[q^{(1)}(Y) > q^{(2)}(Y) | \hat{F}_1, Y]^{k-1}] \\ &= \mathbf{E}_{F_1} [u(\hat{F}_1, Y)^{k-1}] = \mathbf{E}[U^{k-1}]. \end{aligned}$$

108 \square

109 Theorem 3.1 tells us that the problem of extrapolation can be approached by attempting to estimate
 110 the conditional accuracy distribution. The $(t-1)$ th moment of U gives us p_t , which will in turn be a
 111 good estimate of $\text{acc}^{(t)}$.

112 While $U = u(\hat{F}, Y)$ is not directly observed, we can obtain unbiased estimates of $u(\hat{F}_i, y)$ by using
 113 test data. For any $\hat{F}_1, \dots, \hat{F}_k$, and independent test point $Y \sim F_i$, define

$$\hat{u}(\hat{F}_i, Y) = \frac{1}{k-1} \sum_{j \neq i} I(\mathcal{Q}(\hat{F}_i, Y) > \mathcal{Q}(\hat{F}_j, Y)). \quad (2)$$

114 Then $\hat{u}(\hat{F}_i, Y)$ is an unbiased estimate of $u(\hat{F}_i, Y)$, as stated in the following theorem.

115 **Theorem 3.2.** Assume the conditions of theorem 3.1. Then defining

$$V = (k-1)\hat{u}(\hat{F}_i, y), \quad (3)$$

116 we have

$$V \sim \text{Binomial}(k-1, u(\hat{F}_i, y)).$$

117 Hence,

$$\mathbf{E}[\hat{u}(\hat{F}_i, y)] = u(\hat{F}_i, y).$$

118

119 In section 4, we will use this result to estimate the moments of U . Meanwhile, since U is a random
 120 variable on $[0, 1]$, we also conclude that p_t follows a *mixed exponential decay*. Let α be the law of
 121 $-\log(U)$. Then from change-of-variables $\kappa = -\log(u)$, we get

$$\mathbf{E}[\text{acc}^{(t)}] = \mathbf{E}[U^{t-1}] = \int_0^1 u^{t-1} d\nu(u) = \int_0^1 e^{t \log(u)} \frac{1}{u} d\nu(u) = \int_{\mathbb{R}^+} e^{-\kappa t} d\alpha(\kappa).$$

122 This fact immediately suggests the technique of fitting an mixture of exponentials to the test error at
 123 $t = 2, 3, \dots, k$: we explore this idea further in Section 4.1.

124 3.1 Properties of the conditional accuracy distribution

125 The conditional error distribution ν is determined by \mathbb{F} and \mathcal{Q} . What can we say about the the
 126 conditional accuracy distribution without making any assumptions on either \mathbb{F} or \mathcal{Q} ? The answer
 127 is: not much—for an arbitrary probability measure ν' on $[0, 1]$, one can construct \mathbb{F} and \mathcal{Q} such that
 128 $\nu = \nu'$, even if one makes the *perfect sampling assumption* that $\hat{F} = F$.

Theorem 3.3. Let U be defined as in Theorem 3.1, and let ν denote the law of U . Then, for any probability distribution ν' on $[0, 1]$, one can construct a meta-distribution \mathbb{F} and a scoring rule \mathcal{Q} such that $\nu = \nu'$ under perfect sampling (that is, $\hat{F} = F$.)

Proof. Let G be the cdf of ν , $G(x) = \int_0^x d\nu(x)$, and let $H(u) = \sup_x \{G(x) \leq u\}$. Define \mathcal{Q} by

$$\mathcal{Q}(\hat{F}, y) = \begin{cases} 0 & \text{if } \mu(\hat{F}) > y + H(y) \\ 0 & \text{if } y + H(y) > 1 \text{ and } \mu(\hat{F}) \in [H(y) - y, y] \\ 1 + \mu(\hat{F}) - y & \text{if } \mu(\hat{F}) \in [y, y + H(y)] \\ 1 + y + \mu(\hat{F}) & \text{if } \mu(\hat{F}) + H(y) > 1 \text{ and } \mu(\hat{F}) \in [0, H(y) - y]. \end{cases}$$

Let $\theta \sim \text{Uniform}[0, 1]$, and define $F \sim \mathbb{F}$ by $F = \delta_\theta$, and also $\hat{F} = F$. A straightforward calculation yields that $\nu = \nu'$. \square

On the other hand, we can obtain a positive result if we assume that the classifier approximates a *Bayes classifier*. Assuming that F is absolutely continuous with respect to Lebesgue measure Λ with probability one, a Bayes classifier results from assuming perfect sampling ($\hat{F} = F$) and taking $\mathcal{Q}(\hat{F}, y) = \frac{dF}{d\Lambda}(y)$. Theorem 3.4. states that for a Bayes classifier, ν has a density $\eta(u)$ which is monotonically increasing. Since a ‘good’ classifier approximates the Bayes classifier, we intuitively expect that a monotonically increasing density η is a good model for the conditional accuracy distribution of a ‘good’ classifier.

Theorem 3.4. Assume the conditions of theorem 3.1, and further suppose that $\hat{F} = F$, F is absolutely continuous with respect to Λ with probability one, that $\mathcal{Q}(\hat{F}, y) = \frac{dF}{d\Lambda}(y)$, and that $F|Y$ has a regular conditional probability distribution. Let ν denote the law of U . Then ν has a density $\eta(u)$ on $[0, 1]$ which is monotonic in u .

Proof. It suffices to prove that

$$\nu([u, u + \delta]) < \nu([v, v + \delta])$$

for all $0 < u < v < 1$ and $0 < \delta < 1 - v$. Let $\mathcal{P}_{ac}(\mathcal{Y})$ denote the space of distributions supported on \mathcal{Y} which are absolutely continuous with respect to p -dimensional Lebesgue measure Λ . Let \mathbb{Y} denote the marginal distribution of Y for $Y \sim F$ with $F \sim \mathbb{F}$. Define the set

$$J_y(A) = \{F \in \mathcal{P}_{ac}(\mathcal{Y}) : u(F, y) \in A\}.$$

for all $A \subset [0, 1]$. One can verify that for all $y \in \mathcal{Y}$,

$$\Pr_{\mathbb{F}}[J_y([u, u + \delta])|Y = y] \leq \Pr_{\mathbb{F}}[J_y([v, v + \delta])|Y = y],$$

using the fact that \mathbb{F} has no atoms. Hence, we obtain

$$\Pr[U \in [u - \delta, u + \delta]] = \mathbf{E}_{\mathbb{Y}}[\Pr[J_Y([u, u + \delta])|Y]] \leq \mathbf{E}_{\mathbb{Y}}[\Pr[J_Y([v, v + \delta])|Y]] = \Pr[U \in [v - \delta, v + \delta]].$$

Taking $\delta \rightarrow 0$, we conclude the theorem. \square

4 Estimation

Suppose we have m independent test repeats per class, $y^{(i),1}, \dots, y^{(i),m}$. Let us define

$$V_{i,j} = \sum_{\ell \neq i} I(\mathcal{M}_i(\hat{F}_1, \dots, \hat{F}_k, y^{(i,j)}) > \mathcal{M}_\ell(\hat{F}_1, \dots, \hat{F}_k, y^{(i,j)})),$$

which coincides with the definition (3) in the special case that \mathcal{M} is generative.

At a high level, we have a hierarchical model where U is drawn from a distribution ν on $[0, 1]$ and then $V_{i,j} \sim \text{Binomial}(k, U)$. Let us assume that U has a density $\eta(u)$: then the marginal distribution of $V_{i,j}$ can be written

$$\Pr[V_{i,j} = \ell] = \binom{k}{\ell} \int_0^1 u^\ell (1 - u)^{k-\ell} \eta(u) du.$$

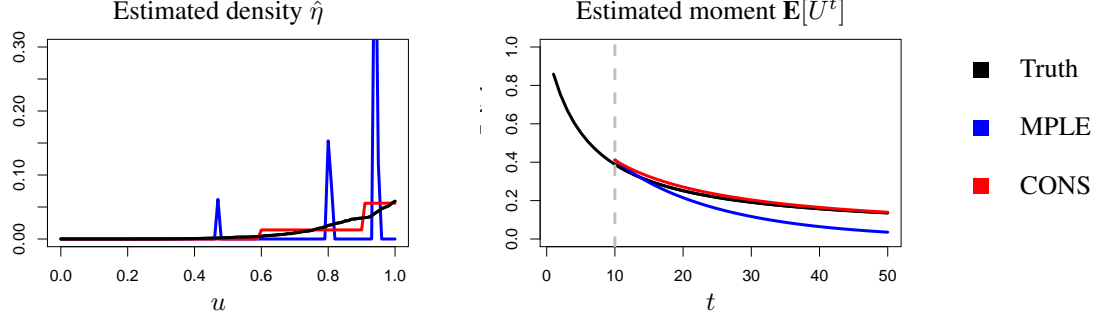


Figure 1: Maximum pseudolikelihood (MPLE) versus constrained pseudolikelihood (CONS).

160 However, the observed $\{V_{i,j}\}$ do *not* comprise an i.i.d. sample.

161 We discuss the following three approaches for estimating $p_t = \mathbf{E}[U^{t-1}]$ based on $V_{i,j}$. The first is
 162 an extension of *unbiased estimation* based on binomial U-statistics, which is discussed in Section
 163 4.1. The second is the *pseudolikelihood* approach. In problems where the marginal distributions
 164 are known, but the dependence structure between variables is unknown, the *pseudolikelihood* is
 165 defined as the product of the marginal distributions. For certain problems in time series analysis
 166 and spatial statistics, the maximum pseudolikelihood estimator (MPLE) is proved to be consistent
 167 (Arnold 1991). We discuss pseudolikelihood-based approaches in Section 4.2. Thirdly, we note that
 168 the high-dimensional theory of Anon 2006 can be applied for prediction accuracy, which we discuss
 169 in Section 4.3.

170 4.1 Extensions of unbiased estimation

171 If $V \sim \text{Binomial}(k, U)$, then an unbiased estimator of U^t exists if and only if $0 \leq t \leq k$.

172 The theory of U-statistics provides the minimal variance unbiased estimator for U^t :

$$U^t = \mathbf{E} \left[\binom{V}{t} \binom{k}{t}^{-1} \right].$$

173 This result can be immediately applied to yield an unbiased estimator of p_t , when $t \leq k$:

$$\hat{p}_t^{UN} = \frac{1}{km} \sum_{i=1}^k \sum_{j=1}^m \binom{V_{i,j}}{t-1} \binom{k}{t-1}^{-1}. \quad (4)$$

174 However, since \hat{p}_t^{UN} is undefined for $k \geq t$, we can use exponential extrapolation to define an
 175 extended estimator \hat{p}_t^{EXP} for $k > t$. Let $\hat{\alpha}$ be a measure defined by solving the optimization problem

$$\text{minimize} \sum_{t=2}^k \left(\hat{p}_t^{UN} - \int_0^\infty \exp[-t\kappa] d\alpha(\kappa) \right)^2$$

176 After discretizing the measure $\hat{\alpha}$, we obtain a convex optimization problem which can be solved
 177 using non-negative least squares (Lawson 1974). Then define

$$\hat{p}_t^{EXP} = \begin{cases} \hat{p}_t^{UN} & \text{for } t \leq k, \\ \int_0^\infty \exp[-t\kappa] d\hat{\alpha}(\kappa) & \text{for } t > k. \end{cases}$$

178 4.2 Maximum pseudolikelihood

179 The pseudolikelihood is defined as

$$\ell(\eta) = \sum_{i=1}^k \sum_{j=1}^m \log \left(\int u^{V_{i,j}} (1-u)^{k-V_{i,j}} \eta(u) du \right), \quad (5)$$

180 and a maximum pseudolikelihood estimator (MPLE) is defined as any density $\hat{\eta}$ such that

$$\ell(\hat{\eta}_{MPLE}) = \sup_{\eta} \ell(\eta).$$

181 The motivation for $\hat{\eta}_{MPLE}$ is that it consistently estimates η in the limit where $k \rightarrow \infty$. However, in
182 finite samples, $\hat{\eta}_{MPLE}$ is not uniquely defined, and if we define the plug-in estimator

$$\hat{p}_t^{MPLE} = \int u^{t-1} \hat{\eta}_{MPLE}(u) du,$$

183 \hat{p}_t^{MPLE} can vary over a large range, depending on which $\hat{\eta} \in \operatorname{argmax}_{\eta} \ell_t(\eta)$ is selected. These
184 shortcomings motivate the adoption of additional constraints on the estimator $\hat{\eta}$.

185 Theorem 3.4. motivates the *monotonicity constraint* that $\frac{d\hat{\eta}}{du} > 0$. A second constraint is to restrict
186 the k th moment of $\hat{\eta}$ to match the unbiased estimate. The addition of these constraints yields the
187 constrained PMLE $\hat{\eta}_{CON}$, which is obtained by solving

$$\text{maximize } \ell(\eta) \text{ subject to } \int u^{k-1} \eta(u) du = \hat{p}_k^{UN} \text{ and } \frac{d\hat{\eta}}{du} > 0.$$

188 By discretizing η , all of the above maximization problems can be solved using a general-purpose
189 convex solver³. As seen in Figure 1, the added constraints can improve estimation of η and thus
190 improve moment estimation.

191 4.3 High-dimensional asymptotics

192 Under a number of conditions on the distribution \mathbb{F} , including (but not limited to) having a large
193 dimension p , Anon 2016 relate the accuracy p_t of the Bayes classifier to the mutual information
194 between the label z and the response y :

$$p_t = \bar{\pi}_t(\sqrt{2I(Z; Y)}).$$

195 where

$$\bar{\pi}_k(c) = \int_{\mathbb{R}} \phi(z - c) \Phi(z)^{k-1} dz.$$

196 While our goal is not to estimate the mutual information, we note that the results of Anon 2016 imply
197 a relationship between p_k and p_K for the Bayes error under the high-dimensional regime:

$$p_K = \bar{\pi}_K(\bar{\pi}_k^{-1}(p_k)).$$

198 Therefore, under the high-dimensional conditions of Anon 2016 and assuming that the classifier
199 approximates the Bayes classifier, we naturally obtain the following estimator

$$\hat{p}_t^{HD} = \bar{\pi}_K(\bar{\pi}_k^{-1}(\hat{p}_k^{UN})).$$

200 5 Results

201 We applied the methods described in Section 4 to predict the 400-class accuracy of multinomial
202 logistic regression, SVM, and ϵ -nearest neighbors⁴ on a Telegu character classification task, using
203 20-class data with 100 examples per class. The results are displayed in Figure 2.

204 Taking the test accuracy on 400 classes (using 50 test examples per class) as a proxy for $\text{acc}^{(400)}$,
205 we compare the performance of the three extrapolation methods. The exponential extrapolation
206 method makes use of the fewest theoretical assumptions, but performs badly on all three problems.
207 Meanwhile, constrained PMLE makes an extra assumption in the monotonicity of $\eta(u)$, which is true
208 if the classifier is sufficiently close to the Bayes classifier, and achieves the best and most consistent
209 results. The high-dimensional estimator \hat{p}^{HD} is the most assumption-heavy; in addition to assuming
210 approximation to the Bayes classifier, it also requires Y to be high-dimensional, and to satisfy a

³ We found that the CVX discipline convex programming language, using the ECOS second-order cone programming solver, succeeds in optimizing the problems where the dimension of the discretized η is as large as 10,000 (Hong 2014, Domahidi 2013.)

⁴ k -nearest neighbors with $k = \epsilon n$ for fixed $\epsilon > 0$

Classifier	Test acc ⁽²⁰⁾	Test acc ⁽⁴⁰⁰⁾	\hat{p}_{400}^{EXP}	\hat{p}_{400}^{CON}	\hat{p}_{400}^{HD}
Logistic	0.922	0.711	0.844	0.721	0.686
SVM	0.860	0.545	0.737	0.575	0.546
ϵ -NN	0.880	0.591	0.903	0.608	0.839

Figure 2: Performance extrapolation: predicting the error on 400 classes using data from 20 classes on a Telugu character dataset. $\epsilon = 0.002$ for ϵ -nearest neighbors.

number of other technical conditions (Anon 2016). Nevertheless, it performs well on the multinomial logistic and SVM classifiers. That said, all three classifiers studied are non-generative, hence violating an assumption common to both the \hat{p}^{EXP} and \hat{p}^{CON} estimators: therefore, it is doubtful if the theory developed so far can do much to explain the relative performance of these methods.

While \hat{p}^{HD} does not explicitly require the classifier to be generative, the assumption that the classifier approximates the Bayes rule is arguably a stronger assumption. It is notably, however, that \hat{p}^{HD} performs best for multinomial logistic regression, which has the highest accuracy and therefore might be the best approximator of the Bayes classifier among the three classifiers studied.

6 Discussion

We have developed a theory of prediction extrapolation for generative classifiers, under the assumption of exchangeable classes. The equivalence between the expected t -class accuracy and the $t - 1$ th moment of the conditional accuracy distribution allows a variety of methods to be applied to the problem.

Empirical results indicate that our methods generalize beyond generative classifiers. A possible explanation is that many non-generative classifiers are *asymptotically generative*, meaning that $\mathcal{M}_i^{(t)}$, the i th margin for the t th classification problem in the sequence, converges to a function of \hat{F}_i and y with probability one:

$$\lim_{t \rightarrow \infty} \mathcal{M}_i^{(t)}(\hat{F}_1, \dots, \hat{F}_t, y) = \mathcal{Q}(\hat{F}_i, y) \text{ w.h.p.} \quad (6)$$

While asymptotically generative classifiers can be decomposed in a similar manner as generative classifiers, they need not share the same practical *limitations* as generative classifiers. A generative classifier requires the user to specify the scoring function \mathcal{Q} in advance: this practically requires some prior knowledge about the marginal distributions of the classes, e.g. that the marginal distributions are approximately multivariate Gaussian. But the scoring function \mathcal{Q} appearing in the definition of an asymptotically generative classifier *need not be specified* by the user: it can (and usually does) depend on the unknown meta-distribution \mathbb{F} . One can show that one-vs-one, one-vs-all, and ϵ -nearest neighbors satisfy the definition (6) after assuming some additional continuity conditions; however, the definition (6) is too weak to extend the result in theorem 3.1. Finding appropriate conditions to obtain a useful definition of an *asymptotically generative* classifier is the subject of current work.

The assumption of exchangeability greatly limits the scope of application for our methods. Many multi-class classification problems have a hierarchical structure (Deng 2010), or have classes distributed according to non-uniform discrete distributions (e.g. power laws) (Garfield 2005); in either case, exchangeability is violated. It would be interesting to extend our theory to the hierarchical setting, or to handle non-hierarchical settings with non-uniform prior class probabilities, but again we leave the subject for future work.

References

- [X] Anonymous, A. "Estimating mutual information in high dimensions via classification error." Submitted to NIPS 2016.
- [X] Ng, Andrew Y., and Michael I. Jordan. "On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes." (2002).
- [X] Arnold, Barry C., and David Strauss. "Pseudolikelihood estimation: some examples." Sankhya: The Indian Journal of Statistics, Series B (1991): 233-243.

- 251 [X] Lawson, Charles L., and Richard J. Hanson. Solving least squares problems. Vol. 161. Englewood Cliffs,
252 NJ: Prentice-hall, 1974.
- 253 [X] Hong, Jenny, Karanveer Mohan, and David Zeng. "CVX. jl: A Convex Modeling Environment in Julia."
254 (2014).
- 255 [X] Domahidi, Alexander, Eric Chu, and Stephen Boyd. "ECOS: An SOCP solver for embedded systems."
256 Control Conference (ECC), 2013 European. IEEE, 2013.
- 257 [X] Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI.
258 *Neuroimage*, 56(2), 400-410.
- 259 [X] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1.
260 Springer, Berlin: Springer series in statistics, 2008.
- 261 [X] Deng, J., Berg, A. C., Li, K., & Fei-Fei, L. (2010). What does classifying more than 10,000 image categories
262 tell us? *Lecture Notes in Computer Science*, 6315 LNCS(PART 5), 71?84.
- 263 [X] Garfield, Sheila, Stefan Wermter, and Siobhan Devlin. "Spoken language classification using hybrid classifier
264 combination." *International Journal of Hybrid Intelligent Systems* 2.1 (2005): 13-33.