Bayesian Estimation for Non-Standard Loss Functions Using a Parametric Family of Estimators

Stefan Uhlich, *Student Member, IEEE,* and Bin Yang, *Senior Member, IEEE,*

Abstract

Bayesian estimation with other loss functions than the standard hit-or-miss loss or the quadratic loss often yields optimal Bayesian estimators (OBE)s that can only be formulated as optimization problems and which have to be solved for each new observation. The contribution of this paper is to introduce a new parametric family of estimators to circumvent this problem. By restricting the estimator to lie in this family, we split the estimation problem into two parts: In a first step, we have to find the best estimator with respect to the Bayes risk for a given non-standard loss function, which has to be done only once. The second step then calculates the estimate for an observation using importance sampling. The computational complexity of this second step is therefore comparable to that of an MMSE estimator if the MMSE estimator also uses Monte Carlo integration. We study the proposed parametric family using two examples and show that the estimator family gives for both a good approximation of the OBE.

Index Terms

Optimal Bayesian estimator, Bayesian estimation, Loss function, Parametric estimator family

I. Introduction

It is well known that the goal of Bayesian estimation is to find the estimator that minimizes the Bayes risk for a given loss function. The loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \geq 0$ assigns a loss to the estimate $\hat{\boldsymbol{\theta}}$ when the

S. Uhlich and B. Yang are with the Chair of System Theory and Signal Processing at the Universität Stuttgart, Pfaffenwaldring 47, 70550 Stuttgart, Germany. E-mail: stefan.uhlich@Lss.uni-stuttgart.de; bin.yang@Lss.uni-stuttgart.de; Parts of this paper appeared in [1].

EDICS: SSP-PARE

correct value is $\boldsymbol{\theta}$ and thereby reflects the cost that is connected to a certain estimation error. It plays a central role in designing the Bayes estimator and should be application-dependent, i.e. should incorporate the knowledge of the specific problem that one wants to solve [2]–[4]. However, the most often used loss functions are the hit-or-miss loss and the quadratic loss

$$L_{\text{MAP}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \begin{cases} 1 & \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\| > \delta \\ 0 & \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\| < \delta \end{cases}, \ \delta \to 0^+ \quad \text{and} \tag{1a}$$

$$L_{\text{MMSE}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbf{W} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}), \ \mathbf{W} \text{ pos. def.} \tag{1b}$$

where it is known that the corresponding optimal Bayesian estimators (OBE)s are the MAP and MMSE estimators [5]. The reason that they are used so widely is often not their suitability to the problem at hand but that the corresponding OBEs are well known and, at least for the MAP estimator, are often computable. They are the maximum and mean of the a posteriori density $p(\boldsymbol{\theta}|\mathbf{x})$. Powerful methods are available to calculate the estimate $\hat{\boldsymbol{\theta}}$ from an observation $\mathbf{x}$, ranging from optimization algorithms [5] and the Expectation-Maximization (EM) algorithm [6] to sampling techniques including Markov chain Monte Carlo (MCMC) methods [7].

In this paper, we will consider Bayesian estimation with other loss functions than those given in (1). This problem is very important for practical applications as the following two examples illustrate:

- Consider the problem of constructing a dam [8]. Underestimating the peak water level from older measurements is clearly more serious than overestimating it and this fact should be reflected in the choice of the loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$. This example motivates the use of an asymmetric loss function, i.e. $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \neq L(-\boldsymbol{\theta}, -\hat{\boldsymbol{\theta}})$, and it is obvious that the two loss functions in (1) are not suited for such an estimation problem.
- Another example that gives rise to other loss functions than those given in (1) can be found in the field of image processing. Traditionally, the mean squared error is used to compare images and therefore many algorithms are optimized for this loss function [9]. The problem with the MSE is that it does not well represent human perception. Images which have a small mean squared error may still look very different and therefore in [9] it is suggested to use other distance measures. One is the structural similarity (SSIM) index, which was introduced by Wang in [10] and e.g. used in [11] for the design of linear equalizers. Fig. 1 compares the MSE with the SSIM index and it can clearly be seen that the SSIM index is a better measure of similarity than the MSE with respect to human perception.

(a) Original image (b) MSE = 150, SSIM = 0.83 (c) MSE = 150, SSIM = 0.66

Fig. 1: Comparison of MSE with SSIM for different image operations
((b) = salt and pepper noise, (c) = Gaussian blurring)

> A related example that discusses the design of loss functions for the reconstruction of images is given by Rue in [12]. He shows how information about the image structure can be used to find a suitable loss function and he proposes the use of MCMC and simulated annealing methods to calculate the Bayesian estimates.

More examples of Bayesian estimation with non-standard loss functions can also be found in cluster analysis [13]–[15] and mixture modeling [16], [17].

However, calculating the OBE for many non-standard loss functions is not trivial and can often only be stated in terms of an optimization problem which has to be solved for each new observation $\mathbf{x}$. Therefore, we propose in this paper a parametric family $\mathcal{F}$ of estimators which are suited for a large variety of loss functions but still have a computational complexity comparable to the MMSE estimator for the same problem. Thus, using the best estimator in $\mathcal{F}$ that has the smallest Bayes risk for a given loss function will be a good approximation of the OBE. Our parametric family of estimators can be viewed as a compromise between the perfect OBE on one side and a (nonlinear) regression approach on the other. It trades off performance against computational complexity as it will have a larger Bayes risk than the OBE but will be easier to learn due to the small and fixed number of parameters compared to a regression approach.

This paper is organized as follows: First, we review in Sec. II the Bayesian estimation problem and introduce the OBE which minimizes the Bayes risk. As the OBE can often not be computed in closed form, we propose in Sec. III and IV two new parametric families of estimators. We start in Sec. III by considering a basic family $\mathcal{F}_{\mathcal{B}}$ of estimators, which includes the MMSE and the MAP estimator. This family, however, has the disadvantage that the underlying loss functions are always symmetric.

Therefore, we generalize the estimator family in Sec. IV. This generalized family $\mathcal{F}$ also includes the OBE for the linear-exponential (LinEx) loss [18] and is thus more versatile. In Sec. V we consider the general approach how to use the estimator family and discuss its computational complexity. We show that we can use importance sampling to efficiently compute an estimate. Finally, two examples in Sec. VI demonstrate the usefulness of our parametric family to approximate the OBE. The first example studies our family of estimators for a bounded LinEx loss problem whereas the second example considers the task of speech enhancement using a perceptual relevant loss function, namely the PESQ measure.

The following notation is used throughout this paper: $\mathbf{x}$ denotes a column vector, $\mathbf{X}$ a matrix and in particular $\mathbf{I}$ the identity matrix. The trace operator, determinant, matrix transpose and Euclidean norm are denoted by $\mathrm{tr}\{.\}$, $\det\{.\}$, $(.)^T$ and $\|.\|$, respectively. $\mathrm{diag}\{\mathbf{x}\}$ returns a diagonal matrix whose diagonal elements are given by $\mathbf{x}$. Finally, $\mathbf{X} \circ \mathbf{Y}$ denotes the elementwise product, also known as Hadamard product.

II. Review of Bayesian Estimation

In this section, we will briefly review the basic elements of Bayesian estimation which we will need throughout this paper. For a more detailed introduction, the interested reader is referred to [2], [19].

Suppose we have an estimator $\hat{\boldsymbol{\theta}}(\mathbf{x})$ that estimates the unknown, random parameter $\boldsymbol{\theta} \in \mathbb{R}^M$ from the observation $\mathbf{x} \in \mathbb{R}^N$. To evaluate the quality of the estimator, we assign a loss $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \geq 0$ to the error of estimating $\hat{\boldsymbol{\theta}}(\mathbf{x})$ although the true value is $\boldsymbol{\theta}$. If $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ exhibits the relationship $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = L(-\boldsymbol{\theta}, -\hat{\boldsymbol{\theta}})$ then it is called symmetric.[1]

Averaging the loss with respect to the joint probability density function (PDF) $p(\boldsymbol{\theta}, \mathbf{x})$ yields an important characteristic value for an estimator. It is called the Bayes risk (BR) and given by [19]

$$\mathrm{BR} = \iint L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{x})) p(\boldsymbol{\theta}, \mathbf{x}) d\boldsymbol{\theta} d\mathbf{x}. \tag{2}$$

The optimal Bayesian estimator (OBE) is now that estimator that minimizes the Bayes risk, i.e.

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\mathrm{OBE}}(\mathbf{x}) &= \arg\min_{\hat{\boldsymbol{\theta}}(\mathbf{x})} \mathrm{BR} = \arg\min_{\hat{\boldsymbol{\theta}}(\mathbf{x})} \iint L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{x})) p(\boldsymbol{\theta}, \mathbf{x}) d\boldsymbol{\theta} d\mathbf{x} \\ &= \arg\min_{\hat{\boldsymbol{\theta}}(\mathbf{x})} \int L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{x})) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \end{aligned} \tag{3}$$

where we used in the last line of (3) the fact that $p(\mathbf{x}) \geq 0$ and therefore it is sufficient to minimize the inner integral for each $\mathbf{x}$. Hence, $\arg\min_{\hat{\boldsymbol{\theta}}} \mathrm{BR}$ is equivalent to minimizing the loss averaged over the a

[1]Besides this property, scale invariance [20], [21] and boundedness [22], [23] are other characteristics of the loss function that may be desired for practical applications.

posteriori distribution. Therefore, we immediately see that all information to find the OBE is included in the a posteriori density $p(\boldsymbol{\theta}|\mathbf{x})$.

Assuming that the loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is differentiable, we can calculate the first-order derivative with respect to the estimate and equate it to zero to obtain a necessary condition[2] to find the OBE, i.e.

$$\frac{\partial}{\partial \hat{\boldsymbol{\theta}}} \int L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{x})) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} = \int \frac{\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \stackrel{!}{=} \mathbf{0}. \tag{4}$$

Solving (4) can often not be done analytically and therefore Bayesian estimation with most loss functions is difficult. We will thus introduce in the next section a parametric family $\mathcal{F}_\mathcal{B}$ of estimators that will transform (3) into an optimization problem to find one parameter. This family is then extended in Sec. IV to asymmetric loss functions.

III. Basic Family of Estimators

The first set of estimators that we consider are all estimators of the form

$$\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda) = \frac{\int \boldsymbol{\theta} p(\boldsymbol{\theta}, \mathbf{x})^\lambda d\boldsymbol{\theta}}{\int p(\boldsymbol{\theta}, \mathbf{x})^\lambda d\boldsymbol{\theta}} \tag{5}$$

and which are parameterized by λ. We call this set the *basic family of estimators* $\mathcal{F}_\mathcal{B}$. Thinking of $p(\boldsymbol{\theta}, \mathbf{x})^\lambda$ as a new (unnormalized) density, we see that (5) calculates the mean of the conditional density $p(\boldsymbol{\theta}, \mathbf{x})^\lambda / \int p(\boldsymbol{\theta}, \mathbf{x})^\lambda d\boldsymbol{\theta}$ and therefore looks similar to the MMSE estimator except for the modified PDF. It is reasonable to restrict λ to positive values, i.e. $\lambda \in [0, \infty)$. Otherwise we average over a new density $p(\boldsymbol{\theta}, \mathbf{x})^\lambda / \int p(\boldsymbol{\theta}, \mathbf{x})^\lambda d\boldsymbol{\theta}$ which is inverted in the sense that it has large values at positions where $p(\boldsymbol{\theta}|\mathbf{x})$ is small, i.e. it emphasizes points $(\boldsymbol{\theta}, \mathbf{x}) \in \mathbb{R}^{M+N}$ that are not likely to occur and we can expect therefore a poor performance for $\lambda < 0$.[3]

We will now show that $\mathcal{F}_\mathcal{B}$ includes three important estimators, namely the uniform a priori MMSE estimator, the MMSE estimator and the MAP estimator. By uniform a priori MMSE estimator, we refer to the estimator where we have no observation $\mathbf{x}$ about $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^M$ and the a priori distribution $p(\boldsymbol{\theta})$ is assumed to be uniform in Θ. The estimator with the minimum MSE is then the "center of gravity" of Θ, i.e. $\hat{\boldsymbol{\theta}} = E[\boldsymbol{\theta}] = \int \boldsymbol{\theta} p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_\Theta \boldsymbol{\theta} d\boldsymbol{\theta} / \int_\Theta 1 d\boldsymbol{\theta}$ which is well defined if Θ is bounded. The following theorem proves that all three estimators are in $\mathcal{F}_\mathcal{B}$.

[2]We assume here that the parameter space is open. Otherwise, the OBE could also lie on the boundary of the parameter space and (4) is not necessary anymore.

[3]For example the loss $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = 1 - L_{\text{MAP}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ results in seeking the minimum of $p(\boldsymbol{\theta}|\mathbf{x})$ which is related (but in general not identical) to $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda)$ for $\lambda \to -\infty$.

Theorem 1. *The estimator family $\mathcal{F}_{\mathcal{B}}$ defined in* (5) *includes the following special cases:*

(a) *If $\Theta \subset \mathbb{R}^M$ is bounded and $p(\boldsymbol{\theta}, \mathbf{x}) \neq 0$, then $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda)$ for $\lambda \to 0$ exists and is equivalent to the uniform a priori MMSE estimator.*

(b) *The case $\lambda = 1$ corresponds to the MMSE estimator.*

(c) *The case $\lambda \to \infty$ corresponds to the MAP estimator.*

Proof:

(a) Assuming Θ to be a bounded set on $\mathbb{R}^M$, we immediately see that $\lim_{\lambda \to 0} p(\boldsymbol{\theta}, \mathbf{x})^\lambda / \int p(\boldsymbol{\theta}, \mathbf{x})^\lambda d\boldsymbol{\theta} =$ const., i.e. it converges pointwise to a uniform distribution on Θ. Therefore, $\hat{\boldsymbol{\theta}}(\mathbf{x}; 0)$ calculates the center of gravity of Θ which is equivalent to the a priori MMSE estimator.

(b) Setting $\lambda = 1$ in (5), we obtain $p(\boldsymbol{\theta}, \mathbf{x}) / \int p(\boldsymbol{\theta}, \mathbf{x}) d\boldsymbol{\theta} = p(\boldsymbol{\theta}|\mathbf{x})$ and thus $\hat{\boldsymbol{\theta}}(\mathbf{x}; 1) = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} = \mathrm{E}[\boldsymbol{\theta}|\mathbf{x}]$, which is the MMSE estimator.

(c) To prove this part, we use a result from Pincus [24]: Given a continuous function $f(\boldsymbol{\theta})$, which attains a global maximum at exactly one point in Θ, then Pincus showed

$$\arg\max_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \lim_{\lambda \to \infty} \frac{\int_{\Theta} \boldsymbol{\theta} f(\boldsymbol{\theta})^\lambda d\boldsymbol{\theta}}{\int_{\Theta} f(\boldsymbol{\theta})^\lambda d\boldsymbol{\theta}}. \tag{6}$$

Using this theorem, we conclude that $\lim_{\lambda \to \infty} \hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda)$ is the MAP estimator.

■

Although it is interesting to see the relationship of this basic family of estimators to other estimators, we also see that the loss functions associated with $\lambda \in \{0, 1, \infty\}$ are all symmetric as they are the hit-or-miss error (1a) and the squared error (1b). In the following, we will prove in Theorem 2 that if there is a continuously differentiable loss function that results in $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda)$ for all PDFs $p(\boldsymbol{\theta}, \mathbf{x})$, then the loss function has to be symmetric.[4] For the proof of Theorem 2, we need the following Lemma. The proofs of the Lemma and Theorem 2 can be found in Appendix A.

Lemma. *The estimator $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda)$ for the PDFs $p(\boldsymbol{\theta}, \mathbf{x}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ and $p(\boldsymbol{\theta}, \mathbf{x}) = P\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + (1 - P)\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_1)$ is given by $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda) = \boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda) = (P^\lambda \boldsymbol{\theta}_0 + (1-P)^\lambda \boldsymbol{\theta}_1)/(P^\lambda + (1-P)^\lambda)$, respectively.*

Theorem 2. *Let $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ be a continuously differentiable loss function that results in the optimal Bayesian estimator $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda)$ for an arbitrary PDF $p(\boldsymbol{\theta}, \mathbf{x})$. Then $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is symmetric, i.e. $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = L(-\boldsymbol{\theta}, -\hat{\boldsymbol{\theta}})$.*

[4]Note that it is difficult to prove the existence of such a loss function for an arbitrary λ and corresponding estimator $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda)$.

From this Theorem, we see that no estimator resulting from an asymmetric, continuously differentiable loss function is included in $\mathcal{F}_{\mathcal{B}}$. However, we would like to cover such estimation problems due to their practical relevance and hence we have to extend $\mathcal{F}_{\mathcal{B}}$. This is done in the next section.

IV. Generalization to asymmetric Loss Functions

In order to extend the basic family of estimators $\mathcal{F}_{\mathcal{B}}$ given in (5), we will now modify its parametric form such that the OBE for LinEx loss is also included. By doing this, we obtain a new family of estimators $\mathcal{F}$ which can deal with the important case of asymmetric loss functions.

The LinEx loss is frequently used in Bayesian estimation, see e.g. [8], [18]. It rises approximately linear on one side and exponential on the other. The univariate LinEx loss function is given by $L_{\text{LinEx}}(\theta, \hat{\theta}) = b\left(e^{a\Delta} - a\Delta - 1\right)$ where $\Delta = \hat{\theta} - \theta$, $a \neq 0$ and $b > 0$. The multivariate LinEx loss is defined as a straightforward extension and given by [8]

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \sum_{m=1}^{M} b_m \left(e^{a_m \Delta_m} - a_m \Delta_m - 1\right) \tag{7}$$

where $\Delta_m = \hat{\theta}_m - \theta_m$, $a_m \neq 0$ and $b_m > 0$. To calculate the OBE, we use (4) with $\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})/\partial \hat{\theta}_m = b_m a_m \left(e^{a_m \Delta_m} - 1\right)$ and finally obtain

$$\hat{\theta}_m = -\frac{1}{a_m} \ln \int e^{-a_m \theta_m} p(\theta_m | \mathbf{x}) d\theta_m, \quad m = 1, \ldots, M. \tag{8}$$

Knowing the OBE for LinEx loss, we can now extend our basic family of estimators $\mathcal{F}_{\mathcal{B}}$. This will be done in such a way that the new family of estimators $\mathcal{F}$ is a kind of "superposition" of both $\mathcal{F}_{\mathcal{B}}$ and the OBE (8). We define this new family of estimators in the following way: Let $\mathcal{F}$ be the set of estimators where each estimator has the form

$$\hat{\boldsymbol{\theta}}(\mathbf{x}; \mathcal{P}) = \mathbf{f}_1\left(\frac{\int \mathbf{f}_2(\boldsymbol{\theta}; \mathcal{P}_2) p(\boldsymbol{\theta}, \mathbf{x})^{\lambda} d\boldsymbol{\theta}}{\int p(\boldsymbol{\theta}, \mathbf{x})^{\lambda} d\boldsymbol{\theta}}; \mathcal{P}_1\right) \tag{9}$$

and depends on the $2M + 4$ parameters $\mathcal{P} = \{\lambda, \mathcal{P}_1, \mathcal{P}_2\}$ with $\mathcal{P}_1 = \{\xi_1, \phi_1, \ldots, \phi_M\}$ and $\mathcal{P}_2 = \{\xi_2, \xi_3, \psi_1, \ldots, \psi_M\}$. The functions $\mathbf{f}_1$ and $\mathbf{f}_2$ are defined as

$$\mathbf{f}_1(\mathbf{z}; \mathcal{P}_1) = \xi_1 \mathbf{z} + \boldsymbol{\phi} \circ \ln|\mathbf{z}|, \tag{10a}$$

$$\mathbf{f}_2(\mathbf{z}; \mathcal{P}_2) = \xi_2 \mathbf{z} + \xi_3 e^{\boldsymbol{\psi} \circ \mathbf{z}} \tag{10b}$$

with $\boldsymbol{\phi} = \left[\phi_1, \ldots, \phi_M\right]^T$ and $\boldsymbol{\psi} = \left[\psi_1, \ldots, \psi_M\right]^T$. Note that $e^{\mathbf{z}}$, $\ln \mathbf{z}$ and $|\mathbf{z}|$ are understood elementwise. λ is again chosen such that $\lambda \in [0, \infty)$ as discussed in the Sec. III.

First, we would like to note that $\mathcal{F}_{\mathcal{B}} \subset \mathcal{F}$ as all estimators $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda)$ from (5) are included in $\hat{\boldsymbol{\theta}}(\mathbf{x}; \mathcal{P})$ for $\xi_1 = \xi_2 = 1$, $\xi_3 = 0$ and $\phi_1 = \cdots = \phi_M = 0$. Therefore, we already know from Theorem 1 that the uniform a priori MMSE, the MMSE and the MAP estimator are included in this family. Furthermore, it is straightforward to see that $\mathcal{F}$ also includes the OBE for LinEx loss as plugging in the values $\xi_1 = \xi_2 = 0$, $\xi_3 = 1$, $\lambda = 1$ and $\psi_m = 1/\phi_m = -a_m$ for $m = 1, \ldots, M$ into (9) results in (8). Thus, we see that the new estimator family $\mathcal{F}$ is more general than $\mathcal{F}_{\mathcal{B}}$ and also includes estimators with asymmetric loss functions as intended.

V. Practical Considerations

This section explains the general approach how to obtain the estimator for a given signal model and loss function and also shows how the estimate $\hat{\boldsymbol{\theta}}(\mathbf{x}; \mathcal{P})$ can be calculated efficiently for a given observation $\mathbf{x}$. In the sequel, we will make the following two assumptions:

- The generation of samples $(\boldsymbol{\theta}_k, \mathbf{x}_k) \sim p(\boldsymbol{\theta}, \mathbf{x})$ is manageable, where $p(\boldsymbol{\theta}, \mathbf{x})$ is the joint PDF of $\boldsymbol{\theta}$ and $\mathbf{x}$. This is often the case as $p(\boldsymbol{\theta}, \mathbf{x})$ can be written as $p(\boldsymbol{\theta}, \mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, where $p(\boldsymbol{\theta})$ is the a priori PDF of $\boldsymbol{\theta}$ and $p(\mathbf{x}|\boldsymbol{\theta})$ is the likelihood PDF. Very often, both are known: $p(\boldsymbol{\theta})$ from expert knowledge and $p(\mathbf{x}|\boldsymbol{\theta})$ through the signal model.
- The generation of samples $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}|\mathbf{x})$ is manageable. This is not a hard restriction as the MMSE estimator is often calculated using Markov chain Monte Carlo (MCMC) methods [2], [7]. MCMC allows the approximate generation of correlated samples from the a posteriori distribution and the MMSE estimator is then simply the average over all samples. Here, we will use importance sampling where the conditional distribution $p(\boldsymbol{\theta}|\mathbf{x})$ is the importance function.

Given the loss function and the signal model, the use of our estimator family for a general estimation problem consists of two steps:

Step 1 – Find the optimal estimator in $\mathcal{F}$

In a first step, we have to find the estimator $\hat{\boldsymbol{\theta}}(\mathbf{x}; \mathcal{P}_0) \in \mathcal{F}$ that has the smallest Bayes risk for the particular loss function and joint PDF $p(\boldsymbol{\theta}, \mathbf{x})$, i.e. we have to solve the optimization problem

$$\mathcal{P}_0 = \arg\min_{\mathcal{P}} \iint L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{x}; \mathcal{P}))p(\boldsymbol{\theta}, \mathbf{x}) d\boldsymbol{\theta} d\mathbf{x}. \tag{11}$$

This optimization has only to be carried out once to learn the optimal values of the parameters $\mathcal{P}$. In the Appendix B, we give the gradient vector of the Bayes risk in (11) with respect to the parameters in $\mathcal{P}$. The knowledge of the gradient vector allows to use a gradient descent method to find the optimal parameter values. As the Bayes risk is in general a multimodal function with respect to $\mathcal{P}$, the gradient descent algorithm should be restarted several times from different initial points.

The integration with respect to $\boldsymbol{\theta}$ and $\mathbf{x}$ can be carried out by a plain Monte Carlo (MC) integration using samples $(\boldsymbol{\theta}_k, \mathbf{x}_k) \sim p(\boldsymbol{\theta}, \mathbf{x})$. The optimization problem (11) becomes then

$$\mathcal{P}_0 = \arg\min_{\mathcal{P}} \frac{1}{K_1} \sum_{k=1}^{K_1} L(\boldsymbol{\theta}_k, \hat{\boldsymbol{\theta}}(\mathbf{x}_k; \mathcal{P})). \tag{12}$$

If the generation of samples from $p(\boldsymbol{\theta}, \mathbf{x})$ is not directly possible, then importance sampling as discussed below is another possibility to obtain an accurate approximation of the integral.

Step 2 – Calculate the estimate $\hat{\boldsymbol{\theta}}(\mathbf{x}; \mathcal{P}_0)$

In a second step, we calculate the estimate for a given observation $\mathbf{x}$. Therefore, we need an efficient method to compute both integrals in (9). Note that (9) can be written as

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\mathbf{x}; \mathcal{P}) &= \mathbf{f}_1\left(\int \mathbf{f}_2(\boldsymbol{\theta}; \mathcal{P}_2) \frac{p(\boldsymbol{\theta}, \mathbf{x})^\lambda}{\int p(\boldsymbol{\theta}, \mathbf{x})^\lambda d\boldsymbol{\theta}} d\boldsymbol{\theta}; \mathcal{P}_1\right) \\ &= \mathbf{f}_1\left(E_{p_\lambda}\left[\mathbf{f}_2(\boldsymbol{\theta}; \mathcal{P}_2)\right]; \mathcal{P}_1\right). \end{aligned} \tag{13}$$

We see that we can write the integrals as the expectation of $\mathbf{f}_2(\boldsymbol{\theta}; \mathcal{P})$ with respect to a new conditional density $p_\lambda(\boldsymbol{\theta}|\mathbf{x}) = p(\boldsymbol{\theta}, \mathbf{x})^\lambda / \int p(\boldsymbol{\theta}, \mathbf{x})^\lambda d\boldsymbol{\theta}$. Assuming that we can generate samples from the a posteriori distribution $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}|\mathbf{x}) = p(\boldsymbol{\theta}, \mathbf{x}) / \int p(\boldsymbol{\theta}, \mathbf{x}) d\boldsymbol{\theta}$, we can use importance sampling [7] for (13).

The importance sampling algorithm is as follows: Suppose we want to calculate $\mathrm{E}\left[h(\boldsymbol{\theta})\right] = \int h(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$. Then we can use the approximation

$$\mathrm{E}\left[h(\boldsymbol{\theta})\right] \approx \frac{\sum\limits_{k=1}^{K} w_k h(\boldsymbol{\theta}_k)}{\sum\limits_{k=1}^{K} w_k} \tag{14}$$

where $\boldsymbol{\theta}_k$ are drawn from a trial distribution $\tilde{p}(\boldsymbol{\theta})$ and the importance weights w_k are defined as $w_k = p(\boldsymbol{\theta}_k)/\tilde{p}(\boldsymbol{\theta}_k)$. Note that w_k has only to be known up to a multiplicative constant in (14). Using importance sampling for our problem, we finally obtain the approximation

$$\hat{\boldsymbol{\theta}}(\mathbf{x}; \mathcal{P}) \approx \mathbf{f}_1\left(\frac{\sum\limits_{k=1}^{K_2} w_k\ \mathbf{f}_2(\boldsymbol{\theta}_k; \mathcal{P}_2)}{\sum\limits_{k=1}^{K_2} w_k}; \mathcal{P}_1\right) \tag{15}$$

with $\tilde{p}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}, \mathbf{x})$ and thus $w_k = p(\boldsymbol{\theta}_k, \mathbf{x})^{\lambda - 1}$. The computational complexity is hence comparable to that of an MMSE estimation if the MMSE estimator also uses MC integration.

VI. EXAMPLES

A. Example 1: BLinEx Loss

The first example is as follows: Given the signal model $x = \theta + z$, we want to estimate θ which is uniformly distributed in $[0, 1]$ from the observation x where we know that the observation is perturbed by additive Gaussian noise $z \sim \mathcal{N}(0, \sigma^2)$. Furthermore, we assume that z and θ are independently distributed. The considered loss function is the bounded LinEx (BLinEx) loss introduced in [23]. The univariate BLinEx loss function is given by

$$L_{\text{BLinEx}}(\theta, \hat{\theta}) = \frac{L_{\text{LinEx}}(\theta, \hat{\theta})}{1 + \rho L_{\text{LinEx}}(\theta, \hat{\theta})}, \;\; \rho > 0. \tag{16}$$

Plugging $L_{\text{LinEx}}(\theta, \hat{\theta})$ from (7) into (16), we obtain

$$L_{\text{BLinEx}}(\theta, \hat{\theta}) = \frac{1}{\rho} \left(1 - \frac{1}{1 + c(e^{a(\hat{\theta} - \theta)} - a(\hat{\theta} - \theta) - 1)} \right) \tag{17}$$

with $c = \rho b$. It differs from the usually used loss functions (1) in two main properties, namely it is (a) asymmetric and (b) bounded:

(a) If $a > 0$ then the positive error $\hat{\theta} > \theta$ results in a larger loss than the corresponding negative error of the same magnitude. If $a < 0$ then negative errors $\hat{\theta} < \theta$ have a larger loss. A case where such an emphasis of negative errors is useful is the dam construction example given in Sec. I as underestimating the peak water level is more severe than overestimating it.

(b) $L_{\text{BLinEx}}(\theta, \hat{\theta})$ is bounded by 0 and $1/\rho$. Such a requirement for a loss function may occur naturally out of the considered problem or may be introduced artificially to improve the robustness of the estimator in the case of outliers.

In our example, we choose $\rho = 0.5$, $a = 10$ and $b = 1$.[5] Fig. 2 shows the graph of the BLinEx loss function for this choice of parameters. Furthermore, the noise variance is $\sigma^2 = 0.25$.

We compare the following five estimators with respect to the squared error loss (1b) and the BLinEx loss (17):

- *MAP estimator*: The MAP estimator is in general given by $\hat{\theta} = \arg\max_{\theta} p(\theta|x)$ with $p(\theta|x) \sim e^{-(x-\theta)^2/(2\sigma^2)} u_{[0,1]}(\theta)$ and $u_{[0,1]}(\theta)$ is the a priori PDF of θ which is uniformly distributed in $[0, 1]$.

[5] We choose these parameter values in order to achieve the following two effects: First, we want to study an asymmetric loss function and therefore a has to be large. Second, we want a loss function which is bounded and therefore different from the LinEx loss. To see this effect, we choose $\rho = 0.5$.

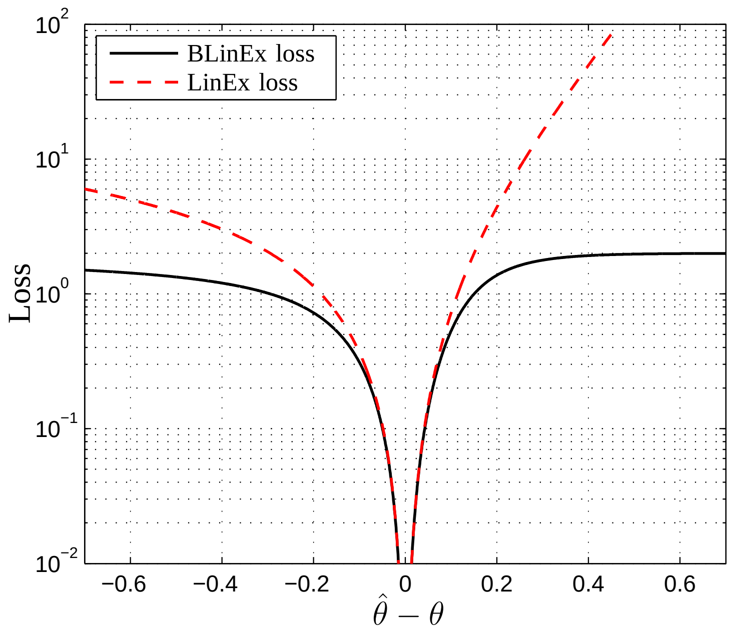

Fig. 2: LinEx and BLinEx loss ($\rho = 0.5$, $a = 10$ and $b = 1$)

This yields

$$\hat{\theta}_{\text{MAP}} = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases} \tag{18}$$

- *MMSE estimator*: The MMSE estimator is given by $\hat{\theta}_{\text{MMSE}} = E[\theta|x]$. For our signal model, the conditional mean can be calculated analytically and one obtains

$$\hat{\theta}_{\text{MMSE}} = x + \sqrt{\frac{2}{\pi}}\sigma \frac{e^{-\frac{x^2}{2\sigma^2}} - e^{-\frac{(x-1)^2}{2\sigma^2}}}{\operatorname{erf}\left(\frac{x}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{x-1}{\sqrt{2}\sigma}\right)} \tag{19}$$

- *OBE for LinEx loss*: The OBE for LinEx loss is given by (8) which can be calculated analytically. It is given by

$$\hat{\theta}_{\text{OBE,LinEx}} = x - \frac{a\sigma^2}{2} - \frac{1}{a}\log\left(\frac{\operatorname{erf}\left(\frac{1+a\sigma^2-x}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{a\sigma^2-x}{\sqrt{2}\sigma}\right)}{\operatorname{erf}\left(\frac{x}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{x-1}{\sqrt{2}\sigma}\right)}\right) \tag{20}$$

- *OBE for BLinEx loss*: The optimization problem (3) for this example can not be carried out analytically and thus (3) has to be solved for each new observation x individually, either by Monte Carlo integration or numerical quadrature. For our simulations, we used the Matlab functions `fminunc` and `quad` to solve (3).

Estimator	**Mean Squared error loss**	**Mean BLinEx loss**
MAP estimator	1.21×10^{-1}	1.02×10^{0}
MMSE estimator	6.28×10^{-2}	9.03×10^{-1}
OBE for LinEx loss	1.23×10^{-1}	8.70×10^{-1}
Optimal estimator $\in \mathcal{F}$	8.16×10^{-2}	8.21×10^{-1}
OBE for BLinEx loss	8.70×10^{-2}	8.12×10^{-1}

TABLE I: Comparison of the Bayes risks

- *Estimator family* (9) *with optimal parameters*: The optimal parameters are found via the Matlab function `fmincon` using 50 random start points for the gradient descent. The found parameters are $\xi_1 \approx 6.77 \times 10^{-1}$, $\xi_2 \approx 4.03 \times 10^{-1}$, $\xi_3 = 1.33 \times 10^{-1}$, $\lambda \approx 8.31$, $\phi \approx 4.02 \times 10^{-3}$ and $\psi \approx 1.91$. $K_1 = 5\,000$ samples are used for the Monte Carlo approximation in (12) and $K_2 = 5\,000$ samples are drawn from the a posteriori density $p(\theta|x)$ for (15) using the sampling method proposed in [25]. The values for K_1 and K_2 were found by simulations to ensure statistical stable results of the MC integral approximations.

Table I shows the results averaged over $10\,000$ trials. Clearly, the MMSE estimator is optimal in terms of the squared error loss as expected. Similarly, the OBE for the BLinEx loss gives the smallest Bayes risk if the BLinEx loss function is used. The optimal estimator $\hat{\boldsymbol{\theta}}(\mathbf{x}; \mathcal{P}_0)$ from the set $\mathcal{F}$ is a good approximation of the OBE for the BLinEx loss as it has a similar Bayes risk. Thus, although the OBE for the BLinEx loss itself is not an element of $\mathcal{F}$, there is an estimator $\hat{\boldsymbol{\theta}}(\mathbf{x}; \mathcal{P}_0)$ in $\mathcal{F}$ which gives nearly the same performance.

In order to study the influence of the noise variance on the simulation results, we rerun the first experiment with varying σ^2 values. Fig. 3 shows the simulation results and it can be concluded that the relative performance of $\hat{\boldsymbol{\theta}}(\mathbf{x}; \mathcal{P}_0)$ with respect to the OBE for BLinEx loss is almost constant.

Finally, the run times to compute the estimates on a standard desktop computer are given in Table II in order to compare the computational costs of the different approaches. It can be observed that the run time of the OBE for BLinEx loss is roughly ten times larger as for our estimator family which justifies to use the approximation given by (15) rather than the OBE itself. Note that the computation of the estimator family according to (15) consists of two steps: First, we have to sample from the a posteriori distribution which in our case is a truncated Gaussian density. We used the sampling algorithm proposed by Robert in [25] for this step. Second, we have to use importance sampling as shown in (15) to find the

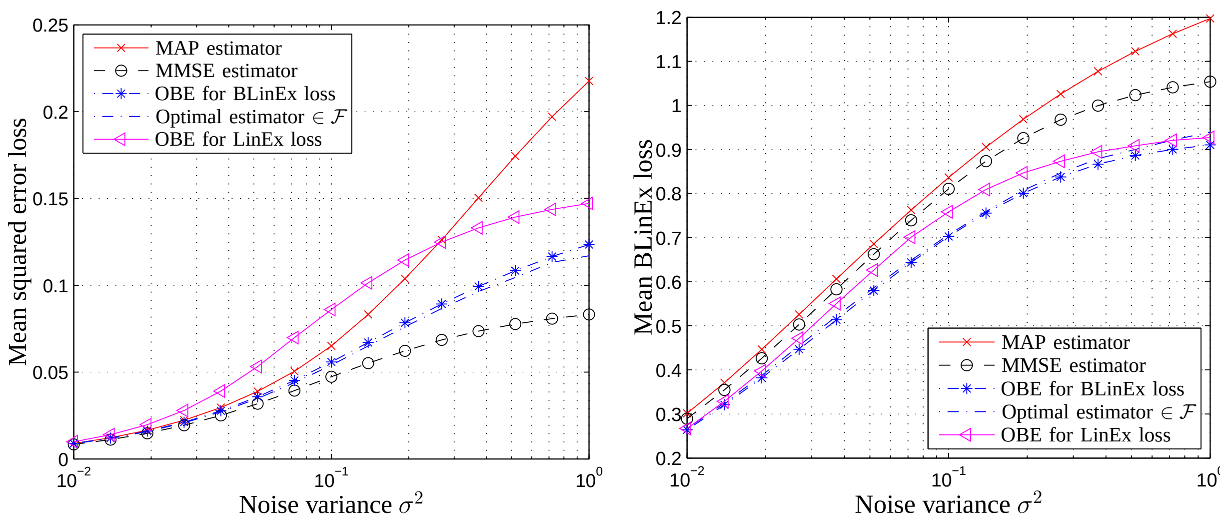

Fig. 3: Squared error and BLinEx loss for a varying noise variance σ^2

Estimator	**Run time**
MAP estimator	$< 1{\times}10^{-3}$ sec.
MMSE estimator	$< 1{\times}10^{-1}$ sec.
OBE for LinEx loss	$< 1{\times}10^{-1}$ sec.
Optimal estimator $\in \mathcal{F}$	$2.3{\times}10^{1}$ sec.
OBE for BLinEx loss	$3.2{\times}10^{2}$ sec.

TABLE II: Comparison of the run times for $10\,000$ trials

estimate. The run time for the first step is 21 seconds and for the second step 2 seconds which results in the 23 seconds that are given in Table II. These numbers show that most of the run time is spent on computing samples from the a posteriori density.

B. Example 2: Speech Enhancement

The second example which we consider is the enhancement of a distorted speech signal. The goal is to suppress an unwanted noise signal while leaving the speech as undistorted as possible, see e.g. [26], [27].

In the time domain, the speech enhancement problem can be written as

$$x(n) = s(n) + z(n), \tag{21}$$

where $s(n)$ is the original (clean) speech signal at time instance n which is distorted by noise $z(n)$ to result in the observed signal $x(n)$. One solution for this problem is the traditional approach of *short-time spectral attenuation* (STSA) which was introduced by [28], [29] and extended in later work [30]–[32]. While [28] is based on the method of *spectral subtraction*, the other papers use a more statistically motivated approach by introducing a suitable loss function and signal model for each frequency bin. The corresponding OBE is then used to perform the STSA operation.

In the following, we will state the speech enhancement problem in the frequency domain where we assume a Gaussian signal model. All necessary elements to use our family of estimators from (9) are derived and this estimator is then compared to the OBE.

1) Problem Formulation and Solution Approach: Using the short-time Fourier transform of (21), the signal model can be written in the frequency domain as

$$X_{k,i} = S_{k,i} + Z_{k,i}, \tag{22}$$

where $X_{k,i} = \mathcal{X}_{k,i} e^{j\theta_{k,i}}$, $S_{k,i} = \mathcal{S}_{k,i} e^{j\phi_{k,i}}$ and $Z_{k,i}$ are the kth spectral component of the noisy signal $x(n)$, clean speech $s(n)$ and noise $z(n)$ in the ith frame. The frequency index k ranges from 0 to $K-1$ where K is the FFT length. In STSA, the speech enhancement problem is solved by using

$$\hat{S}_{k,i} = \hat{\mathcal{S}}_{k,i} e^{j\theta_{k,i}}, \tag{23}$$

i.e. the amplitude $\mathcal{X}_{k,i}$ of the noisy spectral component $X_{k,i}$ is replaced by the estimate $\hat{\mathcal{S}}_{k,i} = \hat{\mathcal{S}}_{k,i}(X_{k,i})$. For convenience, we will drop the dependence of the spectral components on the frame index i in the following.

Using the Gaussian model $S_k = \mathcal{S}_k e^{j\phi_k} \sim \mathcal{CN}(0, \sigma_s^2(k))$, i.e. S_k is *complex Gaussian*, we know that the PDF of $\mathcal{S}_k$ and ϕ_k is given by

$$p(\mathcal{S}_k, \phi_k) = \begin{cases} \frac{1}{2\pi} \frac{\mathcal{S}_k}{\sigma_s^2(k)/2} e^{-\frac{\mathcal{S}_k^2}{\sigma_s^2(k)}} & \mathcal{S}_k \geq 0, 0 \leq \phi_k < 2\pi \\ 0 & \text{otherwise} \end{cases}, \tag{24}$$

i.e. $\mathcal{S}_k$ follows a *Rayleigh* distribution, ϕ_k is uniformly distributed on $[0, 2\pi)$ and they are independent of each other. Assuming furthermore $Z_k \sim \mathcal{CN}(0, \sigma_z^2(k))$ and Z_k is independent of S_k, the a posteriori density $p(\mathcal{S}_k|X_k)$ for $\mathcal{S}_k \geq 0$ is given by

$$p(\mathcal{S}_k|X_k) = \frac{1}{p(X_k)} \int_0^{2\pi} p(X_k|\mathcal{S}_k, \phi_k) p(\mathcal{S}_k, \phi_k) d\phi_k$$
$$= \frac{\mathcal{S}_k(\sigma_z^2(k) + \sigma_s^2(k))}{\pi \sigma_z^2(k) \sigma_s^2(k)} \exp\left\{ -\frac{\sigma_z^2(k) + \sigma_s^2(k)}{\sigma_z^2(k)\sigma_s^2(k)} \mathcal{S}_k^2 - \frac{\sigma_s^2(k)}{\sigma_z^2(k)(\sigma_z^2(k) + \sigma_s^2(k))} \mathcal{X}_k^2 \right\}$$
$$\times \int_0^{2\pi} \exp\left\{ \frac{2\mathcal{S}_k \mathcal{X}_k}{\sigma_z^2(k)} \cos(\phi_k - \theta_k) \right\} d\phi_k. \tag{25}$$

Introducing the *modified Bessel function of the first kind and* n*th order* $I_n(z)$ which has the integral representation [33]

$$I_n(z) = \frac{1}{2\pi} \int_0^{2\pi} \cos(\beta n) \exp\{z \cos(\beta)\} d\beta, \tag{26}$$

and using the shorthand notations $v_k = \frac{\sigma_s^2(k)}{\sigma_z^2(k)(\sigma_z^2(k)+\sigma_s^2(k))} \mathcal{X}_k^2$ and $\lambda_k^{-1} = \frac{\sigma_z^2(k)+\sigma_s^2(k)}{\sigma_z^2(k)\sigma_s^2(k)}$, we can finally write the a posteriori density $p(\mathcal{S}_k|X_k)$ as

$$p(\mathcal{S}_k|X_k) = 2\frac{\mathcal{S}_k}{\lambda_k} \exp\left\{ -\frac{\mathcal{S}_k^2}{\lambda_k} - v_k \right\} I_0\left(2\mathcal{S}_k \sqrt{v_k \lambda_k^{-1}} \right). \tag{27}$$

This density is well known in the literature and shows that $\mathcal{S}_k$ given the observation X_k follows a *Rice distribution* [34]. It is interesting to note that $p(\mathcal{S}_k|X_k)$ only depends on $\mathcal{X}_k$ and therefore, $\hat{\mathcal{S}}_k = \hat{\mathcal{S}}_k(\mathcal{X}_k)$. To derive the OBE in the next Section, we will need to calculate the moments $E[\mathcal{S}_k^m|X_k]$. Interestingly, they can be given analytically using the *Kummer function* $M(a, b, z)$ as shown in [30] and they are

$$E[\mathcal{S}_k^m|X_k] = \lambda_k^{m/2} \Gamma\left(\frac{m}{2} + 1 \right) M\left(-\frac{m}{2}, 1, -v_k \right) \tag{28}$$

for all $m > -2$ where $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ is the *Gamma function*. Eq. (28) results from the identities [33, 11.4.28] and [33, 13.1.27].

2) Loss Function and Corresponding OBE: In the literature, many different loss functions were proposed to perform STSA speech enhancement. The first approach in [29] was to use the squared loss function $L(\mathcal{S}_k, \hat{\mathcal{S}}_k) = (\mathcal{S}_k - \hat{\mathcal{S}}_k)^2$ which results in the *MMSE-STSA* algorithm. Later, other loss functions were proposed in [30]–[32] which show a better performance with respect to perceptual motivated quality measures, e.g. the *perceptual evaluation of speech quality* (PESQ) measure [35]. In [36], these loss functions were combined into a family of loss functions of the form

$$L(\mathcal{S}_k, \hat{\mathcal{S}}_k) = \left(\frac{\mathcal{S}_k^\beta - \hat{\mathcal{S}}_k^\beta}{\mathcal{S}_k^\alpha} \right)^2. \tag{29}$$

This loss function was later generalized in [37] to include even more proposed loss functions. The corresponding OBE for (29) can easily be found by using (4) together with $\partial L/\partial\hat{\mathcal{S}}_k = -2\beta\hat{\mathcal{S}}_k^{\beta-1}(\mathcal{S}_k^\beta - \hat{\mathcal{S}}_k^\beta)/\mathcal{S}_k^{2\alpha}$ and is given by

$$\hat{\mathcal{S}}_k = \left(\frac{\int\limits_0^\infty \mathcal{S}_k^{\beta-2\alpha} p(\mathcal{S}_k|\mathcal{X}_k) d\mathcal{S}_k}{\int\limits_0^\infty \mathcal{S}_k^{-2\alpha} p(\mathcal{S}_k|\mathcal{X}_k) d\mathcal{S}_k}\right)^{\frac{1}{\beta}} = \left(\frac{\mathrm{E}\left[\mathcal{S}_k^{\beta-2\alpha}|\mathcal{X}_k\right]}{\mathrm{E}\left[\mathcal{S}_k^{-2\alpha}|\mathcal{X}_k\right]}\right)^{\frac{1}{\beta}}. \tag{30}$$

3) Simulation Results: In the following, we will compare the OBE for the loss function (29) with the best estimator from the generalized family (9). Two experiments are conducted: In the first experiment, we find the best estimator in $\mathcal{F}$ with respect to the loss function (29) for $\alpha = 0.5$ and $\beta = 1$. This parameter setup was shown in [36] to result in an STSA algorithm with the best PESQ value, which is called *Weighted Euclidean STSA (WE-STSA)*. In contrast, the second experiment optimizes directly on the PESQ measure.

Experiment 1: Fitting of the estimator family to WE-STSA

The following three estimators are considered:

- *Minimum Mean-Squared Error STSA (MMSE-STSA)*: The MMSE-STSA estimator results from the special choice $\alpha = 0$ and $\beta = 1$ in (29). The corresponding OBE is given by [29]

$$\hat{\mathcal{S}}_k = \mathrm{E}\left[\mathcal{S}_k|\mathcal{X}_k\right] = \frac{\sqrt{\pi\lambda_k}}{2} e^{-\frac{v_k}{2}} \left[(1+v_k) I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] \tag{31}$$

  where we used the identities [33, 13.1.27] and [33, 13.3.6] in (28) for $m = 1$.
- *Weighted Euclidean (WE-STSA)*: The WE-STSA estimator is the OBE that corresponds to the choice $\alpha = 0.5$ and $\beta = 1$. It is given by [31]

$$\hat{\mathcal{S}}_k = \left(\mathrm{E}\left[\mathcal{S}_k^{-1}|\mathcal{X}_k\right]\right)^{-1} = \sqrt{\frac{\lambda_k}{\pi}} \frac{e^{v_k/2}}{I_0\left(\frac{v_k}{2}\right)} \tag{32}$$

  where we used the identity $M(\frac{1}{2}, 1, z) = e^{z/2} I_0\left(\frac{z}{2}\right)$ in (28).
- *Estimator Family*: To learn the optimal parameters $\mathcal{P}_0$, $K_1 = 5\,000$ samples from the joint PDF $p(\mathcal{S}_k, \mathcal{X}_k)$ and $K_2 = 5\,000$ samples from the a posteriori PDF $p(\mathcal{S}_k|\mathcal{X}_k)$ are drawn using a uniform (hyper-)prior distribution for $\sigma_z^2(k)$ and $\sigma_s^2(k)$. They were chosen to be uniformly distributed with $\sigma_z^2(k) \sim \mathcal{U}(10^{-2}, 10^0)$ and $\sigma_s^2(k) \sim \mathcal{U}(10^{-12}, 10^3)$.

We used ten female and ten male speakers from the TIMIT database which resulted in a total of 144 utterances. The noise was assumed to be white Gaussian with a SNR of 10dB. The short-time Fourier transform was computed using a Hamming window of length 32ms and an overlap of 50% as in [36]. The noise variance $\sigma_z^2(k)$ was estimated from noise-only segments where those segments were found by a voice activity detector (VAD). $\sigma_s^2(k)$ is estimated from the decision-directed approach as proposed in [29].

The results are shown in Table III. Beside the MMSE loss and the WE loss, we also give the results with respect to the PESQ measure. It can take on values between "1" (bad) and "4.5" (excellent) and was shown to be a good objective quality measure for speech enhancement [38]. From the results we see that WE-STSA gives the best results with respect to the PESQ measure which was already observed in [36]. Furthermore, we also see that the best estimator from $\mathcal{F}$ is a good approximation of the OBE for WE loss. It gives a better PESQ measure than the MMSE-STSA and therefore we could adapt the parametric family to the WE loss function. It is interesting to note that the best estimator from $\mathcal{F}$ has a smaller WE loss than the OBE for this loss function. This stems from the fact that estimates of $\sigma_z^2(k)$ and $\sigma_s^2(k)$ were used during the speech enhancement which influences the performance of the estimators.

Experiment 2: Fitting of the estimator family to PESQ

Instead of using the WE loss as for Experiment 1, we also studied the performance of the estimator family $\mathcal{F}$ if the PESQ measure is directly used as loss function, i.e. we rerun the first experiment with the same setup but this time we search the best estimator in $\mathcal{F}$ that yields the maximum PESQ value. We splitted the 144 files into two sets, a training set consisting of one male and one female speaker, and a disjoint test set which contains the remaining 142 files. The optimization problem (12) was solved using Matlab's `fminsearch` procedure from 50 different randomly chosen starting points.

Table IV shows the results for this new setup. It can be seen that the estimator which is adapted to the PESQ loss has an improved mean PESQ value of 2.87 compared to the estimator we found in the first experiment which had a PESQ loss of 2.80. A difference of 0.07 in the PESQ measure corresponds roughly to a 1dB difference in SNR and hence, we can conclude that the found estimator is capable of fitting to the PESQ loss function. Furthermore, it performs also slightly better than WE-STSA on the 142 utterances of the test set.

VII. Conclusions

In this paper a family of estimators was proposed for the Bayesian estimation with non-standard loss functions. This family has the advantage that it is parameterized by a small number of variables which can be determined offline for a particular loss function. We proved that the family includes many important estimators known from the literature, namely MMSE, MAP, and OBE for the LinEx loss which shows that it is quite versatile. The computational complexity of our approach is comparable to that of an MMSE estimation for the same signal model if we assume that Monte Carlo integration is used for the

	MMSE loss ($\alpha = 0,\ \beta = 1$)	**WE loss** ($\alpha = 0.5,\ \beta = 1$)	**PESQ**
Noisy speech signal	4.13×10^{-2}	8.34×10^{0}	2.26
OBE for $\alpha = 0,\ \beta = 1$ (MMSE-STSA)	1.47×10^{-2}	1.02×10^{0}	2.65
OBE for $\alpha = 0.5,\ \beta = 1$ (WE-STSA)	2.02×10^{-2}	2.01×10^{-1}	2.86
Optimal estimator in $\mathcal{F}$ (WE Loss)	2.10×10^{-2}	1.53×10^{-1}	2.80

TABLE III: Experiment 1: Performance of the STSA estimators

	MMSE loss ($\alpha = 0,\ \beta = 1$)	**WE loss** ($\alpha = 0.5,\ \beta = 1$)	**PESQ**
Noisy speech signal	4.15×10^{-2}	8.28×10^{0}	2.26
OBE for $\alpha = 0,\ \beta = 1$ (MMSE-STSA)	1.48×10^{-2}	1.01×10^{0}	2.65
OBE for $\alpha = 0.5,\ \beta = 1$ (WE-STSA)	2.04×10^{-2}	1.99×10^{-1}	2.85
Optimal estimator in $\mathcal{F}$ (for WE Loss)	2.09×10^{-2}	1.52×10^{-1}	2.80
Optimal estimator in $\mathcal{F}$ (for PESQ)	1.57×10^{-1}	1.36×10^{-1}	2.87

TABLE IV: Experiment 2: Performance of the STSA estimators on the test set

calculation of the MMSE estimator. Please note that a Matlab/MEX implementation of the estimator family is available online [39].

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

APPENDIX A
PROOFS

Proof of the Lemma: First of all, we would like to point out that the delta function can be expressed as a limit of the normal distribution, i.e.

$$g(\boldsymbol{\theta}; a^2) = \frac{1}{a^M \pi^{M/2}} e^{-\|\boldsymbol{\theta}\|^2/a^2} \xrightarrow{a \to 0} \delta(\boldsymbol{\theta}).$$

They are equivalent in the sense that $f(\mathbf{0}) = \int f(\boldsymbol{\theta})\delta(\boldsymbol{\theta})d\boldsymbol{\theta} = \lim_{a\to 0} \int f(\boldsymbol{\theta})g(\boldsymbol{\theta}; a^2)d\boldsymbol{\theta}$. Using this relationship, we can now prove the lemma.

(a) $p(\boldsymbol{\theta}, \mathbf{x}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$:

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda) &= \lim_{a\to 0} \frac{\int \boldsymbol{\theta} g(\boldsymbol{\theta} - \boldsymbol{\theta}_0; a^2)^\lambda d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta} - \boldsymbol{\theta}_0; a^2)^\lambda d\boldsymbol{\theta}} = \lim_{a\to 0} \frac{\int \boldsymbol{\theta} g(\boldsymbol{\theta} - \boldsymbol{\theta}_0; \frac{a^2}{\lambda})d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta} - \boldsymbol{\theta}_0; \frac{a^2}{\lambda})d\boldsymbol{\theta}} \\ &= \lim_{a\to 0} \int \boldsymbol{\theta} g(\boldsymbol{\theta} - \boldsymbol{\theta}_0; \frac{a^2}{\lambda})d\boldsymbol{\theta} = \boldsymbol{\theta}_0\end{aligned}$$

(b) $p(\boldsymbol{\theta}, \mathbf{x}) = P\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + (1-P)\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_1)$:

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda) &= \lim_{a\to 0} \frac{\int \boldsymbol{\theta}[Pg(\boldsymbol{\theta} - \boldsymbol{\theta}_0; a^2) + (1-P)g(\boldsymbol{\theta} - \boldsymbol{\theta}_1, a^2)]^\lambda d\boldsymbol{\theta}}{\int [Pg(\boldsymbol{\theta} - \boldsymbol{\theta}_0; a^2) + (1-P)g(\boldsymbol{\theta} - \boldsymbol{\theta}_1; a^2)]^\lambda d\boldsymbol{\theta}} \\ &= \lim_{a\to 0} \frac{P^\lambda}{P^\lambda + (1-P)^\lambda} \int \boldsymbol{\theta} g(\boldsymbol{\theta} - \boldsymbol{\theta}_0; \frac{a^2}{\lambda})d\boldsymbol{\theta} \\ &\quad + \lim_{a\to 0} \frac{(1-P)^\lambda}{P^\lambda + (1-P)^\lambda} \int \boldsymbol{\theta} g(\boldsymbol{\theta} - \boldsymbol{\theta}_1; \frac{a^2}{\lambda})d\boldsymbol{\theta} \\ &= \frac{P^\lambda \boldsymbol{\theta}_0 + (1-P)^\lambda \boldsymbol{\theta}_1}{P^\lambda + (1-P)^\lambda}\end{aligned}$$

where we used the fact that $[Pg(\boldsymbol{\theta} - \boldsymbol{\theta}_0; a^2) + (1-P)g(\boldsymbol{\theta} - \boldsymbol{\theta}_1; a^2)]^\lambda \to P^\lambda g(\boldsymbol{\theta} - \boldsymbol{\theta}_0; a^2)^\lambda + (1-P)^\lambda g(\boldsymbol{\theta} - \boldsymbol{\theta}_1; a^2)^\lambda$ for $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_1$ and $a \to 0$.

■

Proof of Theorem 2: We will prove this theorem by contradiction. Suppose $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda)$ has a corresponding loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ which is continuously differentiable but not symmetric. Then at least one of the following two cases has to be true:

(a) There is a $\boldsymbol{\theta}_0$ such that

$$\left|\frac{\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}}\right|\Bigg|_{\substack{\boldsymbol{\theta}=\boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\theta}}=\boldsymbol{\theta}_0}} \neq \left|\frac{\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}}\right|\Bigg|_{\substack{\boldsymbol{\theta}=-\boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\theta}}=-\boldsymbol{\theta}_0}}. \qquad (\star)$$

Consider the special PDF $p(\boldsymbol{\theta}, \mathbf{x}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. As $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda)$ from (5) holds for all densities, we can directly use the result of the Lemma and obtain $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda) = \boldsymbol{\theta}_0$. A necessary condition that $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda)$

is the OBE for the loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is (4)

$$\left. \frac{\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \right|_{\substack{\boldsymbol{\theta}=\boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\theta}}=\boldsymbol{\theta}_0}} = \mathbf{0}.$$

Furthermore, consider the special PDF $p(\boldsymbol{\theta}, \mathbf{x}) = \delta(\boldsymbol{\theta} + \boldsymbol{\theta}_0)$ which has the OBE $\hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda) = -\boldsymbol{\theta}_0$. Using again (4), we obtain the necessary condition

$$\left. \frac{\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \right|_{\substack{\boldsymbol{\theta}=-\boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\theta}}=-\boldsymbol{\theta}_0}} = \mathbf{0}$$

which can not be true as we assumed ($\star$).

(b) There is a $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ such that

$$\left| \frac{\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \right| \Bigg|_{\substack{\boldsymbol{\theta}=\boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\theta}}=\boldsymbol{\theta}_1}} \neq \left| \frac{\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \right| \Bigg|_{\substack{\boldsymbol{\theta}=-\boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\theta}}=-\boldsymbol{\theta}_1}}. \tag{$\star\star$}$$

Consider the special PDF $p(\boldsymbol{\theta}, \mathbf{x}) = P\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + (1-P)\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_1)$ which, according to the above Lemma, has the OBE $\mathbf{u} = \hat{\boldsymbol{\theta}}(\mathbf{x}; \lambda) = (P^\lambda \boldsymbol{\theta}_0 + (1-P)^\lambda \boldsymbol{\theta}_1)/(P^\lambda + (1-P)^\lambda)$. A necessary condition that has to be fulfilled is (4) which yields

$$P \left. \frac{\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \right|_{\substack{\boldsymbol{\theta}=\boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\theta}}=\mathbf{u}}} + (1-P) \left. \frac{\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \right|_{\substack{\boldsymbol{\theta}=\boldsymbol{\theta}_1 \\ \hat{\boldsymbol{\theta}}=\mathbf{u}}} = \mathbf{0}.$$

Furthermore, the PDF $p(\boldsymbol{\theta}, \mathbf{x}) = P\delta(\boldsymbol{\theta} + \boldsymbol{\theta}_0) + (1-P)\delta(\boldsymbol{\theta} + \boldsymbol{\theta}_1)$ results in the OBE $-\mathbf{u}$ and the necessary condition (4) is

$$P \left. \frac{\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \right|_{\substack{\boldsymbol{\theta}=-\boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\theta}}=-\mathbf{u}}} + (1-P) \left. \frac{\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \right|_{\substack{\boldsymbol{\theta}=-\boldsymbol{\theta}_1 \\ \hat{\boldsymbol{\theta}}=-\mathbf{u}}} = \mathbf{0}.$$

Without loss of generality, we can assume $\left| \frac{\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \right| \Big|_{\substack{\boldsymbol{\theta}=\boldsymbol{\theta}_1 \\ \hat{\boldsymbol{\theta}}=\boldsymbol{\theta}_1}} = \left| \frac{\partial L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \right| \Big|_{\substack{\boldsymbol{\theta}=-\boldsymbol{\theta}_1 \\ \hat{\boldsymbol{\theta}}=-\boldsymbol{\theta}_1}}$ as we can otherwise use (a) and show that the loss is asymmetric. Taking the limit $P \to 0$ ($P > 0$), we see that both necessary conditions contradict the assumption ($\star\star$) as $\mathbf{u} \to \boldsymbol{\theta}_1$ and ($\star\star$) is also true in a neighbourhood of $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ as the loss is continuously differentiable.

■

Appendix B

Gradient of the Bayes Risk

In this section, we derive the gradient of the Bayes risk with respect to an element $\gamma \in \mathcal{P}$. Using the gradient is advantageous to solve the optimization problem (11) as gradient descent methods can be used. Taking the derivative of BR in (11) with respect to γ, we obtain for the first-order derivative

$$\frac{\partial \mathrm{BR}}{\partial \gamma} = \iint \left(\left. \frac{\partial L(\boldsymbol{\theta}, \mathbf{u})}{\partial \mathbf{u}} \right|_{\mathbf{u}=\hat{\boldsymbol{\theta}}(\mathbf{x};\mathcal{P})} \right)^T \frac{\partial \hat{\boldsymbol{\theta}}(\mathbf{x};\mathcal{P})}{\partial \gamma} p(\boldsymbol{\theta}, \mathbf{x}) d\boldsymbol{\theta} d\mathbf{x}.$$

Using the shorthand notations $p_\lambda(\boldsymbol{\theta}|\mathbf{x}) = p(\boldsymbol{\theta}, \mathbf{x})^\lambda / \int p(\boldsymbol{\theta}, \mathbf{x})^\lambda d\boldsymbol{\theta}$ and $\mathbf{D} = \frac{\partial \mathbf{f}_1}{\partial \mathbf{z}} = \xi_1 \mathbf{I} + \mathrm{diag}\{\phi_1/z_1, \ldots, \phi_M/z_M\}$ evaluated at $\mathbf{z} = \int \mathbf{f}_2(\boldsymbol{\theta}, \mathcal{P}_2) p_\lambda(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}$, we obtain

$$\begin{aligned}
\frac{\partial \hat{\boldsymbol{\theta}}(\mathbf{x};\mathcal{P})}{\partial \xi_1} &= \int \mathbf{f}_2(\boldsymbol{\theta}, \mathcal{P}_2) p_\lambda(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \\
\frac{\partial \hat{\boldsymbol{\theta}}(\mathbf{x};\mathcal{P})}{\partial \xi_2} &= \mathbf{D} \int \boldsymbol{\theta} p_\lambda(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \\
\frac{\partial \hat{\boldsymbol{\theta}}(\mathbf{x};\mathcal{P})}{\partial \xi_3} &= \mathbf{D} \int e^{\boldsymbol{\psi} \circ \boldsymbol{\theta}} p_\lambda(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \\
\frac{\partial \hat{\boldsymbol{\theta}}(\mathbf{x};\mathcal{P})}{\partial \lambda} &= \mathbf{D} \left(\int \mathbf{f}_2(\boldsymbol{\theta}; \mathcal{P}_2) \ln\left(p(\boldsymbol{\theta}, \mathbf{x})\right) p_\lambda(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \right. \\
&\qquad \left. - \int \mathbf{f}_2(\boldsymbol{\theta}; \mathcal{P}_2) p_\lambda(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \int \ln\left(p(\boldsymbol{\theta}, \mathbf{x})\right) p_\lambda(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \right) \\
\frac{\partial \hat{\boldsymbol{\theta}}(\mathbf{x};\mathcal{P})}{\partial \boldsymbol{\phi}} &= \mathrm{diag}\left\{ \ln \left| \int \mathbf{f}_2(\boldsymbol{\theta}; \mathcal{P}_2) p_\lambda(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \right| \right\} \\
\frac{\partial \hat{\boldsymbol{\theta}}(\mathbf{x};\mathcal{P})}{\partial \boldsymbol{\psi}} &= \xi_3 \mathbf{D} \, \mathrm{diag}\left\{ \int \boldsymbol{\theta} \circ e^{\boldsymbol{\psi} \circ \boldsymbol{\theta}} p_\lambda(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \right\}
\end{aligned}$$

Note that all integrals can again be calculated using Monte Carlo integration, especially importance sampling as was shown in Sec. V.

REFERENCES

[1] S. Uhlich and B. Yang, "A parametric family of Bayesian estimators for non-standard loss functions," *Proc. European Signal Processing Conference (EUSIPCO)*, 2010.

[2] C. P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Springer, 2001.

[3] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Chapman & Hall, 2003.

[4] J. G. Norstrom, "The use of precautionary loss functions in risk analysis," *IEEE Trans. on Reliability*, vol. 45, no. 3, pp. 400–403, 1996.

[5] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory*, Prentice-Hall, 1993.

[6] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[7] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, 2008.

[8] A. Zellner, "Bayesian estimation and prediction using asymmetric loss functions," *Journal of the American Statistical Association*, vol. 81, no. 394, pp. 446–451, 1986.

[9] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004.

[11] S. S. Channappayya, A. C. Bovik, C. Caramanis, and R. W. Heath Jr., “Design of linear equalizers optimized for the structural similarity index,” *IEEE Transactions on Image Processing*, vol. 17, no. 6, pp. 857–872, 2008.

[12] H. Rue, “New loss functions in Bayesian imaging,” *Journal of the American Statistical Association*, vol. 90, no. 431, pp. 900–908, 1995.

[13] D. A. Binder, “Bayesian cluster analysis,” *Biometrika*, vol. 65, no. 1, pp. 31–38, 1978.

[14] D. A. Binder, “Approximations to Bayesian clustering rules,” *Biometrika*, vol. 68, no. 1, pp. 275–285, 1981.

[15] J. W. Lau and P. J. Green, “Bayesian model based clustering procedures,” *Journal of Computational and Graphical Statistics*, vol. 16, pp. 526–558, 2007.

[16] G. Celeux, M. Hurn, and C. P. Robert, “Compuational and inferential difficulties with mixture posterior distributions,” *Journal of the American Statistical Association*, vol. 95, no. 451, pp. 957–970, 2000.

[17] M. Hurn, A. Justel, and C. P. Robert, “Estimating mixtures of regressions,” *Journal of Computational and Graphical Statistics*, vol. 12, no. 1, pp. 1, 25 2003.

[18] H. R. Varian, “A Bayesian approach to real estate assessment,” in *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, S. E. Fienberg and A. Zellner, Eds., pp. 195–208. North Holland Press, 1974.

[19] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation and Time Series Analysis*, Addison-Wesley, 1990.

[20] C. Stein, “Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean,” *Annals of the Institute of Statistical Mathematics*, vol. 16, no. 1, pp. 155–160, 1964.

[21] L. D. Brown, “Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters,” *Annals of Mathematical Statistics*, vol. 39, no. 1, pp. 29–48, 1968.

[22] A. Kaminska and Z. Porosinski, “On robust Bayesian estimation under some asymmetric and bounded loss function,” *Statistics*, vol. 43, no. 3, pp. 253–265, 2009.

[23] D. Wen and M. S. Levy, “BLINEX: A bounded asymmetric loss function with application to Bayesian estimation,” *Communications in Statistics - Theory and Methods*, vol. 30, no. 1, pp. 147–153, 2001.

[24] M. Pincus, “A closed form solution of certain programming problems,” *Operations Research*, vol. 16, no. 3, pp. 690–694, 1968.

[25] C. P. Robert, “Simulation of truncated normal variables,” *Statistics and Computing*, vol. 5, no. 2, pp. 121–125, 1995.

[26] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, Springer, 2005.

[27] P. Loizou, “Speech enhancement algorithms: A survey,” *in Recent Advances in Robust Speech Recognition Technology, (J. Ramrez and J. M. Grriz, eds.)*, pp. 60–102, 2011.

[28] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1979.

[29] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error short-time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 1109–1121, 1984.

[30] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error log-spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, pp. 443–445, 1985.

[31] P. C. Loizou, “Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, 2005.

[32] C. H. You, S. N. Koh, and S. Rahardja, “Beta-order MMSE spectral amplitude estimation for speech enhancement,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 475–486, 2005.

[33] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, U.S. Department of Commerce, National Bureau of Standards, 1964.

[34] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 2002.

[35] ITU-T Recommendation P.862, “Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” (02/2001).

[36] E. Plourde and B. Champagne, “Auditory-based spectral amplitude estimators for speech enhancement,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1614–1623, 2008.

[37] E. Plourde and B. Champagne, “Generalized Bayesian estimators of the spectral amplitude for speech enhancement,” *IEEE Signal Processing Letters*, 2009.

[38] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[39] S. Uhlich, “Parametric estimator family for nonstandard loss functions: Matlab implementation,” Available: http://www.iss.uni-stuttgart.de/mitarbeiter/uhlich.en.html.

Stefan Uhlich received the Dipl.-Ing. degree in electrical engineering from the University of Stuttgart, Germany, in 2006. From 2007 to 2011 he was a research assistant at the Chair of System Theory and Signal Processing, University of Stuttgart, Germany. In this time he worked in the area of statistical signal processing, focusing especially on parameter estimation theory and methods. He recently joined Sony Stuttgart Technology Center where he works on problems in data mining and classification.

Bin Yang (SM ’06) received the Dipl.-Ing. and Ph.D. degree in 1986 and 1991 from the Ruhr University Bochum, Germany, all in electrical engineering.

From 1996 to 2001, he was a senior researcher on mobile communications at Infineon Technologies, Germany. Since 2001 he is Professor and head of the Institute of Signal Processing and System Theory at University of Stuttgart, Germany. His research interests include theory and algorithms of statistical signal processing and various applications like array processing, automotive safety systems, localization, blind methods, emotion recognition etc.