Habilitation Thesis

BREGMAN DIVERGENCE AND BAYESIAN ESTIMATION

March 3, 2025

Andrew B. Frigyik
Obuda University
Doctoral School of Applied Informatics
and Applied Mathematics

Contents

List of Figures

Chapter 1

Introduction

We learn by association. Our memory is associative. Perhaps it is not surprising that in statistical machine learning probably the most importan question is about similarity. This is the basis of classification: We group entities by how similar or dissimilar they are to each other.

One way to measure similarity is to create an abstract space out of the entities and define some kind of distance between them. A distance or metric function has certain mathematical properties: The distance between any two entities is always non-negative. If this distance is zero the two entities are the same. The aptly called triangle inequality is valid for this kind of function. Finally, the distance of one

entity from another is the same as the distance of the latter from the former. Distance is symmetric.

There is another way approach the question of similarity: We could measure how dissimilar the entities are. This is the idea of divergence. It is quite similar to distance in many respects but in contrast to it, divergence is not symmetric: The roles of the participating entities cannot be interchanged.

All my results are related to the idea of similarity and all of them has sprung into existence while working in a group at the University of Washington. It was a group of engineers for which I was the mathematician in residence, so to speak.

Bregman divergences are a useful set of distortion functions that are popular in statistical estimation and information theory. Their general mathematical properties were known in case of finite dimensional vector spaces. I've extended the notion to infinite dimensional case using and developing tools from functional analysis.

When the entities in question are probabilistic, like distributions, we run into the problem of finiteness: If we wanted to answer the question: how similar to general distributions are, we would have only finitely many data points to base our answer on. It is natural to ask: Given the amount of data, can we estimate the error we are making by choosing any

particular distribution (typically from a preassigned collection)? My second result is giving an answer to this question using a Bayesian approach.

For classification questions quadratic discriminant analysis is a standard approach. The group used the functional Bregman divergence results mentioned before to derive quite a few results. My results here pertain to the mathematical justification of the results: I've proved results about the convergence rate of the classification process and the completeness of the process, namely that the solution that minimizes the functional that is involved in the process is indeed in the collection of admissible distributions.

Mixture models in machine learning are prevalent. One way to work with them is to consider some kind of distribution over them. The most popular one is the Dirichlet distribution, which is a distribution over the simplex of probability mass functions. In practical applications it is often necessary to restrict the domain of the overarching distribution and work with a conditional one. I have developed a systematic way to deal with this problem and was invited to present these results to Microsoft Research Lab in Redmond, WA.

My last result is related to the well-known Stein para-

dox and was motivated by the results our group got while working on the quadratic discriminant analysis. I have given mathematical explanation for the result, that indeed estimates for independent observations, in certain situation, can nevertheless enhance each other (Stein paradox).

In what follows the named citations are works of others and the simple citations are mine.

Chapter 2

Results

1 Functional Bregman Divergence

1.1 Background

In statistical learning the objects that appear in the models are characterised by probability distributions. The learning process relies heavily on comparing the distributions with each other or with some reference distribution. One way to compare these distributions is to measure the distance (closeness) between them. Another one is to measure the divergence (distorsion) between them.

Bregman divergences are a useful set of distortion func-

tions that include squared error, relative entropy, logistic loss, Mahalanobis distance, and the Itakura-Saito function. Bregman divergences are popular in statistical estimation and information theory. Analysis using the concept of Bregman divergences has played a key role in recent advances in statistical learning [Taskar et al., 2006], [Murata et al., 2004], [Collins et al., 2002], [Azoury and Warmuth, 2001], [Lafferty, 1999], [Della Pietra et al., 2001], [Jones and Byrne, 1990], [Gupta et al., 2006], [1], [Banerjee, 2007], clustering [Banerjee et al., 2005b], [Nock and Nielsen, 2006], inverse problems [Le Besnerais et al., 1999], maximum entropy estimation [Altun and Smola, 2006], and the applicability of the data processing theorem [Pardo and Vajda, 1997]. It was discovered that the mean is the minimizer of the expected Bregman divergence for a set of finite dimensional points [Banerjee et al., 2005b], [Banerjee et al., 2005a].

1.2 Results

In papers [2] (conference version) and [3] (journal version) I have achieved the following (where all the proofs and further details can be found):

1.2.1 Thesis point

I defined a functional Bregman divergence that applies to functions and distributions: Let $\phi : L^p(\nu) \to \mathbb{R}$ a strictly convex, twice-continuously Fréchet-differentiable functional. The Bregman divergence $d_\phi : L^p(\nu) \times L^p(\nu) \to [0, \infty)$ is defined to all (admissible) $f, g \in L^p(\nu)$ as

$$d_\phi[f, g] = \phi[f] - \phi[g] - \delta\phi[g; f - g],$$

where $\delta\phi[g; \cdot]$ is the Fréchet derivative of ϕ at g. The functional definition enabled me to solve functional minimization problems using standard methods from the calculus of variations.

1.2.2 Thesis point

I have shown that this new definition is equivalent to Bregman divergence applied to vectors, i.e. I have extended the divergence from finite dimensional spaces to infinite dimensional ones.

Proposition (Proposition I.2)**.** *Functional Bregman divergence generalizes vector Bregman divergence.*

Proposition (Proposition I.3)**.** *Functional definition gener-*

alizes pointwise definition.

The functional definition generalizes a pointwise Bregman divergence that has been previously defined for measurable functions [Jones and Byrne, 1990],
[Csiszár, 1995], and thus extends the class of distortion functions that are Bregman divergences.

1.2.3 Thesis point

I have extended the result on the expectation of vector Bregman divergence [Jones and Byrne, 1990],
[Banerjee et al., 2005a] to show that the mean minimizes the expected Bregman divergence for a set of functions or distributions.

Theorem (Theorem II.1)**.** *Let $\delta^2\phi[f;a,a]$ be strongly positive and let $\phi \in \mathcal{C}^3(L^1(\nu);\mathbb{R})$ be a three-times continuously Fréchet-differentiable functional on $L^1(\nu)$. Let $\mathcal{M}$ be a set of functions that lie on a manifold M, and have associated measure dM such that integration is well defined. Suppose there is a probability distribution P_F defined over the set $\mathcal{M}$. Let $\mathcal{A}$ be a set of functions that includes $E_{P_F}[F]$, (the expectation of $F \in \mathcal{M}$ with respect to the distribution P_F) if it exists. Suppose the function g^* minimizes the expected*

Bregman divergence between the random function F and any function $g \in \mathcal{A}$ such that

$$g^* = \arg\inf_{g\in\mathcal{A}} E_{P_F}[d_\phi(F, g)].$$

Then, if g^* exists, it is given by

$$g^* = E_{P_F}[F].$$

The finite dimensional case relied heavily on the fact that the unit sphere in a finite dimensional vector space is compact in the usual topology. This is never the case in an infinit dimensional topological vector space.

1.2.4 Thesis point

I have shown how this theorem is connected to Bayesian estimation of distributions. For parametric distributions parameterized by $\theta \in \mathbb{R}^n$, a probability measure $\Lambda(\theta)$, and some risk function $R(\theta, \psi)$, $\psi \in \mathbb{R}^n$, the Bayes estimator is defined as

$$\hat{\theta} = \arg\inf_{\psi\in\mathbb{R}^n} \int R(\theta, \psi) d\Lambda(\theta).$$

The principle of Bayesian estimation can be applied to

the distributions themselves rather than to the parameters:

$$\hat{g} = \arg\inf_{g\in\mathcal{A}} \int_M R(f,g) P_F(f) dM$$

where $P_F(f)$ is a probability measure on the distributions $f \in \mathcal{M}$, dM is a measure for the manifold M, and $\mathcal{A}$ is either the space of all distributions or a subset of the space of all distributions, such as the set $\mathcal{M}$. When the set $\mathcal{A}$ includes the distribution $E_{P_F}[F]$ and the risk function R is a functional Bregman divergence, then Theorem II.1 establishes that $\hat{g} = E_{P_F}[F]$. The functional Bregman definition enables stronger results and a more general application.

1.3 Impact

The two papers mentioned above were cited 54 times, according to MTMT (160 times according to Google Scholar, 128 according to Semantic Scholar). The paper is considered to be the go-to paper when it comes to Bregman divergence on infinite dimensional concrete spaces. There is an operator theoretic version of the divergence introduced by Dénes Petz [Petz, 2007], which considers the same idea on abstract spaces.

2 Bayes Error Estimation

2.1 Background

A standard approach in pattern recognition is to estimate the first two moments of each class-conditional distribution from training samples, and then assume that the unknown distributions are Gaussians. Depending on the exact assumptions, this approach is called linear or quadratic discriminant analysis (QDA) [Hastie et al., 2009], [1]. Gaussians are known to maximize entropy given the first two moments [Cover, 1999] and to have other nice mathematical properties. We asked the question: How robust are they with respect to maximizing the Bayes error? To answer that I have investigate the more general question: "What is the maximum possible Bayes error given moment constraints on the class-conditional distributions?" There are a number of other results regarding the optimization of different functionals given moment constraints (e.g., [Dowson and Wragg, 1973], [Poor, 1980], [Ishwar and Moulin, 2005], [Georgiou, 2006], [Friedlander and Gupta, 2005], [Dukkipati et al., 2006], [Lanckriet et al., 2002], [Csiszár et al., 2001], [Csiszár and Matus, 2009]). However, I was not aware of any previous work bounding the maximum Bayes error given mo-

ment constraints. Some related problems are considered by Antos et al. [Antos et al., 1999]; a key difference to their work is that while I've assumed that the moments are given, they instead take as given some independent identically distributed (i.i.d.) samples from the class-conditional distributions, and they then bound the average error of an estimate of the Bayes error.

2.2 Results

In the paper [4] I have achieved the following:

2.2.1 Thesis point

I have found both a lower and an upper bound for the maximum possible Bayes error. The probability that a Bayes classifier is wrong for a given data x is

$$P_e(x) = 1 - \max_{i \in \mathcal{Y}} p(i|x),$$

where $\mathcal{Y}$ is a set of possible classes and $p(i|x)$ is the probability that the classifier will pick class i given the data x. The Bayes error is the expectation of P_e with respect to a given distribution over the data. The lower bound in this

case means that there exists a set of class-conditional distributions that have the given moments and have a Bayes error above the given lower bound. To be more precise:

Theorem (Theorem I)**.** *Suppose that $\mathcal{X} = \mathbb{R}$ and that there exist G class-conditional measures with first and second moments $\{\gamma_{1,i}, \gamma_{2,i}\}$, $i \in \mathcal{Y}$. Given only this set of moments $\{\gamma_{1,i}, \gamma_{2,i}\}$, $i \in \mathcal{Y}$ a lower bound on the supremum Bayes error is*

$$\sup \mathbb{E}\left[P_e\right] \geq (G-1) \sup_{\Delta \in \mathbb{R}} \min_{i \in \mathcal{Y}} \left\{ p(i) \left(1 - \frac{(\gamma_{1,i} - \Delta)^2}{\gamma_{2,i} + \Delta^2 - 2\Delta\gamma_{1,i}} \right) \right\},$$

where the supremum on the left-hand side is taken over all combinations of G class-conditional measures satisfying the moment constraints.

The upper bound means that no set of class-conditional distributions can exist that have the given moments and have a higher Bayes error than the given upper bound. Even for the case $G = 2$ (binary classification) the result becomes rather technical. The details can be found in the corresponding paper.

2.2.2 Thesis point

I have shown that if we are given only the certainty that two equally likely classes have different means (and no trustworthy estimate of their variances), the Bayes error could be 1/2, i.e., the classes may not be separable at all. Using standard notation for the moments in one dimensional case and for binary classification, for example the bound becomes:

$$\sup \mathbb{E}\left[P_e\right] \leq \min \left\{ \frac{4}{4+\frac{\mu_2^2}{\sigma_1^2}}, \frac{1}{2} \right\}.$$

2.2.3 Thesis point

Given the first two moments, my calculation has shown that the popular Gaussian assumption for the class distributions is fairly optimistic — the true Bayes error could be much worse (see Fig. 2.1).

2.3 Impact

There are only three (3) papers citing our research. Since our finding critiques a very popular and wide-spread method the outcome is not that surprising. One of the papers acknowledges our results but comes up with an alternative approach.

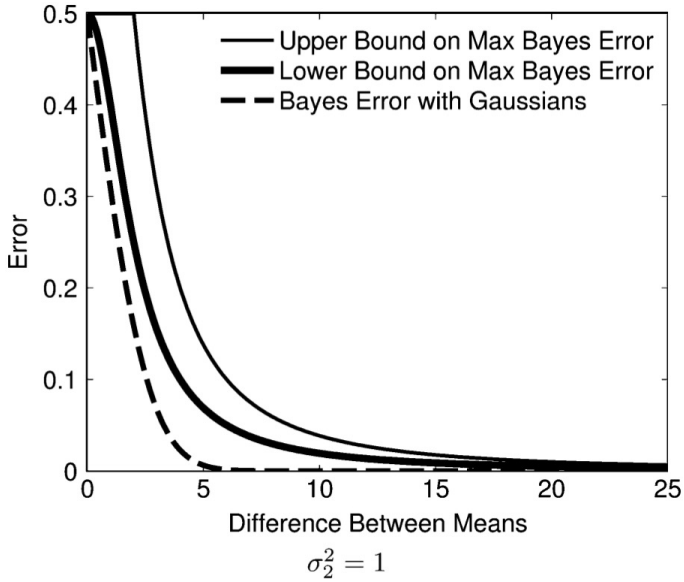

Figure 2.1: Comparison of the given lower bound for the worst case Bayes error with the Bayes error produced by Gaussian class-conditional distributions.

Their method has two orders of magnitude higher impact (measured by citations) then ours. The other two papers are about applications and since the non-parametric model fits them better our approach is considered there.

3 Bayesian Quadratic Discriminant Analysis

3.1 Background

A standard approach to supervised classification problems is quadratic discriminant analysis (QDA), which models the likelihood of each class as a Gaussian distribution, then uses the posterior distributions to estimate the class for a given test point [Hastie et al., 2009]. The Gaussian parameters for each class can be estimated from training points with maximum likelihood (ML) estimation. The simple Gaussian model assumption is best suited to cases when one does not have much information to characterize a class, for example, if there are too few training samples to infer much about the class distributions. Unfortunately, when the number of training samples is small compared to the number of dimensions of each training sample, the ML covariance estimation can be ill-posed.

One approach to resolve the ill-posed estimation is to regularize the covariance estimation; another approach is to use Bayesian estimation. Bayesian estimation for QDA was first proposed by [Geisser, 1964], but this approach has not

become popular, even though it minimizes the expected misclassification cost.
[Ripley, 2007] in his text on pattern recognition states that such predictive classifiers are mostly unmentioned in other texts and that "this may well be because it usually makes little difference with the tightly constrained parametric families." [Geisser, 2017] examines Bayesian QDA in detail, but does not show that in practice it can yield better performance than regularized QDA. In preliminary experiments we found that the performance of Bayesian QDA classifiers is very sensitive to the choice of prior
[Srivastava and Gupta, 2006], and that priors suggested by [Geisser, 1964] and [Keehn, 1965] produce error rates similar to those yielded by ML.

3.2 Results

In the paper [1] (where the detailed explanation and the proof can be found) I have achieved the following:

3.2.1 Thesis point

Using the previously developed functional Bregman divergence I have contributed to the effort of finding the Bayesian

classifier, called **BDA7**, in terms of the Gaussian distributions themselves, as opposed to the standard approach of formulating the problem in terms of the Gaussian parameters. To be more precise:

Theorem. *Let*

$$\mathcal{A}_p = \left\{ a : \mathbb{R}^d \to [0,1] \;\middle|\; a \in L^p(\nu), a > 0, \|a\|_p = 1 \right\}$$

for some measure ν. Let $\phi : \mathcal{A}_1 \to \mathbb{R}$, $\phi \in \mathcal{C}^3$, and $\delta^2\phi[f;\cdot,\cdot]$ be strongly positive. Suppose the function $\hat{f}_h$ minimizes the expected Bregman divergence between a random Gaussian N_h and any probability density function $f \in \mathcal{A}_1$ where the expectation is taken with respect to the distribution $r(\mathcal{N}_h)$, such that

$$\hat{f}_h = \operatorname*{argmin}_{f \in \mathcal{A}_1} E_{N_h}[d_\phi(N_h, f)].$$

Then $\hat{f}_h$ is given by

$$\hat{f}_h = \int_M \mathcal{N}_h r(\mathcal{N}_h) dM = E_{N_h}[N_h(\cdot)].$$

This distribution-based Bayesian discriminant analysis removes the parameter-based Bayesian analysis restriction of requiring more training samples than feature dimensions,

and removes the question of invariance to transformations of the parameters, because the estimate is defined in terms of the Gaussian manifold itself. Essentially I helped to derive a coordinate-free version of the Bayesian classifier.

3.3 Impact

The paper has 137 independent citations, according to MTMT (318 according to Google Scholar and 234 according to Semantic Scholar). The results we obtained are used both in theoretical and application oriented papers. From theory point of view the paper serves as a reference to Bayesian QDA or starting point for models. From application point of view the paper in many cases considered to be a standard which the new result are compared to. In almost all the standard tests the method performed better (see Fig. 2.2).

4 Shadow Dirichlet Distribution

4.1 Background

Modeling probability mass functions as random vectors is useful in solving many real-world problems. A common random model for probability mass functions is the Dirichlet

	BDA7	QB	RDA	EDDA	NM	LDA	QDA
Cover Type	48.3	**44.6**	59.2	83.1	78.5	62.9	62.9
Heart	**30.6**	38.5	38.2	33.9	39.9	35.9	44.5
Image Segmentation	**12.7**	**12.9**	18.3	29.7	29.5	85.7	85.7
Ionosphere	16.9	16.1	**12.5**	26.0	27.1	35.8	35.8
Iris	**6.9**	**7.6**	**8.1**	9.4	9.3	8.8	66.6
Pen Digits	**6.1**	8.1	8.5	7.7	23.9	17.8	23.8
Pima	29.7	32.7	29.4	30.7	35.8	**27.6**	33.4
Sonar	**36.8**	40.4	40.4	39.8	42.6	46.7	46.7
Thyroid	**11.7**	14.8	17.0	14.7	19.2	15.0	30.0
Wine	**9.6**	33.1	34.2	**11.2**	32.0	60.1	60.1

Figure 2.2: Average Error Rate Using Randomly Selected 5% of Training Samples

distribution [5]. The Dirichlet is conjugate to the multinomial and hence mathematically convenient for Bayesian inference, and the number of parameters is conveniently linear in the size of the sample space. However, the Dirichlet is a distribution over the entire probability simplex, and for many problems this is simply the wrong domain if there is application-specific prior knowledge that the probability mass functions come from a restricted subset of the simplex. This was essentially my project, everything here was done by me.

4.2 Results

In the paper [6] I have achieved the following:

4.2.1 Thesis point

I have proposed a simple variant of the Dirichlet distribution whose support is a subset of the simplex, explored its properties, and shown how to learn the model from data.

Definition 1. Let $\mathcal{S}$ be the probability $(d-1)$-simplex, and let $\tilde{\Theta} \in \mathcal{S}$ be a random probability mass function drawn from a Dirichlet distribution with density p_D and un-normalized parameter $\alpha \in \mathbb{R}_+^d$. Then we say the random probability mass function $\Theta \in \mathcal{S}$ is distributed according to a shadow Dirichlet distribution if $\theta = M\tilde{\Theta}$ for some fixed $d \times d$ left-stochastic full-rank matrix M, and we call $\tilde{\Theta}$ the *generating Dirichlet* of Θ, or Θ's *Dirichlet shadow*.

This is essentially an attempt to define conditional probability for certain situations. The motivation for the definition is that it is much easier to draw samples from, and compute maximal likelihood estimates and maximal likelihood compound estimates for a shadow Dirichlet distribution than from and for a regular renormalized Dirichlet distribution.

4.2.2 Thesis point

I have shown that if the restriction is over a set of regularized probability mass functions, i.e.

$$\mathcal{S}_M = \{\theta \;\Big|\; \theta = \lambda\tilde{\theta} + (1-\lambda)\check{\theta}, \tilde{\theta} \in \mathcal{S}\}$$

for some given λ and $\check{\theta}$ then the following left-stochastic matrix

$$M = (1-\lambda)\check{\theta}\mathbf{1}^T + \lambda I,$$

where I is the identity matrix and $\mathbf{1}$ is the vector full of 1s is optimal in certain sense:

Theorem (Corollary 4.1)**.** *Let $\mathcal{M}_{reg}$ be the set of left-stochastic matrices M that parameterize shadow Dirichlet distributions with support in $\mathcal{S}_M = \{\theta \;\Big|\; \theta = \lambda\tilde{\theta} + (1-\lambda)\check{\theta}, \tilde{\theta} \in \mathcal{S}\}$, for a specific choice of λ and $\check{\theta}$. Then the M given above results in the shadow Dirichlet with maximum entropy, that is, it solves*

$$\arg \max_{M \in \mathcal{M}_{reg}} h(p_M),$$

where h is the Shannon entropy.

A similar result holds if the restriction is to monotonic probability mass functions.

4.3 Impact

This paper was cited 5 times, according to Google Scholar and Semantic Scholar, and 0 times according to MTMT. One of the citings recognized the method used in our paper the same as the one they came up with. An other paper cites our model as an inspiration for the one they have proposed. I would like to mention that the course material about the Dirichlet distribution and the related process [5] that was written to accompany the paper is much more popular: According to Semantic Scholar it received 216 citations (277 according to Google Scholar and 39 according to MTMT).

This paper is more like a musing on a standard question of conditional probability in this very specific setting. Not many were interested in this particular question and from the reaction it seems that the classical approach is still much more preferable. Nevertheless, I was invited to give a talk about the idea to Microsoft Research Lab in Redmond, WA, where the audience was enthusiastic.

5 Multi-Task Averaging

5.1 Background

The mean is one of the most fundamental and useful tools in statistics [Salsburg, 2001]. By the 16th century Tycho Brahe was using the mean to reduce measurement error in astronomical investigations [Plackett, 1958]. [Legendre, 1806] noted that the mean minimizes the sum of squared errors to a set of samples. More recently it has been shown that the mean minimizes the sum of any Bregman divergence to a set of samples ([Banerjee et al., 2005b]; [3]). This is what started us on looking further into the mean. The mean is a more subtle quantity than it first appears. In a surprising result popularly called Stein's paradox
([Efron and Morris, 1977]), [Stein, 1955] showed that it is better (in a summed squared error sense) to estimate each of the means of several Gaussian random variables using data sampled from all of them, even if the random variables are independent and have different means. That is, it is beneficial to consider samples from seemingly unrelated distributions to estimate a mean. Stein's result is an early example of the motivating hypothesis behind multi-task learning: that leveraging data from multiple tasks can yield superior per-

formance over learning from each task independently. Simulations (performed by others in the group) shown that multi-task learning indeed can yield better estimates than individual averages can.

5.2 Results

In papers [7] (conference version) and [8] (journal version) I have achieved the following:

5.2.1 Thesis point

I have considered a multi-task regularization approach to the problem of estimating multiple means. The group called it multi-task averaging. The proposed multi-task averaging objective is

$$\{Y_t^*\}_{t=1}^T = \underset{\{\hat{Y}\}_{t=1}^T}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \frac{(Y_{ti} - \hat{Y}_t)^2}{\sigma_t^2} + \frac{\gamma}{T^2} \sum_{r=1}^{T} \sum_{s=1}^{T} A_{rs} (\hat{Y}_r - \hat{Y}_s)^2,$$

where Y_t^* is the multi-task averaging estimate for the tth task, T is the number of tasks, N_t is the number of samples from the tth task, Y_{ti} is the ith sample from the tth task, σ_t^2 is the variance of the tth task, γ is the regularization parameter and $A = (A_{rs})$ is a $T \times T$ matrix that describes how similar the tasks are distribution-wise. Multi-task learning is most intuitive when the multiple tasks are conceptually similar.

5.2.2 Thesis point

I have shown that MTA has provably nice theoretical properties. Among other things, the multi-task averaging estimate has a closed form solution:

$$Y^* = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1} \bar{Y},$$

where I is the identity, $\bar{Y}$ is the task sample average, Σ is the diagonal covariance matrix of $\bar{Y}$ and L is the graph Laplacian of the matrix A. Using Gershgorin's Circle Theorem I showed that the inverse exist if A and γ are non-negative.

5.2.3 Thesis point

I have shown that more can be said about the multi-task averaging estimates:

Theorem. *If* $\gamma \geq 0$, $0 \leq A_{rs} < \infty$ *for all* r, s *and* $0 < \frac{\sigma_t^2}{N_t} < \infty$ *for all* t, *then the multi-task averaging estimates* $\{Y_t^*\}$ *are a convex combination of the task sample averages* $\{\bar{Y}_t\}$.

5.2.4 Thesis point

I have found the optimal task-relatedness matrix $A = [0a; a0]$ in case of $T = 2$, i.e. when we have two tasks. I showed that the optimal value for a is

$$a^* = \frac{2}{\Delta^2},$$

where Δ is the difference between the task sample averages. It is optimal in the sence that the sum of the mean square errors of the individual components of the multi-task averaging estimates is minimal.

5.3 Impact

The two papers were cited 9 times, according to MTMT (37 times according to Google Scholar and 26 times according to Semantic Scholar). Starting with the journal paper: Our work made it into a book and into a highly popular survey paper. Some of the other citations were either using or mod-

ifying the results we published. Others cited our work as alternative approach to the problem at hand. Some of the papers applied the method in multi-task learning.

Considering the conference paper: Some of the papers cite both. After all, the conference paper was almost exclusively practical, while the journal paper has the theoretical underpinning for the method we have proposed. The effect of this paper very similar to the one that was published in the journal. A few papers use our method directly, some use the theoretical results and some talk about the method as a potential alternative.

Own Papers

[1] Santosh Srivastava, Maya R Gupta, and Béla A Frigyik. Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research*, 8(6), 2007.

[2] Bela A Frigyik, Santosh Srivastava, and Maya R Gupta. Functional bregman divergence. In *2008 IEEE Interna-*

tional Symposium on Information Theory, pages 1681–1685. IEEE, 2008.

[3] Béla A Frigyik, Santosh Srivastava, and Maya R Gupta. Functional bregman divergence and bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54(11):5130–5139, 2008.

[4] Bela A Frigyik and Maya R Gupta. Bounds on the bayes error given moments. *IEEE Transactions on Information Theory*, 58(6):3606–3612, 2012.

[5] Bela A Frigyik, Amol Kapila, and Maya R Gupta. Introduction to the dirichlet distribution and related processes. Technical Report 0006, Department of Electrical Engineering, University of Washignton, 2010. UWEETR-2010-0006.

[6] Bela Frigyik, Maya Gupta, and Yihua Chen. Shadow dirichlet for restricted probability modeling. *Advances in Neural Information Processing Systems*, 23, 2010.

[7] Sergey Feldman, Maya Gupta, and Bela Frigyik. Multi-task averaging. *Advances in Neural Information Processing Systems*, 25, 2012.

[8] Sergey Feldman, Maya R Gupta, and Bela A Frigyik. Revisiting Stein's paradox: multi-task averaging. *J. Mach. Learn. Res.*, 15(1):3441–3482, 2014.

Bibliography

[Altun and Smola, 2006] Altun, Y. and Smola, A. (2006). Unifying divergence minimization and statistical inference via convex duality. In *International Conference on Computational Learning Theory*, pages 139–153. Springer.

[Antos et al., 1999] Antos, A., Devroye, L., and Gyorfi, L. (1999). Lower bounds for bayes error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(7):643–645.

[Azoury and Warmuth, 2001] Azoury, K. S. and Warmuth, M. K. (2001). Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246.

[Banerjee, 2007] Banerjee, A. (2007). An analysis of logistic models: Exponential family connections and online performance. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 204–215. SIAM.

[Banerjee et al., 2005a] Banerjee, A., Guo, X., and Wang, H. (2005a). On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669.

[Banerjee et al., 2005b] Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. (2005b). Clustering with bregman divergences. *Journal of machine learning research*, 6(10).

[Collins et al., 2002] Collins, M., Schapire, R. E., and Singer, Y. (2002). Logistic regression, adaboost and bregman distances. *Machine Learning*, 48(1):253–285.

[Cover, 1999] Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

[Csiszár, 1995] Csiszár, I. (1995). Generalized projections for non-negative functions. In *Proceedings of 1995 IEEE International Symposium on Information Theory*, page 6. IEEE.

[Csiszár and Matus, 2009] Csiszár, I. and Matus, F. (2009). On minimization of multivariate entropy functionals. In *2009 IEEE Information Theory Workshop on Networking and Information Theory*, pages 96–100. IEEE.

[Csiszár et al., 2001] Csiszár, I., Tusnady, G., Ispany, M., Verdes, E., Michaletzky, G., and Rudas, T. (2001). Divergence minimization under prior inequality constraints. In *IEEE INTERNATIONAL SYMPOSIUM ON INFORMATION THEORY*, pages 21–21.

[Della Pietra et al., 2001] Della Pietra, S., Della Pietra, V., and Lafferty, J. (2001). Duality and auxiliary functions for bregman distances. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE.

[Dowson and Wragg, 1973] Dowson, D. and Wragg, A. (1973). Maximum-entropy distributions having prescribed first and second moments (corresp.). *IEEE Transactions on Information Theory*, 19(5):689–693.

[Dukkipati et al., 2006] Dukkipati, A., Murty, M. N., and Bhatnagar, S. (2006). Nonextensive triangle equality and other properties of tsallis relative-entropy minimiza-

tion. *Physica A: Statistical Mechanics and its Applications*, 361(1):124–138.

[Efron and Morris, 1977] Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236(5):119–127.

[Friedlander and Gupta, 2005] Friedlander, M. P. and Gupta, M. R. (2005). On minimizing distortion and relative entropy. *IEEE Transactions on Information Theory*, 52(1):238–245.

[Geisser, 1964] Geisser, S. (1964). Posterior odds for multivariate normal classifications. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(1):69–76.

[Geisser, 2017] Geisser, S. (2017). *Predictive inference: an introduction*. Chapman and Hall/CRC.

[Georgiou, 2006] Georgiou, T. T. (2006). Relative entropy and the multivariable multidimensional moment problem. *IEEE Transactions on Information Theory*, 52(3):1052–1066.

[Gupta et al., 2006] Gupta, M., Srivastava, S., and Cazzanti, L. (2006). Optimal estimation for nearest neigh-

bor classifiers. *review by the IEEE Trans. on Information Theory*.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

[Ishwar and Moulin, 2005] Ishwar, P. and Moulin, P. (2005). On the existence and characterization of the maxent distribution under general moment inequality constraints. *IEEE transactions on information theory*, 51(9):3322–3333.

[Jones and Byrne, 1990] Jones, L. K. and Byrne, C. L. (1990). General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE transactions on Information Theory*, 36(1):23–30.

[Keehn, 1965] Keehn, D. (1965). A note on learning for gaussian properties. *IEEE Transactions on Information Theory*, 11(1):126–132.

[Lafferty, 1999] Lafferty, J. (1999). Additive models, boosting, and inference for generalized divergences. In *Proceed-*

ings of the twelfth annual conference on Computational learning theory, pages 125–133.

[Lanckriet et al., 2002] Lanckriet, G. R., Ghaoui, L. E., Bhattacharyya, C., and Jordan, M. I. (2002). A robust minimax approach to classification. *Journal of Machine Learning Research*, 3(Dec):555–582.

[Le Besnerais et al., 1999] Le Besnerais, G., Bercher, J.-F., and Demoment, G. (1999). A new look at entropy for solving linear inverse problems. *IEEE Transactions on Information Theory*, 45(5):1565–1578.

[Legendre, 1806] Legendre, A. M. (1806). *Nouvelles méthodes pour la détermination des orbites des comètes*, chapter Supplément. Courcier.

[Murata et al., 2004] Murata, N., Takenouchi, T., Kanamori, T., and Eguchi, S. (2004). Information geometry of u-boost and bregman divergence. *Neural Computation*, 16(7):1437–1481.

[Nock and Nielsen, 2006] Nock, R. and Nielsen, F. (2006). On weighting clustering. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1223–1235.

[Pardo and Vajda, 1997] Pardo, M. and Vajda, I. (1997). About distances of discrete distributions satisfying the data processing theorem of information theory. *IEEE transactions on information theory*, 43(4):1288–1293.

[Petz, 2007] Petz, D. (2007). Bregman divergence as relative operator entropy. *Acta Mathematica Hungarica*, 116(1-2):127–131.

[Plackett, 1958] Plackett, R. L. (1958). Studies in the history of probability and statistics: Vii. the principle of the arithmetic mean. *Biometrika*, 45(1-2):130–135.

[Poor, 1980] Poor, H. V. (1980). Minimum-distortion functional for one-dimensional quantisation. *Electronics Letters*, 1(16):23–25.

[Ripley, 2007] Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press.

[Salsburg, 2001] Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. Macmillan.

[Srivastava and Gupta, 2006] Srivastava, S. and Gupta, M. R. (2006). Distribution-based bayesian minimum ex-

pected risk for discriminant analysis. In *2006 IEEE international symposium on information theory*, pages 2294–2298. IEEE.

[Stein, 1955] Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal population. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol 1*, pages 197–206. University of California Press Berkeley.

[Taskar et al., 2006] Taskar, B., Lacoste-Julien, S., Jordan, M. I., Bennett, K. P., and Parrado-Hernández, E. (2006). Structured prediction, dual extragradient and bregman projections. *Journal of Machine Learning Research*, 7(7).