Classification Error Bound for Low Bayes Error Conditions in Machine Learning

Zijian Yang*†, Vahe Eminyan*, Ralf Schlüter*†, Hermann Ney*†
*Machine Learning and Human Language Technology Group, Lehrstuhl Informatik 6,
Computer Science Department, RWTH Aachen University, Germany
†AppTek GmbH, Germany

***Abstract*—In statistical classification and machine learning, classification error is an important performance measure, which is minimized by the Bayes decision rule. In practice, the unknown true distribution is usually replaced with a model distribution estimated from the training data in the Bayes decision rule. This substitution introduces a mismatch between the Bayes error and the model-based classification error. In this work, we apply classification error bounds to study the relationship between the error mismatch and the Kullback-Leibler divergence in machine learning. Motivated by recent observations of low model-based classification errors in many machine learning tasks, bounding the Bayes error to be lower, we propose a linear approximation of the classification error bound for low Bayes error conditions. Then, the bound for class priors are discussed. Moreover, we extend the classification error bound for sequences. Using automatic speech recognition as a representative example of machine learning applications, this work analytically discusses the correlations among different performance measures with extended bounds, including cross-entropy loss, language model perplexity, and word error rate.**

***Index Terms*—machine learning, classification error bound, speech recognition, mismatch condition**

I. Introduction

In statistical classification and machine learning tasks, such as automatic speech recognition (ASR), Bayes decision rule is used to minimize the classification error, which is the most critical performance measure for these tasks. However, since the true distribution in Bayes decision rule is unknown, in practice, a probabilistic model trained on the training data is applied to approximate the true distribution in Bayes decision rule. Thus, there is a difference between the true distribution of the data and the probabilistic model [1], [2]. While this difference is not addressed in most of the studies, we will make a mathematically strict distinction between true and model distributions in this work. As the classification error mismatch between the Bayes error and the model-based decision error reflects the proximity of model performance to the optimum, we will study the relationship between error mismatch and other statistical measures in this work.

Kullback–Leibler (KL) divergence is another important statistical measure in machine learning tasks. It is closely associated with the cross-entropy (CE) loss and language model (LM) perplexity (PPL). While the correlation between word error rate (WER) and LM PPL has been observed for a long time [3]–[5], most of the works only empirically demonstrate the correlation. In this work, we aim to examine the relationship from an analytical perspective.

The relationship between error mismatch and KL divergence was first investigated in [1]. There, Ney introduced two error bounds on the error mismatch. Later, Nussbaum et al. derived a generalized statistical bound on error mismatch [2], [6], which included the KL divergence as an implicit upper bound of the error mismatch. The bound derived in [2], [6] was proven to be tight when the true distribution is arbitrary. However, when more information about the true distribution is obtained, the bound can be improved. In practice, many systems/tasks have low Bayes errors. For instance, the WER of human speech recognition, is typically low [7], often dropping below 1% for a wide range of conditions [8], indicating a bound on the Bayes error to be lower. In [9], a refined tight bound between error mismatch and KL divergence is derived when Bayes error is lower than a threshold. In this work, we will revisit classification error bounds within the context of machine learning under the low Bayes error condition. Contributions of this work are as follows:

- Simplify the bound with a linear approximation under the low Bayes error condition
- Propose the bound for class priors and verify its tightness
- Extend classification error bounds for sequences

With the extended bounds, correlations among different performance measures including cross-entropy, language model perplexity and word error rate will be discussed.

II. Statistical Measures

In a statistical classification problem, also known as multiple hypothesis testing in information theory, for a joint event (c, x), where $c \in \mathcal{C}$ is a class and $x \in \mathcal{X}$ is a discrete observation, the expected error is defined as:

$$E[c|x] = \sum_{c'} pr(c'|x)\big(1 - \delta(c, c')\big) = 1 - pr(c|x) \quad (1)$$

where δ is Kroneker-Delta and $pr(c|x)$ is the true posterior distribution. The minimum classification error is obtained by the Bayes decision rule:

$$c_*(x) = \underset{c}{\operatorname{argmax}}\, pr(c|x) \quad (2)$$

In practical applications, the true distribution is unknown. Therefore, a model distribution $q(c, x)$ is employed to estimate the true distribution. Recently, sequence-to-sequence

(Seq2Seq) modeling methods have achieved significant success in machine learning tasks, attaining low classification errors on different tasks. [10]–[13]. Instead of modeling the joint distribution $q(c, x)$, these models directly model the posterior $q(c|x)$. For generalization purposes, we consider the modeling of joint distribution $q(c, x)$ in this paper, unless otherwise specified. All the results can be generalized to the posterior modeling by defining $q(c, x) := q(c|x) \cdot pr(x)$. The model-based decision rule is defined as:

$$c_q(x) = \underset{c}{\operatorname{argmax}}\, q(c|x) \tag{3}$$

In statistical classification, the most important performance criterion is the classification error. The global Bayesian and model-based classification errors are then obtained by computing the expectation across all observations:

$$E_* = \sum_x pr(x)E[c_*(x)|x], \quad E_q = \sum_x pr(x)E[c_q(x)|x] \tag{4}$$

In the mismatch problem, we are interested in the global classification error mismatch Δ_q between E_* and E_q:

$$\Delta_q = E_q - E_* = \sum_x pr(x)\big(pr(c_*(x)|x) - pr(c_q(x)|x)\big) \tag{5}$$

Note that $\Delta_q \geq 0$, i.e. E_q is lower bounded by E_*. Therefore, minimizing Δ_q pushes the model toward achieving the optimal classification error.

KL divergence is another statistical measure used to assess the difference between two distributions, which is defined as:

$$D_{\mathrm{KL}}(pr \parallel q) = \sum_{x,c} pr(c, x) \log \frac{pr(c, x)}{q(c, x)} \tag{6}$$

III. Classification Error Bounds

A. Existing Error Bounds

In the field of information theory, instead of the error mismatch Δ_q, the relationship between total variation distance V and KL divergence has been elucidated in the past years. In [14], Vadja proposed a refinement of *Pinsker's* inequality. In machine learning, [15, p.10] introduced the *Bretagnolle-Huber* bound for density estimation. These bounds were not proposed for the error mismatch Δ_q. However, as pointed out in [1] that V is lower bounded by Δ_q, the bounds for Δ_q can be obtained by replacing the total variation distance V in the inequalities with Δ_q. In [1], starting from V, Ney also derived an error bound for $D_{\mathrm{KL}}(pr \parallel q)$ as a function of Δ_q more directly. Nevertheless, unlike V, the error mismatch Δ_q is asymmetric. As a result, introducing V in the derivation leads to a non-tight bound for $\Delta_q \in (0, 1]$. In [2], [6], started directly from Δ_q, the following bound of Δ_q on $D_{\mathrm{KL}}(pr \parallel q)$ is derived, which is a tight bound for the entire domain of Δ_q when the true and model distributions are unconstrained.

$$\begin{aligned} D_{\mathrm{KL}}(pr \parallel q) \geq \frac{1}{2}\big(&(1+\Delta_q)\log(1+\Delta_q) \\ &+ (1-\Delta_q)\log(1-\Delta_q)\big) \\ := g(\Delta_q)& \end{aligned} \tag{7}$$

B. Error bounds with Constraints on E_*

The bound g introduced in (7) is a tight bound when the true distribution pr is unconstrained. However, if the true distribution is subject to some constraints, the bound can be further improved. In machine learning tasks like pattern and speech recognition, the Bayes error is typically low. For instance, WERs of human speech recognition can be below 1% for a wide range of conditions [8], and the Bayes error is bounded to be lower. Under the constraint that $E_* \leq t < 0.5$, where t is a given threshold, the bound can be refined.

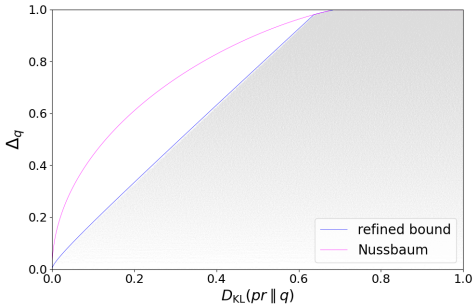

Fig. 1: Comparison of the Nussbaum bound and the refined bound in Theorem 1. The simulation is done under the constraint $E_* \leq 0.01$. The grey dots refer to simulation points.

Theorem 1. *When $E_* \leq t < 0.5$, $D_{KL}(pr \parallel q)$ is lower-bounded by the following function of the mismatch Δ_q,*

$$D_{KL}(pr \parallel q) \geq \underbrace{\begin{cases} (\Delta_q + 2t)g(\frac{\Delta_q}{\Delta_q+2t}), & 0 \leq \Delta_q < 1-2t \\ g(\Delta_q), & 1-2t \leq \Delta_q \leq 1 \end{cases}}_{:=h_t(\Delta_q)} \tag{8}$$

where h_t is the refined bound, and g is defined as in (7).

A detailed proof and the tightness of the bound are derived in our previous work [9]. Figure 1 shows the comparison of the bound derived in [2], [6] and the refined bound in [9], with each grey dot representing the result of a single simulation. The simulation was conducted by generating various distribution pairs (pr, q) until all the reachable areas were covered. All the simulations in this paper are done with $|\mathcal{C}| = 7$ and $|\mathcal{X}| = 15$. The simulation result show that the bound derived in [2], [6] is not tight under the constrained $E_* \leq t$, while the bound in [9] exhibits tightness across the full range of Δ_q.

C. Linear Approximation of the Bound

By computing the derivative of $h_t(\Delta_q)$, it can be observed that when $\Delta_q \gg t$ and within the range $0 \leq \Delta_q \leq 1 - 2t$, the derivative is almost a constant value. Therefore, the bound can be approximated linearly when t is small:

$$D_{\mathrm{KL}}(pr \parallel q) \geq \log(2-2t) \cdot \Delta_q + \beta. \tag{9}$$

where $\beta = t \cdot (\log(1-t) + \log t + 2\log 2)$. Note that since h_t is convex, this linear bound is valid. Figure 2 demonstrates the comparison between the refined bound and its linear approximation. As illustrated in the lower figure for $t = 0.01$,

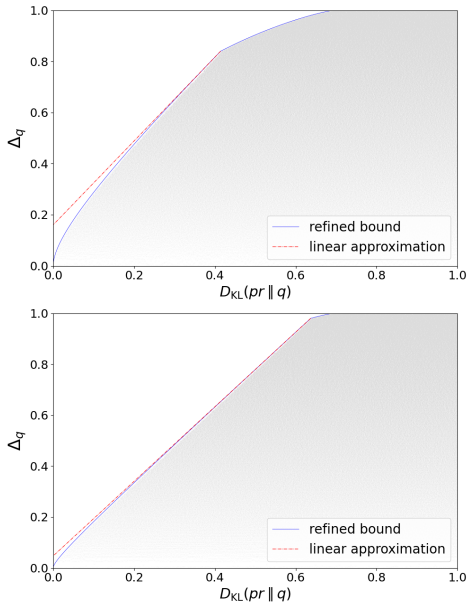

Fig. 2: The linear approximation of the refined bound in Theorem 1. The simulations in the upper figure are under the constraint $E_* \leq 0.08$, and for the lower figure, the constraint is $E_* \leq 0.01$. Grey dots refer to the simulation points.

the exact bound is effectively approximated by the linear bound across nearly the entire domain of Δ_q. However, for $t = 0.08$ in the upper figure, the accuracy of the approximation diminishes when Δ_q is small, as $\Delta_q \gg t$ is not fulfilled.

D. Error Bound for Class Priors

Class prior is important in many statistical classification tasks. For instance, in ASR, an LM (class sequence prior) is usually combined with the acoustic model (AM) to achieve better performance. Therefore, it is crucial to investigate the effect of the class prior. To quantify the discrepancy between the true class prior and the model class prior, the KL divergence between two priors $pr(c)$ and $q(c)$ is employed.

$$D_{\mathrm{KL}}\big(pr(c) \parallel q(c)\big) = \sum_c pr(c) \log \frac{pr(c)}{q(c)}$$

To eliminate the influence of modeling the AM, we assume a perfect acoustic model $q'(x|c) = pr(x|c)$. Effectively, we apply such modeling $q'(c, x) = q'(c)pr(x|c)$ where $q'(c)$ is the model prior for classes. In this case, the KL divergence between joint distributions collapses to between class priors, and the bound proposed in Theorem 1 can be applied:

$$D_{\mathrm{KL}}(pr \parallel q') = \sum_c pr(c) \log \frac{pr(c)}{q'(c)} \geq h_t(\Delta_{q'}) \tag{10}$$

Due to the specific assumption of the joint model distribution $q'(c, x)$, we must reconsider the tightness of the bound. The equality for $\Delta_{q'} \in [0, 1-2t)$ can be achieved with the following parameterized distribution with parameter $\lambda \in [0.5, 1-2t)$:

$$pr(c, x_1) = \begin{cases} 1 - \frac{t}{1-\lambda}, & c = c_1 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

$$pr(c, x_2) = \begin{cases} \frac{t\lambda}{1-\lambda}, & c = c_2 \\ t, & c = c_3 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

$$q'(c) = \lim_{\epsilon \to 0^+} \begin{cases} 1 - \frac{t}{1-\lambda}, & c = c_1 \\ \frac{t}{1-\lambda} \cdot (0.5 - \epsilon), & c = c_2 \\ \frac{t}{1-\lambda} \cdot (0.5 + \epsilon), & c = c_3 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

For $\Delta_{q'} \in [1-2t, 1]$, since the distributions used to achieve equality in [2] meet the condition $q(x|c) = pr(x|c)$, the same distributions can be applied to obtain equality here. Figure 3 presents the simulation result for the KL divergence between class priors and the error mismatch. As shown in the figure, the bound derived for joint distributions also holds for class priors and maintains its tightness.

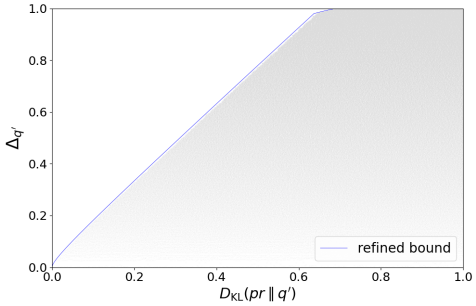

Fig. 3: Simulation results for KL divergence between class priors vs. error mismatch. The simulation is done with $E_* \leq 0.01$. Grey dots refer to the simulation points.

IV. Error Bound for Sequences

Since many machine learning tasks involve Seq2Seq modeling, in this section, we will delve into the application of previously derived bounds within the context of sequence-related scenarios, which bridges the theoretical bound with the practical Seq2Seq machine learning tasks. To simplify the discussion, we assume that all the class sequences have the same length N. The class sequence is defined as c_1^N, while the observation sequence is defined as X. A straightforward way to extend the results from single observations to sequences is to treat the full sequences c_1^N and X as individual events. In this case, Δ_q is computed on sequence level, i.e. sentence error mismatch. However, in Seq2Seq machine learning tasks like ASR, metrics are typically defined at the token or state level for each position. Therefore, a position-wise-defined error function is needed. We consider a position-wise defined error function L for the sequence pair c_1^N and $\tilde{c}_1^N$.

$$L[c_1^N, \tilde{c}_1^N] := \frac{1}{N} \sum_{n=1}^{N} [1 - \delta(c_n, \tilde{c}_n)] \tag{14}$$

The expected error for a given sequence pair (X, c_1^N) is:

$$E[c_1^N|X] = \sum_{\tilde{c}_1^N} pr(\tilde{c}_1^N|X) L[c_1^N, \tilde{c}_1^N] = 1 - \frac{1}{N} \sum_n pr_n(c_n|X) \tag{15}$$

where $pr_n(c|X)$ is the marginal distribution at position n.

$$pr_n(c|X) = \sum_{c_1^N : c_n = c} pr(c_1^N|X) \tag{16}$$

For each position n, the minimum of the expected error is obtained by the following Bayes decision rule:

$$c_*^n(X) = \operatorname*{argmax}_c pr_n(c|X) \tag{17}$$

Consequently, the Bayes decision rule and the corresponding Bayes error for the whole sequence are defined as:

$$\mathbf{c}_*(X) = c_1^N | c_n = c_*^n(X), \overline{E_*} = \sum_X pr(X) E[\mathbf{c}_*(X)|X] \tag{18}$$

The model-based decision rule and classification error can be defined similarly:

$$c_q^n(X) = \operatorname*{argmax}_c q_n(c|X), \ \mathbf{c}_q(X) = c_1^N | c_n = c_q^n(X) \tag{19}$$

$$\overline{E_q} = \sum_X pr(X) E[\mathbf{c}_q(X)|X] \tag{20}$$

The error mismatch for the whole sequence is then computed as follows:

$$\overline{\Delta_q} = \overline{E_q} - \overline{E_*} = \frac{1}{N} \sum_n \Delta_q^n \tag{21}$$

where

$$\Delta_q^n = \sum_X pr(X) \Big[pr_n\big(c_*^n(X)|X\big) - pr_n\big(c_q^n(X)|X\big) \Big] \tag{22}$$

By employing Ineq. (25) in [9] and log-sum inequality, the relationship between the sequence-level KL divergence $D_{\text{KL}}(pr \parallel q)$ and $\overline{\Delta_q}$ can be derived as follows:

$$\begin{aligned}
&\underbrace{D_{\text{KL}}(pr \parallel q) \geq \sum_X pr(X) \sum_{c_1^N} pr(c_1^N|X) \log \frac{pr(c_1^N|X)}{q(c_1^N|X)}}_{\text{Ineq. (25) in [9]}} \\
&= \sum_X pr(X) \frac{1}{N} \sum_n \sum_c \underbrace{\sum_{c_1^N : c_n = c} pr(c_1^N|X) \log \frac{pr(c_1^N|X)}{q(c_1^N|X)}}_{\text{apply log-sum inequality}} \\
&\geq \sum_X pr(X) \sum_c \frac{1}{N} \sum_n pr_n(c|X) \log \frac{pr_n(c|X)}{q_n(c|X)} \\
&= \frac{1}{N} \sum_n \underbrace{\sum_X pr(X) \sum_c pr_n(c|X) \log \frac{pr_n(c|X)}{q_n(c|X)}}_{\text{apply Theorem 1}} \\
&\geq \underbrace{\frac{1}{N} \sum_n h_t(\Delta_q^n)}_{h_t \text{ is convex}} \geq h_t\Big(\frac{1}{N} \sum_n \Delta_q^n\Big) = h_t(\overline{\Delta_q})
\end{aligned} \tag{23}$$

A. Error Bound for the Model Classification Error

KL-divergence can be reformulated in terms of cross-entropy $H(pr, q)$ and entropy $H(pr)$.

$$D_{\text{KL}}(pr \parallel q) = H(pr, q) - H(pr) \tag{24}$$

For a given task, the true distribution, $\overline{E_*}$ and $H(pr)$ are fixed. Therefore, by applying (9) and (23), the model error $\overline{E_q}$ is linear bounded by $H(pr, q)$:

$$\begin{aligned}
&D_{\text{KL}}(pr \parallel q) \geq \log(2 - 2t) \cdot (\overline{E_q} - \overline{E_*}) + \beta \\
&\Rightarrow H(pr, q) \geq \log(2 - 2t) \cdot \overline{E_q} + \text{const}
\end{aligned} \tag{25}$$

B. Error Bound and CE Training Loss

When there is enough data, as discussed in [1], the true distribution can be approximated by the empirical distribution:

$$pr(c_1^N, X) \approx \frac{1}{M} \sum_{m=1}^M \delta(c_1^N, \mathbf{c}_m) \cdot \delta(X, X_m), \tag{26}$$

$$\begin{aligned}
H(pr, q) &= -\sum_{X, c_1^N} pr(c_1^N, X) \log q(c_1^N, X) \\
&\approx -\frac{1}{M} \sum_{m=1}^M \log q(\mathbf{c}_m, X_m)
\end{aligned} \tag{27}$$

where $(\mathbf{c}_m, X_m)$ are sequence pairs in the training data and M is the number of sequences. Substituting the true distribution with the empirical distribution transforms $H(pr, q)$ into the standard CE training loss. The error bound (25) implies that the model error is linearly upper bounded by the CE loss.

C. The Correlation between Word Error Rate and Perplexity

In this section, we investigate the correlation between WER and LM PPL via the derived error bound. Since the exact WER computation involves the alignment problem, which makes the problem much more complicated, we study the averaged Hamming distance instead, which is an upper bound to the WER, if the hypothesis is not longer than the reference. By definition, $\overline{E_q}$ is the error rate when applying Hamming distance as the metric. Similar to the discussion in Section III-D, we assume a perfect acoustic model to eliminate the influence of modeling the AM. In this case, the cross-entropy is $H\big(pr(c_1^N), q'(c_1^N)\big)$, which is effectively the logarithm of the LM PPL. By applying (25), the relationship between LM PPL and WER can be approximately derived as:

$$\log \text{PPL} \geq \log(2 - 2t) \bar{E}_{q'} + \text{const} \geq \log(2 - 2t) \text{WER} + \text{const}$$

This inequality indicates that the WER is linearly upper-bounded by the logarithm of PPL. In [5], a log-linear relationship is observed between PPL and WER. To verify this relationship in theory, refined lower/upper bounds with further constraints on true/model distributions would be needed.

V. Conclusion

In this work, bounds on the mismatch between the Bayes and model classification error based on Kullback–Leibler divergence were discussed for low Bayes error conditions. A linear approximation of the bound was proposed for these low Bayes error conditions. Following discussions on the bound for class priors, classification error bounds were extended for sequences. Based on extended bounds, linear bounds between different performance metrics including cross-entropy loss, language model perplexity, and word error rate were derived.

REFERENCES

[1] H. Ney, "On the relationship between classification error bounds and training criteria in statistical pattern recognition," in *Pattern Recognition and Image Analysis: First Iberian Conference, IbPRIA 2003, Puerto de Andratx, Mallorca, Spain, JUne 4-6, 2003. Proceedings 1*. Springer, 2003, pp. 636–645.

[2] R. Schlüter, M. Nussbaum-Thom, E. Beck, T. Alkhouli, and H. Ney, "Novel tight classification error bounds under mismatch conditions based on f-divergence," in *2013 IEEE Information Theory Workshop (ITW)*. IEEE, 2013, pp. 1–5.

[3] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 179–190, 1983.

[4] J. Makhoul and R. Schwartz, "State of the art in continuous speech recognition." *Proceedings of the National Academy of Sciences*, vol. 92, no. 22, pp. 9956–9963, 1995.

[5] D. Klakow and J. Peters, "Testing the correlation of word error rate and perplexity," *Speech Communication*, vol. 38, no. 1-2, pp. 19–28, 2002.

[6] M. Nussbaum-Thom, E. Beck, T. Alkhouli, R. Schlüter, and H. Ney, "Relative error bounds for statistical classifiers based on the f-divergence." in *Interspeech*, 2013, pp. 2197–2201.

[7] T. Wesker, B. T. Meyer, K. Wagener, J. Anemüller, A. Mertins, and B. Kollmeier, "Oldenburg logatome speech corpus (ollo) for speech recognition experiments with humans and machines." in *Interspeech*. Citeseer, 2005, pp. 1273–1276.

[8] R. P. Lippmann, "Speech recognition by machines and humans," *Speech communication*, vol. 22, no. 1, pp. 1–15, 1997.

[9] Z. Yang, V. Eminyan, R. Schlüter, and H. Ney, "Refined Statistical Bounds for Classification Error Mismatches with Constrained Bayes Error," *arXiv preprint arXiv:2409.01309*, 2024.

[10] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," *arXiv preprint arXiv:2303.03329*, 2023.

[11] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[12] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[13] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[14] A. A. Fedotov, P. Harremoës, and F. Topsoe, "Refinements of pinsker's inequality," *IEEE Transactions on Information Theory*, vol. 49, no. 6, pp. 1491–1498, 2003.

[15] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998.