

## Asymptotic Properties of the Fisher Kernel

**Koji Tsuda**

*koji.tsuda@tuebingen.mpg.de*

*Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany, and  
AIST Computational Biology Research Center, Koto-ku, Tokyo, 135-0064, Japan*

**Shotaro Akaho**

*s.akaho@aist.go.jp*

*AIST Neuroscience Research Institute, Tsukuba, 305-8568, Japan*

**Motoaki Kawanabe**

*nabe@first.fhg.de*

*Fraunhofer FIRST, 12489 Berlin, Germany*

**Klaus-Robert Müller**

*klaus@first.fhg.de*

*Fraunhofer FIRST, 12489 Berlin, Germany, and University of Potsdam,  
14482 Potsdam, Germany*

This letter analyzes 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 analyze 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., it contains the true distribution). Realizability 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 preferable—because both yield consistent estimators—to other losses such as the exponential loss, when a linear classifier is used together with the Fisher kernel. Therefore, this letter underlines that the Fisher kernel should be viewed not as a heuristics but as a powerful statistical tool with well-controlled statistical properties.

### 1 Introduction ---

Recently, the Fisher kernel (Jaakkola & Haussler, 1999) has been successfully applied as a feature extractor in supervised classification (Jaakkola & Haussler, 1999; Tsuda, Kawanabe, Rätsch, Sonnenburg, & Müller, 2002; Son-

nenburg, Rätsch, Jagota, & Müller, 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. First 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 hidden Markov models (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). A second feature of the Fisher kernel is that it allows incorporating 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, Mika, Rätsch, Tsuda, & Schölkopf, 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).<sup>1</sup> In spite of its practical success, a theoretical analysis of the Fisher kernel has not been sufficiently explored so far, with the exceptions being Jaakkola, Meila, and Jebara (1999), Tsuda and Kawanabe (2002), Seeger (2002), and Tsuda, Kawanabe, and Müller, in press). 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. And Seeger (2002) pointed out that the Fisher kernel can be perceived as an approximation of the mutual information kernel.

This article presents 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. Let us assume that a probabilistic model  $q(x | \theta)$ ,  $x \in \mathcal{X}$ ,  $\theta \in \mathfrak{N}^d$  is available. Given a parameter estimate  $\hat{\theta}$  from training samples, the feature vector (i.e., the Fisher score) is obtained as

$$f_{\hat{\theta}}(x) = \left( \frac{\partial \log q(x | \hat{\theta})}{\partial \theta_1}, \dots, \frac{\partial \log q(x | \hat{\theta})}{\partial \theta_d} \right)^\top. \quad (1.1)$$

<sup>1</sup> Of course, a brute force search over all possible kernels can be pursued using cross-validation procedures or bounds from learning theory to select the “best” kernel (cf. Müller et al., 2001).

The Fisher kernel refers to the inner product in this space. When used in supervised classification, the Fisher kernel is commonly combined with a linear classifier such as 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).

First, we will show the sufficient conditions that the obtained posterior model is realizable—that it contains the true posterior distribution, which then immediately leads to consistency. Once realizability is ensured, 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, we point out that when a linear classifier is combined with the Fisher kernel, then the log loss and the squared loss are theoretically preferable to other loss functions. This result recommends using a classifier based on the log loss or the squared loss.

## 2 Realizability 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  $\partial_\alpha f = \partial f / \partial \alpha$ ,  $\nabla_\theta f = (\partial_{\theta_1} f, \dots, \partial_{\theta_d} f)^\top$ , and  $\nabla_\theta^2 f$  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, equation 1.1. The posterior probability is described by a linear function followed by an activation function  $h$ ,

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

where  $\mathbf{w} \in \mathfrak{R}^d$ ,  $b \in \mathfrak{R}$ , parameters are summarized as  $\boldsymbol{\eta} = (\mathbf{w}^\top, b, \boldsymbol{\theta}^\top)^\top$ , and  $h$  is a linear activation function:<sup>2</sup>

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

In the following, we will investigate the conditions that  $Q(y | x, \boldsymbol{\eta})$  is realizable—that 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,

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

<sup>2</sup> Compared with sigmoid functions, the linear activation function is not so common in literature. However, it allows us to perform statistical analysis, as we show later.

which we call the core model. 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

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

Substituting equations 2.1 and 2.2, equation 2.4 holds if and only if

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

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

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

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  $b = 2\alpha^* - 1$ , equation 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  $\mathcal{M} = \{q_0 | q_0(x | \alpha), \alpha \in [0, 1]\}$ . According to information geometry (Amari & Nagaoka, 2001),  $\mathcal{M}$  is regarded as a manifold in a Riemannian space. Let  $\mathcal{Q}$  denote the manifold of  $q(x | \boldsymbol{\theta})$ :  $\mathcal{Q} = \{q | q(x | \boldsymbol{\theta}), \boldsymbol{\theta} \in \Re^d\}$ . Now the question is how to determine a manifold  $\mathcal{Q}$  such that equation 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 (see Figure 1), then the Fisher score  $\boldsymbol{f}$  derived from  $q(x | \boldsymbol{\theta}^*)$  gives a realizable posterior model, equation 2.1.*

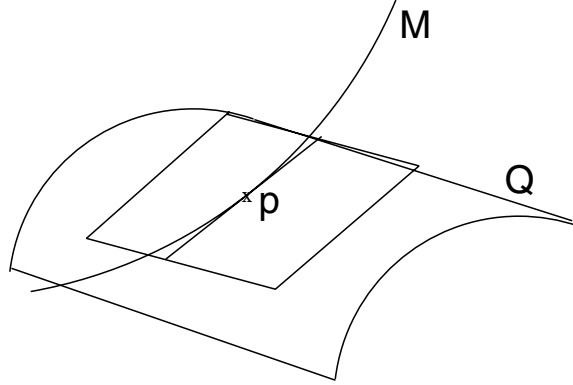


Figure 1: Information-geometric picture of a probabilistic model whose Fisher kernel leads to a realizable posterior model. The important point is that the tangent space of manifold  $\mathcal{M}$  is contained in that of manifold  $\mathcal{Q}$ . Details are explained in the text.

**Proof.** To prove the theorem, it is sufficient to show the existence of  $w \in \mathfrak{N}^d$  and  $b \in \mathfrak{N}$  such that

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

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:

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

Let  $u$  be the  $d$ -dimensional vector where the  $i$ th element is

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

Then equation 2.7 holds when

$$w = 2\alpha^*(1 - \alpha^*)u, \quad b = 2\alpha^* - 1.$$

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 ensure realizability while keeping the number of parameters small. To this end, 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. (in press), 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, Györfi, & Lugosi, 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  $z_n = \{x_i, y_i\}_{i=1}^n$  the set of  $n$  independently and identically distributed (i.i.d.) labeled samples derived from  $p(x, y)$ . Denote by  $x_m^u = \{x_j^u\}_{j=1}^m$  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,  $\theta$  is obtained as

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

Then, in the second step,  $w$  and  $b$  are obtained as

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

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

$$[\hat{w}, \hat{b}] = \operatorname{argmin}_{w, b} \sum_{i=1}^n \ell(w^\top f_{\hat{\theta}}(x_i) + b, y_i), \quad (3.3)$$

where the loss function in equation 3.2 corresponds to

$$\ell(y, y') = -\log h(y y'). \quad (3.4)$$

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

**Lemma 2.** *Assume that  $q(x | \theta)$  satisfies the realizability conditions in theorem 1. The two-step estimator with the log loss (see equation 3.4) is consistent.*

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

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

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

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

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

$$\mathcal{G} = \{g \mid g(x) = w^\top f_{\theta^*}(x) + b, w \in \mathfrak{R}^d, b \in \mathfrak{R}\}. \quad (3.6)$$

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 equation 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 the 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  $\delta g$ ; thus, we have

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

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

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

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

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

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

$$\ell(y, y') = (y - y')^2, \quad (3.9)$$

and as we will see in the following, consistency is ensured 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 equation 3.5, we have

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

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

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

which is solved as

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

Since this solution is the same as equation 3.8, this lemma is proved by following the same procedure as lemma 2.

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

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

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.<sup>3</sup> 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 (Baum & Haussler, 1989; Amari & Murata, 1993; Müller, Finke, Schulten, Murata, & Amari, 1996; Haussler, Kearns, Seung, & Tishby, 1996; Malzahn & Opper, 2002). Studies about learning curves have been playing an important role in elucidating the behavior of learning machines. Research 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, Sompolinsky, & Tishby, 1992; Watkin, Rau, & Biehl, 1993; Haussler et al., 1996; Malzahn & Opper, 2002) has contributed to the study of generalization errors apart from bounds derived in a statistical learning theory framework (e.g., Devroye et al., 1996; Vapnik, 1998). In this section, we will adopt asymptotic statistical techniques following Barndorff-Nielsen and Cox (1989).

The generalization error is defined as  $R(\boldsymbol{\eta}) := E_{x,y}[r(x, y, \boldsymbol{\eta})]$  with a risk function  $r(x, y, \boldsymbol{\eta})$ . Here,  $E_{x,y}[\cdot]$  denotes the expectation with respect to  $p(x, y)$ . In the following, the risk function is determined as the Kullback-Leibler divergence:

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

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 equation 3.1 and (2) fixing these parameters and estimating the parameters of the linear model in equation 3.2.

The Cramér-Rao bound (Barndorff-Nielsen & Cox, 1989) effectively determines the theoretical limit of learning. We will especially elucidate how

<sup>3</sup> As discussed in the remark in section 2, the lack of consistency does not necessarily mean that they are not Bayes optimal. Further analyses are needed to clarify this point.

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 the 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

$$v(z_n, \mathbf{x}_m^u, \hat{\boldsymbol{\eta}}) = \mathbf{0}, \quad (4.2)$$

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

$$\mathbb{E}[v(z_n, \mathbf{x}_m^u, \boldsymbol{\eta}^*)] = \mathbf{0}, \quad (4.3)$$

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

$$\frac{1}{n} \nabla_{\boldsymbol{\eta}} v(z_n, \mathbf{x}_m^u, \boldsymbol{\eta}^*) \rightarrow \Gamma, \quad \text{in probability,} \quad (4.4)$$

$$\frac{1}{\sqrt{n}} v(z_n, \mathbf{x}_m^u, \boldsymbol{\eta}^*) \rightarrow \mathcal{N}(\mathbf{0}, \Lambda), \quad \text{in distribution,} \quad (4.5)$$

where

$$\Gamma = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}[\nabla_{\boldsymbol{\eta}} v(z_n, \mathbf{x}_m^u, \boldsymbol{\eta}^*)], \quad (4.6)$$

$$\Lambda = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}[v(z_n, \mathbf{x}_m^u, \boldsymbol{\eta}^*) \{v(z_n, \mathbf{x}_m^u, \boldsymbol{\eta}^*)\}^\top]. \quad (4.7)$$

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

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

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

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

and it is asymptotically gaussian distributed,

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

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  $R(\hat{\eta})$  over labeled and unlabeled samples  $z_n$  and  $x_m^u$ ,

$$\begin{aligned} \mathbb{E}[R(\hat{\eta})] &= R(\eta^*) + \nabla_{\eta}^{\top} R(\eta^*) \mathbb{E}[\hat{\eta} - \eta^*] \\ &\quad + \frac{1}{2} \text{tr}\{\nabla_{\eta}^2 R(\eta^*) \mathbb{E}[(\hat{\eta} - \eta^*)(\hat{\eta} - \eta^*)^{\top}]\} + O(n^{-3/2}) \\ &= R(\eta^*) + \frac{1}{2n} \text{tr}\{\nabla_{\eta}^2 R(\eta^*) \Gamma^{-1} \Lambda \Gamma^{-\top}\} + O(n^{-3/2}). \end{aligned} \quad (4.10)$$

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

$$\nabla_{\eta}^2 R(\eta^*) = -\mathbb{E}_{x,y}[\nabla_{\eta}^2 \log q(x, y | \eta^*)] = G,$$

where

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

Therefore, the generalization error is described as

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

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

$$\mathbf{v}_{\theta}(z_n, x_m^u, \hat{\theta}) = \sum_{i=1}^n \nabla_{\theta} \log q(x_i | \hat{\theta}) + \sum_{j=1}^m \nabla_{\theta} \log q(x_j^u | \hat{\theta}) = \mathbf{0}, \quad (4.12)$$

$$\mathbf{v}_{\xi}(z_n, x_m^u, \hat{\xi}, \hat{\theta}) = \sum_{i=1}^n \nabla_{\xi} \log h\{y_i (\hat{w}^{\top} \mathbf{f}_{\hat{\theta}}(x_i) + \hat{b})\} = \mathbf{0}, \quad (4.13)$$

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

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

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

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

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

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

$$G_{\theta\theta}^E := S_{\theta\theta}^{-1} = G_{\theta\theta} - G_{\theta\xi}G_{\xi\xi}^{-1}G_{\xi\theta}, \quad (4.14)$$

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

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

Then the generalization error is derived as follows:

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

$$\mathbb{E}[R(\hat{\eta})] = \frac{1}{2n} \left\{ d + 1 + \frac{1}{1+r} \text{tr} \left( G_{\theta\theta}^E U_{\theta\theta}^{-1} \right) \right\} + O(n^{-3/2}). \quad (4.15)$$

The proof is described in appendix A.

**4.4 Cramér-Rao Bound.** Now that we have derived the generalization error (see equation 4.15), the next question is how it compares to other estimators. In order to answer this question, we will consider the lower bound of the generalization errors among a reasonable set of estimators.

It is well known that the parameter variance of any asymptotically unbiased estimator<sup>4</sup> is lower-bounded by means of the Fisher information (e.g., Barndorff-Nielsen & Cox, 1989).

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

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

*The covariance matrix of the estimator is asymptotically lower-bounded as*

$$\lim_{n \rightarrow \infty} nV[\hat{\eta} - \eta^*] \geq J^{-1}, \quad (4.16)$$

<sup>4</sup> One could consider asymptotically biased estimators, but typically such estimators are too tricky to be used in practice.

where  $J$  is the Fisher information matrix,

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

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

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

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

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

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 lower-bounded as*

$$\lim_{n \rightarrow \infty} n\mathbb{E}[R(\hat{\eta})] \geq \frac{1}{2} \text{tr}(I + rG^{-1}U)^{-1}. \quad (4.17)$$

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

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

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

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

**4.5 Effect of Unlabeled Data.** As we compare the generalization error 4.15 with the lower bound 4.17, it is obvious that the generalization error does not achieve the lower bound 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  $\theta$  at the first step, discarding the information of  $y$ .

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

**Lemma 4.** *When expanded with respect to  $r$ , the lower bound in equation 4.17 is described as*

$$\begin{aligned} \lim_{n \rightarrow \infty} n \mathbb{E}[R(\hat{\eta})] &\geq \frac{1}{2} \text{tr}(I + rG^{-1}U)^{-1} \\ &= \frac{1}{2} \left\{ d + 1 + \frac{1}{r} \text{tr}(G_{\theta\theta}^E U_{\theta\theta}^{-1}) \right\} + O(r^{-2}). \end{aligned} \quad (4.19)$$

The proof is in appendix B. The  $n^{-1}$  coefficient of the generalization error 4.15 is described as

$$\frac{1}{2} \left\{ d + 1 + \frac{1}{r} \text{tr}(G_{\theta\theta}^E U_{\theta\theta}^{-1}) \right\} + O(r^{-2}). \quad (4.20)$$

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

In order to illustrate this result, we 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:

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

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 lower bound 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 theoretical analysis and suggests the importance of unlabeled samples in learning with the Fisher kernel.

## 5 Conclusion

In this article, we have investigated several theoretical aspects of the Fisher kernel. One contribution is that we have put the Fisher kernel into the framework of statistical inference by showing the realizability conditions. This allows for subsequent analysis of 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). We examined in this article only one option for feature extraction from marginal models: the combination of the Fisher kernel, a linear classi-

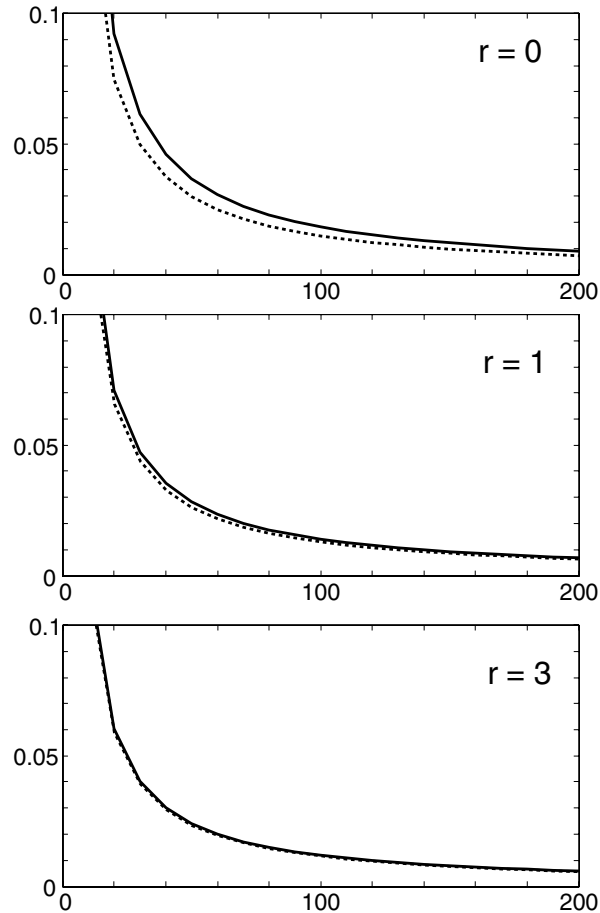


Figure 2: Theoretical learning curves of the Fisher kernel classifier. The horizontal axis shows the number of labeled samples  $n$ , and the vertical axis shows the generalization error  $\mathbb{E}[R(\hat{\eta})]$ . The solid and broken curves correspond to the generalization error of the two-step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

fier, 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, which allows 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.

**Appendix A: Proof of Theorem 2**


---

Let us decompose the matrices  $\Gamma$ ,  $\Lambda$  as

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

The submatrices are computed as follows:

$$\begin{aligned} \Gamma_{\xi\xi} &= \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} \left[ \nabla_{\xi} \sum_{i=1}^n \nabla_{\xi} \log Q(y_i | x_i, \boldsymbol{\eta}^*) \right] \\ &= \mathbb{E}[\nabla_{\xi}^2 \log q(x, y | \boldsymbol{\eta}^*) - \nabla_{\xi}^2 \log q(x | \boldsymbol{\theta}^*)] = -G_{\xi\xi}, \end{aligned}$$

$$\begin{aligned} \Gamma_{\xi\theta} &= \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} \left[ \nabla_{\theta} \sum_{i=1}^n \nabla_{\xi} \log Q(y_i | x_i, \boldsymbol{\eta}^*) \right] \\ &= -G_{\xi\theta}, \end{aligned}$$

$$\begin{aligned} \Gamma_{\theta\xi} &= \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} \left[ \nabla_{\xi} \left\{ \sum_{i=1}^n \nabla_{\theta} \log q(x_i | \boldsymbol{\theta}^*) + \sum_{j=1}^m \nabla_{\theta} \log q(x_j^u | \boldsymbol{\theta}^*) \right\} \right] \\ &= 0, \end{aligned}$$

$$\begin{aligned} \Gamma_{\theta\theta} &= \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} \left[ \nabla_{\theta} \left\{ \sum_{i=1}^n \nabla_{\theta} \log q(x_i | \boldsymbol{\theta}^*) + \sum_{j=1}^m \nabla_{\theta} \log q(x_j^u | \boldsymbol{\theta}^*) \right\} \right] \\ &= -(1+r)U_{\theta\theta}, \end{aligned}$$

$$\begin{aligned} \Lambda_{\xi\xi} &= \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^n \nabla_{\xi} \log Q(y_i | x_i, \boldsymbol{\eta}^*) \sum_{i=1}^n \nabla_{\xi} \log Q(y_i | x_i, \boldsymbol{\eta}^*)^{\top} \right] \\ &= G_{\xi\xi}, \end{aligned}$$

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

$$\Lambda_{\theta\xi} = 0,$$



$$\begin{aligned}\Lambda_{\theta\theta} &= \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} \left[ \left\{ \sum_{i=1}^n \nabla_{\theta} \log q(x_i | \theta^*) + \sum_{j=1}^m \nabla_{\theta} \log q(x_j^u | \theta^*) \right\} \right. \\ &\quad \left. \times \left\{ \sum_{i=1}^n \nabla_{\theta} \log q(x_i | \theta^*) + \sum_{j=1}^m \nabla_{\theta} \log q(x_j^u | \theta^*) \right\}^{\top} \right] \\ &= (1+r)U_{\theta\theta}.\end{aligned}$$

In summary, we have the following:

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

The inverse matrix of  $\Gamma$  becomes

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

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

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

Therefore,

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

By substituting it to equation 4.11, we get the asymptotic expansion of the generalization error as

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

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

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

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

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

where

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

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

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

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

where

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

#### Appendix B: Proof of Lemma 4

In order to prove equation 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:

$$(I + rZ)^{-1} = (I - ZZ^{\dagger}) - \sum_{j=1}^k \left( -\frac{1}{r} Z^{\dagger} \right)^j$$

$$-\left(-\frac{1}{r}Z^\dagger\right)^{k+1}\left(I+\frac{1}{r}Z^\dagger\right)^{-1}, \quad (\text{A.1})$$

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

The proof is described in appendix C. The lower bound in equation 4.17 is rewritten as

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

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

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

where the coefficients are described as

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

Equation 4.19 is proved because the coefficients are derived as follows:

**Lemma 6.** *The coefficients  $\xi_0$  and  $\xi_1$  are described as*

$$\xi_0 = d + 1, \quad (\text{A.4})$$

$$\xi_1 = \text{tr}(G_{\theta\theta}^E U_{\theta\theta}^{-1}), \quad (\text{A.5})$$

respectively.

**Proof.** Since  $q(x | \theta)$  does not depend on  $w$  and  $b$ ,  $U$  is described as

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

Then  $Z$  is rewritten as

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

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

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

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

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

The coefficient  $\xi_0$  is rewritten as

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

where

$$\begin{aligned} \text{tr}(ZZ^\dagger) &= \text{tr}(BB^\top B(B^\top B)^{-2} B^\top) = \text{tr}(B(B^\top B)^{-1} B^\top) \\ &= \text{tr}(B^\top B(B^\top B)^{-1}) = d. \end{aligned}$$

We thus have  $\xi_0 = d + 1$ .  $\xi_1 = \text{tr}(Z^\dagger)$  is rewritten as

$$\begin{aligned} \text{tr}(Z^\dagger) &= \text{tr}[B(B^\top B)^{-2} B^\top] \\ &= \text{tr}(B^\top B)^{-1} \\ &= \text{tr} \left( \begin{bmatrix} 0 & U_{\theta\theta}^{1/2} \end{bmatrix} G^{-1} \begin{bmatrix} 0 \\ U_{\theta\theta}^{1/2} \end{bmatrix} \right)^{-1} \\ &= \text{tr}[U_{\theta\theta}^{-1/2} S_{\theta\theta}^{-1} U_{\theta\theta}^{-1/2}] \\ &= \text{tr}(S_{\theta\theta}^{-1} U_{\theta\theta}^{-1}) = \text{tr}(G_{\theta\theta}^E U_{\theta\theta}^{-1}). \end{aligned}$$

### Appendix 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  $Z$ :

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

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

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

By repeatedly applying equation A.2 to equation A.1, we have

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

## Acknowledgments

---

K.R.M. acknowledges partial financial support by DFG (MU 987/1-1) and and B.M.B.F. under contract FKZ 01IBB02A.

## References

---

- Albert, A. (1972). *Regression and the Moore-Penrose pseudoinverse*. Orlando, FL: 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*. Providence, RI: American Mathematical Society.
- Barndorff-Nielsen, O., & Cox, D. (1989). *Asymptotic techniques for use in statistics*. London: 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*. New York: Pitman Publishing.
- Cox, D., & Hinkley, D. (1974). *Theoretical statistics*. London: Chapman & Hall.
- Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge: Cambridge University Press.
- Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. New York: Springer-Verlag.
- 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*. Englewood Cliffs, NJ: 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 M. Kearns, S. Solla, & D. Cohn (Eds.), *Advances in neural information processing systems*, *11* (pp. 487–493). Cambridge, MA: MIT Press.
- Jaakkola, T., Meila, M., & Jebara, T. (1999). *Maximum entropy discrimination* (Tech. Rep. No. AITR-1668). Cambridge, MA: MIT Press.
- Kawanabe, M., & Amari, S. (1994). Estimation of network parameters in semi-parametric stochastic perceptron. *Neural Computation*, *6*, 1244–1261.
- Malzahn, D., & Oppner, M. (2002). A variational approach to learning curves. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems*, *14* (pp. 463–469). Cambridge, MA: 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*. Cambridge, MA: MIT Press.
- Seeger, M. (2001). *Learning with labeled and unlabeled data* (Tech. Rep.). Edinburgh: Institute for Adaptive and Neural Computation, University of Edinburgh. Available on-line: <http://www.dai.ed.ac.uk/homes/seeger/papers/review.ps.gz>.
- Seeger, M. (2002). Covariance kernels from Bayesian generative models. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems*, *14*, (pp. 905–912). Cambridge, MA: MIT Press.
- 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 T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems*, *14*, (pp. 1197–1204). Cambridge, MA: MIT Press.
- Sonnenburg, S., Rätsch, G., Jagota, A., & Müller, K.-R. (2002). New methods for splice site recognition. In J. Dorransoro (Ed.), *Artificial neural networks—ICANN 2002* (pp. 329–336). New York: Springer-Verlag.
- Sugiyama, M. (2001). *A theory of model selection and active learning for supervised learning*. Unpublished doctoral dissertation, Tokyo Institute of Technology.
- Tsuda, K., & Kawanabe, M. (2002). The leave-one-out kernel. In J. Dorransoro (Ed.), *Artificial neural networks—ICANN 2002* (pp. 727–732). New York: Springer-Verlag.
- Tsuda, K., Kawanabe, M., & Müller, K.-R. (in press). Clustering with the Fisher score. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems*, *15*. Cambridge, MA: MIT Press.

- 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: Cambridge University Press.
- Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.
- Vinokourov, A., & Girolami, M. (2002). A probabilistic framework for the hierarchical 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 P. Langley (Ed.), *Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 1191–1198). San Mateo, CA: Morgan Kaufmann.
- 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.

---

Received February 3, 2003; accepted June 3, 2003.