Robust PAC^m: Training Ensemble Models Under Misspecification and Outliers

Matteo Zecchin, *Student Member, IEEE,* Sangwoo Park, *Member, IEEE,* Osvaldo Simeone, *Fellow, IEEE,* Marios Kountouris, *Fellow, IEEE*, David Gesbert, *Fellow, IEEE*

***Abstract*—Standard Bayesian learning is known to have suboptimal generalization capabilities under misspecification and in the presence of outliers. PAC-Bayes theory demonstrates that the free energy criterion minimized by Bayesian learning is a bound on the generalization error for Gibbs predictors (i.e., for single models drawn at random from the posterior) under the assumption of sampling distributions uncontaminated by outliers. This viewpoint provides a justification for the limitations of Bayesian learning when the model is misspecified, requiring ensembling, and when data is affected by outliers. In recent work, PAC-Bayes bounds – referred to as PAC^m – were derived to introduce free energy metrics that account for the performance of ensemble predictors, obtaining enhanced performance under misspecification. This work presents a novel robust free energy criterion that combines the generalized logarithm score function with PAC^m ensemble bounds. The proposed free energy training criterion produces predictive distributions that are able to concurrently counteract the detrimental effects of misspecification – with respect to both likelihood and prior distribution – and outliers.**

***Index Terms*—Bayesian learning, robustness, outliers, misspecification, ensemble models, machine learning.**

I. Introduction

Key assumptions underlying Bayesian inference and learning are that the adopted probabilistic model is well specified and that the training data set does not include outliers, so that training and testing distributions are matched [1]. Under these favorable conditions, the Bayesian posterior distribution provides an optimal solution to the inference and learning problems. In contrast, optimality does not extend to scenarios characterized by misspecification [2], [3] *or* outliers [4]. This work aims at addressing *both* problems by integrating the use of ensemble predictors [5], generalized logarithm score functions [6], and generalized prior-dependent information-theoretic regularizers [7] in Bayesian learning.

The proposed learning framework – termed (m, t)-robust Bayesian learning – is underpinned by a novel *free energy* learning criterion parameterized by integer $m \geq 1$ and scalar $t \in [0, 1]$. The parameter m controls robustness to misspecification by determining the size of the ensemble used for prediction. In contrast, parameter t controls robustness to outliers by dictating the degree to which the loss function penalizes low predictive probabilities. The proposed learning criterion generalizes the standard free energy criterion underlying Bayesian learning, which is obtained for $m = 1$ and $t = 1$ [8], [9]; as well as the m-free energy criterion, obtained for $t = 1$, which was recently introduced in [10]. A further generalization of (m, t)-robust Bayesian learning is also introduced, which aims at ensuring robustness to prior misspecification by modifying the information-theoretic regularizer present in the free energy [7].

To illustrate the shortcomings of conventional Bayesian learning and the advantages of the proposed (m, t)-robust Bayesian learning, consider the example in Figure 1. In it, the in-distribution (ID) data generating measure (dashed line) is multimodal, while the probabilistic model is Gaussian, and hence misspecified. Furthermore, the training data set, represented as crosses, comprises an outlying data point depicted in red. In these conditions, the predictive distribution resulting from standard Bayesian learning (see the gray curve labeled as $m = 1,\ t = 1$) is unimodal, and it poorly approximates the underlying ID measure. The predictive distribution resulting from the minimization of the m-free energy criterion [10], which corresponds to $(m, 1)$-robust Bayesian learning, mitigates misspecification, being multimodal, but it is largely affected by the outlying data point (see the dark blue curve associated to $m = 10,\ t = 1$). Conversely, (m, t)-robust Bayesian learning for $t < 1$ mitigates both misspecification and the presence of the outlier: the predictive distribution resulting from the minimization of the proposed robust (m, t)-free energy criterion (see light blue, pink and red curves) is not only multimodal, but it can also better suppress the effect of outliers.

The work of M. Zecchin is funded by the Marie Curie action WINDMILL (Grant agreement No. 813999), while O. Simeone and S. Park have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731). The work of M. Zecchin and O. Simeone was also supported by the European Union's Horizon Europe project CENTRIC (Grant agreement No. 101096379). The work of O. Simeone has also been supported by an Open Fellowship of the EPSRC with reference EP/W024101/1. M. Kountouris has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation programme (Grant agreement No. 101003431). The work of D. Gesbert was supported by the 3IA artificial intelligence interdisciplinary project French funded by ANR, No. ANR-19-P3IA-0002.

Marios Kountouris and David Gesbert are with the Communication Systems Department, EURECOM, Sophia-Antipolis, France (e-mail: kountour@eurecom.fr,gesbert@eurecom.fr).

Matteo Zecchin, Sangwoo Park and Osvaldo Simeone are with the King's Communications, Learning & Information Processing (KCLIP) lab, Department of Engineering, King's College London, London WC2R 2LS, U.K. (e-mail: matteo.1.zecchin@kcl.ac.uk, sangwoo.park@kcl.ac.uk; osvaldo.simeone@kcl.ac.uk).

The second author has contributed to the definition of technical tools and to the experiments. The third author has had an active role in defining problem and tools, as well as in writing the text, while the last two authors have had a supervisory role.

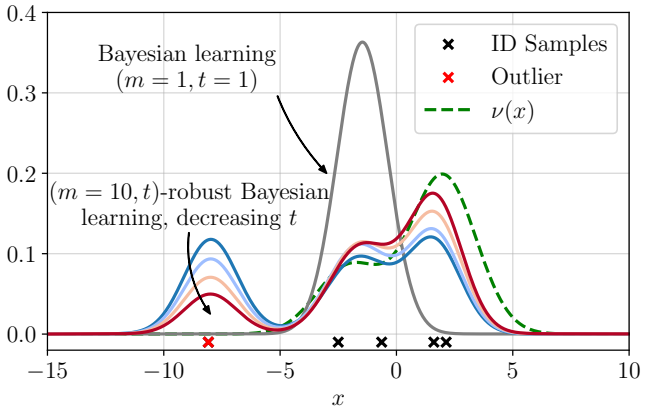


Fig. 1: Ensemble predictive distributions $p_q(x)$ obtained via conventional and robust Bayesian learning based on the data set represented as crosses. The underlying in-distribution (ID) measure $\nu(x)$ (dashed line – a mixture of Gaussians) produces the data points in black, while the contaminating out-of-distribution (OOD) measure $\xi(x)$ produces the data point in red. Conventional Bayesian learning ($m = 1$, $t = 1$) is shown in gray; the $(m, 1)$-robust Bayesian learning approach in [10] with $m = 10$ and $t = 1$ is in dark blue; and the proposed (m, t)-robust Bayesian learning with $m = 10$ and $t = \{0.9, 0.7, 0.5\}$ is displayed in light blue, pink and red.

A. Related Work

Recent work has addressed the problem of model misspecification for Bayesian learning, using tighter approximations of the ensemble risk [10], [11], using pseudo-likelihoods [12] or modeling aleatoric uncertainty [13]. In particular, references [10], [11] have argued that the minimization of the standard free energy criterion – which defines Bayesian learning [8], [9] – yields predictive distributions that do not take advantage of ensembling, and thus have poor generalization capabilities for misspecified models.

To mitigate this problem, references [10], [11] introduced alternative free energy criteria that account for misspecification. The author of [11] leveraged a second-order Jensen's inequality to obtain a tighter bound on the cross entropy loss; while the work [10] proposed an m-free energy criterion that accounts for the performance of an ensemble predictor with m constituent models. Both optimization criteria were shown to be effective in overcoming the shortcomings of Bayesian learning under misspecification, by yielding posteriors that make better use of ensembling.

The free energy metrics introduced in [10], [11] are defined by using the standard log-loss, which is known to be sensitive to outliers. This is because the log-loss grows unboundedly on data points that are unlikely under the model [14]. Free energy criteria metrics based on the log-loss amount to Kullback–Leibler (KL) divergence measures between data and model distributions. A number of papers have proposed to mitigate the effect of outliers by replacing the classical criteria based on the KL divergence in favor of more robust divergences, such as the β-divergences [15], [16] and the γ-divergence [17], [18]. These criteria can be interpreted as substituting the log-loss with generalized logarithmic scoring rules. To optimize such criteria, variational methods have been proposed that were shown to be robust to outliers, while not addressing model misspecification [19].

A separate line of work is focused on addressing the problem of prior misspecification. Prior misspecification refers to learning scenarios in which the true underlying distribution may not be well-represented by the chosen prior, causing Bayesian methods to produce biased or erroneous predictions. To address this issue, generalized formulations of Bayesian learning were introduced that aim to mitigate the influence of misspecified priors by generalizing the standard Bayesian learning criterion to alternative classes of prior-dependent information-theoretic regularizers [7], [8], [20]–[22].

B. Contributions

This work extends standard Bayesian learning by concurrently tackling model misspecification, with respect to both likelihood and prior, and the presence of outliers. Specifically, the contributions of this paper are as follows.

- We introduce the (m, t)-robust Bayesian learning framework, which is underpinned by a novel free energy criterion based on ensemble-based loss measures and generalized logarithmic scoring rules. The predictive distribution resulting from the minimization of the proposed objective takes full advantage of ensembling, while at the same time reducing the effect of outliers.
- We generalize the (m, t)-robust Bayesian learning framework to encompass Rényi divergence-based prior regularizers, which allow to mitigate the detrimental effect of misspecified priors.
- We theoretically justify and analyze the proposed robust m-free energy criterion within the PAC-Bayesian framework, and we prove its enhanced robustness through the lens of the influence function [23].
- We present a wide array of experiments that corroborate the theoretical results, while also highlighting the enhanced generalization capabilities and calibration performance of the proposed learning criterion under model misspecification, prior misspecification and with data sets corrupted by outliers.

C. Paper Organization

The rest of the paper is organized as follows. In Section II, we review the generalized logarithm function, the associated entropy and divergence measures. We also formally describe the learning setup, providing the definition of model misspecification and introducing the contamination model under consideration. After reviewing the standard free energy criterion and its multi-sample version proposed in [10], we provide a toy example highlighting the shortcoming of these two learning criteria when the model class is misspecified and the training data contains outliers. Then, in Section III, we introduce the (m, t)-robust Bayesian learning framework that tackles both model misspecification and the presence of

outliers, and that overcomes the limitations of the standard Bayesian learning rule. We theoretically analyze the proposed learning criterion, providing PAC-Bayesian guarantees for the ensemble model with respect to the contaminated and the in-distribution measures. In Sec. IV, we introduce a generalized form of robust Bayesian learning that addresses also pior misspecification. Finally, in Section V, we provide regression and classification experiments to quantitatively and qualitatively measure the performance of the proposed learning criterion.

II. Preliminaries

A. Generalized Logarithms

The t-logarithm function, also referred to as generalized or tempered logarithm is defined as

$$\log_t(x) := \frac{1}{1-t}\left(x^{1-t}-1\right) \quad \text{for } x>0, \tag{1}$$

for $t \in [0,1)\cup(1,\infty)$, and

$$\log_1(x) := \log(x) \quad \text{for } x>0 \tag{2}$$

where the standard logarithm (2) is recovered from (1) in the limit $\lim_{t\to 1}\log_t(x) = \log(x)$. As shown in Figure 2, for $t \in [0,1)$, the t-logarithms is a concave function, and for $t<1$ is lower bounded as $\log_t(x) \geq -(1-t)^{-1}$.

Largely employed in classical and quantum physics, the t-logarithm has also been applied to machine learning problems. Specifically, t-logarithms have been used to define alternatives to the log-loss as a score function for probabilistic predictors with the aim of enhancing robustness to outliers [6], [24], [25]. Accordingly, the loss associated to a probabilistic model $q(x)$ is measured as $-\log_t q(x)$ instead of the standard log-loss $-\log q(x)$. Note that we have the upper bound $-\log_t q(x) \leq (1-t)^{-1}$ for $t<1$.

In information theory, the t-logarithm was used by [26] to define the t-Tsallis entropy

$$H_t(p(x)) := -\int p(x)^t \log_t p(x) dx, \tag{3}$$

and the t-Tsallis divergence

$$D_t(p(x)||q(x)) := -\int p(x)^t [\log_t p(x) - \log_t q(x)] dx. \tag{4}$$

When using the Tsallis divergence (4) as an optimization criterion in machine learning, the concept of escort distribution is often useful [27]. Given a probability density $p(x)$, the associated t-escort distribution is defined as

$$\mathcal{E}_t(p(x)) = \frac{p(x)^t}{\int p(x)^t dx}. \tag{5}$$

Another popular divergence related to the Tsallis divergence, and hence also to the t-logarithm, is the t-Rényi entropy. For $t \in [0,1)\cup(1,\infty)$, it is defined as

$$H_t^R(p(x)) := \frac{1}{1-t}\log\int p(x)^t dx, \tag{6}$$

which can be obtained as a monotonically increasing function of the t-Tsallis entropy as

$$H_t^R(p(x)) = \frac{1}{1-t}\log\left(1+(1-t)H_t(p(x))\right). \tag{7}$$

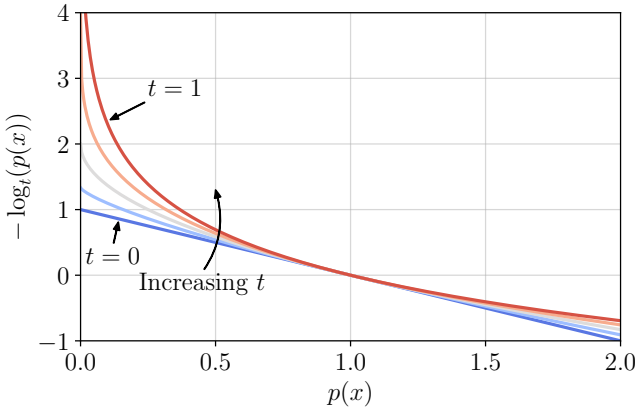


Fig. 2: t-logarithm loss, or $\log_t$-loss, of a predictive distribution $p(x)$ for different values of t. For $t=1$, the samples x corresponding to low predictive probability $p(x)\to 0$ have a potentially unbounded loss value. On the contrary, for $t<1$, the t-logarithm loss is bounded by $(1-t)^{-1}$ and it limits their influence.

The associated t-Rényi divergence is defined as

$$D_t^R(p(x)||q(x)) := \frac{1}{t-1}\log\int p(x)^t q(x)^{1-t} dx, \tag{8}$$

which is related via a monotonically increasing function to the t-Tsallis divergence as

$$D_t^R(p(x)||q(x)) = \frac{1}{t-1}\log\left(1+(t-1)D_t(p(x)||q(x))\right). \tag{9}$$

In the limit $t\to 1$, both t-Tsallis and t-Rényi entropies recover the Shannon (differential) entropy, i.e.,

$$\lim_{t\to 1} H_t(p(x)) = \lim_{t\to 1} H_t^R(p(x)) = \mathbb{E}_{p(x)}[-\log p(x)]. \tag{10}$$

Furthermore, under the same limit, both t-Tsallis and t-Rényi divergences recover the Kullback–Leibler (KL) divergence, i.e.,

$$\lim_{t\to 1} D_t(p(x)||q(x)) = \lim_{t\to 1} D_t^R(p(x)||q(x)) = \mathbb{E}_{p(x)}\left[\log\frac{p(x)}{q(x)}\right]. \tag{11}$$

We finally note that t-logarithm does not satisfy the distributive property of the logarithm, i.e., $\log(xy) = \log(x)+\log(y)$. Instead, we have the equalities [28]

$$\log_t(xy) = \log_t x + \log_t y + (1-t)\log_t x \log_t y \tag{12}$$

and

$$\log_t\left(\frac{x}{y}\right) = y^{t-1}\left(\log_t x - \log_t y\right). \tag{13}$$

TABLE I: Total variation (TV) distance between the ID measure $\nu(x)$ and the predictive distribution $p_q(x)$ obtained from the optimization of the different free energy criteria described in the text.

	$m=1$ $t=1$	$m=10$ $t=1$	$m=10$ $t=0.9$	$m=10$ $t=0.7$	$m=10$ $t=0.5$
$\mathrm{TV}(\nu(x)\|\|p_q(x))$	0.59	0.35	0.30	0.24	0.19

B. Assumptions and Motivation

We consider a learning setup in which the data distribution is contaminated by outliers [29], and the assumed parametric model is misspecified [11]. As in [29], the presence of outliers is modelled by assuming that the observed vector $x \in \mathcal{X}$ follows a sampling distribution $\tilde{\nu}(x)$ given by the contamination of an in-distribution (ID) measure $\nu(x)$ by an out-of-distribution (OOD) measure $\xi(x)$.

Assumption 1 (Outliers)**.** *The sampling distribution follows the gross error model proposed by [29],*

$$\tilde{\nu}(x) = (1-\epsilon)\nu(x) + \epsilon\xi(x) \tag{14}$$

where $\nu(x)$ is the ID measure; $\xi(x)$ is the OOD measure accounting for outliers; and $\epsilon \in [0,1]$ denotes the contamination ratio.

In order for model (14) to be meaningful, one typically assumes that the OOD measure $\xi(x)$ is large for values of x at which the ID measure $\nu(x)$ is small.

The learner is assumed to have access to a data set $\mathcal{D} = \{(x_i)\}_{i=1}^n \sim \tilde{\nu}(x)^{\otimes n}$ drawn from the sampling distribution, and it assumes a uniformly upper bounded parametric model family $p_\theta(x)$ defined by a model parameter vector $\theta \in \Theta$, which is generally misspecified.

Assumption 2 (Misspecification)**.** *The model class $\{p_\theta(\cdot) : \theta \in \Theta\}$ is misspecified with respect to the ID measure $\nu(x)$, in the sense that there is no model parameter vector $\theta \in \Theta$ such that $\nu(x) = p_\theta(x)$. Furthermore, it is uniformly upper bounded, in the sense that there exists a finite constant C such that $p_\theta(x) \leq C$ for all $\theta \in \Theta$ and values of $x \in \mathcal{X}$.*

In order to account for misspecification, as in [11], we adopt ensemble models of the form

$$p_q(x) := \mathbb{E}_{q(\theta)}[p_\theta(x)], \tag{15}$$

where $q(\theta)$ is the ensembling distribution on the model parameter space Θ. The rationale for considering ensemble models is that the average (15), which accounts from the combinations of multiple models $p_\theta(\cdot)$, may better represent the ID measure $\nu(x)$ in the presence of misspecification (see Assumption 2).

The $\log_t$-loss of the ensemble model (15) is given as

$$\mathcal{R}_t(q,x) := -\log_t p_q(x) = -\log_t \mathbb{E}_{q(\theta)}[p_\theta(x)]. \tag{16}$$

This contrasts with the average $\log_t$-loss obtained by drawing a model parameter vector $\theta \sim q(\theta)$ and then using the resulting model $p_\theta(x)$ — an approach known as Gibbs model. The corresponding $\log_t$-loss is

$$\hat{\mathcal{R}}_t(q,x) := \mathbb{E}_{q(\theta)}[-\log_t p_\theta(x)]. \tag{17}$$

Unlike the standard log-loss with $t=1$, the $\log_t$-loss is upper bounded by $(1-t)^{-1}$ for any $t \in (0,1)$. This constrains the impact of anomalous data points to which the model — either $p_\theta(x)$ or $p_q(x)$ for Gibbs and ensemble predictors, respectively — assigns low probability.

Since the sampling distribution $\tilde{\nu}(x)$ is not known, the risk

$$\mathcal{R}_t(q) := \mathbb{E}_{\tilde{\nu}(x)}[-\log_t \mathbb{E}_{q(\theta)}[p_\theta(x)]] \tag{18}$$

of the ensemble model cannot be computed by the learner. However, for $t=1$, using Jensen's inequality and standard PAC-Bayes arguments, the risk (18) can be upper bounded using the data set $\mathcal{D}$ (and neglecting inessential constants) by the free energy criterion [1], [30]

$$\mathcal{J}(q) := \frac{1}{n}\sum_{x\in\mathcal{D}} \hat{\mathcal{R}}_1(q,x) + \frac{D_1(q(\theta)||p(\theta))}{\beta} \tag{19}$$

where we recall that $D_1(q(\theta)||p(\theta))$ is the KL divergence with respect to a prior distribution $p(\theta)$, while $\beta > 0$ is a constant, also known as inverse temperature.

The criterion (19), for $\beta = n$, is minimized by the standard Bayesian posterior i.e.,

$$q_{Bayes}(\theta) = \arg\min_q \sum_{x\in\mathcal{D}} \hat{\mathcal{R}}_1(q,x) + D_1(q(\theta)||p(\theta)) \tag{20}$$

$$\propto p(\mathcal{D}|\theta)p(\theta). \tag{21}$$

Even disregarding outliers, in the misspecified setting, the resulting ensemble predictor $p_{q_{Bayes}}(x) = \mathbb{E}_{q_{Bayes}(\theta)}[p_\theta(x)]$ is known to lead to poor performance, as the criterion (19) neglects the role of ensembling to mitigate misspecification [11].

Example: Consider a Gaussian model class $p_\theta(x) = \mathcal{N}(x|\theta,1)$ and a prior $p(\theta) = \mathcal{N}(\theta|0,9)$. We obtain the standard Bayesian posterior $q_{Bayes}(\theta)$ by minimizing the free energy (19) with $n=5$ data points sampled from the ID measure $\nu(x) = 0.7\mathcal{N}(x|2,2) + 0.3\mathcal{N}(x|-2,2)$, contaminated by an OOD measure $\xi(x) = \mathcal{N}(x|-8,1)$ with a contamination ratio $\epsilon = 0.1$. Note that the model is misspecified, since it assumes a single Gaussian, while the ID measure $\nu(x)$ is a mixture of Gaussians. Therefore, the resulting predictive ensemble distribution $p_{q_{Bayes}}(x)$ (gray line) is not able to capture the multimodal nature of the sampling distribution. Furthermore, the presence of outliers leads to a predictive distribution that deviates from the ID distribution $\nu(x)$ (green dashed line). ■

C. PAC^m-Bayes

The limitations of the standard PAC-Bayes risk bound (19) as a learning criterion in the presence of model misspecification have been formally investigated by [11] and [10]. These works do not consider the presence of outliers, hence setting the sampling distribution $\tilde{\nu}(x)$ to be equal to the ID measure $\nu(x)$ (or $\epsilon = 0$ in 14). Here we review the PAC^m bound introduced by [10] with the goal of overcoming the outlined limitations of (19) in the presence of misspecification.

In [10], the free energy criterion (19) is modified by replacing the Gibbs risk $\hat{\mathcal{R}}_1(q,x)$ with a sharper bound on

the ensemble log-loss $\mathcal{R}_1(q,x)$. For $m \geq 1$, this bound is defined as

$$\hat{\mathcal{R}}_1^m(q,x) := \mathbb{E}_{\theta_1,\ldots,\theta_m \sim q(\theta)^{\otimes m}} \left[-\log \mathbb{E}_{j\sim U[1:m]}[p(x|\theta_j)] \right] \tag{22}$$

where the inner expectation is over an index j uniformly distributed in the set $[1:m] = \{1, 2, \ldots, m\}$.

By leveraging the results of [31] and [32], the multi-sample criterion $\hat{\mathcal{R}}_1^m(q,x)$ can be shown to provide a sharper bound to the ensemble risk $\mathcal{R}_1(q,x)$ in (16) as compared to the Gibbs risk $\hat{\mathcal{R}}_1^1(q,x)$ in (17), i.e.,

$$\mathcal{R}_1(q,x) \leq \hat{\mathcal{R}}_1^m(q,x) \leq \hat{\mathcal{R}}_1^1(q,x) = \hat{\mathcal{R}}_1(q,x) \tag{23}$$

Furthermore, the first inequality in (23) becomes asymptotically tight as $m \to \infty$, i.e.,

$$\lim_{m\to\infty} \hat{\mathcal{R}}_1^m(q,x) = \mathcal{R}_1(q,x). \tag{24}$$

Using PAC-Bayes arguments, [10] show that the log-risk (18) with $t = 1$ and $\tilde{\nu}(x) = \nu(x)$ can be upper bounded, for $\beta > 0$, (neglecting inessential constants) by the m-free energy criterion

$$\mathcal{J}^m(q) := \frac{1}{n} \sum_{x\in\mathcal{D}} \hat{\mathcal{R}}_1^m(q,x) + \frac{m}{\beta} D_1(q(\theta)||p(\theta)). \tag{25}$$

The minimization of the m-free energy $\mathcal{J}^m(q)$ produces a posterior

$$q^m(\theta) := \arg\min_q \mathcal{J}^m(q), \tag{26}$$

which can take better advantage of ensembling, resulting in predictive distributions that are more expressive than the ones obtained following the standard Bayesian approach based on (19).

Example (continued): This is shown in Figure 1, in which we plot the predictive distribution $p_{q^m}(\theta)$ obtained by minimizing the m-free energy $\mathcal{J}^m(q)$ in (25) for $m = 10$ for the same example described in the previous subsection. The optimized predictive distribution $p_{q^m}(x)$ is multimodal; it covers all data samples; and, as shown in Table I, it reduces the total variational distance from the ID measure $\nu(x)$ as compared to the predictive distribution obtained minimizing $\mathcal{J}(q)$. ■

III. (m,t)-Robust Bayesian Learning

In the previous section, we reviewed the m-free energy criterion introduced by [10], which was argued to produce predictive distributions that are more expressive, providing a closer match to the underlying sampling distribution $\nu(x)$. However, the approach is not robust to the presence of outliers. In this section, we introduce (m,t)-robust Bayesian learning and the associated novel free energy criterion that addresses both expressivity in the presence of misspecification and robustness in setting with outliers. To this end, we study the general setting described in Section II-B in which the sampling distribution $\tilde{\nu}(x)$ satisfies both Assumption 1 and Assumption 2, and we investigate the use of the $\log_t$-loss with $t \in [0,1)$ as opposed to the standard log-loss as assumed in [10].

A. Robust m-free Energy

For a proposal posterior $q(\theta)$, generalizing (22), we define the multi-sample empirical $\log_t$-loss evaluated at a data point x as

$$\hat{\mathcal{R}}_t^m(q,x) := \mathbb{E}_{\theta_1,\ldots,\theta_m \sim q(\theta)^{\otimes m}} \left[-\log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j) \right]. \tag{27}$$

From the concavity of the t-logarithm with $t \in [0,1)$, in a manner similar to (23), the loss (27) provides an upper bound on the original $\log_t$-loss $\mathcal{R}_t(q,x)$ in (16)

$$\mathcal{R}_t(q,x) \leq \hat{\mathcal{R}}_t^m(q,x). \tag{28}$$

Furthermore, the bound becomes increasingly tighter as m increases, and we have the limit

$$\lim_{m\to\infty} \hat{\mathcal{R}}_t^m(q,x) = \mathcal{R}_t(q,x) \tag{29}$$

for $t \in [0,1)$. The m-sample $\log_t$-loss (27) is used to define, for $\beta > 0$, the robust m-free energy as

$$\mathcal{J}_t^m(q) := \frac{1}{n} \sum_{x\in\mathcal{D}} \hat{\mathcal{R}}_t^m(q,x) + \frac{m}{\beta} D_1(q(\theta)||p(\theta)). \tag{30}$$

The proposed free energy generalizes the standard free energy criterion (19), which corresponds to the training criterion of (m,t)-robust Bayesian learning for $m = 1$ and $t = 1$, and the m-free energy criterion (25), which corresponds to the training criterion of (m,t)-robust Bayesian learning for $t = 1$.

Following similar steps as in [10], the robust m-free energy can be proved to provide an upper bound on the $\log_t$-risk in (18), as detailed in the following lemma.

Lemma 1. *With probability $1-\sigma$, with $\sigma \in (0,1)$, with respect to the random sampling of the data set $\mathcal{D}$, for all distributions $q(\theta)$ that are absolutely continuous with respect the prior $p(\theta)$, the following bound on the risk (18) of the ensemble model holds*

$$\mathcal{R}_t(q) \leq \mathcal{J}_t^m(q) + \psi(\tilde{\nu}, n, m, \beta, p, \sigma) \tag{31}$$

where

$$\psi(\tilde{\nu}, n, m, \beta, p, \sigma) := \frac{1}{\beta} \left(\log \mathbb{E}_{\mathcal{D},p(\theta)} \left[e^{\beta \Delta_{m,n}} \right] - \log \sigma \right) \tag{32}$$

and

$$\begin{aligned} \Delta_{m,n} :=& \frac{1}{n} \sum_{x\in\mathcal{D}} \log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j) \\ & - \mathbb{E}_{\tilde{\nu}(x)} \left[\log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j) \right]. \end{aligned} \tag{33}$$

Furthermore, the risk with respect to the ID measure $\nu(x)$ can be bounded as

$$\begin{aligned} \mathbb{E}_{\nu(x)}[\mathcal{R}_t(q,x)] \leq & \frac{1}{1-\epsilon} \left(\mathcal{J}_t^m(q) + \psi(\tilde{\nu}, n, m, \beta, p, \sigma) \right) \\ & + \frac{\epsilon(C^{1-t}-1)}{(1-\epsilon)(1-t)}, \end{aligned} \tag{34}$$

if the contamination ratio satisfies the inequality $\epsilon < 1$.

Lemma 1 provides an upper bound on the $\log_t$-risk (18), which is defined with respect to the sampling distribution $\tilde{\nu}(x)$

corrupted by outliers, as well as on the ensemble $\log_t$-risk (16) evaluated with respect to the ID measure $\nu(x)$. Reflecting that the data set $\mathcal{D}$ contains samples from the corrupted measure $\tilde{\nu}(x)$, while the bound (31) vanishes as $n \to \infty$, a non-vanishing term appears in the bound (34).

B. Minimizing the Robust m-free Energy

Using standard tools from calculus of variations, it is possible to express the minimizer of the robust m-free energy

$$q_t^m(\theta) := \arg\min_q \mathcal{J}_t^m(q) \tag{35}$$

as fixed-point solution of an operator acting on the ensembling distribution $q(\theta)$.

Theorem 1. *The minimizer (35) of the robust m-free energy objective (30) is the fixed point of the operator*

$$T(q) := p(\theta_j) \exp\left(\beta \sum_{x\in\mathcal{D}} \mathbb{E}_{\{\theta_i\}_{i\neq j}}\left[\log_t\left(\frac{\sum_{i=1}^m p_{\theta_i}(x)}{m}\right)\right]\right) \tag{36}$$

where the average in (36) is taken with respect to the i.i.d. random vectors $\{\theta_i\}_{i\neq j} \sim q(\theta)^{\otimes m-1}$.

Theorem 1 is useful to develop numerical solutions to problem (35) for non-parametric posteriors, and it resembles standard mean-field variational inference iterations [33].

Alternatively, we can tackle the problem (35) over a parametric family of distribution using standard tools from variational inference [34].

To further characterize the posterior minimizing the robust m-free energy criterion, and to showcase the beneficial effect of the generalized logarithm, we now consider the asymptotic regime in which $m \to \infty$ and then $n \to \infty$. In this limit, the robust m-free energy (30) coincides with the $\log_t$-risk $\mathcal{R}_t(q)$. From the definition of t-Tsallis divergence (4), the $\log_t$-risk can be shown in turn to be equivalent to the minimization of the divergence

$$D_t\left(\mathcal{E}_t(\tilde{\nu}(x))||p_{q(\theta)}(x)\right) \tag{37}$$

between the t-escort distribution (5) associated to the sampling distribution $\tilde{\nu}(x)$ and the ensemble predictive distribution $p_{q(\theta)}(x)$. Therefore, unlike the standard Bayesian setup with $t = 1$, the minimizer of the robust m-free energy does not seek to approximate the sampling distribution $\tilde{\nu}(x)$. Instead, the minimizing ensembling posterior $q(\theta)$ aims at matching the t-escort version of the sampling distribution $\tilde{\nu}(x)$. In the case of corrupted data generation procedures, i.e., when $\nu(x) \neq \tilde{\nu}(x)$, recovering the sampling distribution $\tilde{\nu}(x)$ is not always the end goal, and, as shown by [6], escort distributions are particularly effective at reducing the contribution of OOD measures.

Example (continued): Consider again the example in Figure 1. The minimization of the proposed robust m-free energy $\mathcal{J}_t^m(q)$ for $m = 10$ and $t = \{0.9, 0.7, 0.5\}$ is seen to lead to expressive predictive distributions (35) that are also able to downweight the contribution of the outlying data point. This is quantified by reduced total variation distances as seen in Table I.

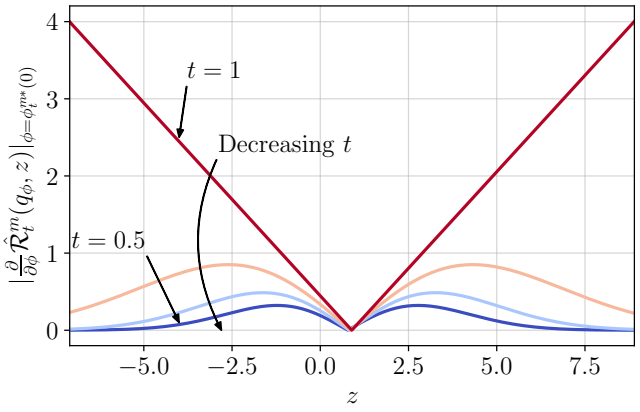


Fig. 3: Absolute value of the contamination dependent term $\frac{\partial}{\partial\phi}\hat{\mathcal{R}}_t^m(q_\phi, z)$ evaluated at $\phi_t^{m*}(0)$ for different values of t. The predictive distribution of the ensemble model concentrates around 1.

C. Influence Function Analysis

In this section, we study the robustness of the proposed free energy criterion by using tools from classical statistics. The robustness of an estimator is typically measured by the means of its influence function [23]. The influence function quantifies the extent to which an estimator derived from a data set $\mathcal{D}$ changes when a data point z is added to $\mathcal{D}$. We are specifically interested in quantifying the effect of data contamination, via the addition of a point z, on the ensembling distribution $q_t^m(\theta)$ that minimizes the proposed robust m-free energy objective (30). To this end, given a set $\mathcal{D}$ of n data points $\{x_1, \ldots, x_n\} \in \mathcal{X}^n$, we define the empirical measure

$$P^n(x) = \frac{1}{n}\sum_{i=1}^n \delta(x - x_i) \tag{38}$$

where $\delta(\cdot)$ denotes the Dirac function, and we introduce its γ-contaminated version for an additional data point $z \in \mathcal{X}$ as

$$P_{\gamma,z}^n(x) = \frac{(1-\gamma)}{n}\sum_{i=1}^n \delta(x - x_i) + \gamma\delta(x - z) \tag{39}$$

with $\gamma \in [0, 1]$.

The following analysis is inspired by [19], which considered Gibbs models trained using generalized free energy criteria based on the β-divergence and γ-divergence.

To compute the influence function we consider parametric ensembling distributions $q_\phi(\theta)$ defined by the parameter vector $\phi \in \Phi \subseteq \mathbb{R}^d$. We denote the robust m-free energy (30) evaluated using the empirical distribution (39) as

$$\mathcal{J}_t^m(\gamma, \phi) = \mathbb{E}_{P_{\gamma,z}^n(x)}\left[\hat{\mathcal{R}}_t^m(q_\phi, x)\right] + \frac{m}{\beta} D_1(q_\phi(\theta)||p(\theta)), \tag{40}$$

and its minimizer as

$$\phi_t^{m*}(\gamma) = \arg\min_{\phi\in\Phi} \mathcal{J}_t^m(\gamma, \phi). \tag{41}$$

The influence function is then defined as the derivative

$$IF_t^m(z,\phi,P^n) = \left.\frac{d\phi_t^{m*}(\gamma)}{d\gamma}\right|_{\gamma=0} \tag{42}$$

$$= \lim_{\gamma\to 0}\frac{\phi_t^{m*}(\gamma)-\phi_t^{m*}(0)}{\gamma}. \tag{43}$$

Accordingly, the influence function measures the extent to which the minimizer $\phi_t^{m*}(\gamma)$ changes for an infinitesimal perturbation of the data set.

Theorem 2. *The influence function of the robust m-free energy objective (40) is*

$$IF_t^m(z,\phi,P^n) = -\left[\frac{\partial^2\mathcal{J}_t^m(\gamma,\phi)}{\partial\phi^2}\right]^{-1}\times\left.\frac{\partial^2\mathcal{J}_t^m(\gamma,\phi)}{\partial\gamma\partial\phi}\right|_{\substack{\gamma=0\\ \phi=\phi_t^{m*}(0)}}, \tag{44}$$

where

$$\frac{\partial^2\mathcal{J}_t^m(\gamma,\phi)}{\partial\phi^2} = \mathbb{E}_{P^n_{\gamma,z}(x)}\frac{\partial^2}{\partial\phi^2}\left[\hat{\mathcal{R}}_t^m(q_\phi,x)\right] \tag{45}$$

$$+\frac{\partial^2}{\partial\phi^2}\left[\frac{m}{\beta}KL(q_\phi(\theta)||p(\theta))\right] \tag{46}$$

and

$$\frac{\partial^2\mathcal{J}_t^m(\gamma,\phi)}{\partial\gamma\partial\phi} = \frac{\partial}{\partial\phi}\left[\mathbb{E}_{P^n(x)}\left[\hat{\mathcal{R}}_t^m(q_\phi,x)\right]-\hat{\mathcal{R}}_t^m(q_\phi,z)\right]. \tag{47}$$

Theorem 2 quantifies the impact of the data point z through the contamination dependent term $\frac{\partial}{\partial\phi}\hat{\mathcal{R}}_t^m(q_\phi,z)$. We study the magnitude of this term to illustrate the enhanced robustness deriving from the proposed robust m-free energy objective. For ease of tractability, we consider the limit $m\to\infty$. In this case, the contamination dependent term can be expressed as

$$\frac{\partial}{\partial\phi}\lim_{m\to\infty}\hat{\mathcal{R}}_t^m(q_\phi,z) = \frac{\partial}{\partial\phi}\log_t\mathbb{E}_{q_\phi(\theta)}[p(z|\theta)] \tag{48}$$

$$= \left[\mathbb{E}_{q_\phi(\theta)}[p(z|\theta)]\right]^{-t}\frac{\partial\mathbb{E}_{q_\phi(\theta)}[p(z|\theta)]}{\partial\phi}. \tag{49}$$

The effect of the t-logarithm function thus appears in the first multiplicative term, and it is the one of reducing the influence of anomalous data points to which the ensemble predictive distribution $p_q(x)$ assigns low probability.

Example: To illustrate how the t-logarithm improves the robustness to outlying data points, we consider again the example of Figure 1 and we assume a parametrized ensembling posterior $q_\phi(\theta)=\mathcal{N}(\theta|\phi,1)$. In Figure 3, we plot the magnitude of the contamination dependent term evaluated at the parameter $\phi_t^{m*}(0)$ that minimizes the robust m-free energy $\mathcal{J}_t^m(0,\phi)$ for $m=\infty$ and different values of t. For all values of t, the optimized predictive distribution concentrates around 0, where most of sampled data points lie. However, as the value of the contaminated data point z becomes smaller and moves towards regions where the ensemble assign low probability, the contamination dependent term grows linearly for $t=1$, while it flattens for $t\in(0,1)$. This showcases the role of the robust m-free energy criterion as a tool to mitigate the influence of outlying data points by setting $t<1$.

TABLE II: Total variation (TV) distance between the ID measure $\nu(x)$ and the predictive distribution $p_q(x)$ obtained from the optimization of the different free energy criteria for the setting in Figure 4 (the TV values are scaled by 10^4).

	$t=1$ $\epsilon=0$	$t=1$ $\epsilon=0.1$	$t=0.9$ $\epsilon=0.1$	$t=0.8$ $\epsilon=0.1$
$\text{TV}(\nu(x)\|\|p_q(x))$	1.38	2.15	1.88	1.79

IV. Generalized (m,t)-Robust Bayesian Learning

So far, we have addressed the problem of model misspecification with respect to the likelihood function $p_\theta(\cdot)$ as defined in Assumption 2. When applying Bayesian learning, a further common concern with regards to misspecification has to do with the choice of the prior distribution $p(\theta)$. In Bayesian learning, as well as robust Bayesian learning as presented in this paper, the prior distribution $p(\theta)$ is accounted for in the design problem by including a regularizer $D_1(q(\theta)||p(\theta))$ on the ensembling distribution $q(\theta)$ under optimization on the free energy objective (see (19) for conventional Bayesian learning and (30) for robust Bayesian learning). It has been recently argued that the KL divergence $D_1(q(\theta)||p(\theta))$ may not offer the best choice for the regularizer when the prior is not well specified due to its mode-seeking behavior [7], [9], [35]. In this section, we extend the generalized Bayesian learning framework in [7] to incorporate robustness to likelihood misspecification and outliers.

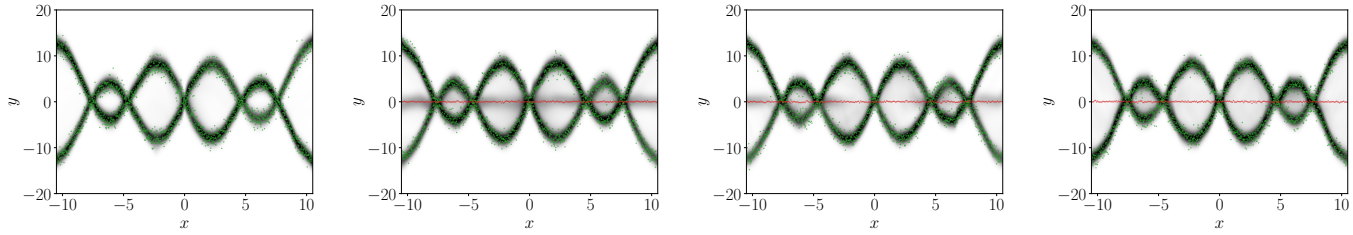
To this end, we extend the (m,t)-robust Bayesian learning criterion (30) by replacing the KL divergence $D_1(q(\theta)||p(\theta))$ with the more general t-Rényi divergence $D_t^R(q(\theta)||p(\theta))$ in (8). Recall that the Rényi divergence tends to the KL divergence as t approaches to 1. This extension is motivated by the fact that, for $t<1$, the Rényi divergence exhibits a mass-covering behavior that has been shown to improve robustness against ill-specified prior distributions [7], [20], [21].

Accordingly, the generalized (m,t)-robust Bayesian learning criterion is defined as

$$\mathcal{J}_{t,t_p}^m(q) := \frac{1}{n}\sum_{x\in\mathcal{D}}\hat{\mathcal{R}}_t^m(q,x)+\frac{m}{\beta}D_{t_p}^R(q(\theta)||p(\theta)). \tag{50}$$

We emphasize that the parameter t_p specifying the Rényi regularizer need not equal the parameter t used for the loss function. In fact, parameter t_p accounts for the degree of robustness that the designer wishes to enforce with respect to the choice of the prior, while parameter t controls robustness to outliers. The (m,t)-robust Bayesian learning criterion is a special case of the generalized criterion (50) for $t_p=1$. Furthermore, the Bayesian learning objective in [7], [20] is recovered by setting $m=1$ and $t=1$.

The results presented in previous section extend to the generalized (m,t)-robust learning criterion as follows. First, the criterion (50) can be obtained as a bound on the risk (18), extending Lemma 1, as briefly elaborated in Appendix A. Furthermore, the influence function analysis developed in Section III-C applies directly also to the generalized criterion (50), since the derivation therein is only reliant on the properties of the t-logarithm and it does not depend on the choice of the prior regularization.



(a) $\epsilon = 0$, $t = 1$ (b) $\epsilon = 0.1$, $t = 1$ (c) $\epsilon = 0.1$, $t = 0.9$ (d) $\epsilon = 0.1$, $t = 0.8$

Fig. 4: Ensemble predictive distribution obtained minimizing different free energy criteria. The samples from the ID measure are represented as green dots, while data points sampled from the OOD component are in red. The optimized predictive distributions are displayed in shades of gray. In (a), we plot the predictive distribution associated to $(m, 1)$-robust Bayesian learning obtained minimizing the m-free energy criterion $\mathcal{J}^m$ of [10] with $m = 20$ by using only samples from the ID measure (i.e., there are no outliers). In (b), we show the predictive distribution obtained by minimizing the same criterion when using samples from the ID measure and OOD measure with a contamination ratio $\epsilon = 0.1$. In (c) and (d) we consider the same scenario as in (b), but we consider the proposed (m, t)-robust Bayesian based on the robust m-free energy criterion $\mathcal{J}_t^m$ with $m = 20$, when setting $t = 0.9$ and $t = 0.8$, respectively.

V. Experiments

In this section, we first describe a simple regression task with an unimodal likelihood, and then we present results for larger-scale classification and regression tasks. The main aim of these experiments is to provide qualitative and quantitative insights into the performance of $(m, 1)$-robust Bayesian learning of [10] and the proposed robust (m, t)-robust Bayesian learning. All examples are characterized by misspecification and outliers.

A. Multimodal Regression

For the first experiment, we modify the regression task studied by [11] and [10] in order to capture not only model misspecification but also the presence of outliers as in the contamination model (14). To this end, we assume that the ID distribution $\nu(x)$, with $x = (a, b)$, is given by $\nu(a, b) = p(a)\nu(b|a)$, where the covariate a is uniformly distributed in the interval $[-10.5, 10.5]$ – i.e., $p(a) = 1/21$ in this interval and $p(a) = 0$ otherwise – and by a response variable b that is conditionally distributed according to the two-component mixture

$$\nu(b|a) = \mathcal{N}(b|\alpha\mu_a, 1), \tag{51}$$

$$\alpha \sim \text{Rademacher}, \tag{52}$$

$$\mu_a = 7 \sin\left(\frac{3a}{4}\right) + \frac{a}{2}. \tag{53}$$

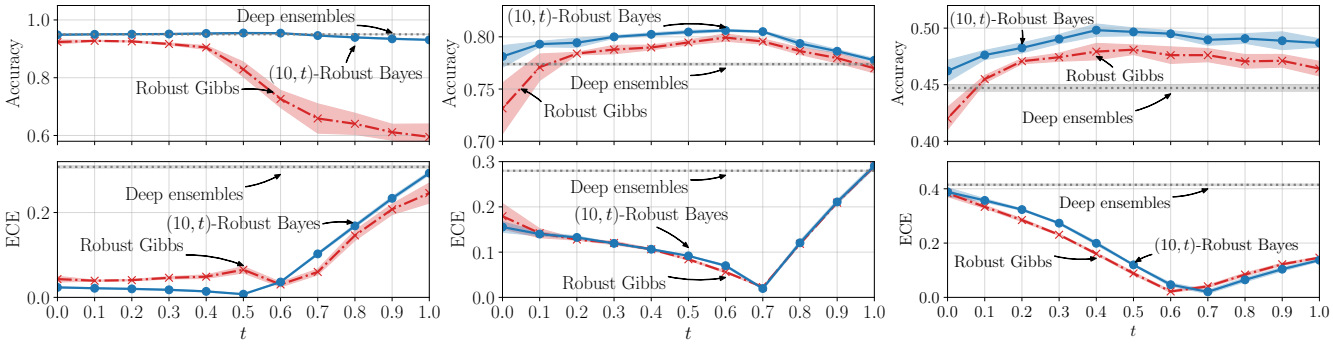
The OOD component $\xi(x) = \xi(a, b) = p(a)\xi(b)$ also has a uniformly distributed covariate a in the interval $[-10.5, 10.5]$, but, unlike the ID measure, the response variable b is independent of a, with a distribution concentrated around $b = 0$ as

$$\xi(b) = \mathcal{N}(b|0, 0.1). \tag{54}$$

The parametric model is given by $p(x|\theta) = p(a, b|\theta) = p(a)\mathcal{N}(b|f_\theta(a), 1)$, where $f_\theta(a)$ is the output of a three-layer fully connected Bayesian neural network with 50 neurons and Exponential Linear Unit (ELU) activation functions [36] in the two hidden layers. We consider a Gaussian prior $p(\theta) = \mathcal{N}(0, I)$ over the neural network weights and use a Monte Carlo estimator of the gradient based on the reparametrization trick [37] as in [38].

Consider first only the effect of misspecification. The parametric model assumes a unimodal likelihood $\mathcal{N}(b|f_\theta(a), 1)$ for the response variable, and is consequently misspecified with respect to the ID measure (51). As a result, the standard Bayesian learning leads to a unimodal predictive distribution that approximates the mean value of the response variable, while $(m, 1)$-robust Bayesian learning can closely reproduce the data distribution [10], [11]. This is shown in Figure 4a, which depicts the predictive distribution obtained by minimizing the m-free energy criterion $\mathcal{J}^m$ with $m = 20$ when using exclusively samples from the ID measure (green dots). In virtue of ensembling, the resulting predictive distribution becomes multimodal, and it is seen to provide a good fit to the data from the ID measure.

Let us evaluate also the effect of outliers. To this end, in Figure 4b we consider $(m, 1)$-robust Bayesian learning and minimize again the m-free energy criterion, but this time using a data set contaminated with samples from the OOD component (red points) and with a contamination ratio $\epsilon = 0.1$. The predictive distribution is seen to cover not only the ID samples but also the outlying data points. In Figure 4c and 4d, we finally plot the predictive distributions obtained by (m, t)-robust Bayesian learning with $m = 20$, when setting $t = \{0.9, 0.8\}$, respectively. The proposed approach is able to mitigate the effect of the outlying component for $t = 0.9$, and, for $t = 0.8$, it almost completely suppresses it. As a result, the proposed energy criterion produces predictive distributions that match more closely the ID measure. This qualitative behavior is quantified in Table II, where we report the total variation distance from the ID measure for the setting and predictors considered in Figure 4.



(a) MNIST data set (b) Extended MNIST data set (c) CIFAR-10 data set

Fig. 5: Test accuracy (top) and expected calibration error (ECE) (bottom) as a function of t under the contamination ratio $\epsilon = 0.3$ for: *(i)* deep ensembles [39]; *(ii)* robust Gibbs predictor, which minimizes the free energy criterion $\mathcal{J}_t^1$ [25]; and *(iii)* (m, t)-robust Bayesian learning, which minimizes the free energy criterion $\mathcal{J}_t^{10}$.

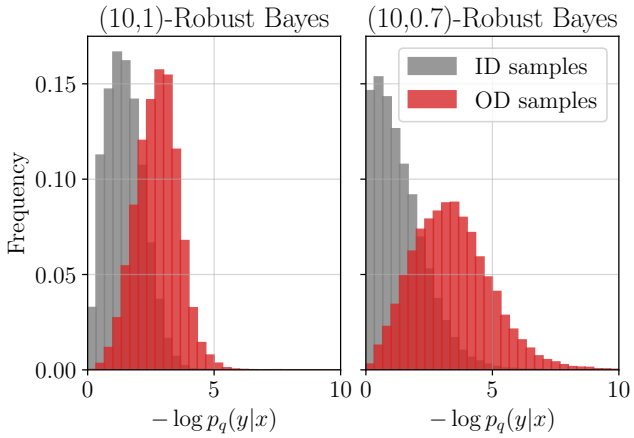


Fig. 6: Distribution of the negative log-likelihood of ID and OD training data samples for an ensemble model minimizing (on the left) the log-loss based criterion $\mathcal{J}_1^{10}$, and (on the right) the proposed robust objective $\mathcal{J}_{0.7}^{10}$ based on the $\log_t$-loss with $t = 0.7$.

B. MNIST and CIFAR-10 Classification Tasks

We now address the problem of training Bayesian neural network classifiers in the presence of misspecification and outliers. We consider three different experimental setups entailing distinct data sets and model architectures:

- Classification of MNIST digits [40] based on a fully connected neural network comprising a single hidden layer with 25 neurons.
- Classification of Extended MNIST characters and digits [41] based on a fully connected neural network with two hidden layers with 25 neurons each.
- Classification of CIFAR-10 [42] images using a convolutional neural network (CNN) with two convolutional layers, the first with 8 filters of size 3×3 and the second with 4 filters of size 2×2, followed by a hidden layer with 25 neurons each.

All hidden units use ELU activations [36] except the last, classifying, layer that implements the standard softmax function. Model misspecification is enforced by adopting neural network architectures with small capacity. As in [25], outliers are obtained by randomly modifying the labels for fraction ϵ of the data points in the training set. Additional details for the experiments can be found in the supplementary material.

We measure the accuracy of the trained models, as well as their calibration performance. Calibration refers to the capacity of a model to quantify uncertainty (see, e.g., [39]). We specifically adopt the expected calibration error (ECE) [43], a standard metric that compares model confidence to actual test accuracy (see supplementary material for the exact definition). We train the classifiers using corrupted data sets with a contamination ratio $\epsilon = 0.3$, and then we evaluate their accuracy and ECE as a function of $t \in [0, 1]$ based on a clean ($\epsilon = 0$) holdout data set. We compare the performance of (m, t)-robust Bayesian learning based on the minimization of the robust m-free energy $\mathcal{J}_t^m$, with $m = 10$, to: *(i) deep ensembles* [39], also with 10 models in the ensembles; and *(ii)* the robust Gibbs predictor of [25], which optimizes over a single predictor (not an ensemble) by minimizing the free energy metric $\mathcal{J}_t^1$. The inverse temperature parameter β is set to 0.1 in the (m, t)-robust Bayesian and the Gibbs predictor objectives.

In Figure 5 we report the performance metrics attained by the trained models in the three different setups listed above. From the top panels we conclude that (m, t)-robust Bayesian learning is able to mitigate model misspecification by improving the final accuracy as compared to the robust Gibbs predictor and the deep ensemble models. Furthermore, the use of the robust loss for a properly chosen value of t leads to a reduction of the detrimental effect of outliers and to an increase in the model accuracy performance as compared to the standard log-loss ($t = 1$). In terms of calibration performance, the lower panels demonstrate the capacity of robust ensemble predictors with $t < 1$ to drastically reduce the ECE as compared to deep ensembles. In this regard,

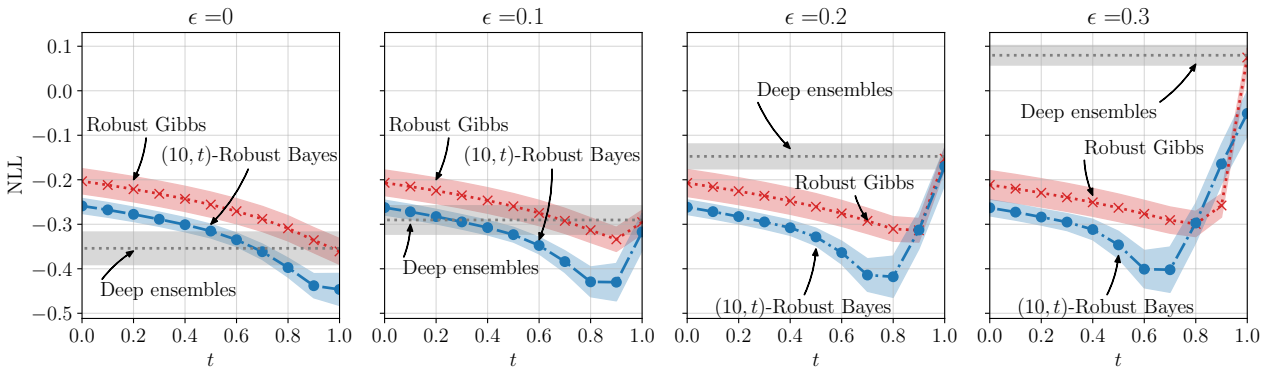


Fig. 7: Negative log-likelihood computed on a uncorrupted data set for: *(i)* deep ensembles [39]; *(ii)* robust Gibbs predictor, which minimizes $\mathcal{J}_t^1$ [25]; and *(iii)* the (m,t)-robust Bayesian learning, which minimizes $\mathcal{J}_t^{10}$. The models are trained on ϵ-contaminated data set for $\epsilon \in \{0, 0.1, 0.2, 0.3\}$

it is also observed that the accuracy and ECE performance levels depend on the choice of parameter t. In practice, the selection of t may be addressed using validation or meta-learning methods in a manner akin to [44]. Additional results on calibration in the form of reliability diagrams [45] can be found in supplementary material.

As shown shown theoretically in Section III-C, the effect of the $\log_t$-loss is to reduce the influence of outliers during training for $t < 1$. We empirically investigate the effect of the robust loss in Figure 6, in which we compare the distribution of the negative log-likelihood for ID and OD training data samples. We focus on the CIFAR-10 data set, and we compare the histogram of the negative log-likelihood under a CNN model trained based on the m-free energy $\mathcal{J}_1^m$, with $m = 10$ and standard logarithmic loss, and a CNN minimizing the proposed robust m-free energy $\mathcal{J}_t^m$, with $m = 10$ and $t = 0.7$. The $(m, 1)$-robust Bayesian based on the standard log-loss tries to fit both ID and OD samples and, as a result, the two components have similar likelihoods. In contrast, (m, t)-robust Bayesian learning is able to downweight the influence of outliers and to better fit the ID component.

C. California Housing Regression Task

We consider the problem of training a robust regressor based on training data sets corrupted by outliers and in the presence of model misspecification. We consider the California housing dataset, which is characterized by response variables y normalized in the $[0, 1]$ interval, and we fix a unimodal likelihood $p(y|x, \theta) = \mathcal{N}(y|f_\theta(x), 0.1)$, where $f_\theta(x)$ is the output of a three-layer neural network with hidden layers comprising 10 units with ELU activation functions [36]. We consider a Gaussian prior $p(\theta) = \mathcal{N}(\theta|0, I)$. The model class is misspecified since the response variable is bounded and hence not Gaussian. Outliers are modeled by replacing the label of fraction ϵ of the training sample with random labels picked uniformly at random within the $[0, 1]$ interval.

We consider training based on data sets with different contamination ratios $\epsilon \in \{0, 0.1, 0.2, 0.3\}$, and measure the trained model ability to approximate the ID data by computing the negative log-likelihood on a clean holdout data set ($\epsilon = 0$). As in the previous subsection, we compare models trained using (m, t)-robust Bayesian learning, with $m = 5$, to: *(i) deep ensembles* [39], also with 5 models in the ensembles; and *(ii)* the robust Gibbs predictor of [25] minimizing the free energy metric $\mathcal{J}_t^1$. The inverse temperature parameter β is set to 0.1 in the (m, t)-robust Bayesian and the Gibbs predictor objectives.

In Figure 7 we report the negative log-likelihood of an uncontaminated data set for models trained according to the different learning criteria. The leftmost panel ($\epsilon = 0$) corresponds to training based on an uncontaminated data set. For this case, the best performance is obtained for $t = 1$ – an expected result due to the absence of outliers – and the proposed criterion outperforms both the Gibbs predictor and deep ensembles, as it is capable of counteracting misspecification by the means of ensembling. In the remaining panels, training is performed based on ϵ-contaminated data sets, with the contamination ϵ increasing from left to right. In these cases, learning criteria based on robust losses are able to retain similar performance to the uncontaminated case for suitable chosen values of t. Furthermore, the optimal value of t is observed to increase with the fraction of outliers in the training data set.

D. Robustness to Prior Misspecification

We finally turn to exploring the robustness of the generalized (m, t)-robust criterion (50) with respect to the choice of the prior distribution. To this end, we consider the same regression problem and likelihood model of the previous subsection, but we allow for a Gaussian prior distribution $p(\theta|\Delta\mu_p) = \mathcal{N}(\theta|\Delta\mu_p I, 0.1I)$ with a generally non-zero mean $\Delta\mu_p$. In Bayesian neural network training, it is customary to set $\Delta\mu_p = 0$, favoring posteriors with a small expected norm, which are expected to generalize better [46], [47]. In order to study the impact of misspecification, similarly to [22], we evaluate the performance obtained with different prior regularizers when choosing a non-zero prior mean $\Delta\mu_p$. Non-zero values of the prior may be considered to be misspecified as they do not comply with the actual expectation on the best model parameters for this problem.

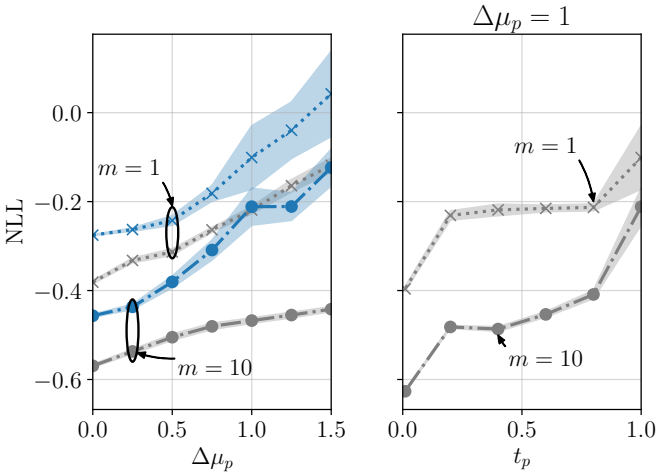


Fig. 8: Negative log-likelihood obtained minimizing the free energy with the Rényi entropy and different different values of t_p. In the left panel, we consider the generalized (m, t)-robust Bayesian learning for $t = 1$ and $m \in \{1, 10\}$ using the standard KL regularizer (in blue) and the Rényi regularizer for $t_p = 0.5$ (in gray). In the right panel we fix the parameter $\Delta\mu_p = 1$ and evaluate the performance as a function of t_p.

In the leftmost panel of Figure 8, we show the negative log-likelihood obtained by (m, t)-robust Bayesian learning for $t = 1$ and $m \in \{1, 10\}$, which uses the standard KL regularizer (in blue), as well as by generalized robust Bayesian learning with the Rényi regularizer with $t_p = 0.5$ (in gray). The advantage of the generalized approach is particularly apparent for $m = 10$, in which case generalized (m, t)-robust Bayesian learning with $t_p = 0.5$ shows a more graceful performance degradation for increasing values of $\Delta\mu_p$.

To further elaborate on the role of the choice of the parameter t_p, in the rightmost panel, we fix the prior parameter as $\Delta\mu_p = 1$, and plot the negative log-likelihood as a function of t_p for $m = 1$ and $m = 10$. In both cases, we find that the robustness to a misspecified prior increases as t_p decreases, demonstrating the advantages of the generalized robust Bayesian learning framework.

VI. Conclusion

In this work, we addressed the problem of training ensemble models under model misspecification and in the presence of outliers. We proposed the (m, t)-robust Bayesian learning framework that leverages generalized logarithm score functions in combination with multi-sample bounds, with the goal of deriving posteriors that are able to take advantage of ensembling, while at the same time being robust with respect to outliers. The proposed learning framework is shown to lead to predictive distributions characterized by better generalization capabilities and calibration performance in scenarios in which the standard Bayesian posterior fails.

The proposed robust Bayesian learning framework can find application to learning scenarios that can benefit from uncertainty quantification in their decision making processes and are characterized by the presence of outliers and model misspecification. Examples include inference in wireless communication systems [48], medical imaging [49] and text sentiment analysis [50], [51].

We conclude by suggesting a number of directions for future research. The (m, t)-robust Bayesian learning has been shown to lead to the largest performance gains for properly chosen values of t. The optimal values of t depend on the particular task at hand, and deriving rules to automate the tuning of these parameters represents a practical and important research question. Furthermore, (m, t)-robust Bayesian learning can be extended to reinforcement learning, as well as to meta-learning, for which Bayesian methods have recently been investigated (see, e.g., [52], [53] and references therein).

References

[1] S. Theodoridis, *Machine learning: a Bayesian and optimization perspective*. Academic Press, 2015.

[2] S. G. Walker, "Bayesian inference with misspecified models," *Journal of Statistical Planning and Inference*, vol. 143, no. 10, pp. 1621–1633, 2013.

[3] P. Grünwald and T. Van Ommen, "Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it," *Bayesian Analysis*, vol. 12, no. 4, pp. 1069–1103, 2017.

[4] R. Martinez-Cantin, K. Tee, and M. McCourt, "Practical Bayesian optimization in the presence of outliers," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 1722–1731.

[5] D. Madigan, A. E. Raftery, C. Volinsky, and J. Hoeting, "Bayesian model averaging," in *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, Portland, OR*, 1996, pp. 77–83.

[6] T. Sypherd, M. Diaz, J. K. Cava, G. Dasarathy, P. Kairouz, and L. Sankar, "A loss function for robust classification: Calibration, landscape, and generalization," *arXiv preprint arXiv:1906.02314*, 2019.

[7] J. Knoblauch, J. Jewson, and T. Damoulas, "An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference," *Journal of Machine Learning Research*, vol. 23, no. 132, pp. 1–109, 2022.

[8] ——, "Generalized variational inference: Three arguments for deriving new posteriors," *arXiv preprint arXiv:1904.02063*, 2019.

[9] O. Simeone, *Machine Learning for Engineers*. Cambridge University Press, 2022.

[10] W. R. Morningstar, A. A. Alemi, and J. V. Dillon, "PAC^m-Bayes: narrowing the empirical risk gap in the misspecified Bayesian regime," *arXiv preprint arXiv:2010.09629*, 2020.

[11] A. R. Masegosa, "Learning under model misspecification: Applications to variational and ensemble methods," *arXiv preprint arXiv:1912.08335*, 2019.

[12] B.-E. Chérief-Abdellatif and P. Alquier, "Mmd-bayes: Robust bayesian estimation via maximum mean discrepancy," in *Symposium on Advances in Approximate Bayesian Inference*. PMLR, 2020, pp. 1–21.

[13] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[14] J. Jewson, J. Q. Smith, and C. Holmes, "Principles of Bayesian inference using general divergence criteria," *Entropy*, vol. 20, no. 6, p. 442, 2018.

[15] A. Basu, I. R. Harris, N. L. Hjort, and M. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998.

[16] A. Ghosh and A. Basu, "Robust Bayes estimation using the density power divergence," *Annals of the Institute of Statistical Mathematics*, vol. 68, no. 2, pp. 413–437, 2016.

[17] H. Fujisawa and S. Eguchi, "Robust parameter estimation with a small bias against heavy contamination," *Journal of Multivariate Analysis*, vol. 99, no. 9, pp. 2053–2081, 2008.

[18] T. Nakagawa and S. Hashimoto, "Robust Bayesian inference via γ-divergence," *Communications in Statistics-Theory and Methods*, vol. 49, no. 2, pp. 343–360, 2020.

[19] F. Futami, I. Sato, and M. Sugiyama, "Variational inference based on robust divergences," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 813–822.

[20] Y. Li and R. E. Turner, "Rényi divergence variational inference," *Advances in neural information processing systems*, vol. 29, 2016.

[21] X. Yue and R. Kontar, "The Rényi gaussian process: Towards improved generalization," *arXiv preprint arXiv:1910.06990*, 2019.

[22] J. Knoblauch, "Frequentist consistency of generalized variational inference," *arXiv preprint arXiv:1912.04946*, 2019.

[23] F. R. Hampel, "The influence curve and its role in robust estimation," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 383–393, 1974.

[24] E. Amid and M. K. Warmuth, "A more globally accurate dimensionality reduction method using triplets." *arXiv preprint arXiv:1803.00854*, 2018.

[25] E. Amid, M. K. Warmuth, R. Anil, and T. Koren, "Robust bi-tempered logistic loss based on bregman divergences," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[26] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *Journal of statistical physics*, vol. 52, no. 1, pp. 479–487, 1988.

[27] T. Sears *et al.*, "Generalized maximum entropy, convexity and machine Learning," 2008.

[28] S. Umarov, C. Tsallis, and S. Steinberg, "On a q-central limit theorem consistent with nonextensive statistical mechanics," *Milan Journal of Mathematics*, vol. 76, no. 1, pp. 307–328, 2008.

[29] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964. [Online]. Available: http://www.jstor.org/stable/2238020

[30] O. Catoni, "A PAC-Bayesian approach to adaptive classification," *preprint*, vol. 840, 2003.

[31] Y. Burda, R. B. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," in *4th International Conference on Learning Representations, ICLR 2016*, 2016.

[32] A. Mnih and D. Rezende, "Variational inference for Monte Carlo objectives," in *International Conference on Machine Learning*. PMLR, 2016, pp. 2188–2196.

[33] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[34] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[35] T. Minka *et al.*, "Divergence measures and message passing," Citeseer, Tech. Rep., 2005.

[36] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[37] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[38] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1613–1622.

[39] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[40] Y. LeCun, "The MNIST database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[41] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "Emnist: Extending mnist to handwritten letters," in *2017 international joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 2921–2926.

[42] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[43] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.

[44] R. Zhang, Y. Li, C. De Sa, S. Devlin, and C. Zhang, "Meta-learning divergences for variational inference," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 4024–4032.

[45] M. H. DeGroot and S. E. Fienberg, "The comparison and evaluation of forecasters," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 32, no. 1-2, pp. 12–22, 1983.

[46] M. Dusenberry, G. Jerfel, Y. Wen, Y. Ma, J. Snoek, K. Heller, B. Lakshminarayanan, and D. Tran, "Efficient and scalable bayesian neural nets with rank-1 factors," in *International conference on machine learning*. PMLR, 2020, pp. 2782–2792.

[47] J. M. Hernández-Lobato and R. Adams, "Probabilistic backpropagation for scalable learning of bayesian neural networks," in *International conference on machine learning*. PMLR, 2015, pp. 1861–1869.

[48] M. Zecchin, S. Park, O. Simeone, M. Kountouris, and D. Gesbert, "Robust bayesian learning for reliable wireless ai: Framework and applications," *arXiv preprint arXiv:2207.00300*, 2022.

[49] S. Liu, R. Cao, Y. Huang, T. Ouypornkochagorn, and J. Jia, "Time sequence learning for electrical impedance tomography using bayesian spatiotemporal priors," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 6045–6057, 2020.

[50] A. Onan, S. Korukoğlu, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Information Processing & Management*, vol. 53, no. 4, pp. 814–833, 2017.

[51] A. Onan, "Biomedical text categorization based on ensemble pruning and optimized topic modelling," *Computational and Mathematical Methods in Medicine*, vol. 2018, 2018.

[52] J. Yoon, T. Kim, O. Dia, S. Kim, Y. Bengio, and S. Ahn, "Bayesian model-agnostic meta-learning," *Advances in neural information processing systems*, vol. 31, 2018.

[53] S. T. Jose, S. Park, and O. Simeone, "Information-theoretic analysis of epistemic uncertainty in bayesian meta-learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 9758–9775.

[54] A. Banerjee, "On Bayesian bounds," in *Proceedings of the 23rd International Conference on Machine learning*, 2006, pp. 81–88.

[55] L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy, "PAC-bayesian bounds based on the Rényi divergence," in *Artificial Intelligence and Statistics*. PMLR, 2016, pp. 435–444.

APPENDIX

A. Proofs

Lemma. *With probability $1-\sigma$, with $\sigma \in (0,1)$, with respect to the random sampling of the data set $\mathcal{D}$, for all distributions $q(\theta)$ that are absolutely continuous with respect the prior $p(\theta)$, the following bound on the risk (18) of the ensemble model holds*

$$\mathcal{R}_t(q) \leq \mathcal{J}_t^m(q) + \psi(\tilde{\nu}, n, m, \beta, p, \sigma), \tag{55}$$

where

$$\psi(\tilde{\nu}, n, m, \beta, p, \sigma) := \frac{1}{\beta}\left(\log \mathbb{E}_{\mathcal{D}, p(\theta)}\left[e^{\beta \Delta_{m,n}}\right] - \log \sigma\right) \tag{56}$$

and

$$\begin{aligned}\Delta_{m,n} :=& \frac{1}{n}\sum_{x\in\mathcal{D}} \log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j) \\ &- \mathbb{E}_{\tilde{\nu}(x)}\left[\log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j)\right].\end{aligned} \tag{57}$$

Furthermore, the risk with respect to the ID measure $\nu(x)$ can be bounded as

$$\begin{aligned}\mathbb{E}_{\nu(x)}[\mathcal{R}_t(q,x)] \leq& \frac{1}{1-\epsilon}\left(\mathcal{J}_t^m(q) + \psi(\tilde{\nu}, n, m, \beta, p, \sigma)\right) \\ &+ \frac{\epsilon(C^{1-t}-1)}{(1-\epsilon)(1-t)},\end{aligned} \tag{58}$$

if the contamination ratio satisfies the inequality $\epsilon < 1$.

Proof: The proof follows in a manner similar to [10]. For a data set size n, and for an ensemble of models $\Theta = \{\theta\}_{i=1}^m$, we define the quantity

$$\begin{aligned}\Delta_{m,n}(\Theta, \mathcal{D}) :=& \frac{1}{n}\sum_{x\in\mathcal{D}} \log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j) \\ &- \frac{1}{n}\sum_{x\in\mathcal{D}} \mathbb{E}_{\tilde{\nu}(x)}\left[\log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j)\right].\end{aligned} \tag{59}$$

From the compression lemma [54], we have that for any distribution $q(\theta)$ which is absolutely continuous with respect to the prior $p(\theta)$, and for any $\beta < 0$, the following holds

$$\begin{aligned}\mathbb{E}_{q(\theta)^{\otimes m}}\left[\beta\Delta_{m,n}\right] \leq & D_1(q(\theta)^{\otimes m}||p(\theta)^{\otimes m}) \\ & + \log \mathbb{E}_{p(\theta)^{\otimes m}}\left[e^{\beta\Delta_{m,n}}\right] \qquad (60)\\ = & mD_1(q(\theta)||p(\theta)) \\ & + \log \mathbb{E}_{p(\theta)^{\otimes m}}\left[e^{\beta\Delta_{m,n}}\right], \qquad (61)\end{aligned}$$

where we have used the simplified notation $\Delta_{m,n} = \Delta_{m,n}(\Theta, \mathcal{D})$, and the equality follows from the basic properties of the KL divergence.

A direct application of Markov's inequality is then used to bound the last term of (61) with high probability. Namely, with probability greater then $1-\sigma$ with respect to the random drawn of the data set $\mathcal{D} \sim \tilde{\nu}(x)^{\otimes n}$, the following holds

$$\mathbb{E}_{p(\theta)^{\otimes m}}\left[e^{\Delta_{m,n}}\right] \leq \frac{\mathbb{E}_{\tilde{\nu}(x)^{\otimes n}, p(\theta)^{\otimes m}}\left[e^{\Delta_{m,n}}\right]}{\sigma}, \qquad (62)$$

or, equivalently,

$$\log \mathbb{E}_{p(\theta)^{\otimes m}}\left[e^{\Delta_{m,n}}\right] \leq \log \mathbb{E}_{\tilde{\nu}(x)^{\otimes n}, p(\theta)^{\otimes m}}\left[e^{\Delta_{m,n}}\right] - \log \sigma. \qquad (63)$$

Combining (61) with (63), the following upper bound on the predictive risk holds with probability $1-\sigma$

$$\begin{aligned}\mathcal{R}_t(q) \leq & \mathbb{E}_{\tilde{\nu}(x), q(\theta)^{\otimes m}}\left[-\log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j)\right] \qquad (64)\\ \leq & \mathbb{E}_{q(\theta)^{\otimes m}}\left[\frac{1}{n}\sum_{x\in\mathcal{D}} \log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j)\right] \\ & + \frac{m}{\beta} D_1(q(\theta)||p(\theta)) \\ & + \frac{\log \mathbb{E}_{\tilde{\nu}(x)^{\otimes n}} \mathbb{E}_{p(\theta)^{\otimes m}}\left[e^{\Delta_{m,n}}\right] - \log\sigma}{\beta}. \qquad (65)\end{aligned}$$

Finally, the result above can be translated to a guarantee with respect to the ID measure $\nu(x) = \frac{\tilde{\nu}(x)}{1-\epsilon} - \frac{\epsilon}{1-\epsilon}\xi(x)$ via the sequence of inequalities

$$\begin{aligned}\mathbb{E}_{\nu(x), q(\theta)^{\otimes m}}\left[-\log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j)\right] = & \\ = & \frac{\mathbb{E}_{\tilde{\nu}(x), q(\theta)^{\otimes m}}\left[-\log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j)\right]}{1-\epsilon} \\ & + \epsilon \frac{\mathbb{E}_{\epsilon(x), q(\theta)^{\otimes m}}\left[-\log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j)\right]}{1-\epsilon} \qquad (66)\\ \leq & \frac{\mathbb{E}_{\tilde{\nu}(x), q(\theta)^{\otimes m}}\left[-\log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j)\right]}{1-\epsilon} \\ & + \epsilon \frac{\left(C^{1-t}-1\right)}{(1-\epsilon)(1-t)}, \qquad (67)\end{aligned}$$

where the last inequality follows by having assumed the probabilistic model being uniformly upper bounded by C (Assumption 2).

■

The above result can readily be extended to generalized robust Bayesian learning by applying the change of measure inequality presented in [55], namely

$$\begin{aligned}\frac{t_p}{1-t_p}\log \mathbb{E}_{q(\theta)^{\otimes m}} \phi(\Theta) \leq & D^R_{t_p}(q(\theta)^{\otimes m}||p(\theta)^{\otimes m}) \qquad (68)\\ & + \log \mathbb{E}_{p(\theta)^{\otimes m}}\left[\phi(\Theta)^{\frac{t_p}{1-t_p}}\right], \qquad (69)\end{aligned}$$

with $\phi(\Theta)$ being the function

$$\phi(\Theta) := e^{\frac{t_p}{1-t_p}\Delta_{m,n}(\Theta,\mathcal{D})}, \qquad (70)$$

and by exploiting the tensorization of the Rényi divergence

$$D^R_{t_p}(q(\theta)^{\otimes m}||p(\theta)^{\otimes m}) = mD^R_{t_p}(q(\theta)||p(\theta)). \qquad (71)$$

Finally, with regard to the comparison between the PAC^m bound in Theorem 1 in [10] and the guarantee with respect to the ID measure, we observe that it is not in general possible to translate a guarantee on the $\log_t$-risk to one on the $\log$-risk. This can be illustrated by the following counter-example. Consider the following discrete target distribution parametrized by integer k, which defines the size of its support, as

$$\nu_k(x) = \begin{cases} 1-\frac{1}{k}, & \text{for } x = 0 \\ \frac{1}{k}2^{-k^2}, & \text{for } x = 1, \dots, 2^{k^2}, \end{cases} \qquad (72)$$

and the optimization of the $\log_t$-loss over a predictive distribution $p(x)$. The following limit holds

$$\lim_{k\to\infty} \min_p \mathbb{E}_{\nu_k(x)}[\log_t p(x)] = \begin{cases} 0, & \text{for } t \in [0,1) \\ \infty, & \text{for } t = 1 \end{cases}, \qquad (73)$$

and therefore that an ensemble optimized for a value of t in the range $[0,1)$ can incur in an unboundedly large loss when scored using the $\log$-loss.

Theorem. *The minimizer of the robust m-free energy objective*

$$\mathcal{J}^m_t(q) := \frac{1}{n}\sum_{x\in\mathcal{D}} \hat{\mathcal{R}}^m_t(q,x) + \frac{m}{\beta} D_1(q(\theta)||p(\theta)). \qquad (74)$$

is the fixed point of the operator

$$T(q) := p(\theta_j) \exp\left(\beta \sum_{x\in\mathcal{D}} \mathbb{E}_{\{\theta_i\}_{i\neq j}}\left[\log_t\left(\frac{\sum_{i=1}^m p_{\theta_i}(x)}{m}\right)\right]\right) \qquad (75)$$

where the average in (36) is taken with respect to the i.i.d. random vectors $\{\theta_i\}_{i\neq j} \sim q(\theta)^{\otimes m-1}$.

Proof: The functional derivative of the multi-sample risk is instrumental to computation of the minimizer of the robust m-free energy objective (30). This is given as

$$\frac{d\hat{\mathcal{R}}_t^m(q,x)}{dq} =$$
$$= \frac{d}{dq}\mathbb{E}_{\theta_1,\dots,\theta_m\sim q(\theta)^{\otimes m}}\left[-\log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j)\right] \tag{76}$$
$$= -\frac{d}{dq}\int_{\Theta^m} \log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j) \prod_{i=1}^m q(\theta_i) d\theta_i \tag{77}$$
$$\overset{(a)}{=} -\sum_{k=1}^m \int_{\Theta^{m-1}} \log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j) \prod_{i\neq k} q(\theta_i) d\theta_i \tag{78}$$
$$\overset{(b)}{=} -m\int_{\Theta^{m-1}} \log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j) \prod_{i=1}^{m-1} q(\theta_i) d\theta_i, \tag{79}$$
$$= -m\mathbb{E}_{\theta_1,\dots,\theta_{m-1}\sim q(\theta)^{\otimes m-1}}\left[\log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j)\right], \tag{80}$$

where (a) follows from the derivative of a nonlocal functional of m functions, and (b) holds since the integrand is invariant under the permutation of $\{\theta_i\}_{i\neq k}$.

The functional derivative of the robust m-free energy then follows as

$$\frac{d\mathcal{J}_t^m(q)}{dq} = \tag{81}$$
$$= \frac{d\hat{\mathcal{R}}_t^m(q,x)}{dq} + \frac{m}{\beta}\frac{dD_1(q(\theta)||p(\theta)}{dq} \tag{82}$$
$$= -m\mathbb{E}_{\theta_1,\dots,\theta_{m-1}\sim q(\theta)^{\otimes m-1}}\left[\log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j)\right] \tag{83}$$
$$+ \frac{m}{\beta}\left(1+\log(q(\theta)) - \log(p(\theta))\right). \tag{84}$$

Imposing the functional derivative equals to zero function it follows that the optimized posterior must satisfy

$$q(\theta_m) = p(\theta_m)\cdot \tag{85}$$
$$\cdot \exp\left\{\beta \mathbb{E}_{\theta_1,\dots,\theta_{m-1}\sim q(\theta)^{\otimes m-1}}\left[\log_t \mathbb{E}_{j\sim U[1:m]} p(x|\theta_j)\right]\right\}. \tag{86}$$

■

Theorem. *The influence function of the robust m-free energy objective (40) is*

$$IF_t^m(z,\phi,P^n) = -\left[\frac{\partial^2 \mathcal{J}_t^m(\gamma,\phi)}{\partial\phi^2}\right]^{-1} \times \frac{\partial^2 \mathcal{J}_t^m(\gamma,\phi)}{\partial\gamma\partial\phi}\Bigg|_{\substack{\gamma=0\\ \phi=\phi_t^{m*}(0)}}, \tag{87}$$

where

$$\frac{\partial^2 \mathcal{J}_t^m(\gamma,\phi)}{\partial\phi^2} = \mathbb{E}_{P^n_{\gamma,z}(x)}\frac{\partial^2}{\partial\phi^2}\left[\hat{\mathcal{R}}_t^m(q_\phi,x)\right] + \frac{\partial^2}{\partial\phi^2}\left[\frac{m}{\beta}KL(q_\phi(\theta)||p(\theta))\right] \tag{88}$$

and

$$\frac{\partial^2 \mathcal{J}_t^m(\gamma,\phi)}{\partial\gamma\partial\phi} = \frac{\partial}{\partial\phi}\left[\mathbb{E}_{P^n(x)}\left[\hat{\mathcal{R}}_t^m(q_\phi,x)\right] - \hat{\mathcal{R}}_t^m(q_\phi,z)\right]. \tag{89}$$

The proof of Theorem 2 directly follows from the Cauchy implicit function theorem stated below.

TABLE III: Total variation (TV) distance between the ID measure $\nu(x)$ and the predictive distribution $p_q(x)$ obtained from the optimization of the different free energy criteria.

	$t=1$	$t=0.9$	$t=0.7$	$t=0.5$	$t=0.3$	$t=0.1$
$m=1$	0.59	0.42	0.27	0.18	**0.16**	0.18
$m=2$	0.44	0.32	0.22	0.17	**0.15**	0.15
$m=5$	0.34	0.32	0.23	0.18	0.15	**0.14**
$m=10$	0.34	0.30	0.24	0.19	**0.15**	0.16

Theorem 3 (Cauchy implicit function theorem)**.** *Given a continuously differentiable function $F:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}^m$, with domain coordinates (x,y), and a point $(x^*,y^*)\in\mathbb{R}^n\times\mathbb{R}^m$ such that $F(x^*,y^*)=0$, if the Jacobian $J_{F,y}(x^*,y^*) = \left[\frac{\partial F_1(x^*,y^*)}{\partial y_1},\dots,\frac{\partial F_m(x^*,y^*)}{\partial y_m}\right]$ is invertible, then there exists an open set U that contains x^* and a function $g:U\to Y$ such that $g(x^*)=y^*$ and $F(x,g(x))=0$, $\forall x\in U$. Moreover the partial derivative of $g(x)$ in U are given by*

$$\frac{\partial g}{\partial x_i}(x) = -\left[J_{F,y}(x,g(x))\right]^{-1}\left[\frac{\partial F}{\partial x_i}(x,g(x))\right] \tag{90}$$

Proof: Replacing $F(x,y)$ with $\frac{\partial \mathcal{J}_t^m(\gamma,\phi)}{\partial\phi}$ and $g(x)$ with $\phi_t^{m*}(\gamma)$ and accordingly rewriting (90), we obtain

$$\frac{d\phi_t^{m*}(\gamma)}{d\gamma} = -\left[\frac{\partial^2 \mathcal{J}_t^m(\gamma,\phi_t^{m*}(\gamma))}{\partial\phi^2}\right]^{-1} \times \frac{\partial^2 \mathcal{J}_t^m(\gamma,\phi_t^{m*}(\gamma))}{\partial\gamma\partial\phi}. \tag{91}$$

The influence function (87) is then obtained evaluating (91) at $\gamma=0$.

■

B. Details on the Toy Example of Figure 1

In the toy example of Figure 1, the ID distribution $\nu(x)$ is a two component Gaussian mixture with means $\{-2,2\}$, variance equal to 2, and mixing coefficients $\{0.3,0.7\}$, respectively. The OOD distribution $\xi(x)$ is modelled using a Gaussian distribution with mean -8 and variance equal to 1.

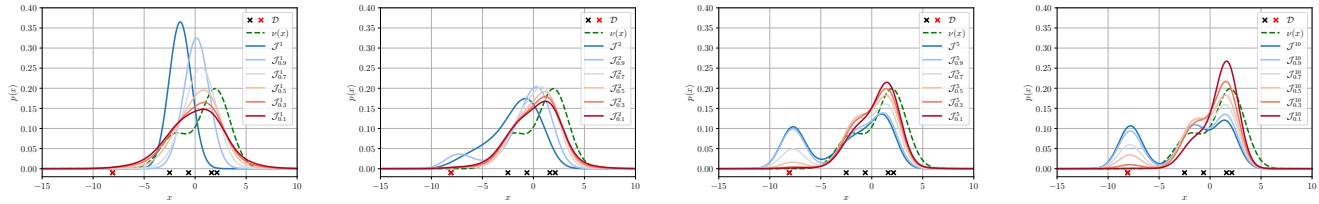
The probabilistic model is a Gaussian unit variance $p_\theta(x)=\mathcal{N}(x|\theta,1)$, the ensembling distribution $q(\theta)$ is represented by a discrete probability supported on 500 evenly spaced values in the interval $[-30,30]$, and the prior is $p(\theta)=\mathcal{N}(\theta|0,9)$. For a given m, β and t, the optimized ensembling distribution is obtained applying the fixed-point iteration in Theorem 1, i.e.,

$$q^+(\theta) = p(\theta)\exp\left\{\beta \sum_{\theta_1,\dots,\theta_{m-1}}\prod_{i=1}^{m-1} q^t(\theta_i)\cdot \cdot \log_t\left(\frac{\sum_{j=1}^{m-1} p(x|\theta_j) + p(x|\theta)}{m}\right)\right\}, \tag{92}$$

$$q^{t+1}(\theta) = (1-\alpha)q^t(\theta) + \alpha\frac{q^+(\theta)}{\sum_\theta q^+(\theta)}, \tag{93}$$

for $\alpha\in(0,1)$.

In Figure 9 we report the optimized predictive distributions produced by the above procedure for $\beta=1$, $m=\{1,2,5,20\}$ and $t=\{1,0.9,0.7,0.5,0.3,0.1\}$. As m grows larger, the



(a) $m = 1$ (b) $m = 2$ (c) $m = 5$ (d) $m = 20$

Fig. 9: Ensemble predictive distribution obtained minimizing different free energy criteria and different values of m. The samples from the ID measure are represented as green dots, while data points sampled from the OOD component are in red. The optimized predictive distributions. The predictive distribution obtained minimizing the standard m-free energy is denoted by $\mathcal{J}^m$, while the predictive distribution yielded by the minimization of the robust m-free energy are denoted by $\mathcal{J}^m_{0.9}, \mathcal{J}^m_{0.7}, \mathcal{J}^m_{0.5}, \mathcal{J}^m_{0.3}$ and $\mathcal{J}^m_{0.1}$ for $t = \{1, 0.9, 0.7, 0.5, 0.3, 0.1\}$ respectively.

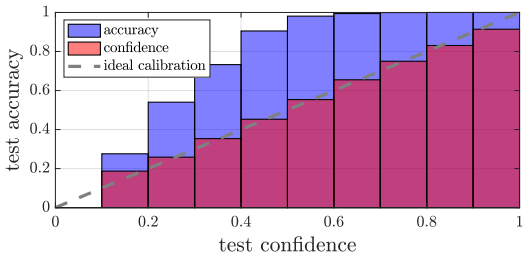


Fig. 10: Reliability diagram of deep ensembles [39].

multi-sample bound on the predictive risk becomes tighter. As a result, the predictive distribution becomes more expressive, and it covers all the data points. The use of generalized logarithms offers increased robustness against the outlier data point, and leads to predictive distributions that are more concentrated around the ID measure. In Table III we report the total variation distance between the ID measure and the predictive distribution $p_q(x)$. The proposed robust m-free energy criterion consistently outperforms the standard criterion by halving the total variation distance form the ID measure for $t = 0.3$.

C. Details and Further Results for the Classification Example in Sec. V-C

In Figure 5, we used expected calibration error (ECE) [43] to assess the quality of uncertainty quantification of the classifier. In this section, we formally define the ECE, along with the related visual tool of reliability diagrams [45], and present additional results using reliability diagrams.

Consider a probabilistic parametric classifier $p(b|a, \theta)$, where $b \in \{1, \dots, C\}$ represents the label and a the covariate. The *confidence* level assigned by the model to the predicted label

$$\hat{b}(a) = \arg\max_b p(b|a, \theta) \tag{94}$$

given the covariate a is given as [43]

$$\hat{p}(a) = \max_b p(b|a, \theta). \tag{95}$$

Perfect calibration corresponds to the equality [43]

$$\mathbb{P}(\hat{b}(a) = b|\hat{p}(a) = p) = p, \ \forall p \in [0, 1], \tag{96}$$

where the probability is taken over the ID sampling distribution $\nu(a, b)$. This equality expresses the condition that the probability of a correct decision for inputs with confidence level p equals p for all $p \in [0, 1]$. In words, confidence equals accuracy.

The ECE and reliability diagram provide means to quantify the extent to which the perfect calibration condition (96) is satisfied. To start, the probability interval $[0, 1]$ is divided into K bins, with the k-th bin being interval $(\frac{k-1}{K}, \frac{k}{K}]$. Assume that we have access to test data from the ID distribution. Denote as $\mathcal{B}_k$ the set of data points (a, b) in such test set for which the confidence $\hat{p}(a)$ lies within the k-th bin, i.e., $\hat{p}(a) \in (\frac{k-1}{K}, \frac{k}{K}]$. The average accuracy of the predictions for data points in $\mathcal{B}_k$ is defined as

$$\text{acc}(\mathcal{B}_k) = \frac{1}{|\mathcal{B}_k|} \sum_{a \in \mathcal{B}_k} \mathbf{1}(\hat{b}(a) = b), \tag{97}$$

with $\mathbf{1}(\cdot)$ being indicator function, b being the label corresponding to a in the given data point (a, b), and $|\mathcal{B}_k|$ denoting the number of total samples in the k-th bin $\mathcal{B}_k$. Similarly, the average confidence of the predictions for covariates in $\mathcal{B}_k$ can be written as

$$\text{conf}(\mathcal{B}_k) = \frac{1}{|\mathcal{B}_k|} \sum_{a \in \mathcal{B}_k} \hat{p}(a). \tag{98}$$

Note that perfectly calibrated model $p(b|a, \theta)$ would have $\text{acc}(\mathcal{B}_k) = \text{conf}(\mathcal{B}_k)$ for all $k \in \{1, \dots, K\}$ in the limit of a sufficiently large data set.

1) Expected Calibration Error (ECE) [43]: ECE quantifies the amount of miscalibration by computing the weighted average of the differences between accuracy and confidence levels across the bins, i.e.,

$$\text{ECE} = \sum_{k=1}^{K} \frac{|\mathcal{B}_k|}{\sum_{k=1}^{K} |\mathcal{B}_k|} \Big| \text{acc}(\mathcal{B}_k) - \text{conf}(\mathcal{B}_k) \Big|. \tag{99}$$

2) Reliability Diagrams: Since the ECE quantifies uncertainty by taking an average over the bins, it cannot provide insights into the individual calibration performance per bin. In contrast, reliability diagrams plot the accuracy $\text{acc}(\mathcal{B}_k)$ versus

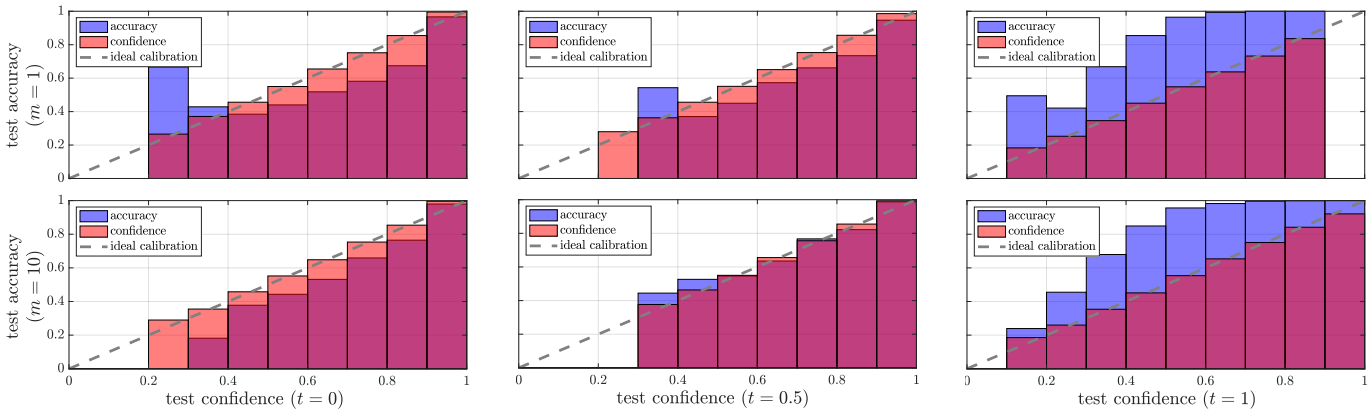


Fig. 11: Reliability diagrams of robust Gibbs predictor that optimizes $\mathcal{J}_t^1$ (top); and proposed robust ensemble predictor that optimizes $\mathcal{J}_t^{10}$ (bottom) under contamination ratio $\epsilon = 0.3$ for different $t = 0, 0.5, 1$.

the confidence $\text{conf}(\mathcal{B}_k)$ as a function of the bin index k, hence offering a finer-grained understanding of the calibration of the predictor.

3) Additional Results: For the MNIST image classification problem considered in Section V-C, Figure 10 plots for reference the reliability diagrams for deep ensembles [39], while Figure 11 reports reliability diagrams for the proposed classifiers with different values of m and t. The figures illustrate that using the standard log-loss ($t = 1$) tends to yield poorly calibrated decisions (Figure 10 and Figure 11 (right)), while the proposed robust ensemble predictor can accurately quantify uncertainty using $t = 0.5$ (Figure 11 (bottom, middle)). It is also noted that setting $t = 1$ is seen to yield underconfident predictions due to the presence of outliers, while a decrease in t leads to overconfident decision due to the reduced expressiveness of t-logarithms. A proper choice of t leads to well-calibrated, robust prediction.