Asymptotic Properties of the Fisher Kernel*

Koji Tsuda[†+], Shotaro Akaho[&],
Motoaki Kawanabe[*] and Klaus-Robert Müller[*‡]

[†]Max Planck Institute for Biological Cybernetics
Spemannstr. 38, 72076 Tübingen, Germany
[+]AIST Computational Biology Research Center
2-43, Aomi, Koto-ku, Tokyo, 135-0064, Japan
[&]AIST Neuroscience Research Institute
1-1-4, Umezono, Tsukuba, 305-8568, Japan
[*]Fraunhofer FIRST
Kekuléstr. 7, 12489 Berlin, Germany
[‡]University of Potsdam
August-Bebelstr. 89, 14482 Potsdam, Germany
koji.tsuda@tuebingen.mpg.de, s.akaho@aist.go.jp
{nabe,klaus}@first.fhg.de

June 10, 2003

Abstract

This paper analyses the Fisher kernel from a statistical point of view. The Fisher kernel is a particularly interesting method for constructing a model of the posterior probability that makes intelligent use of unlabeled data, i.e. of the underlying data density. It is important to analyse and ultimately understand the statistical properties of the Fisher kernel. To this end, we first establish sufficient conditions that the constructed posterior model is realizable, i.e. that it contains the true distribution. Realizability then immediately leads to consistency results. Subsequently we focus on an asymptotic analysis of the generalization error, which elucidates the learning curves of the Fisher kernel and how unlabeled data contribute to learning. We also point out that the squared or log loss is theoretically more preferable—as they yield consistent estimators—than other losses such as the exponential loss, when a linear classifier is used together with the Fisher kernel. Therefore this paper underlines that the Fisher kernel should not be viewed as a heuristics but as a powerful statistical tool with well controlled statistical properties.

[*]To appear in *Neural Computation*

1 Introduction

Recently, the Fisher kernel (Jaakkola & Haussler, 1999) has been successfully applied as a feature extractor in supervised classification (Jaakkola & Haussler, 1999; Tsuda et al., 2002; Sonnenburg et al., 2002; Smith & Gales, 2002; Vinokourov & Girolami, 2002). The original intuition (Jaakkola & Haussler, 1999) for the Fisher kernel was to construct a probabilistic model of the data in order to induce a metric for a subsequent discriminative training. Two problems could be addressed simultaneously: (1) it became possible to compare "apples and oranges"—as the Fisher kernel approach measures distances in the space of the respective probabilistic model parameters. So for example a DNA sequence of length, say, 100 and another one of length 1000 can be easily compared by using a representation in the respective HMM parameter space. Thus, the Fisher kernel is very much in contrast to alignment methods that compare directly by essentially using dynamic programming techniques (e.g., Gotoh, 1982). (2) A further feature of the Fisher kernel is that it allows to incorporate prior knowledge about the data distribution into the classification process in a highly principled manner.

In the practical use of support vector machines (SVM) (e.g., Vapnik, 1998; Cristianini & Shawe-Taylor, 2000; Müller et al., 2001; Schölkopf & Smola, 2002), where the choice of the kernel is of crucial importance, either the kernel can be engineered using all available prior knowledge (e.g., Zien et al., 2000) or it can be derived as the Fisher kernel from a probabilistic model (e.g., Jaakkola & Haussler, 1999; Tsuda et al., 2002; Sonnenburg et al., 2002; Smith & Gales, 2002).[1] In spite of its practical success, a theoretical analysis of the Fisher kernel has not been sufficiently explored so far, with the current exceptions being e.g., Jaakkola et al. (1999); Tsuda & Kawanabe (2002); Seeger (2002); Tsuda et al. (2003). For example, Jaakkola et al. (1999) showed how to determine the prior distribution of parameters to recover the Fisher kernel in the framework of maximum entropy discrimination. Also Seeger (2002) pointed out that the Fisher kernel can be perceived as an approximation of the mutual information kernel.

This paper will aim to present theoretical results from a statistical point of view. In particular we perceive the Fisher kernel as a method of constructing a model of the *posterior* probability of the class labels.

The Fisher kernel can be derived as follows: Let $\mathcal{X}$ denote the domain of objects, which can be discrete or continuous. Also let us assume that a probabilistic model $q(x|\boldsymbol{\theta}),\ x \in \mathcal{X}, \boldsymbol{\theta} \in \Re^d$ is available. Given a parameter estimate $\hat{\boldsymbol{\theta}}$ from training samples, the feature vector (i.e. the Fisher score) is obtained as

$$\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(x) = \left(\frac{\partial \log q(x|\hat{\boldsymbol{\theta}})}{\partial \theta_1}, \ldots, \frac{\partial \log q(x|\hat{\boldsymbol{\theta}})}{\partial \theta_d}\right)^\top. \tag{1.1}$$

The Fisher kernel refers to the inner product in this space. When used in

[1] Of course also brute force search over all possible kernels can be pursued using cross-validation procedures or bounds from learning theory to finally select the 'best' kernel (cf. Müller et al., 2001).

supervised classification, the Fisher kernel is commonly combined with a linear classifier such as support vector machines (SVMs) (Vapnik, 1998), where a linear function is trained to discriminate two classes. Since the Fisher kernel can efficiently make use of prior knowledge about the marginal distribution $p(x)$ (which can be estimated rather well using *unlabeled samples*), it is especially attractive in vision, text classification and bioinformatics where we can expect a lot of unlabeled samples (Zhang & Oles, 2000; Seeger, 2001).

As the first analysis, we will show the sufficient conditions that the obtained posterior model is *realizable*, i.e. that it contains the true posterior distribution, which then immediately leads to consistency. Once realizability is assured, we can evaluate the expected generalization error in large sample situations by means of asymptotic statistics (Barndorff-Nielsen & Cox, 1989). This enables us to elucidate learning curves and how *unlabeled samples* contribute in reducing the generalization error. In addition, it is pointed out that, when a linear classifier is combined with the Fisher kernel, then the log loss and the squared loss are theoretically more preferable than other loss functions. This result recommends us to use a classifier based on the log loss or the squared loss.

2 Realizablity Conditions

Let $y \in \{+1, -1\}$ be the set of class labels. Denote by $p(x)$, $P(y|x)$ and $p(x, y)$ the true underlying marginal, posterior and joint distributions, respectively. Let $\partial_\alpha f = \partial f / \partial \alpha$, $\nabla_{\boldsymbol{\theta}} f = (\partial_{\theta_1} f, \ldots, \partial_{\theta_d} f)^\top$, and $\nabla^2_{\boldsymbol{\theta}} f$ denote the $d \times d$ matrix, the Hessian, whose (i, j)-th element is $\partial^2 f / (\partial \theta_i \partial \theta_j)$.

For statistical learning, we construct a model of posterior probability $P(y|x)$ out of the Fisher score (1.1). The posterior probability is described via a linear function followed by an activation function h:

$$Q(y|x, \boldsymbol{\eta}) = h(y[\boldsymbol{w}^\top \boldsymbol{f}_{\boldsymbol{\theta}}(x) + b]), \tag{2.1}$$

where $\boldsymbol{w} \in \Re^d$, $b \in \Re$, parameters are summarized as $\boldsymbol{\eta} = (\boldsymbol{w}^\top, b, \boldsymbol{\theta}^\top)^\top$, and h is a linear activation function[2]

$$h(t) = \frac{1}{2} t + \frac{1}{2}. \tag{2.2}$$

In the following, we will investigate the conditions that $Q(y|x, \boldsymbol{\eta})$ is realizable, i.e., there is a parameter value $\boldsymbol{\eta}^*$ such that $Q(y|x, \boldsymbol{\eta}^*) = P(y|x)$.

2.1 Core Model

First, a trivial example is shown to give a realizable model. Denote by $q_0(x|\boldsymbol{\theta})$ a mixture model of the true class distributions:

$$q_0(x|\alpha) = \alpha p(x|y = +1) + (1 - \alpha) p(x|y = -1), \quad \alpha \in [0, 1], \tag{2.3}$$

[2]Compared with sigmoid functions, the linear activation function is not so common in literature. However it allows us to perform statistical analysis as will shown later.

which we call the *core model*. Obviously this model realizes the true marginal distribution $p(x)$, when $\alpha = p(y = +1) := \alpha^*$.

Lemma 1. *When the Fisher score is determined as* $f_{\alpha^*}(x) = \partial_\alpha \log q_0(x|\alpha^*)$, *the posterior model (2.1) is realizable.*

(proof) The posterior model $Q(y|x, \boldsymbol{\eta})$ is realizable, if there is a parameter value $\boldsymbol{\eta}^*$ such that

$$Q(y|x, \boldsymbol{\eta}^*) = P(y|x), \ \ \forall x \in \mathcal{X}, \ y \in \{+1, -1\}. \tag{2.4}$$

Substituting (2.1) and (2.2), (2.4) holds if and only if

$$(\boldsymbol{w}^*)^\top \boldsymbol{f}_{\boldsymbol{\theta}^*}(x) + b^* = P(y = +1|x) - P(y = -1|x). \tag{2.5}$$

To prove the lemma for q_0, it is sufficient to show the existence of $w, b \in \Re$ such that

$$w\partial_\alpha \log q_0(x|\alpha^*) + b = P(y = +1|x) - P(y = -1|x). \tag{2.6}$$

The Fisher score for $q_0(x|\alpha^*)$ can be written as

$$\partial_\alpha \log q_0(x|\alpha^*) = \frac{P(y = +1|x)}{\alpha^*} - \frac{P(y = -1|x)}{1 - \alpha^*}.$$

When $w = 2\alpha^*(1 - \alpha^*)$ and $b = 2\alpha^* - 1$, (2.6) holds. □

2.2 Deriving Realizability Conditions

Since we do not know the true class distributions $p(x|y)$, the core model $q_0(x|\alpha)$ in Lemma 1 is never available. In the following, the result of Lemma 1 is therefore relaxed to a more general class of probability models.

Denote by $\mathcal{M}$ a set of probability distributions $\mathcal{M} = \{q_0 \mid q_0(x|\alpha), \alpha \in [0, 1]\}$. According to the information geometry (Amari & Nagaoka, 2001), $\mathcal{M}$ is regarded as a manifold in a Riemannian space. Let $\mathcal{Q}$ denote the manifold of $q(x|\boldsymbol{\theta})$: $\mathcal{Q} = \{q \mid q(x|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Re^d\}$. Now the question is how to determine a manifold $\mathcal{Q}$ such that (2.1) is realizable, which is answered by the following theorem.

Theorem 1. *Assume that the true distribution* $p(x)$ *is contained in* $\mathcal{Q}$*:*

$$p(x) = q(x|\boldsymbol{\theta}^*) = q_0(x|\alpha^*), \quad x \in \mathcal{X},$$

where $\boldsymbol{\theta}^*$ *is the true parameter. If the tangent space of* $\mathcal{Q}$ *at* $p(x)$ *contains the tangent space of* $\mathcal{M}$ *at the same point (Figure 1), then the Fisher score* $\boldsymbol{f}$ *derived from* $q(x|\boldsymbol{\theta}^*)$ *gives a realizable posterior model (2.1).*

(proof) To prove the theorem, it is sufficient to show the existence of $\boldsymbol{w} \in \Re^d$ and $b \in \Re$ such that

$$\boldsymbol{w}^\top \nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta}^*) + b = P(y = 1|x) - P(y = -1|x). \tag{2.7}$$

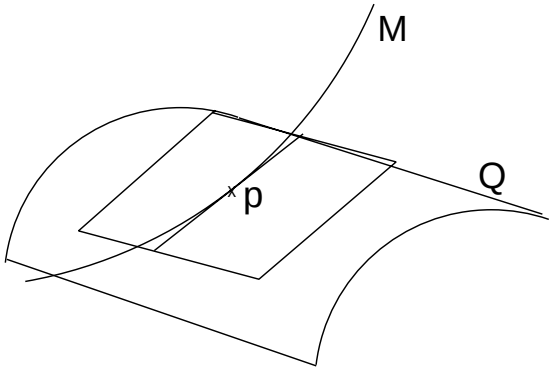

Figure 1: Information geometric picture of a probabilistic model whose Fisher kernel leads to a realizable posterior model. The important point is that the tangent space of manifold $\mathcal{M}$ is contained in that of manifold $\mathcal{Q}$. Details explained in the text.

When the tangent space of $\mathcal{M}$ is contained in that of $\mathcal{Q}$ around $p(x)$, we have the following by the chain rule:

$$\frac{\partial \log q_0(x|\alpha^*)}{\partial \alpha} = \sum_{j=1}^{d} \frac{\partial \log q(x|\boldsymbol{\theta}^*)}{\partial \theta_j} \left. \frac{\partial \theta_j}{\partial \alpha} \right|_{\alpha=\alpha^*}. \tag{2.8}$$

Let $\boldsymbol{u}$ be the d dimensional vector where the i-th element is

$$u_i = \left. \frac{\partial \theta_j}{\partial \alpha} \right|_{\alpha=\alpha^*}.$$

Then (2.7) holds when

$$\boldsymbol{w} = 2\alpha^*(1-\alpha^*)\boldsymbol{u}, \qquad b = 2\alpha^* - 1.$$

□

This theorem indicates that realizability depends on local geometry of manifold $\mathcal{Q}$ around the true distribution. In order to have a good posterior model, we have to assure realizability while keeping the number of parameters small. To this aim, we should make the manifold $\mathcal{Q}$ as low dimensional as possible, while capturing the tangent space of $\mathcal{M}$. If $\mathcal{Q}$ completely contains $\mathcal{M}$, the realizability condition is satisfied. One example of this case was shown by Tsuda et al. (2003), where each class distribution is the mixture of shared Gaussian components.

Remark A classifier is called *Bayes optimal*, if it achieves the Bayes error in the limit that the number of samples goes to infinity (Devroye et al., 1996). Realizability is only a sufficient condition for Bayes optimality. It would be an interesting research topic to derive the conditions for Bayes optimality as well.

3 Consistency Results

Denote by $\boldsymbol{z}_n = \{x_i, y_i\}_{i=1}^n$ the set of n i.i.d. labeled samples derived from $p(x, y)$. Denote by $\boldsymbol{x}_m^u = \{x_j^u\}_{i=1}^m$ the set of m unlabeled i.i.d. samples derived from $p(x)$. In learning with the Fisher kernel from these samples, the learning procedure is typically separated into two steps (e.g., Jaakkola & Haussler, 1999). First, $\boldsymbol{\theta}$ is obtained as

$$\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^n \log q(x_i|\boldsymbol{\theta}) + \sum_{j=1}^m \log q(x_j^u|\boldsymbol{\theta}). \tag{3.1}$$

Then, in the second step, $\boldsymbol{w}$ and b are obtained as

$$[\hat{\boldsymbol{w}}, \hat{b}] = \mathrm{argmax}_{\boldsymbol{w},b} \sum_{i=1}^n \log h(y_i[\boldsymbol{w}^\top \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(x_i) + b]). \tag{3.2}$$

The second step maximizes the conditional likelihood $Q(y|x, \boldsymbol{\eta})$. Let $\ell(y, y')$ denote a loss function. Then (3.2) is generalized as follows:

$$[\hat{\boldsymbol{w}}, \hat{b}] = \mathrm{argmin}_{\boldsymbol{w},b} \sum_{i=1}^n \ell(\boldsymbol{w}^\top \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(x_i) + b, y_i), \tag{3.3}$$

where the loss function in (3.2) corresponds to

$$\ell(y, y') = -\log h(yy'). \tag{3.4}$$

First we prove that the consistency is assured for the log loss (3.4), that is, in the limit that n goes to infinity, the estimator $\hat{\boldsymbol{\eta}}$ converges to the true one $\boldsymbol{\eta}^*$. Further it will be shown that the consistency can be proved for the squared loss, which has advantages from the practical viewpoint.

Lemma 2. *Assume that $q(x|\boldsymbol{\theta})$ satisfies the realizablity conditions in Theorem 1. The two step estimator with the log loss (3.4) is consistent.*

(Proof) In the two step scheme, $\boldsymbol{\theta}$ is estimated separately as maximum likelihood (3.1), so obviously $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}^*$. When we have infinite samples, (3.3) is written as

$$[\boldsymbol{w}^+, b^+] = \mathrm{argmin}_{\boldsymbol{w},b} \sum_{j\in\{1,-1\}} P(y=j) \int \ell(\boldsymbol{w}^\top \boldsymbol{f}_{\boldsymbol{\theta}^*}(x) + b, j) p(x|y=j) dx.$$

Therefore, we should prove $\boldsymbol{w}^+ = \boldsymbol{w}^*$ and $b^+ = b^*$. In other words, this problem is rewritten as a constrained variation problem (Gelfand & Fomin, 1963), where we find a function $g : \mathcal{X} \to \Re$ that minimizes the following functional

$$L(g) = \sum_{j\in\{1,-1\}} P(y=j) \int \ell(g(x), j) p(x|y=j) dx, \tag{3.5}$$

subject to the constraint $g \in \mathcal{G}$ where

$$\mathcal{G} = \{g \mid g(x) = \boldsymbol{w}^\top \boldsymbol{f}_{\boldsymbol{\theta}^*}(x) + b,\ \boldsymbol{w} \in \Re^d, b \in \Re\}. \tag{3.6}$$

If the optimal solution of the variation problem without the above constraint is eventually contained in $\mathcal{G}$, it is the solution of the constrained problem (3.5) as well. So let us consider the unconstrained problem first: When the log loss is substituted into (3.5), we have

$$L(g) = -\sum_{j\in\{1,-1\}} P(y=j) \int \log\left\{\frac{jg(x)+1}{2}\right\} p(x|y=j)dx.$$

The variation of L with respect to small increment of g is written as

$$\delta L = -\sum_{j\in\{1,-1\}} P(y=j) \int \delta g \, \frac{j}{jg(x)+1}\, p(x|y=j)dx.$$

In order that g is an minimum, it is necessary that $\delta L = 0$ holds for any δg, thus we have

$$\frac{1}{1+g(x)}p(x,y=1) - \frac{1}{1-g(x)}p(x,y=-1) = 0, \tag{3.7}$$

$$\Leftrightarrow \quad \hat{g}(x) = P(y=1|x) - P(y=-1|x). \tag{3.8}$$

Since realizability is assured by assumption, $Q(y|x,\boldsymbol{\eta}^*) = P(y|x)$. According to (2.1) and (2.2), it holds that

$$(\boldsymbol{w}^*)^\top \boldsymbol{f}_{\boldsymbol{\theta}^*}(x) + b^* = P(y=+1|x) - P(y=-1|x).$$

Thus $\hat{g}$ is contained in $\mathcal{G}$, and the true parameters are obtained by solving the constrained problem (3.5). □

Note that Lemma 2 holds even if $m = 0$. Here $\hat{\boldsymbol{w}}$ and $\hat{b}$ cannot be obtained in closed form for the log loss. This turns out to be possible for the squared loss

$$\ell(y,y') = (y-y')^2, \tag{3.9}$$

and as we will see in the following, the consistency is assured as well in this case.

Lemma 3. *The two step estimator with the squared loss (3.9) is consistent.*

(proof) When the squared loss is substituted into (3.5), we have

$$L(g) = \sum_{j\in\{1,-1\}} P(y=j) \int (g(x)-j)^2 p(x|y=j)dx. \tag{3.10}$$

In this case, the variational equation (3.7) turns out that

$$g(x)p(x) - \sum_{j\in\{1,-1\}} jp(x,y=j) = 0,$$

which is solved as

$$\hat{g}(x) = \frac{p(x, y = +1) - p(x, y = -1)}{p(x)} = P(y = +1|x) - P(y = -1|x). \quad (3.11)$$

Since this solution is the same as (3.8), this lemma is proved by following the same procedure as Lemma 2. □

Interestingly, one *cannot* assure consistency for general loss functions. For example, when we use the exponential loss

$$\ell(y, y') = \exp(-\frac{1}{2}yy')$$

or the logistic loss

$$\ell(y, y') = \log(1 + \exp(-yy')),$$

the unconstrained variational solution is obtained as follows (Eguchi & Copas, 2001):

$$\hat{g}(x) = \log P(y = +1|x) - \log P(y = -1|x).$$

Since this solution may not be included in $\mathcal{G}$, such losses do not necessarily achieve consistency.[3] The squared loss is the appropriate choice for the Fisher kernel because from the theoretical viewpoint it achieves consistency and because from the practical point of view the solution can be obtained analytically in closed form.

4 Generalization Errors

In this section, the generalization error of Fisher kernel classifiers is investigated. Specifically, we will study the behavior of the generalization error when the number of training samples is sufficiently large. Such an analysis is often called "learning curve analysis", where a learning curve describes the relation of the generalization error against the number of training samples (e.g., Baum & Haussler, 1989; Amari & Murata, 1993; Müller et al., 1996; Haussler et al., 1996; Malzahn & Opper, 2002). The studies about learning curves have been playing an important role in elucidating the behavior of learning machines. For studying generalization errors apart from bounds derived in a statistical learning theory framework (e.g., Devroye et al., 1996; Vapnik, 1998), researches in asymptotic statistics (e.g., Cox & Hinkley, 1974; Barndorff-Nielsen & Cox, 1989; Amari & Murata, 1993; Müller et al., 1996; van der Vaart, 1998; Amari & Nagaoka, 2001) and statistical mechanics approaches (e.g., Seung et al., 1992; Watkin et al., 1993; Haussler et al., 1996; Malzahn & Opper, 2002) have contributed. In this section, we will adopt asymptotic statistical techniques following (Barndorff-Nielsen & Cox, 1989).

The generalization error is defined as $R(\boldsymbol{\eta}) := \mathrm{E}_{x,y}[r(x, y, \boldsymbol{\eta})]$ with a risk function $r(x, y, \boldsymbol{\eta})$. Here $\mathrm{E}_{x,y}[\cdot]$ denotes the expectation with respect to $p(x, y)$.

[3] As discussed in the Remark of Section 2, the lack of consistency does not necessarily mean that they are not Bayes optimal. Further analyses are needed to clarify this point.

In the following, the risk function is determined as the Kullback-Leibler divergence:

$$\begin{aligned} R(\boldsymbol{\eta}) &= \mathrm{E}_{x,y}\left[\log \frac{p(x,y)}{q(x,y|\boldsymbol{\eta})}\right] \\ &= \sum_{y\in\{-1,1\}} \int p(x,y)\, \log \frac{p(x,y)}{q(x,y|\boldsymbol{\eta})}\, dx. \end{aligned} \tag{4.1}$$

We will study the asymptotic generalization error of the two step estimator typically used in the context of the Fisher kernel: (1) estimating the parameter of the marginal model using (3.1) and (2) fixing these parameters and estimating the parameters of the linear model in (3.2).

The Cramér-Rao bound (Barndorff-Nielsen & Cox, 1989) effectively determines the theoretical limit of learning. We will especially elucidate how the generalization error is reduced to the theoretical limit as the number of unlabeled samples increases.

4.1 Asymptotics of M-estimators

Before getting into details, we briefly review how to derive the generalization error of a general *M-estimator* (Barndorff-Nielsen & Cox, 1989). An M-estimator is calculated from an equation like

$$\boldsymbol{v}(\boldsymbol{z}_n, \boldsymbol{x}^u_m, \hat{\boldsymbol{\eta}}) = \mathbf{0}, \tag{4.2}$$

where $\boldsymbol{z}_n$ is the labeled data of size n, $\boldsymbol{x}^u_m$ is the unlabeled data of size m, $\boldsymbol{\eta}$ is an s-dimensional parameter and $\boldsymbol{v}$ is an s-dimensional vector valued function (i.e. an estimating function). The function $\boldsymbol{v}$ is assumed to satisfy the unbiasedness condition

$$\mathbb{E}\left[\boldsymbol{v}(\boldsymbol{z}_n, \boldsymbol{x}^u_m, \boldsymbol{\eta}^*)\right] = \mathbf{0}, \tag{4.3}$$

and other regularity conditions which guarantee the consistency of the estimator $\hat{\boldsymbol{\eta}}$. Here $\mathbb{E}[\cdot]$ denotes the expectation with respect to training samples (both $\boldsymbol{z}_n$ and $\boldsymbol{x}^u_m$). When n goes to infinity ($r = m/n$ is fixed), we have

$$\frac{1}{n}\nabla_{\boldsymbol{\eta}}\boldsymbol{v}(\boldsymbol{z}_n, \boldsymbol{x}^u_m, \boldsymbol{\eta}^*) \;\to\; \Gamma, \qquad \text{in probability,} \tag{4.4}$$

$$\frac{1}{\sqrt{n}}\boldsymbol{v}(\boldsymbol{z}_n, \boldsymbol{x}^u_m, \boldsymbol{\eta}^*) \;\to\; \mathcal{N}(\mathbf{0}, \Lambda), \qquad \text{in distribution,} \tag{4.5}$$

where

$$\Gamma = \lim_{n\to\infty} \frac{1}{n}\, \mathbb{E}\left[\nabla_{\boldsymbol{\eta}}\boldsymbol{v}(\boldsymbol{z}_n, \boldsymbol{x}^u_m, \boldsymbol{\eta}^*)\right], \tag{4.6}$$

$$\Lambda = \lim_{n\to\infty} \frac{1}{n}\, \mathbb{E}\left[\boldsymbol{v}(\boldsymbol{z}_n, \boldsymbol{x}^u_m, \boldsymbol{\eta}^*)\left\{\boldsymbol{v}(\boldsymbol{z}_n, \boldsymbol{x}^u_m, \boldsymbol{\eta}^*)\right\}^\top\right]. \tag{4.7}$$

We calculate the asymptotic distribution of the M-estimator $\hat{\boldsymbol{\eta}}$. From the estimating equation and (4.4), we have

$$\begin{aligned}\mathbf{0} &= \frac{1}{n}\boldsymbol{v}(\boldsymbol{z}_n, \boldsymbol{x}_m^u, \boldsymbol{\eta}^*) + \frac{1}{n}\nabla_{\boldsymbol{\eta}}\boldsymbol{v}(\boldsymbol{z}_n, \boldsymbol{x}_m^u, \boldsymbol{\eta}^*)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) + O_p(\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|^2) \\ &= \frac{1}{\sqrt{n}}\boldsymbol{\zeta} + \Gamma(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) + O_p(n^{-1}),\end{aligned}$$

where $\boldsymbol{\zeta} = \boldsymbol{v}(\boldsymbol{z}_n, \boldsymbol{x}_m^u, \boldsymbol{\eta}^*)/\sqrt{n}$. Therefore, the estimator $\hat{\boldsymbol{\eta}}$ can be approximated as

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) = -\Gamma^{-1}\boldsymbol{\zeta} + O_p(n^{-1/2}) \tag{4.8}$$

and it is asymptotically Gaussian distributed,

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \sim \mathcal{N}(\mathbf{0}, \Gamma^{-1}\Lambda\Gamma^{-\top}), \tag{4.9}$$

where $\Gamma^{-\top} = (\Gamma^\top)^{-1}$.

4.2 Asymptotic Expansion of the Generalization Errors

Next, let us consider the asymptotic expansion of generalization error (4.1). By Taylor expansion, we can calculate the expectation of $R(\hat{\boldsymbol{\eta}})$ over labeled and unlabeled samples $\boldsymbol{z}_n$ and $\boldsymbol{x}_m^u$,

$$\begin{aligned}\mathbb{E}[R(\hat{\boldsymbol{\eta}})] &= R(\boldsymbol{\eta}^*) + \nabla_{\boldsymbol{\eta}}^\top R(\boldsymbol{\eta}^*)\,\mathbb{E}[\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*] \\ &\qquad + \frac{1}{2}\operatorname{tr}\left\{\nabla_{\boldsymbol{\eta}}^2 R(\boldsymbol{\eta}^*)\;\mathbb{E}[(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)\,(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^\top]\right\} + O(n^{-3/2}) \\ &= R(\boldsymbol{\eta}^*) + \frac{1}{2n}\operatorname{tr}\left\{\nabla_{\boldsymbol{\eta}}^2 R(\boldsymbol{\eta}^*)\,\Gamma^{-1}\Lambda\Gamma^{-\top}\right\} + O(n^{-3/2}).\end{aligned} \tag{4.10}$$

When we adopt the KL divergence (4.1), it turns out that $R(\boldsymbol{\eta}^*) = 0$, and the Hessian is equal to the Fisher information matrix (Barndorff-Nielsen & Cox, 1989):

$$\nabla_{\boldsymbol{\eta}}^2 R(\boldsymbol{\eta}^*) = -\operatorname{E}_{x,y}\left[\nabla_{\boldsymbol{\eta}}^2 \log q(x, y|\boldsymbol{\eta}^*)\right] \;=\; G.$$

where

$$G = \operatorname{E}_{x,y}[\nabla_{\boldsymbol{\eta}} \log q(x, y|\boldsymbol{\eta}^*)\nabla_{\boldsymbol{\eta}}^\top \log q(x, y|\boldsymbol{\eta}^*)].$$

Therefore, the generalization error is described as

$$\mathbb{E}[R(\hat{\boldsymbol{\eta}})] = \frac{1}{2n}\operatorname{tr}\left\{G\Gamma^{-1}\Lambda\Gamma^{-\top}\right\} + O(n^{-3/2}). \tag{4.11}$$

Notice that the derivation of the generalization error (4.11) relies heavily on the regularity conditions governing the limiting properties of M-estimators (Barndorff-Nielsen & Cox, 1989). When the regularity conditions do not hold, we need a different mathematical machinery to analyze the generalization error (Watanabe, 2001).

4.3 Generalization Error of the Two Step Estimator

The two step estimator (3.1) and (3.2) is regarded as a special case of M-estimators, where the estimating function is

$$\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_n, \boldsymbol{x}^u_m, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log q(x_i|\hat{\boldsymbol{\theta}}) + \sum_{j=1}^{m} \nabla_{\boldsymbol{\theta}} \log q(x^u_j|\hat{\boldsymbol{\theta}}) = \mathbf{0}, \quad (4.12)$$

$$\boldsymbol{v}_{\boldsymbol{\xi}}(\boldsymbol{z}_n, \boldsymbol{x}^u_m, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\xi}} \log h \left\{ y_i \left(\hat{\boldsymbol{w}}^\top \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(x_i) + \hat{b} \right) \right\} = \mathbf{0}, \quad (4.13)$$

where we will write $\boldsymbol{\xi} = (\boldsymbol{w}^\top, b)^\top$ for convenience.

Following the general recipe presented in Section 4.1, the generalization error can be derived. To this aim, let us define important notations first. Let us decompose the Fisher information matrix G as

$$G = \begin{pmatrix} G_{\boldsymbol{\xi}\boldsymbol{\xi}} & G_{\boldsymbol{\xi}\boldsymbol{\theta}} \\ G_{\boldsymbol{\theta}\boldsymbol{\xi}} & G_{\boldsymbol{\theta}\boldsymbol{\theta}} \end{pmatrix},$$

where $G_{\boldsymbol{\xi}\boldsymbol{\xi}}$,$G_{\boldsymbol{\xi}\boldsymbol{\theta}}, G_{\boldsymbol{\theta}\boldsymbol{\theta}}$ are the matrices of size $(d+1) \times (d+1)$, $(d+1) \times d$ and $d \times d$, respectively, and $G_{\boldsymbol{\theta}\boldsymbol{\xi}} = G^\top_{\boldsymbol{\xi}\boldsymbol{\theta}}$. Then its inverse is written as

$$G^{-1} = \begin{pmatrix} S_{\boldsymbol{\xi}\boldsymbol{\xi}} & S_{\boldsymbol{\xi}\boldsymbol{\theta}} \\ S_{\theta\boldsymbol{\xi}} & S_{\boldsymbol{\theta}\boldsymbol{\theta}} \end{pmatrix},$$

where $S_{\boldsymbol{\theta}\boldsymbol{\theta}} = (G_{\boldsymbol{\theta}\boldsymbol{\theta}} - G_{\boldsymbol{\theta}\xi} G^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} G_{\boldsymbol{\xi}\boldsymbol{\theta}})^{-1}$ (others not shown for brevity). From these sub matrices, we define the effective Fisher information (Kawanabe & Amari, 1994) as

$$G^E_{\boldsymbol{\theta}\boldsymbol{\theta}} := S^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} = G_{\boldsymbol{\theta}\boldsymbol{\theta}} - G_{\boldsymbol{\theta}\boldsymbol{\xi}} G^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} G_{\boldsymbol{\xi}\boldsymbol{\theta}}, \quad (4.14)$$

which is the net information of $\boldsymbol{\theta}$ after subtracting the amount shared with the other parameter $\boldsymbol{\xi}$. We also define

$$U_{\boldsymbol{\theta}\boldsymbol{\theta}} = \mathrm{E}_x[\nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta}^*)^\top].$$

Then, the generalization error is derived as follows:

Theorem 2. *The generalization error of the two step estimator is*

$$\mathbb{E}[R(\hat{\boldsymbol{\eta}})] = \frac{1}{2n} \left\{ d + 1 + \frac{1}{1+r} \operatorname{tr} \left(G^E_{\boldsymbol{\theta}\boldsymbol{\theta}} U^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} \right) \right\} + O(n^{-3/2}). \quad (4.15)$$

The proof is described in Appendix A.

4.4 Cramér-Rao Bound

As we have derived the generalization error (4.15), the next question would be how it compares to other estimators. In order to answer this question, we will

consider the lowerbound of the generalization errors among a reasonable set of estimators.

It is well known that the parameter variance of any asymptotically unbiased estimator[4] is lowerbounded by means of the Fisher information (e.g., Barndorff-Nielsen & Cox, 1989).

Theorem 3 (Asymptotic Cramér-Rao bound). *Assume that there are n samples $x_1, \ldots, x_n$ derived i.i.d. from $p(x|\boldsymbol{\eta}^*)$. Also assume that an estimator $\hat{\boldsymbol{\eta}}(x_1, \ldots, x_n)$ is asymptotically unbiased, that is,*

$$\mathbb{E}[\boldsymbol{\eta}(x_1, \ldots, x_n)] = \boldsymbol{\eta}^* + o(n^{-1/2}).$$

The covariance matrix of the estimator is asymptotically lowerbounded as

$$\lim_{n\to\infty} nV[\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*] \geq J^{-1}, \tag{4.16}$$

where J is the Fisher information matrix,

$$J = \mathrm{E}_x[\nabla_{\boldsymbol{\eta}} \log p(x|\boldsymbol{\eta}^*) \nabla_{\boldsymbol{\eta}} \log p(x|\boldsymbol{\eta}^*)^\top],$$

and $A \geq B$ means that $A - B$ is positive semidefinite.

Our problem is slightly more complicated than stated in this Theorem, because we have both n labeled and m unlabeled samples. In this case, the total Fisher information is simply the sum of Fisher information of labeled and unlabeled data (e.g., Zhang & Oles, 2000; Seeger, 2001). Therefore, fixing the ratio $r = m/n$, the bound (4.16) is rewritten as follows:

$$\lim_{n\to\infty} nV[\boldsymbol{\eta} - \boldsymbol{\eta}^*] \geq (G + rU)^{-1},$$

where U is the Fisher information of the marginal model $q(x|\boldsymbol{\theta})$:

$$U = \mathrm{E}_x[\nabla_{\boldsymbol{\eta}} \log q(x|\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\eta}}^\top \log q(x|\boldsymbol{\theta}^*)].$$

Once the parameter variance is bounded, we can bound the generalization error asymptotically as follows:

Theorem 4. *The generalization error of any asymptotically unbiased estimator is lowerbounded as*

$$\lim_{n\to\infty} n\,\mathbb{E}[R(\hat{\boldsymbol{\eta}})] \geq \frac{1}{2} \operatorname{tr}(I + rG^{-1}U)^{-1}. \tag{4.17}$$

(proof) Let us abbreviate $V[\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*]$ as V. As seen in (4.11), the generalization error is asymptotically expanded as

$$\mathbb{E}[R(\hat{\boldsymbol{\eta}})] = \frac{1}{2n} \operatorname{tr}\{GV\} + O(n^{-3/2}). \tag{4.18}$$

[4]One could consider asymptotically *biased* estimators, but typically such estimators are too tricky to be used in practice.

Since $\lim_{n\to\infty} nV \geq (G + rU)^{-1}$, we derive (4.17) as

$$\begin{aligned}\lim_{n\to\infty} n\,\mathbb{E}[R(\hat{\boldsymbol{\eta}})] &\geq \frac{1}{2}\operatorname{tr} G(G+rU)^{-1} \\ &= \frac{1}{2}\operatorname{tr}(I + rG^{-1}U)^{-1}.\end{aligned}$$

□

4.5 Effect of Unlabeled Data

As we compare the generalization error (4.15) with the lowerbound (4.17), it is obvious that the generalization error does not achieve the lowerbound by equality. This means that the two step estimator fails to exploit all the Fisher information provided by the samples. Intuitively, it is because we only use x's in estimating $\boldsymbol{\theta}$ at the first step, throwing away the information of y.

However, we will show that the difference to the lowerbound gets smaller as the number of unlabeled samples increases. In order to compare the generalization error and the lowerbound, the lowerbound is expanded as follows:

Lemma 4. *When expanded with respect to r, the lowerbound in (4.17) is described as*

$$\lim_{n\to\infty} n\,\mathbb{E}[R(\hat{\boldsymbol{\eta}})] \geq \frac{1}{2}\operatorname{tr}(I + rG^{-1}U)^{-1} = \frac{1}{2}\left\{d + 1 + \frac{1}{r}\operatorname{tr}\left(G^E_{\boldsymbol{\theta\theta}} U^{-1}_{\boldsymbol{\theta\theta}}\right)\right\} + O(r^{-2}). \tag{4.19}$$

The proof is described in Appendix B. On the other hand, the n^{-1} coefficient of the generalization error (4.15) is described as

$$\frac{1}{2}\left\{d + 1 + \frac{1}{r}\operatorname{tr}\left(G^E_{\boldsymbol{\theta\theta}} U^{-1}_{\boldsymbol{\theta\theta}}\right)\right\} + O(r^{-2}). \tag{4.20}$$

Thus the difference to the lowerbound is within the order of r^{-2}, which becomes very small when r is large.

In order to illustrate this result, we actually calculate the learning curves for a simple model. The Fisher score is derived from the core model (2.3), where class distributions are one-dimensional unit Gaussians centered on -1 and 1, respectively:

$$x \mid y = +1 \sim \mathcal{N}(-1, 1), \quad x \mid y = -1 \sim \mathcal{N}(1, 1).$$

The learning curves at $r = 0, 1$ and 3 are shown in Figure 2. When there are no unlabeled samples ($r = 0$), the difference between the learning curve and the lowerbound is substantially large. However, the difference gets smaller quickly as r increases, and the two curves become almost identical at $r = 3$. This illustrative result underlines our theorerical analysis, and suggests the importance of unlabeled samples in learning with the Fisher kernel.

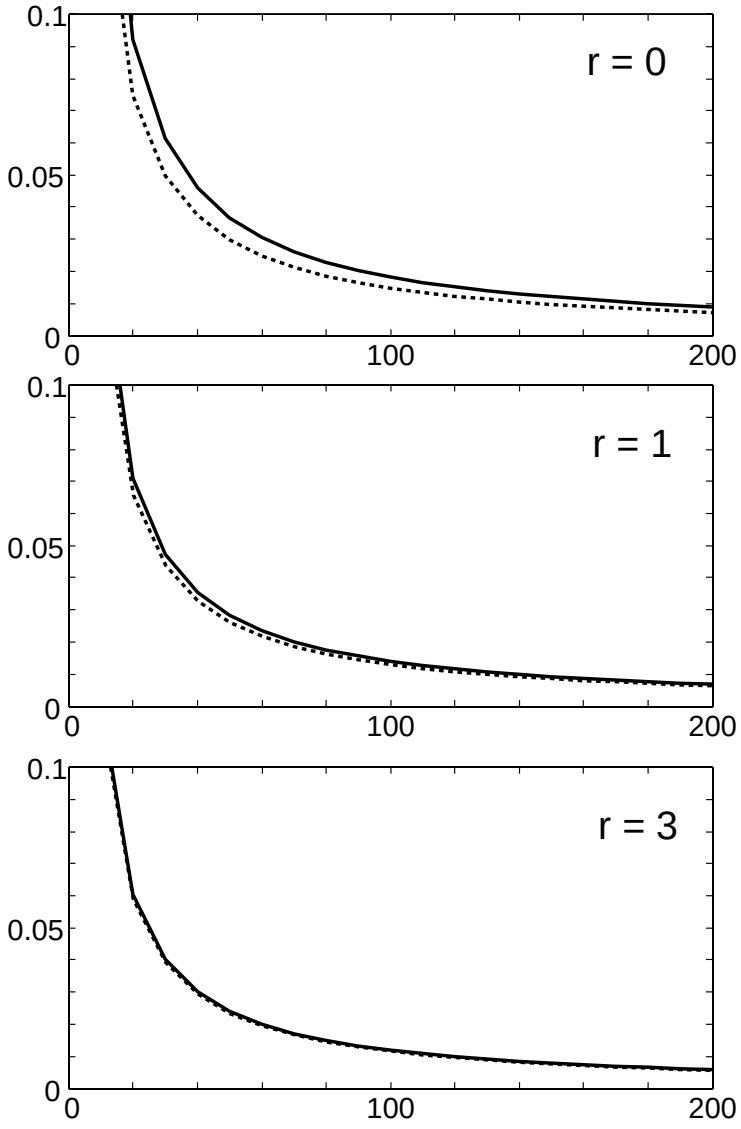

Figure 2: Theoretical learning curves of the Fisher kernel classifier. The horizontal axis shows the number of labeled samples n and the vertical axis shows the generalization error $\mathbb{E}[R(\hat{\boldsymbol{\eta}})]$. The solid and broken curves correspond to the generalization error of the two step estimator and the lowerbound determined by the Cramér-Rao bound, respectively. As the unlabeled/labeled ratio r increases, the two curves get closer.

5 Conclusion

In this paper, we have investigated several theoretical aspects of the Fisher kernel. One contribution is that we have put the Fisher kernel into the framework of statistic inference by showing the realizability conditions. This allows for a subsequent analysis about discriminative classifiers, consistency and learning curves of the generalization error (including unlabeled data). Thus our study has put the Fisher kernel approach on a more solid statistical basis, from which new algorithmic directions can be explored (e.g. the Bayes inference). In this paper only one option for feature extraction from marginal models was pursued: the combination of the Fisher kernel, a linear classifier and a linear activation function. In practice, it makes sense to consider alternative combinations. Ultimately our goal is to construct a universal statistical theory of feature extraction from marginal models, that allows an even wider practical use and a better inclusion of prior knowledge (e.g. hidden in unlabeled data or in industrial domain knowledge) into kernel based learning methods.

Acknowledgements. KRM acknowledges partial financial support by DFG (MU 987/1-1) and and BMBF under contract FKZ 01IBB02A.

References

Albert, A. (1972). Regression and the Moore-Penrose Pseudoinverse. Academic Press.

Amari, S. & Murata, N. (1993). Statistical theory of learning curves under entropic loss criterion. Neural Computation 5, 140–153.

Amari, S. & Nagaoka, H. (2001). Methods of Information Geometry. American Mathematical Society.

Barndorff-Nielsen, O. & Cox, D. (1989). Asymptotic Techniques for Use in Statistics. Chapman and Hall.

Baum, E. & Haussler, D. (1989). What size net gives valid generalization? Neural Computation 1, 151–160.

Campbell, S. & Meyer, C. (1979). Generalized Inverse of Linear Transformations. Pitman Publishing.

Cox, D. & Hinkley, D. (1974). Theoretical Statistics. Chapman & Hall, London, UK.

Cristianini, N. & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK.

Devroye, L., Györfi, L. & Lugosi, G. (1996). A Probabilistic Theory of Pattern Recognition. Springer.

Eguchi, S. & Copas, J. (2001). Information geometry on discriminant analysis and recent development. Journal of the Korean Statistical Society 27, 101–117.

Gelfand, I. & Fomin, S. (1963). Calculus of Variations. Prentice-Hall.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. Journal of Molecular Biology 162, 705–708.

Haussler, D., Kearns, M., Seung, H. & Tishby, N. (1996). Rigorous learning curve bounds from statistical mechanics. Machine Learning 25, 195–236.

Jaakkola, T. & Haussler, D. (1999). Exploiting generative models in discriminative classifiers. In Advances in Neural Information Processing Systems 11, (Kearns, M., Solla, S. & Cohn, D., eds), pp. 487–493, MIT Press.

Jaakkola, T., Meila, M. & Jebara, T. (1999). Maximum entropy discrimination. Technical Report AITR-1668 MIT.

Kawanabe, M. & Amari, S. (1994). Estimation of network parameters in semiparametric stochastic perceptron. Neural Computation 6, 1244–1261.

Malzahn, D. & Opper, M. (2002). A variational approach to learning curves. In Advances in Neural Information Processing Systems 14, (Dietterich, T. G., Becker, S. & Ghahramani, Z., eds), MIT Press.

Müller, K.-R., Finke, M., Schulten, K., Murata, N. & Amari, S. (1996). A numerical study on learning curves in stochastic multi-layer feed-forward networks. Neural Computation 8, 1085–1106.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K. & Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. IEEE Trans. Neural Networks 12, 181–201.

Schölkopf, B. & Smola, A. (2002). Learning with Kernels. MIT Press, Cambridge, MA.

Seeger, M. (2001). Learning with labeled and unlabeled data. Technical report Institute for Adaptive and Neural Computation, University of Edinburgh. http://www.dai.ed.ac.uk/homes/seeger/papers/review.ps.gz.

Seeger, M. (2002). Covariance kernels from Bayesian generative models. In Advances in Neural Information Processing Systems 14, (Dietterich, T. G., Becker, S. & Ghahramani, Z., eds), MIT Press, Cambridge, MA.

Seung, S., Sompolinsky, H. & Tishby, N. (1992). Statistical mechanics of learning from examples. Physical Review A 45, 6056.

Smith, N. & Gales, M. (2002). Speech recognition using SVMs. In Advances in Neural Information Processing Systems 14, (Dietterich, T., Becker, S. & Ghahramani, Z., eds), MIT Press.

Sonnenburg, S., Rätsch, G., Jagota, A. & Müller, K.-R. (2002). New methods for splice site recognition. In Artificial Neural Networks – ICANN 2002, (Dorronsoro, J., ed.), pp. 329–336, Springer Verlag.

Sugiyama, M. (2001). A Theory of Model Selection and Active Learning for Supervised Learning. PhD thesis, Department of Computer Science, Tokyo Institute of Technology.

Tsuda, K. & Kawanabe, M. (2002). The leave-one-out kernel. In Artificial Neural Networks – ICANN 2002, (Dorronsoro, J., ed.), pp. 727–732, Springer Verlag.

Tsuda, K., Kawanabe, M. & Müller, K.-R. (2003). Clustering with the Fisher score. In Advances in Neural Information Processing Systems 15, (Becker, S., Thrun, S. & Obermayer, K., eds), MIT Press. to appear.

Tsuda, K., Kawanabe, M., Rätsch, G., Sonnenburg, S. & Müller, K.-R. (2002). A new discriminative kernel from probabilistic models. Neural Computation 14, 2397–2414.

van der Vaart, A. (1998). Asymptotic Statistics. Cambridge University Press, Cambridge, UK.

Vapnik, V. (1998). Statistical Learning Theory. Wiley, New York.

Vinokourov, A. & Girolami, M. (2002). A probabilistic framework for the hierarchic organization and classification of document collections. Journal of Intelligent Information Systems 18, 153–172.

Watanabe, S. (2001). Algebraic analysis for non-identifiable learning machines. Neural Computation 13, 899–933.

Watkin, T., Rau, A. & Biehl, M. (1993). The statistical mechanics of learning a rule. Reviews of Modern Physics 65, 499.

Zhang, T. & Oles, F. (2000). The value of unlabeled data for classification problems. In Proceedings of the Seventeenth International Conference on Machine Learning, (Langley, P., ed.), pp. 1191–1198, Morgan Kaufmann, Stanford, CA.

Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T. & Müller, K.-R. (2000). Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites. Bioinformatics 16, 799–807.

A Proof of Theorem 2

Let us decompose the matrices Γ, Λ as

$$\Gamma = \begin{pmatrix} \Gamma_{\boldsymbol{\xi\xi}} & \Gamma_{\boldsymbol{\xi\theta}} \\ \Gamma_{\boldsymbol{\theta\xi}} & \Gamma_{\boldsymbol{\theta\theta}} \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} \Lambda_{\boldsymbol{\xi\xi}} & \Lambda_{\boldsymbol{\xi\theta}} \\ \Lambda_{\boldsymbol{\theta\xi}} & \Lambda_{\boldsymbol{\theta\theta}} \end{pmatrix}.$$

The submatrices are computed as follows:

$$\begin{aligned}
\Gamma_{\boldsymbol{\xi\xi}} &= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\nabla_{\boldsymbol{\xi}} \sum_{i=1}^{n} \nabla_{\boldsymbol{\xi}} \log Q(y_i|x_i, \boldsymbol{\eta}^*) \right] \\
&= \mathbb{E}\left[\nabla_{\boldsymbol{\xi}}^2 \log q(x, y|\boldsymbol{\eta}^*) - \nabla_{\boldsymbol{\xi}}^2 \log q(x|\boldsymbol{\theta}^*) \right] = -G_{\boldsymbol{\xi\xi}}, \\
\Gamma_{\boldsymbol{\xi\theta}} &= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} \nabla_{\boldsymbol{\xi}} \log Q(y_i|x_i, \boldsymbol{\eta}^*) \right] \\
&= -G_{\boldsymbol{\xi\theta}}, \\
\Gamma_{\boldsymbol{\theta\xi}} &= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\nabla_{\boldsymbol{\xi}} \left\{ \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log q(x_i|\boldsymbol{\theta}^*) + \sum_{j=1}^{m} \nabla_{\boldsymbol{\theta}} \log q(x_j^u|\boldsymbol{\theta}^*) \right\} \right] \\
&= 0, \\
\Gamma_{\boldsymbol{\theta\theta}} &= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log q(x_i|\boldsymbol{\theta}^*) + \sum_{j=1}^{m} \nabla_{\boldsymbol{\theta}} \log q(x_j^u|\boldsymbol{\theta}^*) \right\} \right] \\
&= -(1+r) U_{\boldsymbol{\theta\theta}}, \\
\Lambda_{\boldsymbol{\xi\xi}} &= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^{n} \nabla_{\boldsymbol{\xi}} \log Q(y_i|x_i, \boldsymbol{\eta}^*) \sum_{i=1}^{n} \nabla_{\boldsymbol{\xi}} \log Q(y_i|x_i, \boldsymbol{\eta}^*)^\top \right] \\
&= G_{\boldsymbol{\xi\xi}}, \\
\Lambda_{\boldsymbol{\xi\theta}} &= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^{n} \nabla_{\boldsymbol{\xi}} \log Q(y_i|x_i, \boldsymbol{\eta}^*) \right. \\
&\qquad \left. \times \left\{ \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log q(x_i|\boldsymbol{\theta}^*) + \sum_{j=1}^{m} \nabla_{\boldsymbol{\theta}} \log q(x_j^u|\boldsymbol{\theta}^*) \right\}^\top \right] = 0, \\
\Lambda_{\boldsymbol{\theta\xi}} &= 0, \\
\Lambda_{\boldsymbol{\theta\theta}} &= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\left\{ \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log q(x_i|\boldsymbol{\theta}^*) + \sum_{j=1}^{m} \nabla_{\boldsymbol{\theta}} \log q(x_j^u|\boldsymbol{\theta}^*) \right\} \right. \\
&\qquad \left. \times \left\{ \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log q(x_i|\boldsymbol{\theta}^*) + \sum_{j=1}^{m} \nabla_{\boldsymbol{\theta}} \log q(x_j^u|\boldsymbol{\theta}^*) \right\}^\top \right] \\
&= (1+r) U_{\boldsymbol{\theta\theta}}.
\end{aligned}$$

In summary, we have the following:

$$\Gamma = -\begin{pmatrix} G_{\boldsymbol{\xi\xi}} & G_{\boldsymbol{\xi\theta}} \\ 0 & (1+r)U_{\boldsymbol{\theta\theta}} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} G_{\boldsymbol{\xi\xi}} & 0 \\ 0 & (1+r)U_{\boldsymbol{\theta\theta}} \end{pmatrix}.$$

The inverse matrix of Γ becomes

$$\Gamma^{-1} = \begin{pmatrix} \Gamma_{\boldsymbol{\xi\xi}}^{-1} & -\Gamma_{\boldsymbol{\xi\xi}}^{-1}\Gamma_{\boldsymbol{\xi\theta}}\Gamma_{\boldsymbol{\theta\theta}}^{-1} \\ 0 & \Gamma_{\boldsymbol{\theta\theta}}^{-1} \end{pmatrix} = -\begin{pmatrix} G_{\boldsymbol{\xi\xi}}^{-1} & -\frac{1}{1+r}G_{\boldsymbol{\xi\xi}}^{-1}G_{\boldsymbol{\xi\theta}}U_{\boldsymbol{\theta\theta}}^{-1} \\ 0 & \frac{1}{1+r}U_{\boldsymbol{\theta\theta}}^{-1} \end{pmatrix}.$$

The asymptotic covariance of $\hat{\boldsymbol{\eta}}$ is

$$\begin{aligned} \Gamma^{-1}\Lambda\Gamma^{-\top} &= \begin{pmatrix} A & B \\ B^\top & D \end{pmatrix}, \\ A &= \Gamma_{\boldsymbol{\xi\xi}}^{-1}\Lambda_{\boldsymbol{\xi\xi}}\Gamma_{\boldsymbol{\xi\xi}}^{-1} + \Gamma_{\boldsymbol{\xi\xi}}^{-1}\Gamma_{\boldsymbol{\xi\theta}}\Gamma_{\boldsymbol{\theta\theta}}^{-1}\Lambda_{\boldsymbol{\theta\theta}}\Gamma_{\boldsymbol{\theta\theta}}^{-1}\Gamma_{\boldsymbol{\xi\theta}}^{\top}\Gamma_{\boldsymbol{\xi\xi}}^{-1} \\ &= G_{\boldsymbol{\xi\xi}}^{-1} + \frac{1}{1+r}G_{\boldsymbol{\xi\xi}}^{-1}G_{\boldsymbol{\xi\theta}}U_{\boldsymbol{\theta\theta}}^{-1}G_{\boldsymbol{\theta\xi}}G_{\boldsymbol{\xi\xi}}^{-1}, \\ B &= -\Gamma_{\boldsymbol{\xi\xi}}^{-1}\Gamma_{\boldsymbol{\xi\theta}}\Gamma_{\boldsymbol{\theta\theta}}^{-1}\Lambda_{\boldsymbol{\theta\theta}}\Gamma_{\boldsymbol{\theta\theta}}^{-1} = -\frac{1}{1+r}G_{\boldsymbol{\xi\xi}}^{-1}G_{\boldsymbol{\xi\theta}}U_{\boldsymbol{\theta\theta}}^{-1}, \\ D &= \Gamma_{\boldsymbol{\theta\theta}}^{-1}\Lambda_{\boldsymbol{\theta\theta}}\Gamma_{\boldsymbol{\theta\theta}}^{-1} = \frac{1}{1+r}U_{\boldsymbol{\theta\theta}}^{-1}. \end{aligned}$$

Therefore,

$$\begin{aligned} &G\Gamma^{-1}\Lambda\Gamma^{-\top} \\ &= \begin{pmatrix} G_{\boldsymbol{\xi\xi}} & G_{\boldsymbol{\xi\theta}} \\ G_{\boldsymbol{\theta\xi}} & G_{\boldsymbol{\theta\theta}} \end{pmatrix}\begin{pmatrix} G_{\boldsymbol{\xi\xi}}^{-1} + \frac{1}{1+r}G_{\boldsymbol{\xi\xi}}^{-1}G_{\boldsymbol{\xi\theta}}U_{\boldsymbol{\theta\theta}}^{-1}G_{\boldsymbol{\theta\xi}}G_{\boldsymbol{\xi\xi}}^{-1} & -\frac{1}{1+r}G_{\boldsymbol{\xi\xi}}^{-1}G_{\boldsymbol{\xi\theta}}U_{\boldsymbol{\theta\theta}}^{-1} \\ -\frac{1}{1+r}U_{\boldsymbol{\theta\theta}}^{-1}G_{\boldsymbol{\theta\xi}}G_{\boldsymbol{\xi\xi}}^{-1} & \frac{1}{1+r}U_{\boldsymbol{\theta\theta}}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} I & 0 \\ G_{\boldsymbol{\theta\xi}}G_{\boldsymbol{\xi\xi}}^{-1} - \frac{1}{1+r}G_{\boldsymbol{\theta\theta}}^{E}U_{\boldsymbol{\theta\theta}}^{-1}G_{\boldsymbol{\theta\xi}}G_{\boldsymbol{\xi\xi}}^{-1} & \frac{1}{1+r}G_{\boldsymbol{\theta\theta}}^{E}U_{\boldsymbol{\theta\theta}}^{-1} \end{pmatrix}. \end{aligned}$$

By subsutituting it to (4.11), we get the asymptotic expansion of the generalization error as

$$\mathbb{E}[R(\hat{\boldsymbol{\eta}})] = \frac{1}{2n}\left\{d + 1 + \frac{1}{1+r}\operatorname{tr}\left(G_{\boldsymbol{\theta\theta}}^{E}U_{\boldsymbol{\theta\theta}}^{-1}\right)\right\} + O(n^{-3/2}). \tag{A.1}$$

□

We remark that $G_{\boldsymbol{\theta\theta}}^{E} > U_{\boldsymbol{\theta\theta}}$ in this case. This can be shown as follows. The conditional information matrix

$$\begin{aligned} J(Y|X) &= \mathrm{E}_{x,y}\left[\nabla_{\boldsymbol{\eta}}\log Q(y|x,\boldsymbol{\eta}^*)\nabla_{\boldsymbol{\eta}}\log Q(y|x,\boldsymbol{\eta}^*)\right] \\ &= -\mathrm{E}_{x,y}\left[\nabla_{\boldsymbol{\eta}}\nabla_{\boldsymbol{\eta}}\log Q(y|x,\boldsymbol{\eta}^*)\right] \\ &= \begin{pmatrix} G_{\boldsymbol{\xi\xi}} & G_{\boldsymbol{\xi\theta}} \\ G_{\boldsymbol{\theta\xi}} & G_{\boldsymbol{\theta\theta}} - U_{\boldsymbol{\theta\theta}} \end{pmatrix} \end{aligned}$$

is positive definite, if the probabilistic model is regular. Let us transform the information matrix as

$$FJ(Y|X)F^\top = \begin{pmatrix} G_{\boldsymbol{\xi\xi}} & 0 \\ 0 & G_{\boldsymbol{\theta\theta}}^{E} - U_{\boldsymbol{\theta\theta}} \end{pmatrix},$$

where

$$F = \begin{pmatrix} I & 0 \\ -G_{\boldsymbol{\theta\xi}} G_{\boldsymbol{\xi\xi}}^{-1} & I \end{pmatrix}.$$

Since the matrix $FJ(Y|X)F^\top$ is positive definite, $G^E_{\boldsymbol{\theta\theta}} - U_{\boldsymbol{\theta\theta}}$ must be positive definite, too.

Although we only showed the result in the case of log likelihood loss, it is possible to calculate the generalization error for general loss functions. The fomula becomes

$$\mathbb{E}[R(\hat{\boldsymbol{\eta}})] = \frac{1}{2n}\left\{ \mathrm{tr}(G_{\boldsymbol{\xi\xi}} L_{\boldsymbol{\xi\xi}}^{-1} \Lambda_{\boldsymbol{\xi\xi}} L_{\boldsymbol{\xi\xi}}^{-1}) + \frac{1}{1+r}\,\mathrm{tr}(H U_{\boldsymbol{\theta\theta}}^{-1}) \right\} + O(n^{-3/2}), \quad \text{(A.2)}$$

where

$$\begin{aligned} H &= G_{\boldsymbol{\theta\theta}} - G_{\boldsymbol{\theta\xi}} L_{\boldsymbol{\xi\xi}}^{-1} L_{\boldsymbol{\xi\theta}} - L_{\boldsymbol{\theta\xi}} L_{\boldsymbol{\xi\xi}}^{-1} G_{\boldsymbol{\xi\theta}} + L_{\boldsymbol{\theta\xi}} L_{\boldsymbol{\xi\xi}}^{-1} G_{\boldsymbol{\xi\xi}} L_{\boldsymbol{\xi\xi}}^{-1} L_{\boldsymbol{\xi\theta}} \\ L_{\boldsymbol{\eta\eta}} &= -\mathrm{E}_{x,y}\left[\nabla_{\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}} \ell(\boldsymbol{w}^\top \boldsymbol{f}_{\boldsymbol{\theta}}(x) + b,\, y) \right] \\ \Lambda_{\boldsymbol{\xi\xi}} &= \mathrm{E}_{x,y}\left[\nabla_{\boldsymbol{\xi}} \ell(\boldsymbol{w}^\top \boldsymbol{f}_{\boldsymbol{\theta}}(x) + b,\, y)\, \nabla_{\boldsymbol{\xi}} \ell(\boldsymbol{w}^\top \boldsymbol{f}_{\boldsymbol{\theta}}(x) + b,\, y) \right]. \end{aligned}$$

B Proof of Lemma 4

In order to prove (4.19), we will use the following expansion (Sugiyama, 2001):

Lemma 5. *For any symmetric matrix Z and $r \neq 0$, $(I + rZ)^{-1}$ is expanded as follows:*

$$(I + rZ)^{-1} = (I - ZZ^\dagger) - \sum_{j=1}^{k} (-\frac{1}{r} Z^\dagger)^j - (-\frac{1}{r} Z^\dagger)^{k+1} (I + \frac{1}{r} Z^\dagger)^{-1}, \quad \text{(A.1)}$$

where † indicates the Moore-Penrose pseudo inverse (Campbell & Meyer, 1979) and k is an arbitrary positive integer.

The proof is described in Appendix C. The lowerbound in (4.17) is rewritten as

$$\begin{aligned} \frac{1}{2}\,\mathrm{tr}(I + rG^{-1}U)^{-1} &= \frac{1}{2}\,\mathrm{tr}\left[(G^{1/2} + rG^{-1/2}U)^{-1} G^{1/2} \right] \\ &= \frac{1}{2}\,\mathrm{tr}\left[G^{1/2} (G^{1/2} + rG^{-1/2}U)^{-1} \right] \\ &= \frac{1}{2}\,\mathrm{tr}(I + rG^{-1/2}UG^{-1/2})^{-1}. \end{aligned}$$

Setting $Z = G^{-1/2}UG^{-1/2}$, it is expanded as follows:

$$\frac{1}{2}\,\mathrm{tr}(I + rG^{-1}U)^{-1} = \frac{1}{2}\left\{ \xi_0 + \frac{\xi_1}{r} \right\} + O(r^{-2}), \quad \text{(A.2)}$$

where the coefficients are described as

$$\xi_0 = \mathrm{tr}(I - ZZ^\dagger), \quad \xi_1 = \mathrm{tr}(Z^\dagger). \quad \text{(A.3)}$$

The equation (4.19) is proved because the coefficients are derived as follows:

Lemma 6. *The coefficients ξ_0 and ξ_1 are described as*

$$\xi_0 = d+1, \tag{A.4}$$
$$\xi_1 = \mathrm{tr}(G^E_{\boldsymbol{\theta\theta}} U^{-1}_{\boldsymbol{\theta\theta}}), \tag{A.5}$$

respectively.

(proof) Since $q(x|\boldsymbol{\theta})$ does not depend on $\boldsymbol{w}$ and b, U is described as

$$U = \begin{pmatrix} 0 & 0 \\ 0 & U_{\boldsymbol{\theta\theta}} \end{pmatrix}.$$

Then Z is rewritten as

$$Z = G^{-1/2} U G^{-1/2} = BB^\top,$$

where B is a $(2d+1) \times d$ matrix

$$B = G^{-1/2} \begin{bmatrix} 0 \\ U^{1/2}_{\boldsymbol{\theta\theta}} \end{bmatrix}.$$

In terms of B, the pseudo inverse of Z is written as

$$Z^\dagger = B(B^\top B)^{-2} B^\top.$$

The coefficient ξ_0 is rewritten as

$$\xi_0 = \mathrm{tr}(I - ZZ^\dagger) = (2d+1) - \mathrm{tr}(ZZ^\dagger),$$

where

$$\mathrm{tr}(ZZ^\dagger) = \mathrm{tr}(BB^\top B(B^\top B)^{-2} B^\top) = \mathrm{tr}(B(B^\top B)^{-1} B^\top) = \mathrm{tr}(B^\top B(B^\top B)^{-1}) = d,$$

we thus have $\xi_0 = d+1$. On the other hand, $\xi_1 = \mathrm{tr}(Z^\dagger)$ is rewritten as

$$\begin{aligned}
\mathrm{tr}(Z^\dagger) &= \mathrm{tr}\left[B(B^\top B)^{-2} B^\top\right] \\
&= \mathrm{tr}(B^\top B)^{-1} \\
&= \mathrm{tr}\left(\begin{bmatrix} 0 & U^{1/2}_{\boldsymbol{\theta\theta}} \end{bmatrix} G^{-1} \begin{bmatrix} 0 \\ U^{1/2}_{\boldsymbol{\theta\theta}} \end{bmatrix}\right)^{-1} \\
&= \mathrm{tr}\left[U^{-1/2}_{\boldsymbol{\theta\theta}} S^{-1}_{\boldsymbol{\theta\theta}} U^{-1/2}_{\boldsymbol{\theta\theta}}\right] \\
&= \mathrm{tr}(S^{-1}_{\boldsymbol{\theta\theta}} U^{-1}_{\boldsymbol{\theta\theta}}) = \mathrm{tr}(G^E_{\boldsymbol{\theta\theta}} U^{-1}_{\boldsymbol{\theta\theta}}).
\end{aligned}$$

□

C Proof of Lemma 5

This expansion was originally derived in Lemma 4.8 of (Sugiyama, 2001). In the following, we quote his proof for readers' convenience. Let us define $\alpha = 1/r$. According to Theorem 4.8 in (Albert, 1972), the following holds for a symmetric matrix Z:

$$(I + \alpha^{-1}Z)^{-1} = (I - ZZ^\dagger) + \alpha Z^\dagger (I + \alpha Z^\dagger)^{-1}. \tag{A.1}$$

Also for any matrix B, we have the following equation:

$$(I+B)^{-1} = I(I+B)^{-1} = (I+B-B)(I+B)^{-1} = I - B(I+B)^{-1}.$$

Defining $B = \alpha Z^\dagger$, we have

$$(I + \alpha Z^\dagger)^{-1} = I - \alpha Z^\dagger (I + \alpha Z^\dagger)^{-1}. \tag{A.2}$$

By repeatedly applying (A.2) to (A.1), we have

$$\begin{aligned}
(I + \alpha^{-1}Z)^{-1} &= (I - ZZ^\dagger) + \alpha Z^\dagger [I - \alpha Z^\dagger (I + \alpha Z^\dagger)^{-1}] \\
&= (I - ZZ^\dagger) + \alpha Z^\dagger - (\alpha Z^\dagger)^2 (I + \alpha Z^\dagger)^{-1} \\
&= (I - ZZ^\dagger) + \alpha Z^\dagger - (\alpha Z^\dagger)^2 + (\alpha Z^\dagger)^3 (I + \alpha Z^\dagger)^{-1} \\
&\vdots \\
&= (I - ZZ^\dagger) - \sum_{j=1}^{k} (-\alpha Z^\dagger)^j - (-\alpha Z^\dagger)^{k+1} (I + \alpha Z^\dagger)^{-1}.
\end{aligned}$$

□