Robust Bayesian Learning for Reliable Wireless AI: Framework and Applications

Zecchin Matteo

joint work with S. Park, O. Simeone, M. Kountouris and D. Gesbert

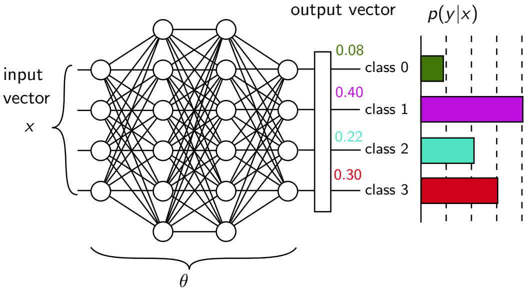
FAAS Seminar
29/9/2022

The Role of AI in 6G & Beyond

- AI is playing an increasingly significant role in **engineering**.
- **Next-generation communication systems** will leverage AI at all layers of the protocol stack.
- This imposes **new requirements** on the performance of AI.
- **Accuracy** should be weighted against:
  - **reliability**, or **calibration**, providing a faithful quantification of the uncertainty of the AI's decisions, e.g., for **monitoring**;
  - **robustness** to deviations from design assumptions

Predictive Uncertainty

- Discriminative **probabilistic models** $p(y|x,\theta)$ output hard decisions and confidence levels.



- **Hard decision**: $\hat{y}(x|\theta) = 1$ (class with largest score)
- **Confidence level**: $\mathrm{conf}(x|\theta) = p(\hat{y}(x|\theta)|x,\theta) = 0.4$ (self reported)
- How **reliable** is the estimate of predictive uncertainty reported by the model?

Quantifying Calibration

- Assume that the data is generated from some ground-truth **population distribution** $P(x, y)$.
- In practice, this can be estimated based on validation/ test data.
- The **accuracy** of a probabilistic model $p(y|x, \theta)$ on input x is

$$\mathrm{acc}(x|\theta) = P(\hat{y}(x|\theta)|x)$$

- A probabilistic model $p(y|x, \theta)$ is **reliable**, or **well calibrated**, if

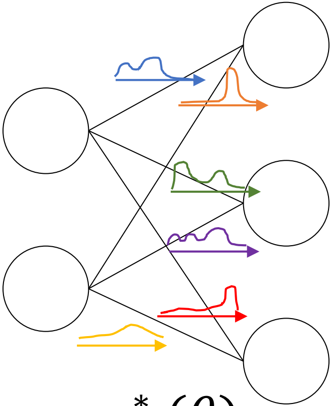
$$\mathrm{conf}(x|\theta) \approx \mathrm{acc}(x|\theta),$$

or

$$\text{confidence level} \approx \text{ accuracy}$$

Bayesian Learning

- **Bayesian** learning:
  - Optimization of a distribution $q(\theta)$ in the model parameter space
  - Distribution $q(\theta)$ encodes **epistemic uncertainty**.



$$q^*(\theta)$$

(Generalized) Bayesian Learning

- **Generalized Bayesian learning** obtain $q^*(\theta)$ by minimizing the free energy[1,2]

$$F_{\mathcal{D}}(q(\theta)) = N \underbrace{\mathrm{E}_{\theta\sim q(\theta)}[f(\theta,\mathcal{D})]}_{\text{average training loss}} + \beta \underbrace{\mathsf{KL}\left(q(\theta)||p(\theta)\right)}_{\text{information-theoretic}}$$

- With $\beta = 0$, the problem reduces to frequentist learning, which outputs a single model parameter vector θ^*.
- This criterion is well justified by **PAC Bayes theory**, which derives it as an upper bound on the population loss.[3]

---

1 J. Knoblauch, et al, "Generalized variational inference: Three arguments for deriving new posteriors," arXiv:1904.02063, 2019.

2 O. Simeone, "Machine Learning for Engineers", Cambridge University Press, 2022.

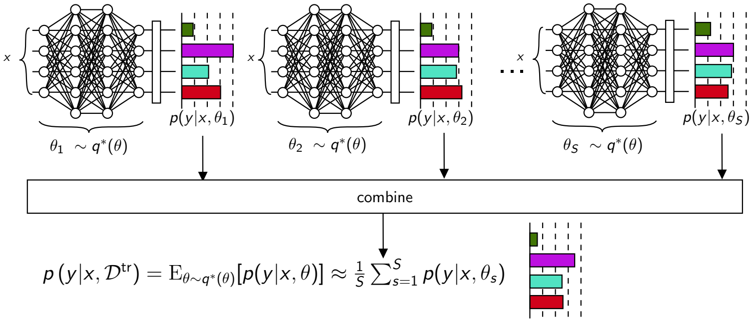
3 P. Alquier, "User-friendly introduction to PAC-Bayes bounds," arXiv preprint, 2021.

Bayesian Learning

- Decision obtained via **ensembling**, i.e., via

$$\mathrm{E}_{\theta \sim q^*(\theta)}\left[p(y|x,\theta)\right],$$

  accounting for the "opinions" of multiple models.
- In practice, the average is done over S i.i.d. model parameters $\theta \sim q^*(\theta)$.



Bayesian Learning

- At test time, we have
  - **Hard decision**: $\hat{y}(x|q^*) = \arg\max_y \mathrm{E}_{\theta \sim q^*(\theta)} [p(y|x,\theta)]$ (class with largest average score)
  - **Confidence level**:

$$\mathrm{conf}(x|q^*) = \mathrm{E}_{\theta \sim q^*(\theta)}[p(\hat{y}(x|q^*)|x,\theta)]$$

- The confidence level accounts for **epistemic uncertainty** via the **disagreement** among models.[4]

[4] N. Houlsby, et al, "Bayesian active learning for classification and preference learning," arXiv:1112.5745, 2011.

Limitations of Bayesian Learning: Model Misspecification

- A design assumption in Bayesian learning is that prior and likelihood reflect the population data distribution.
- When this is not the case, e.g., when we are forced to choose a "simple" model class, the model is said to be **misspecified**.
- Choosing $\beta \neq 1$ in generalized Bayesian learning can partly address this problem:
  - This is related to the "**cold posterior** problem"[5]

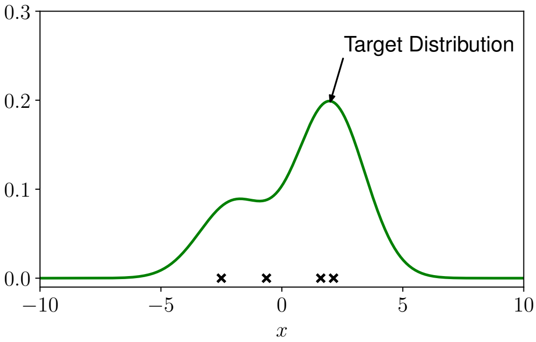
[5] K. Pitas and J. Arbel, "Cold Posteriors through PAC-Bayes," arXiv:2206.11173.

A Toy Example

- Consider a density estimation problem with an underlying data distribution that is a mixture of Gaussians (e.g., a fading channel with blocking for mmwave or THz communications):

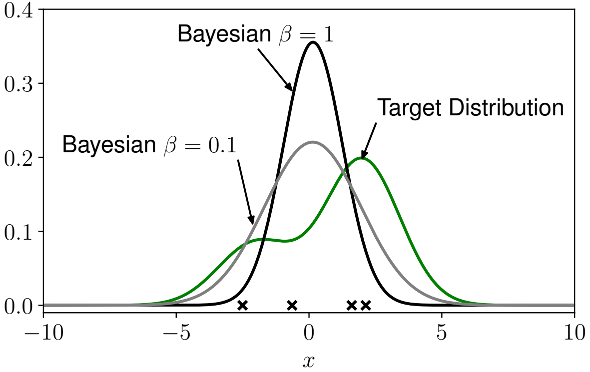
$$P(x) = 0.7\mathcal{N}(x|2,2) + 0.3\mathcal{N}(x|-2,2).$$

- Assume a Gaussian likelihood function: $p(x|\theta) = \mathcal{N}(x|\theta,1)$



A Toy Example

- The model class is **misspecified** since it is not possible to capture both modes of the data distribution using a *single* Gaussian model.
- In this scenario, generalized Bayesian learning presents poor generalization, even with $\beta \neq 1$.



Generalized Bayesian Learning and Misspecification

- As we have seen, Bayesian learning leverages ensembling, producing the average across multiple models

$$\mathrm{E}_{\theta \sim q^*(\theta)}[p(x|\theta)],$$

with $q^*(\theta)$ being the distribution obtained via training.

- However, generalized Bayesian learning does not capture the performance of ensemble predictors as it merely include the *average* training loss $\mathrm{E}_{\theta \sim q^*(\theta)}[L_{\mathcal{D}}(\theta)]$.

$(m,1)$-Robust (Generalized) Bayesian Learning

- To overcome the limitations of (generalized) Bayesian learning, it was recently proposed to use a multi-sample version of the free energy: $(m,1)$-**robust Bayesian learning**.
- The m-**sample free energy** is defined as[6]

$$F_{\mathcal{D}}^{m}(q(\theta)) = N\mathrm{E}_{\theta_1,\ldots,\theta_m \sim q(\theta)^{\otimes m}}\left[L_{\mathcal{D}}^{m}(\theta)\right] + \beta \mathrm{KL}\left(q(\theta)||p_0(\theta)\right)$$

where the training loss

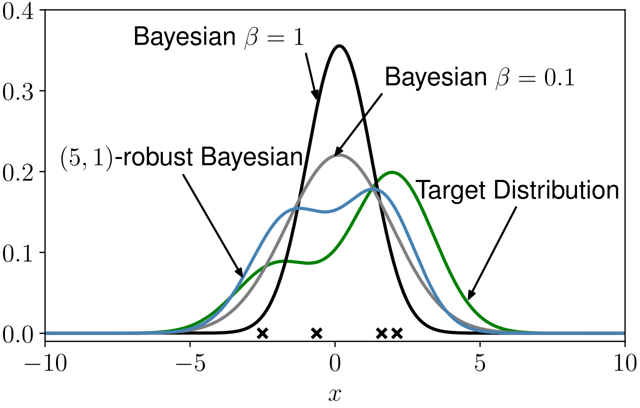
$$L_{\mathcal{D}}^{m}(\theta) = -\frac{1}{N}\sum_{x\in\mathcal{D}} \log\left(\frac{1}{m}\sum_{i=1}^{m} p(x|\theta_i)\right)$$

explicitly captures the log-loss of a mixture of m models drawn from $q(\theta)$.
- Like the free energy, it can be justified via the analysis of generalization based on **PAC Bayes** theory.

6 W. Morningstar, et al "PAC ^{m}-Bayes: Narrowing the Empirical Risk Gap...", NeurIPS 2021.

Toy Example (Continued)

- $(m, 1)$-robust Bayesian learning is clearly better able to capture the multi-modal properties of the ground-truth distribution $P(x)$.



Limitations of Bayesian Learning: Outliers

- Training data often contains **outliers** – anomalous data points that do not follow the same distribution of test data
  - Errors due to human labeling, measuring tools failures or interference, adversarial examples, ...

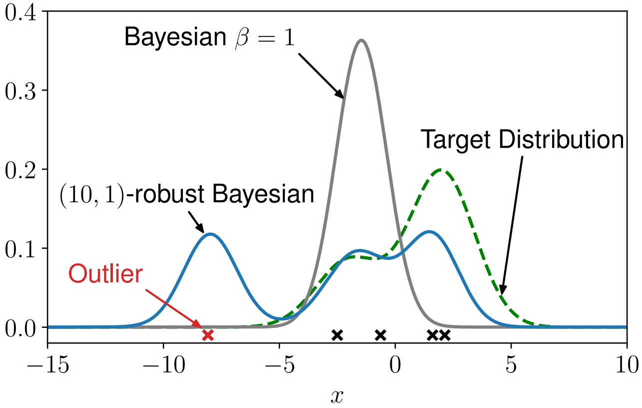
- Outliers can be modelled via the gross-error model: Given a contamination ratio $\epsilon \in (0, 1]$, the sampling distribution is[7]

$$\tilde{P}(x) = \epsilon \underbrace{Q(x)}_{\substack{\text{out-of-distribution} \\ \text{measure (OOD)}}} + (1-\epsilon) \underbrace{P(x)}_{\substack{\text{in-distribution} \\ \text{measure (ID)}}}$$

[7] P. J. Huber, "Robust estimation of a location parameter," The Annals of Mathematical Statistics, 1964.

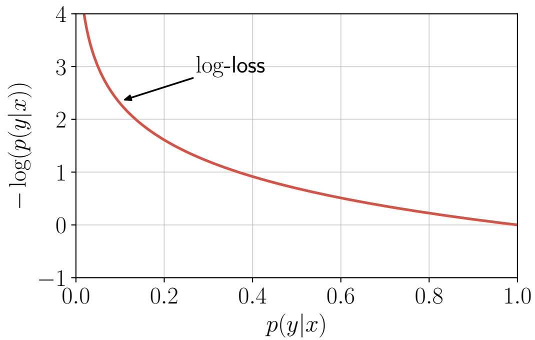
Toy Example (Continued)

- While more robust to misspecification, $(m, 1)$-robust Bayesian learning is significantly affected by outliers.



Reconsidering the Log-Loss

- What is the cause of the lack of robustness of existing free energy metrics?
- The free energy relies on the standard log-loss $-\log p(x|\theta)$, which penalizes very strongly models that do not cover well all data points, including outliers.

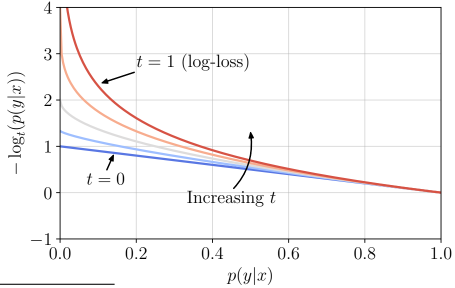


Beyond the Log-Loss: t-Log-Loss

- The ***t*-log-loss**, for $t \in [0,1)$, is defined as[8],

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

  - for $t \to 1$ recovers the standard log-loss
  - Since we have $-\log_t(p) \leq (1-t)^{-1}$, outliers have a bounded influence when t is small.



8 C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," Journal of Statistical Physics, 1988.

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

- (m, t)-**robust Bayesian learning** minimizes the (m, t)-**free energy criterion**:[9]

$$F_{\mathcal{D}}^{m,t}(q(\theta)) = N\mathrm{E}_{\theta_1,\ldots,\theta_m \sim q(\theta)^{\otimes m}}\left[f^{m,t}(\theta, \mathcal{D})\right] + \mathrm{KL}\left(q(\theta)||p(\theta)\right)$$

  where

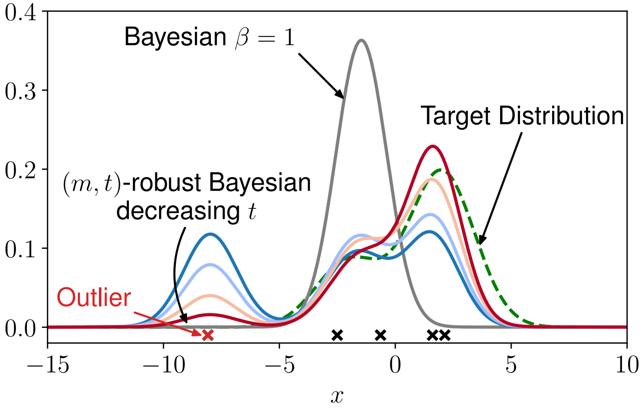
$$f^{m,t}(\theta, \mathcal{D}) = \sum_{x \in \mathcal{D}} \log_t \left(\frac{1}{m} \sum_{i=1}^{m} p(x|\theta_i) \right)$$

  replaces the log-loss with the t-log-loss.
- The criterion has two tuning knobs:
  - the generalized logarithm parameter $t \in [0, 1)$, which determines the robustness to outliers;
  - and the number constituent models $m \geq 1$ in the ensemble, which determines the robustness to misspecification.

[9] M. Zecchin, et al, "Robust PAC^m..." arXiv:2203.01859, 2022.

Toy Example (Continued)

- (m, t)-robust Bayesian learning is able to tackle both model misspecification and the presence of outliers.



Properties of (m,t)-robust Bayesian learning

- The *population* risk can be bounded w.r.t to the ID and contaminated measures[10].

Theorem (Population Risk Bound)

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 of the ensemble model holds*

$$\mathbb{E}_{q(\theta),\tilde{P}(x)}[-\log_t p_\theta(x)] \leq F^{m,t}_{\mathcal{D}}(q)) + \psi(\tilde{P}, n, m, \beta, p, \sigma) \quad (1)$$

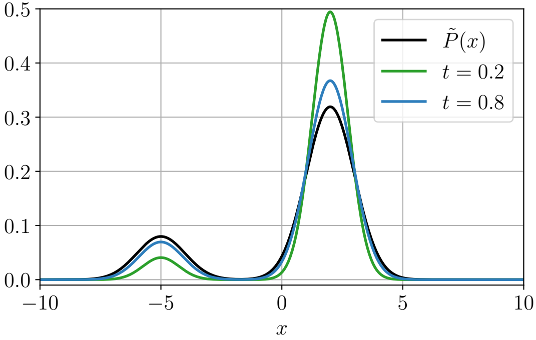
Furthermore, the risk with respect to the ID measure $P(x)$ *can be bounded as*

$$\mathbb{E}_{q(\theta),P(x)}[-\log_t p_\theta(x)] \leq \frac{1}{1-\epsilon}\left(F^{m,t}_{\mathcal{D}}(q) + \psi(\tilde{P}, n, m, \beta, p, \sigma)\right) + \frac{\epsilon(C^{1-t}-1)}{(1-\epsilon)(1-t)}, \quad (2)$$

[10] Zecchin, Park, Simeone, Kountouris and Gesbert. *Robust PAC ^{m}: Training Ensemble Models Under Model Misspecification and Outliers.*

Properties of (m, t)-robust Bayesian learning

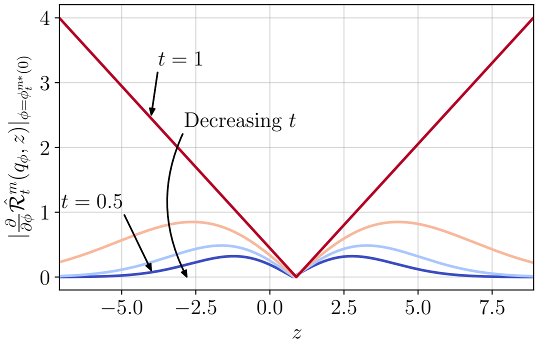
- For the number of samples $n \to \infty$ and the number of ensemble components $m \to \infty$, robust Bayesian learning minimize the $t-$Tsallis divergence between the predictive distribution $p_{q(\theta)}(x)$ and the $t-$escort version of $\tilde{P}(x)$.



- For $t = 1$ we recover the standard KL divergence minimization and the mode seeking behaviour of standard Bayesian learning.

Properties of (m, t)-robust Bayesian learning

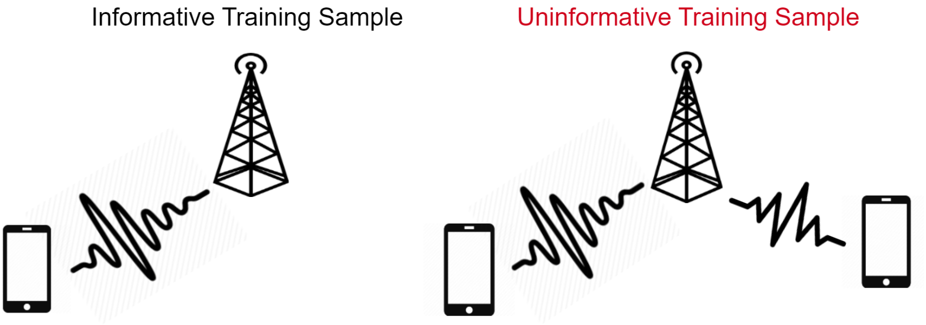
- The $\log_t$ loss effectively bounds the effect of anomalous data points. We study the influence function, measure changes of an estimator by the means of perturbation of a training data point.



Robust Bayesian Learning for Wireless Communications

- Many wireless communication applications are characterized by:
  - Training data affected by exogenous noise (e.g., interference and malicious reporting): **outliers**
  - Light-weight models deployed on resource constrained devices: **misspecification**
- We now review some specific applications of robust Bayesian learning to wireless systems.

Robust Bayesian Learning: Automatic Modulation Classification



- Determine the modulation type y associated to a received based-band signal vector x.
- Interference leads to uninformative training samples with ambiguous labels, i.e., outliers.

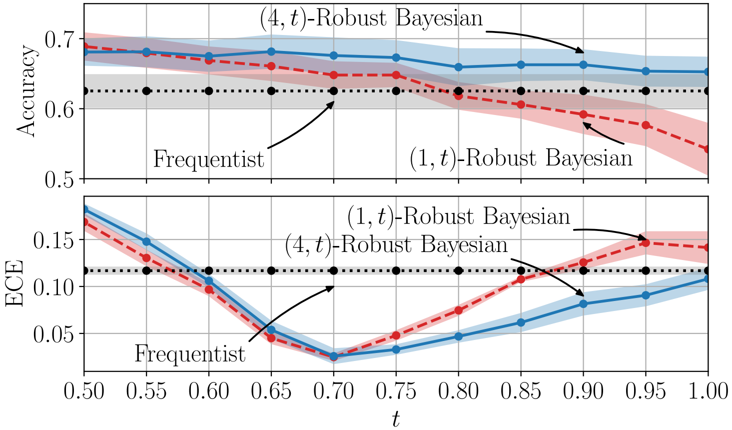
Robust Bayesian Learning: Automatic Modulation Classification

- The model is a neural network classifier comprising two convolutional layers and two linear layers.
- The dataset is the *DeepSIG: RadioML 2016.10A*[11] data set with 30% of the samples affected by interference.
- Testing is done on a clean data set.
- We evaluate the final model in terms of *accuracy* and *calibration*.

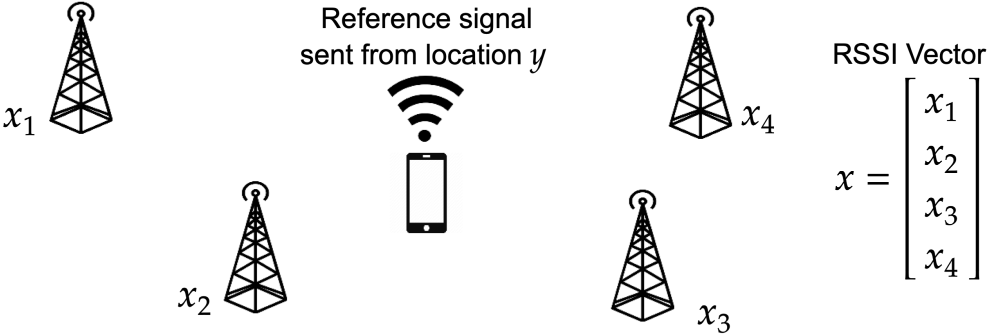
[11] T. j O'Shea, et al, "Convolutional Radio Modulation Recognition Networks" arXiv:1602.04105, 2016.

Robust Bayesian Learning: Automatic Modulation Classification

- Robust Bayesian learning can improve calibration for $t < 1$, while also enhancing accuracy with $m > 1$ ($\beta = 0.01$).

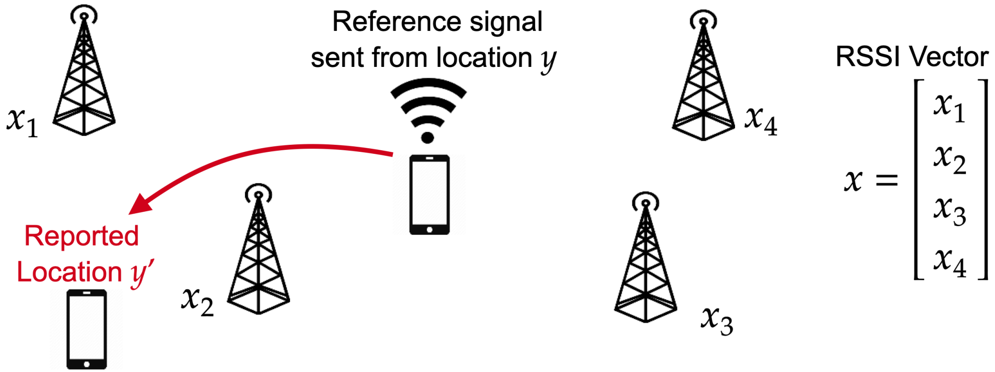


Robust Bayesian Learning: RSSI Based Localization



- Determine the location y of a transmitter based on received signal strength indicator (RSSI) vector x measured at different base stations.

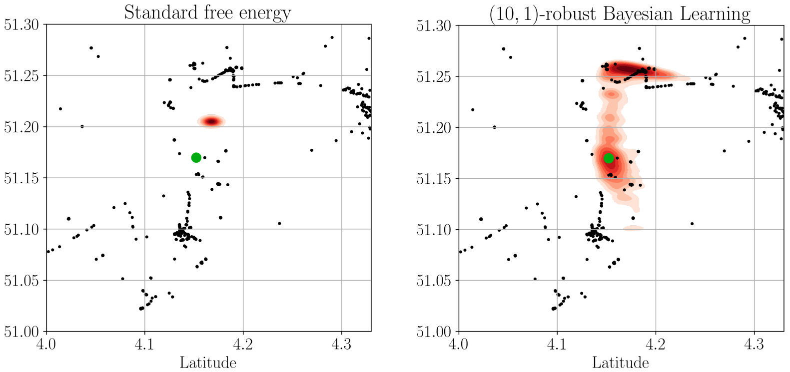
Robust Bayesian Learning: RSSI Based Localization



- **Outliers** are modelled by replacing an ϵ-fraction of the true labels y with a random location (e.g., malicious or inaccurate reporting).

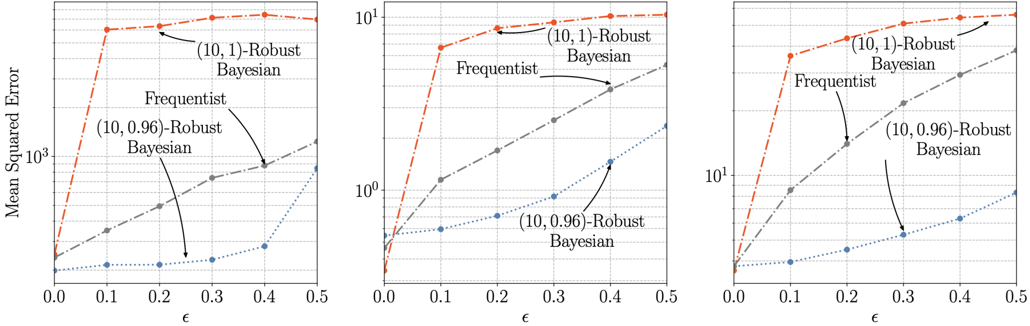
Robust Bayesian Learning: RSSI Based Localization

- We consider a model class $p(y|x,\theta) = \mathcal{N}(y|f_\theta(x), 0.01)$ where $f_\theta(x)$ is the output of a neural network.
- The model class is misspecified whenever the device location conditioned on the RSSI vector is not Gaussian distributed.
- $(m,1)$-robust Bayesian learning mitigates model misspecification.



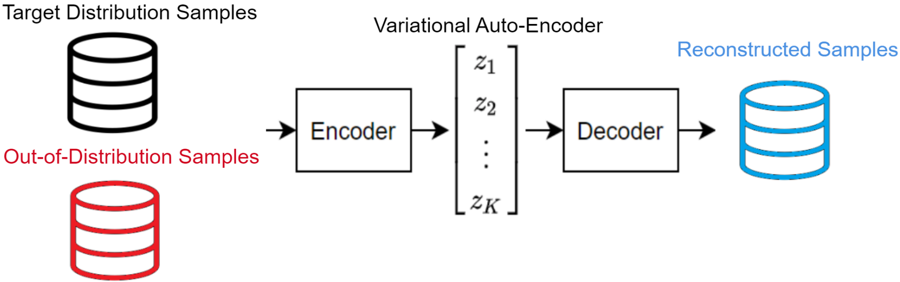
Robust Bayesian Learning: RSSI Based Localization

- (m, t)-robust Bayesian learning with $t < 1$ mitigates performance degradation due to outliers.



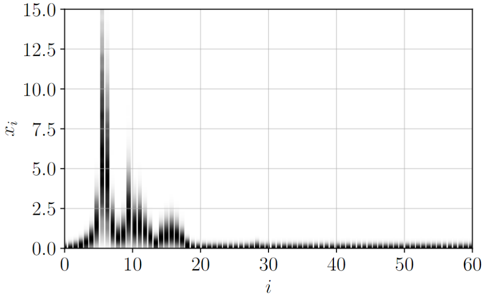
(a) *SigfoxRural* (b) *UTSIndoor* (c) *UJIIndoor*

Robust Bayesian Learning: Channel Simulation

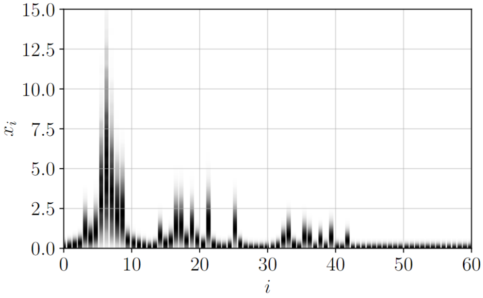


- Given a training dataset of channel responses x, train a generative model that is able to simulate new samples approximately distributed as the target channel model.
- We consider a training dataset comprising **outliers** from a different channel model.

Robust Bayesian Learning: Channel Simulation



(a) TDL-A $\tau = 100$ns

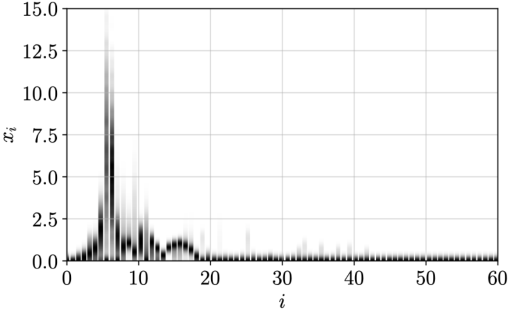


(b) TDL-A $\tau = 300$ns

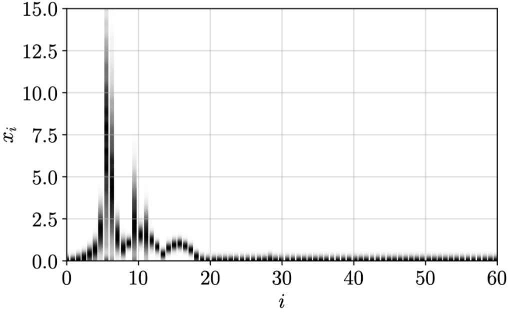
- Target (ID) distribution: TDL-A channel model with delay $\tau = 100$ ns
- Outliers (OOD) distribution: TDL-A channel model with a longer delay spread $\tau = 300$ ns

Robust Bayesian Learning: Channel Simulation

- We train a **variational autoencoder (VAE)** using the corrupted data set with $\epsilon = 0.2$, and use the generative model to generate new samples.



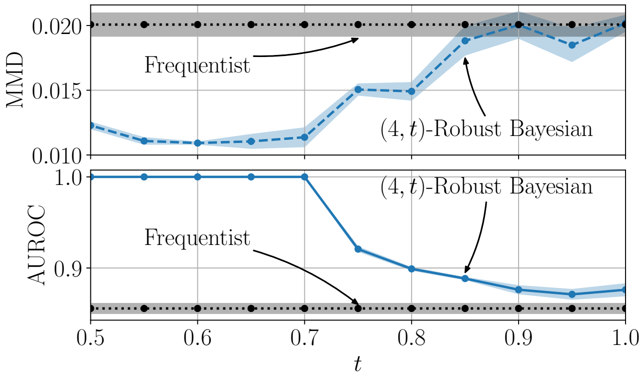
(c) Frequentist Learning



(d) $(4, 0.7)$-Robust Bayesian Learning

Robust Bayesian Learning: Channel Simulation

- Performance in terms of maximum mean discrepancy (MMD) between true and generated distributions, and in terms of area under the receiver operating curve (AUROC):
  - (m, t)-robust Bayesian learning with $t < 1$ yields higher accuracy in the generative model and better out-of-distribution detection capabilities.



Conclusion

- Standard Bayesian learning does not cater reliability under pratical conditions in wireless communication systems.
- (m, t)-robust Bayesian learning is an alternative learning criterion based on multi-sample estimators and generalized logarithmic losses that counteracts model misspecification and outliers.
- (m, t)-robust Bayesian learning enjoys nices mathematically properties and its merits have been shown over a range of wireless communication problems.