

---

# Robust uncertainty estimates with out-of-distribution pseudo-inputs training

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 Probabilistic models often use neural networks to control their predictive uncer-  
2 tainty. However, when making *out-of-distribution (OOD)* predictions, the often-  
3 uncontrollable extrapolation properties of neural networks yield poor uncertainty  
4 predictions. Such models then don't *know what they don't know*, which directly lim-  
5 its their robustness w.r.t unexpected inputs. To counter this, we propose to explicitly  
6 train the uncertainty predictor where we are not given data to make it reliable. As  
7 one cannot train without data, we provide mechanisms for generating *pseudo-inputs*  
8 in informative low-density regions of the input space, and show how to leverage  
9 these in a practical Bayesian framework that casts a prior distribution over the  
10 model uncertainty. With a holistic evaluation, we demonstrate that this yields  
11 robust and interpretable predictions of uncertainty while retaining state-of-the-art  
12 performance on diverse tasks such as regression and generative modeling.

## 13 1 Introduction

14 Neural networks generally extrapolate arbitrarily [Xu et al., 2020], and high quality predictions are  
15 limited to regions of the input space where the networks have been trained. This is to be expected and  
16 is only problematic if the associated predictions are not accompanied with a well-calibrated measure  
17 of uncertainty. If a neural network is used for estimating such a measure of uncertainty, we, however,  
18 quickly run into trouble, as the reported uncertainty then exhibits arbitrary behaviour in regions with  
19 no training data. Alarmingly, these are exactly the regions where evaluating the uncertainty is most  
20 important to the safe deployment of machine learning models in real world applications [Amodei  
21 et al., 2016]. One potential solution is to avoid using directly the output of neural networks for  
22 predicting uncertainty, and let it emerge from another mechanism, e.g. an *ensemble* [Hansen and  
23 Salamon, 1990, Lakshminarayanan et al., 2017] or some notion of *Monte Carlo* [MacKay, 1992, Gal  
24 and Ghahramani, 2016]. Here we explore the alternative view that the networks should simply be  
25 trained where there is no data.

26 But can we train without data? The Bayesian formalism often  
27 does so implicitly: most *conjugate priors* can be seen as addi-  
28 tional training data [Bishop, 2006], e.g. in Gaussian models,  
29 a mean prior  $\mathcal{N}(\mu_0, \sigma_0^2)$  can be realised by additional training  
30 data of  $\mu_0$  with  $\sigma_0^2$  setting the amount of observations. Placing  
31 a prior over the output of a neural network can, thus, be inter-  
32 preted as additional training data. Unfortunately, this view is  
33 not practical as it implies additional data *for all* possible inputs  
34 to a neural network, resulting in infinite data. Our approach  
35 is simple: we locate regions of low data density in *input space*  
36 and implicitly place observations here in *output space* by min-

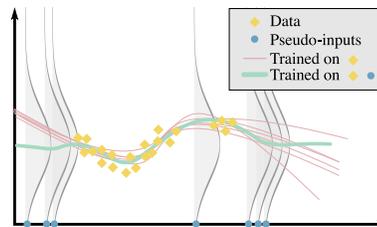


Figure 1: Pseudo-inputs are generated out of distribution, and there we train towards a prior (grey density).

37 imizing an appropriate KL divergence towards a prior (see Fig. 1). The result is a remarkably simple  
38 algorithm that drastically improves uncertainty estimates in both regression and generative modeling.

### 39 1.1 Background and related work

40 The predictive performance of machine learning models has drastically increased in the past decade,  
41 but the quality of the accompanying uncertainties have not followed. Uncertainties are reported as be-  
42 ing miscalibrated [Guo et al., 2017] and overconfident [Lakshminarayanan et al., 2017, Hendrycks and  
43 Gimpel, 2016]. Some models even see higher likelihoods of out-of-distribution than in-distribution  
44 data [Nalisnick et al., 2019, Nguyen et al., 2015, Louizos and Welling, 2017].

45 **Neural networks** commonly output distributions which gives a notion of predictive uncertainty. Clas-  
46 sifiers trained with *soft-max* is an ever-present example of such. These predictions are generally ob-  
47 served to be *overconfident* [Lakshminarayanan et al., 2017, Hendrycks and Gimpel, 2016] and to carry  
48 little meaning outside the support of the training data [Skaftel et al., 2019, Lee et al., 2017]. The latter  
49 is an artifact of the hard-to-control extrapolation that comes with neural networks [Xu et al., 2021].  
50 In general, since extrapolation is difficult to control, uncertainties predicted by neural networks will  
51 exhibit seemingly arbitrary behavior outside the support of the data, yielding untrustworthy results.

52 **Mean-variance networks** for regression [Nix and Weigend, 1994] model the conditional target  
53 density as a normal  $p(y|x) = \mathcal{N}(y|\mu(x), \sigma(x)^2)$  with mean and variance predicted by neural  
54 networks. The reported predictive uncertainty is generally accurate in regions near training data,  
55 but otherwise unreliable [Hauberg, 2019]. To counter this, Arvanitidis et al. [2017] and Skaftel et al.  
56 [2019] proposed variance network architectures to enforce a specified extrapolation value, but these  
57 heuristics tend to be difficult to tune, and lack principle. Mean-variance networks have seen a recent  
58 uptake within generative modeling, where they are applied as an *encoder* distribution in *variational*  
59 *autoencoders (VAEs)* [Kingma and Welling, 2013, Rezende et al., 2014].

60 **Which uncertainty?** A commonly called-upon dichotomy [Der Kiureghian and Ditlevsen, 2009]  
61 is that the uncertainty of a model’s *prediction* can be decomposed into the uncertainty of the *model*  
62 (*epistemic*) and of the *data* (*aleatoric*). The epistemic uncertainty can be lowered by increasing  
63 the amount of data, simplifying the model or otherwise reducing the complexity of the learning  
64 problem. The aleatoric uncertainty, on the other hand, is a property of the world, and cannot be  
65 changed; no prediction should ever be more certain than the uncertainty displayed by the associated  
66 data. Depending on the task at hand, we may be interested in different types of uncertainty: In *active*  
67 *learning* [Settles, 2012] and *Bayesian optimization* [Moćkus, 1975] we request data for which we  
68 have high epistemic, but low aleatoric uncertainty to ensure maximal information gain; while for  
69 classification and regression we often just want to minimize the overall predictive uncertainty.

70 **Bayesian methods** are often used to quantify uncertainty due to their explicit formulation of  
71 uncertainty. *Gaussian processes (GPs)* [Rasmussen and Williams, 2005] provide an elegant  
72 framework that provide state-of-the-art uncertainty estimates, but, alas, the corresponding mean  
73 predictions are often not up to the standards of neural networks. GPs are tightly linked to *Bayesian*  
74 *neural networks (BNNs)* [MacKay, 1992] that place a prior over the network weights and seek the  
75 corresponding posterior. Despite advances in *variational approximations* [Graves, 2011, Kingma  
76 and Welling, 2013, Blundell et al., 2015], *expectation propagation* [Hernández-Lobato and Adams,  
77 2015, Hasenclever et al., 2017], or *Monte Carlo* methods [Welling and Teh, 2011, Springenberg et al.,  
78 2016], training BNNs remains difficult. Furthermore, the predictive uncertainty seems dependent  
79 on the degree of approximation and is thus controlled by the available compute power.

80 **Ensemble methods** have long been used to produce aggregated predictions with uncertainty estimates  
81 [Hansen and Salamon, 1990, Breiman, 1996]. *Deep ensembles* [Lakshminarayanan et al., 2017],  
82 a collection of differently initialized networks trained on the same data, are generally reported as  
83 state-of-the-art for uncertainty quantification in deep models [Thagaard et al., 2020, Ovadia et al.,  
84 2019]. As the models in the ensemble are trained on overlapping data, they are correlated, which  
85 influence the ensemble uncertainty in ways that remains unclear [Breiman, 2001]. *Monte-Carlo*  
86 *dropout* [Gal and Ghahramani, 2016] casts dropout training [Srivastava et al., 2014] as an ensemble  
87 model. It is computationally cheap, but experiments [Ovadia et al., 2019, Skaftel et al., 2019] show  
88 that the increased correlation of ensemble elements amplifies the method’s overconfidence.

89 **Robustness to distribution shift** is paramount to a well-behaved uncertainty predictor [Ovadia  
90 et al., 2019] and must be evaluated accordingly. For out-of-distribution detection, Liang et al.

91 [2017] proposes a pre-processing perturbation step inspired by adversarial attacks [Goodfellow et al.,  
 92 2014a] that helps the model distinguish in-distribution and out-of-distribution inputs. Hendrycks  
 93 et al. [2018] used a *Generative Adversarial Network (GAN)* [Goodfellow et al., 2014b] to generate  
 94 out-of-distribution pseudo-inputs whose inclusion in the training under an additional regularizing  
 95 term in the loss function, called *outlier exposure*, enhances the predictor’s ability to discriminate  
 96 out-of-distribution inputs [Lee et al., 2017, Dai et al., 2017].

## 97 1.2 Robust uncertainty estimates

98 Our work is strongly inspired by the critical assessment of the issues that undermine variance estima-  
 99 tion ran by Skaftø et al. [2019] and by the proposal of Stirn and Knowles [2020] which we detail here.

100 **Notation.** Let the observed variable  $x \in \mathcal{X}$  follow the data generating distribution  $p_{\text{data}}(x)$ , only  
 101 known through the training dataset of  $N$  i.i.d samples  $\mathcal{D}_{\text{train}} = \{x_n\}_{n=1}^N$ . In the case of supervised  
 102 learning, the observed variables  $x = (x, y)$ , with  $x \in \mathbb{R}^d$  being the input and  $y \in \mathbb{R}^{d'}$  the target  
 103 for the model, follow the joint decomposition  $p_{\text{data}}(x, y) = p_{\text{data}}(y|x)p_{\text{data}}(x)$ . The proposed  
 104 probabilistic model  $p_{\theta}(x)$ , whose weights are indicated by  $\theta$ , aims to accurately emulate  $p_{\text{data}}(x)$ .

105 **Practical problems in variance estimation.** Gaussian likelihoods in the form of  $p_{\theta}(x) =$   
 106  $\mathcal{N}(x|\mu_{\theta}(x), \sigma_{\theta}(x)^2)$  are widely adopted to model continuous covariates. Real world data cannot be  
 107 expected to be *homoscedastic*, i.e constant throughout input space, and thus the predictive uncertainty,  
 108  $\sigma_{\theta}(x)$ , most often uses neural networks to map continuously the observed  $x$  onto the parameter space.  
 109 Beyond the well-known unreliable extrapolation properties of neural networks, this parametrisation  
 110 of predictive uncertainty is hamstrung by serious defects. Firstly, the predictive variance scales the  
 111 learning rates of the mean and variance updates by  $1/2\sigma_{\theta}(x)^2$ , resulting in a bias for data regions with  
 112 low uncertainty [Nix and Weigend, 1994]. Secondly, the maximisation of the modeled likelihood is  
 113 particularly sensitive to scarce data, as local gradient updates for the variance point towards the then  
 114 undefined *maximum likelihood estimate (MLE)* [Skaftø et al., 2019]. Lastly, and more worryingly, such  
 115 model’s likelihood is ill-defined [Mattei and Frelsen, 2018a], as it can arbitrarily and without bound  
 116 increase when the variance estimates collapse towards a detrimental 0. Overall, the naive maximisa-  
 117 tion of model likelihood seems insufficient to generate robust and well-behaved uncertainty estimates.

118 **Student-t likelihood.** The Bayesian formalism, by imposing to learn a parametrised distribution  
 119 over the predictive uncertainty, offers an attractive view to approaching the problem of uncertainty  
 120 estimation. Skaftø et al. [2019] notably adopts a Gamma distributed precision,  $1/\sigma^2 = \lambda \sim \Gamma(\alpha, \beta)$ ,  
 121 as the conjugate of an unknown precision for a Gaussian likelihood, to yield a non-standard Student-t  
 122 distributed marginal likelihood<sup>1</sup>. It is known to offer a more robust likelihood, especially in the scarce  
 123 data regime [Gelman et al., 2013],

$$p_{\theta}(x) = \int \mathcal{N}(x|\mu, \lambda)\Gamma(\lambda|\alpha, \beta)d\lambda = T\left(x|\nu = 2\alpha, \hat{\mu} = \mu, \hat{\sigma} = \sqrt{\beta/\alpha}\right). \quad (1)$$

124 Interestingly, its variance  $\text{Var}[x] = \beta/(\alpha - 1) = (\beta/\alpha) \cdot (\alpha/(\alpha - 1))$  can be explicitly decomposed  
 125 to an aleatoric term  $\beta/\alpha$  and an epistemic term<sup>1</sup>  $\alpha/(\alpha - 1)$  [Jørgensen, 2020, p16], and offers a  
 126 direct verification of whether a model knows what it knows.

127 **Variational variance.** Stirn and Knowles [2020] assumes a latent model precision  $\lambda$ . This is  
 128 generated by a prior  $p(\lambda)$  and its posterior is approximated variationally by the family of Gamma  
 129 distributions, conditioned on the inputs to reflect heteroscedasticity. Through *amortized variational*  
 130 *inference (AVI)* [Kingma and Welling, 2013] neural networks  $f_{\phi}$  map to the posterior parameters from  
 131 data,  $q(z|f_{\phi}(x))$ . As such, variational variance preserves the modelling capacity and robustness of  
 132 the non-standard Student-t marginal likelihood, without modifying its parameter architecture, while  
 133 the definition of a prior over the latent precision induces a more robust training objective. Assuming  
 134 the likelihood precision is the unique latent code, the *evidence lower bound (ELBO)*,

$$\begin{aligned} \mathcal{L}(q; x) &= \mathbb{E}_{q(\lambda)} [\log p(x|\lambda)] - D_{\text{KL}}(q(\lambda|x) || p(\lambda)) \\ &= \frac{1}{2} \left( \psi(\alpha) - \log \beta - \log(2\pi) - \frac{\alpha}{\beta}(x - \mu)^2 \right) - D_{\text{KL}}(q(\lambda|x) || p(\lambda)), \end{aligned} \quad (2)$$

135 takes the form of a regularised log-likelihood, exposing the benefits of the prior regularisation. It  
 136 penalises predicted variances that would unrealistically get arbitrarily close to either the detrimental

<sup>1</sup>See Section I. of the supplementary materials.

137 limits of 0 or  $\infty$ , reducing the concerns regarding the ill-definition of the objective. Additionally,  
 138 the scaling effect of the learning rates of the likelihood parameters is reduced. Naturally, the effect  
 139 of the regularisation will be highly dependent on the prior selected. Here, because we are mostly  
 140 interested in enforcing a constant desired uncertainty extrapolation, we adopt an homoscedastic  
 141 Gamma distributed prior,  $p(\lambda) = \Gamma(\lambda|a, b)$ , that matches the level of uncertainty observed in data,  
 142 and leave it for future practitioners to adopt the most adequate prior for the task at hand.

## 143 2 Out-of-distribution pseudo-inputs training

### 144 2.1 Dissipative loss

145 In the variational variance formalism, due to AVI, the predictive uncertainty is controlled by  $\alpha$  and  $\beta$ ,  
 146 the independent neural networks parametrising the posterior distribution,  $\text{Var}[x] = \beta(x)/(\alpha(x) - 1)$ .  
 147 The unreliable extrapolation properties of neural networks therefore directly challenge the robustness  
 148 of the method’s uncertainty estimates outside of its training support, limiting the applicability of the  
 149 method. We consider that this flawed extrapolation is not inevitable.

150 Inspired by outlier exposure [Hendrycks et al., 2018], we propose to include deliberately generated  
 151 out-of-distribution *pseudo-inputs*,  $\{\hat{x}_k\}_{k=1}^K$  where  $\hat{x}_k \sim p_{\text{out}}(x)$ , in the training of our variational  
 152 objective to constrain the extrapolation of the posterior parametrisation. The optimal variational  
 153 objective  $q^*$  is chosen such that it minimises our proposed *dissipative loss* over the consolidated  
 154 dataset  $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{out}}$ , where  $\mathcal{D}_{\text{out}} = \{\hat{x}_k\}_{k=1}^K$ ,

$$\text{Loss}(q; \mathcal{D}) = -\left[\mathcal{L}_{\text{in}}(q; \mathcal{D}_{\text{train}}) + \mathcal{L}_{\text{out}}(q; \mathcal{D}_{\text{out}})\right]. \quad (3)$$

155 The in-distribution component of the loss function  $\mathcal{L}_{\text{in}}(q; \mathcal{D})$  naturally arises as the standard ELBO  
 156 over the training set. The out-of-distribution component  $\mathcal{L}_{\text{out}}(q; \mathcal{D})$  operates on a fundamentally  
 157 different source of data. As the only information available regarding the pseudo-inputs is that they  
 158 are out-of-distribution, we assert for them a constant, non-informative likelihood  $p(\hat{x}|\lambda) = c$ , that  
 159 has thus no influence on optimisation. This is similar to the strategy of *censoring* [Lee and Wang,  
 160 2003] where different likelihoods are used for observations with different properties. As a result, the  
 161 dissipative loss becomes,

$$\text{Loss}(q; \mathcal{D}) = -\left[\sum_{x \in \mathcal{D}_{\text{train}}} \mathbb{E}_{q(\lambda|x)} [p_{\theta}(x|\lambda)] - D_{\text{KL}}(q(\lambda|x} || p(\lambda)) - \sum_{\hat{x} \in \mathcal{D}_{\text{out}}} D_{\text{KL}}(q(\lambda|\hat{x}) || p(\lambda))\right]. \quad (4)$$

162 It share the same motivating intuition as the *confidence loss* of Lee et al. [2017] and completes the  
 163 variational variance formalism with a principled mechanism to learn robust variance estimates with  
 164 the desired extrapolation properties. It indeed explicitly forces the predictor to match our high-entropy  
 165 prior expectations on out-of-distribution samples while learning the low-entropy covariate dependent  
 166 distribution, hence the name of dissipative. The reliance of the model’s predictive uncertainty  
 167 on its mean predictions implies that it is primordial here to safeguard its generative performance.  
 168 We guarantee it with the implementation of a split training procedure [Skafte et al., 2019]; the  
 169 out-of-distribution regularisation is only applied after the model’s mean has been trained.

### 170 2.2 Pseudo-input generators (PIGs)

171 Minimising the posterior KL divergence out-of-distribution requires an efficient sampling procedure  
 172 of pseudo-inputs. As exposed in Fig. 2, their generation should leverage a-priori knowledge about  
 173  $p_{\text{data}}(x)$  to resolve the undefined nature of  $p_{\text{out}}(x)$ . In this simple regression case, we show the  
 174 predictive uncertainty of variational variance models trained on artificial heteroscedastic data. We  
 175 use a prior uncertainty level that matches the maximum of the data uncertainty. As anticipated,  
 176 without pseudo-inputs, the model extrapolates uncertainty to a constant, arbitrary level, and only the  
 177 introduction of pseudo-inputs near the training data results in the desired uncertainty extrapolation.  
 178 Reassuringly, this suggests that we do not need to regularise our model’s extrapolation in the entire  
 179 out-of-distribution space. Instead, we can focus on the simpler task of generating pseudo-inputs  
 180 in low-density regions of the input space that neighbours training data, as they can enforce correct  
 181 extrapolation in the rest of the out-of-distribution space. Lee et al. [2017] gives supporting arguments.

182 Recent contributions have relied on GANs for generating a useful representation of  $p_{\text{out}}(x)$  [Lee  
 183 et al., 2017, Dai et al., 2017]. Although conceptually intuitive, GANs incur a heavy computational

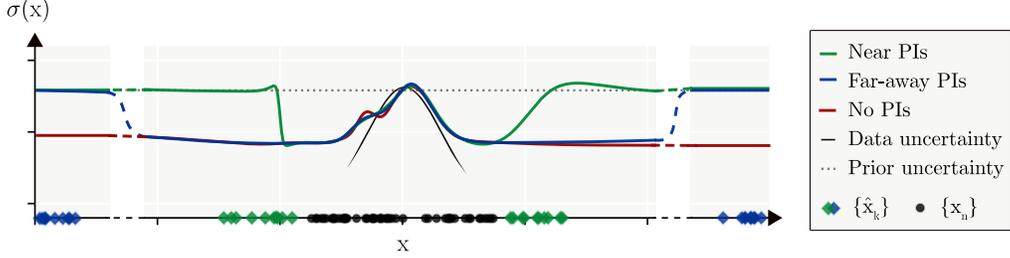


Figure 2: Effect of different pseudo-input distributions on the predictive uncertainty of variational variance models. Training data (black points) is generated uniformly on  $[-5, 5]$ , with a variance that scales as  $\exp(-0.5(\|x\|/s)^2)$ . The *near* pseudo-inputs (green diamonds) are generated uniformly in  $[-10, -5] \cup [5, 10]$ , while the *far-away* (blue diamonds) are on  $[-200, -190] \cup [190, 200]$ . Dashes amount for the empty space that separates far away pseudo-inputs.

184 burden and most likely induce serious practical challenges as a result of the instability of their training  
 185 [Shrivastava et al., 2017]. Furthermore, as one need to understand what is in-distribution to model  
 186 what it is not, we instead propose to directly leverage the information at hand about the data.

187 Algorithm 1 gives a simple procedure for generating pseudo-inputs using the data density. Pseudo-inputs are originally sampled from  $p_{\text{data}}(x)$ , and their positions iteratively updated with gradient descent, with step size  $\delta$ , by following the directions that minimise their likelihood under  $p_{\text{data}}(x)$ , similarly to reversed adversarial steps [Goodfellow et al., 2014a].

---

**Algorithm 1:** Pseudo-Input Generator (PIG)

---

```

 $\forall k \in [1, K], \hat{x}_k \sim p_{\text{data}}(x). \text{ iterations} = 0. \epsilon = \infty;$ 
while ( $\text{iterations} < \text{max\_iterations}$ ) & ( $\epsilon > \text{tolerance}$ ) do
  compute  $\forall k \in [1, K], \nabla_x p(x)(\hat{x}_k);$ 
   $\epsilon = \max_{k \in [1, K]} (\delta \nabla_x p(x)(\hat{x}_k));$ 
   $\forall k \in [1, K], \hat{x}_k = \hat{x}_k - \delta \nabla_x p(x)(\hat{x}_k);$ 
   $\text{iterations} = \text{iterations} + 1;$ 
end

```

---

197 The procedure can run prior to training, in parallel for all  $\hat{x}_k$  with automatic differentiation, and  
 198 thus results in limited additional complexity for the optimisation<sup>2</sup>. It relies on the availability of  
 199 a differentiable density estimate of the data, which is, depending on the use case, either directly  
 200 available (see Sec. 3.2), or can be approximated through a variety of methods such as *Bayesian*  
 201 *Gaussian mixture models* [Bishop, 2006], or various *normalising flows* [Rezende and Mohamed,  
 202 2015] based methods such as *masked autoregressive flows* [Papamakarios et al., 2017] (see Sec. 3.1).  
 203 A caveat here is that depending on the PIG’s parameters, and on the quality of the density estimate  
 204 available, pseudo-inputs might be generated in undesired regions of the input space, e.g uninformative  
 205 density minima. In practice, we adopted conservative density estimates and parameters and did not  
 206 observe any significant degradation of the predictive uncertainty due to the addition of pseudo-inputs.

### 207 3 Experiments

208 **Holistic evaluation of uncertainty estimates.** The ground truth for uncertainty estimates is usually  
 209 unknown, making their evaluation non-trivial. Similarly as in Stirn and Knowles [2020], we propose  
 210 to assess them using a collection of metrics. Calibration, which evaluates probabilistic predictions  
 211 w.r.t the long-run frequencies that actually occur [Dawid, 1982] can be measured by *proper scoring*  
 212 *rules* [Lakshminarayanan et al., 2017] such as the model log-likelihood  $\log p_\theta(x|\lambda)$ . Additionally,  
 213 the *root mean squared error (RMSE)* between the predictive and empirical variance,  $\text{Var}[x] -$   
 214  $(\mathbb{E}_{q(z|x)} [p_\theta(x|\lambda)] - x)^2$ , offers a quantification of the model’s awareness of its own uncertainty. It  
 215 nevertheless requires an understanding of the model’s mean predictive performance, as commonly  
 216 measured by the RMSE of the mean residuals,  $\mathbb{E}_{q(\lambda|x)} [p_\theta(x|\lambda)] - x$ . We further propose to evaluate  
 217 the cooperation of mean and uncertainty estimates for the generation of credible samples, which  
 218 constitutes a consistency check for the learned precision distribution [Gelman et al., 2013], by  
 219 measuring the RMSE of sample residuals  $x^* - x$ , with  $x^* \sim p_\theta(x)$ . Finally, The ELBO, despite the  
 220 absence of theoretical grounding for it [Blei et al., 2017], is commonly reported as an approximation  
 221 of the marginal likelihood, and thus of the overall model’s predictive performance.

---

<sup>2</sup>Running times are reported in Sec. IV. of the supplementary materials.

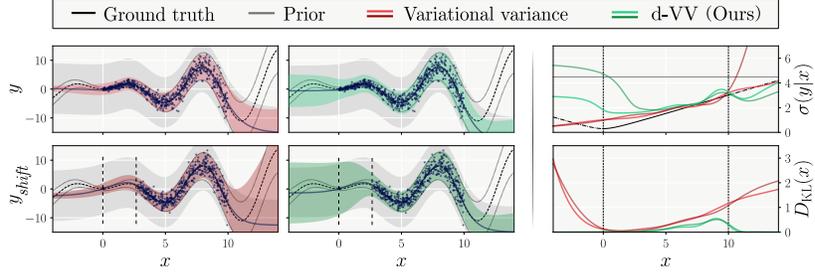


Figure 4: Toy regression results. On the left, the mean predictions are surrounded by  $\pm 2$  standard deviations, with the training data of the bottom row presenting a shift. On the right are displayed the predictive uncertainty fit and the prior KL divergence.

222 A complete assessment of a model’s uncertainty estimates further requires their evaluation under  
 223 distributional shift [Ovadia et al., 2019], which we either introduce voluntarily through deliberate  
 224 splitting of the training and test sets, as in Sec. 3.1, or by using test data from a different dataset  
 225 altogether, as in Sec. 3.2.

### 226 3.1 Regression

227 In a regression setting where the proposed model must capture the conditioning between targets and  
 228 inputs  $y|x$ , the precision  $\lambda$  of a Gaussian likelihood is the only assumed latent code.

229 Faithfully to variational variance [Stirn and Knowles, 2020] we  
 230 adopt a Gamma heteroscedastic variational posterior  $q_\phi(\lambda|x) =$   
 231  $\Gamma(\lambda|\alpha_\phi(x), \beta_\phi(x))$  parametrised by the independent  $\alpha_\phi$  and  $\beta_\phi$   
 232 networks, with weights  $\phi$ , uniquely conditioned on the inputs (see  
 233 Fig. 3). This approximate posterior, independent of the targets, gives  
 234 up on the dependency of the true posterior on both covariates to  
 235 guarantee heteroscedasticity<sup>3</sup>.

236 For strictly more than 2 degrees of freedom, or equivalently,  
 237  $\alpha_\phi(x) > 1$ , the marginal predictive probability  $p_{\theta,\phi}(y|x) =$   
 238  $T\left(y|2\alpha_\phi(x), \mu_\theta(x), \sqrt{\beta_\phi(x)/\alpha_\phi(x)}\right)$ , has its first two moments

239 defined,  $\mathbb{E}[y|x] = \mu_\theta(x)$  and  $\text{Var}[y|x] = \beta_\phi(x)/(\alpha_\phi(x) - 1)$ , pro-  
 240 viding explicit mean and uncertainty estimates with a single forward  
 241 pass in the single layered, fully connected,  $\alpha_\phi$ ,  $\beta_\phi$  and  $\mu_\theta$  networks used here. To ensure definition of  
 242 both the posterior distribution and of the marginal distribution’s variance, the parameter maps use a  
 243 soft-plus activation on their last layer to ensure positivity, and the  $\alpha_\phi$  network is further shifted by 1.

244 The unique dependence of the posterior on the inputs implies that the generation of pseudo-inputs  
 245 should only rely on the input density. In a general regression setting, it is unknown, and we estimate  
 246 it here prior to training with a Bayesian Gaussian mixture model [Bishop, 2006]. We refer to it  
 247 henceforth as *dissipative variational variance (d-VV)*. The specific implementation details are listed  
 248 in Section II. of the supplementary materials.

#### 249 3.1.1 Toy regression

250 The desiderata for our method are clear: capture of the data heteroscedasticity, extrapolation to a  
 251 higher uncertainty level, no underestimation of the predictive uncertainty, and posterior extrapolation  
 252 to the prior out-of-distribution. Skafte et al. [2019] first showed on the toy regression task,  $y =$   
 253  $x \sin(x) + 0.3 \epsilon_1 + 0.3 x \epsilon_2$ , where  $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, 1)$ , that amongst a collection of methods, only their  
 254 proposed variance network architecture could realise our first three expectations. Fig. 4 demonstrates  
 255 that our more principled approach also fulfills all of our requirements, without the need for arbitrarily  
 256 enforcing the desired extrapolation in our architecture. The importance of out-of-distribution training  
 257 is also revealed as the standard variational variance approach fails to produce uncertainty estimates  
 258 that extrapolate correctly and are robust to distributional shift (bottom row of Fig. 4).

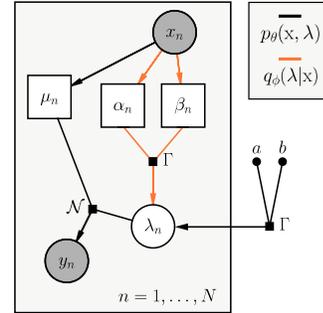


Figure 3: PGM for regression

<sup>3</sup>See Section II. of the supplementary materials for the expression of the true posterior.

Table 1: UCI benchmarks. Each square shows the performance of a given model (rows) on a given dataset (columns). The intensity of the colouring represents the certitude that the associated model performed best on the given dataset. Grey rows mean impossible evaluation for a metric.

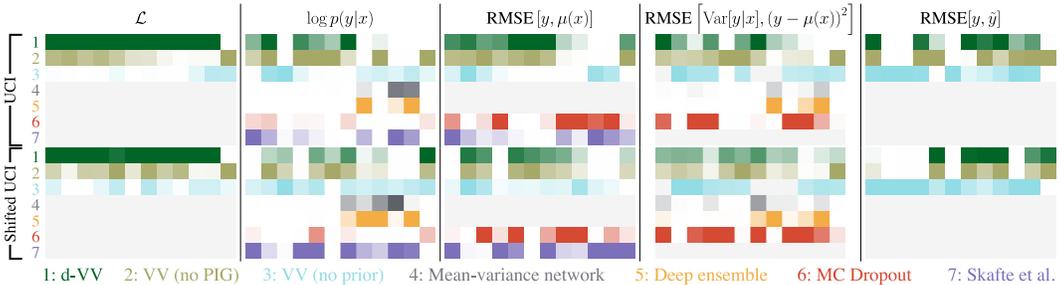


Table 2: Evaluation of the generative modeling. For each dataset, we report mean  $\pm$  std over 5 trials.

		FashionMNIST	SVHN	CIFAR
$\log p(x)$	VAE	2215.54 $\pm$ 68.81	<b>4304.90 <math>\pm</math> 58.45</b>	<b>2930.64 <math>\pm</math> 14.82</b>
	d-V3AE	<b>2349.71 <math>\pm</math> 11.80</b>	4133.41 $\pm$ 64.28	2668.85 $\pm$ 13.23
RMSE(x, $\tilde{x}$ )	VAE	0.171 $\pm$ 0.003	0.097 $\pm$ 7e-4	0.154 $\pm$ 5e-4
	d-V3AE	<b>0.158 <math>\pm</math> 0.003</b>	<b>0.087 <math>\pm</math> 0.002</b>	<b>0.129 <math>\pm</math> 7e-4</b>

259 **Decomposition of the model and data uncertainty.** Fig. 5  
 260 presents the decomposition of the predictive uncertainty. The  
 261 aleatoric component captures the heteroscedastic increase of  
 262 uncertainty in the training data while the epistemic uncertainty,  
 263 constant in distribution, extrapolates to higher values. The  
 264 proposed method therefore demonstrates, to the best of our  
 265 knowledge, a principled decomposition of uncertainty factors.

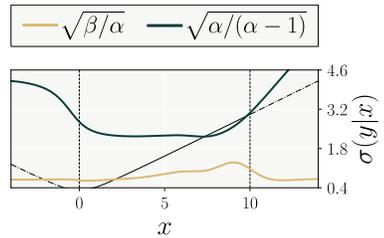


Figure 5: Aleatoric (yellow) and epistemic (dark) uncertainties.

### 266 3.1.2 UCI Benchmarks

267 Real world regression datasets from the UCI repository<sup>4</sup> are used to evaluate our model against  
 268 curated baselines, analogically to the setup from Hernández-Lobato and Adams [2015] and Skafte  
 269 et al. [2019]<sup>5</sup>. As revealed by the summarising Tab. 1<sup>6</sup>, our method retains the mean predictive power  
 270 of the variational variance method. The log-likelihood and RMSE of variance and sample residuals  
 271 further show the improved calibration resulting from the imposition of a prior on the variance, as  
 272 the VV methods generally outperform the MLE Student-t (VV (no prior)) that shares the same  
 273 architecture. The table thus proves that holistically, the dissipative loss strengthens the variational  
 274 variance model’s performance, which itself generally surpasses the chosen baselines.

275 The robustness of the methods to distributional shift is further evaluated as in Foong et al. [2019]. For  
 276 each input feature, a hole is created in the training data by assigning the middle third of observations  
 277 to the test set, when sorted w.r.t that feature. Interestingly, we see that our method’s calibration  
 278 slightly improves under the shift, highlighting the robustness benefits of the OOD prior regularisation.

279 We note that both MC Dropout and the combined method of Skafte et al. [2019] generally perform  
 280 well, confirming their interest for regression tasks requiring uncertainty quantification, but the  
 281 former’s calibration is not robust to data shifts, as is also reported in Ovadia et al. [2019], and the  
 282 latter is in practice difficult to tune and lacks principle.

<sup>4</sup><https://archive.ics.uci.edu/ml/index.php>

<sup>5</sup>See Sec. II. for details about the chosen baselines, datasets and implementations specifics.

<sup>6</sup>The full numbers are included in Sec. II. of the supplements.

283 **3.2 Generative models**

284 We extend the evaluation of our proposal to the case of generative models through the lens of VAEs  
 285 [Kingma and Welling, 2013, Rezende et al., 2014]. Variational auto-encoders infer a low dimensional  
 286 latent encoding of the data  $z \in \mathbb{R}^D$ , on which is conditioned the generative process  $p_\theta(x|z)$ . Its  
 287 predictive uncertainty, which evaluates the confidence of the model in its ability to adequately  
 288 reconstruct inputs is known to be untrustworthy.

289 In the case of continuous or seemingly continuous inputs,  
 290 the adoption of a Gaussian decoder  $p_\theta(x|z) = \mathcal{N}(x|\mu_\theta(z), \sigma_\theta(z)^2)$  results in an ill-defined model likelihood  
 291 [Mattei and Frellsen, 2018a] that encourages decoder variance  
 292 collapse, making the training of the model notoriously harder  
 293 [Skafte et al., 2019]. Most implementations therefore choose  
 294 to fix the variance to a set level e.g  $\sigma_\theta(z) = 0.1$ , or elude the  
 295 challenge by adopting a Bernoulli likelihood.  
 296

297 Motivated by our previous results, we now aim to demonstrate  
 298 that VAEs, whose decoder is fitted with our method, are  
 299 able to provide robust uncertainty estimates. Assuming a  
 300 latent generative precision, the latent variables of the model  
 301 are decomposed into  $z = \{z, \lambda\}$ , with  $z$  the latent input  
 302 representations. The marginalisation of the Gamma distributed  
 303 latent variance results in a Student-T decoder, as detailed in  
 304 Eq. 1. The overall architecture of the *variational variance variational auto encoder (V3AE)* [Stirn  
 305 and Knowles, 2020] is shown in Fig. 6, and yields, with the addition of our out-of-distribution  
 306 pseudo-inputs training, the dissipative loss function<sup>7</sup>,

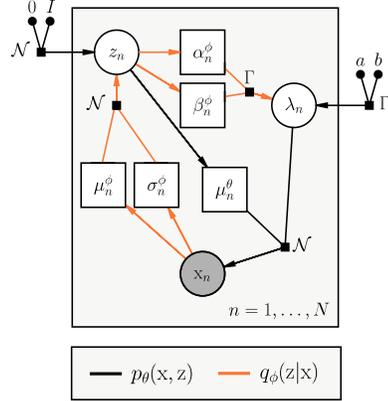


Figure 6: PGM for V3AE

$$\text{Loss}(q_\phi, \theta; \mathcal{D}_{\text{train}}) = - \left[ \sum_{x \in \mathcal{D}_{\text{train}}} \mathcal{L}(q_\phi, \theta; x) - \mathbb{E}_{q_{\text{out}}(z)} [D_{\text{KL}}(q_\phi(\lambda|z) || p(\lambda))] \right]. \quad (5)$$

307 Because only the decoder is regularised, the pseudo-inputs lie in the space of latent representations,  
 308  $\mathcal{D}_{\text{out}} = \{\hat{z}_k\}_{k=1}^K \in \mathbb{R}^D$ . The distribution of training inputs is therefore readily accessible as the  
 309 aggregate variational posterior  $q_\phi(z|\mathcal{D}_{\text{train}}) = q_\phi(z|x_1) \cdot \dots \cdot q_\phi(z|x_N)$ . Here again, we rely on a split  
 310 training procedure to leverage this perk; the encoder parameter maps  $\mu_\theta$  and  $\sigma_\theta$ , as well as the  
 311 decoder mean  $\mu_\phi$  are first trained until convergence, allowing the generation of the out-of-distribution  
 312 pseudo-inputs and subsequently, the training of the decoder variance.

313 **Image data.** We evaluate the performance of our proposed *dissipative*  
 314 *V3AE (d-V3AE)* against a fully Gaussian VAE on image data, coming from  
 315 FashionMNIST, SVHN and CIFAR10. For both models, all parameter  
 316 maps share the same underlying architecture, with the addition of either  
 317 a softplus and/or a shifting last layer to ensure definition of both the  
 318 variational and the generative distribution’s moments<sup>8</sup>.

319 Tab. 2 compares model performance on two metrics, the log-likelihood  
 320 and the RMSE between the original inputs  $x$  and reconstructed samples  $\tilde{x}$ ,  
 321 where  $\tilde{x} \sim p_\theta(x|\lambda, z)$ ,  $(\lambda, z) \sim q_\phi(\lambda, z|x)$ . Unlike most previous imple-  
 322 mentations, we focus on actual samples, and not the mean, of the generative  
 323 distributions. This comparison emphasize the cooperation between the  
 324 decoder’s mean and variance, allowing evaluation of the models’ uncer-  
 325 tainty estimates. Our method both qualitatively (Fig. 7), and quantitatively  
 326 improves on a Gaussian VAE’s sampling ability. The prior smoothens the  
 327 uncertainty estimates, resulting in more realistic and less crisp samples.  
 328 The log-likelihoods, evaluated at test time using truncation, i.e.  $p_{\text{trunc}}(x) =$   
 329  $p_\theta(x)/(F_x(1) - F_x(0))$ , to account for the finite support of data, reveal  
 330 that our model can achieve a better fit, if the prior is selected correctly. In  
 331 SVHN and CIFAR10, the presence of color channels complicates the selection process and challenges  
 332 our choice of a single homoscedastic prior for all pixels and channels. We note that the dissipative loss  
 333 also applies to classic VAEs with Bernoulli-only decoders; see Sec. III. of the supplements for details.

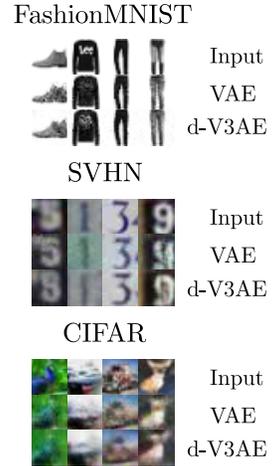


Figure 7: Generated samples

<sup>7</sup>The derivation of the dissipative loss function is provided in Sec. III. of the supplementary materials.

<sup>8</sup>Implementation details are provided in Sec. III. of the supplementary materials.

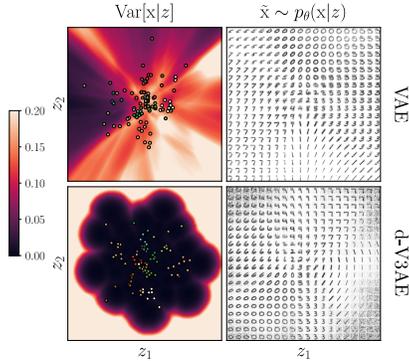


Figure 8: Decoder’s aggregated variance (left) and generated samples (right) from the latent space. Coloured points correspond to latent representations of test data, with per-class colours.

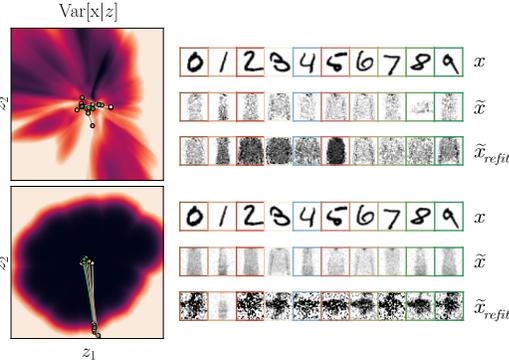


Figure 9: Effect of encoder refitting on the latent representations (left) and corresponding samples (right). OOD inputs (first rows,  $x$ ) initially result in in-distribution samples (second rows,  $\tilde{x}$ ). The refitted encoder displaces the encodings (coloured trajectories), modifying the generated samples (third rows,  $\tilde{x}_{\text{refit}}$ ).

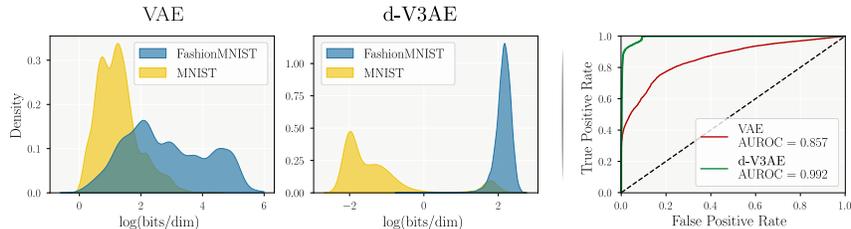


Figure 10: Empirical densities of likelihoods for FashionMNIST (ID) and MNIST (OOD). The clear separation of distributions offered by our method is reflected in the high AUROC shown on the right.

334 **Applications of robust generative uncertainty.** In Figs. 8 & 9, the colouring of the 2D latent space  
 335 represent the aggregated decoder variance  $\sum_{i=1}^d (\sigma_{\theta}(z)^2)_i$ . It is clear that our method displays more  
 336 regular uncertainty estimates, and provides the extrapolation guarantees we strove for. Beyond in-  
 337 creased robustness and better generative power, this unlocks meaningful out-of-distribution detection,  
 338 beating previous state-of-the-art [Havtorn et al., 2021]. For Figs. 9 & 10, as argued in Mattei and  
 339 Frelsen [2018b], we refit at test time the encoder of models trained on FashionMNIST on MNIST.  
 340 The regularity and structure of the decoder variance rewards the encoder for learning to place represen-  
 341 tations of OOD data outside of the region of in-distribution latent encodings, resulting in a model that  
 342 is aware of its own inability to reconstruct plausible data, as displayed by the row  $\tilde{x}_{\text{refit}}$  of d-V3AE.

## 343 4 Conclusion

344 We have introduced a novel loss, the dissipative loss, that leverages artificial out-of-distribution  
 345 pseudo-inputs for learning robust uncertainty estimates. We demonstrate through a Bayesian approach  
 346 that casts a prior distribution over the model’s variance a principled mechanism for controlling the  
 347 extrapolation properties of neural networks governing the predictive uncertainty. Our experimental  
 348 results reflect the benefits of our principled and scalable approach, displaying better calibrated and  
 349 more robust uncertainty estimates, while matching the predictive power of known baselines. Finally,  
 350 and most interestingly, our approach can instill into probabilistic models a notion of their own  
 351 ignorance, increasing their ability to *know what they don’t know*.

352 **Limitations.** The largest limitation of our approach is that it depends on an estimate of the density of  
 353 the input data. In our experience, even coarse-grained densities are sufficient to significantly improve  
 354 upon current approaches. However, as one rarely has guaranteed good estimates of the input density,  
 355 our method cannot be approached as a black-box. One exception seems to be the application to VAEs,  
 356 where the aggregated posterior, in our experience, always provide a suitable density estimate.

357 **Negative societal impact.** Improving the ability of predictive models to assess their own confidence  
 358 is solely a positive contribution as it can help alleviate potential consequences of incorrect predictions.  
 359 We are therefore not aware of any potential negative impacts of our work.

## 360 References

- 361 Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka.  
362 How neural networks extrapolate: From feedforward to graph neural networks. *arXiv preprint*  
363 *arXiv:2009.11848*, 2020.
- 364 Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané.  
365 Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
- 366 Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern*  
367 *analysis and machine intelligence*, 12(10):993–1001, 1990.
- 368 Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive  
369 uncertainty estimation using deep ensembles. In *Advances in neural information processing*  
370 *systems*, pages 6402–6413, 2017.
- 371 David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computa-*  
372 *tion*, 4(3):448–472, 1992.
- 373 Yarín Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model  
374 uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059,  
375 2016.
- 376 Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, August 2006. ISBN  
377 0387310738.
- 378 Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural  
379 networks. *arXiv preprint arXiv:1706.04599*, 2017.
- 380 Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution  
381 examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
- 382 Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-  
383 distribution inputs to deep generative models using typicality. *arXiv preprint arXiv:1906.02994*,  
384 2019.
- 385 Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence  
386 predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision*  
387 *and pattern recognition*, pages 427–436, 2015.
- 388 Christos Louizos and Max Welling. Multiplicative normalizing flows for variational bayesian neural  
389 networks. *arXiv preprint arXiv:1703.01961*, 2017.
- 390 Nicki Skafté, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance  
391 networks. In *Advances in Neural Information Processing Systems*, pages 6326–6336, 2019.
- 392 Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for  
393 detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.
- 394 Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S. Du, Ken ichi Kawarabayashi, and Stefanie Jegelka.  
395 How neural networks extrapolate: From feedforward to graph neural networks, 2021.
- 396 David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability  
397 distribution. In *Proceedings of 1994 IEEE international conference on neural networks (ICNN'94)*,  
398 volume 1, pages 55–60. IEEE, 1994.
- 399 Søren Hauberg. Only bayes should learn a manifold (on the estimation of differential geometric  
400 structure from data), 2019.
- 401 Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of  
402 deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.
- 403 Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint*  
404 *arXiv:1312.6114*, 2013.

- 405 Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and  
406 approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.
- 407 Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*,  
408 31(2):105–112, 2009.
- 409 Burr Settles. Active learning. *Synthesis lectures on artificial intelligence and machine learning*, 6(1):  
410 1–114, 2012.
- 411 Jonas Moćkus. On bayesian methods for seeking the extremum. In *Optimization techniques IFIP*  
412 *technical conference*, pages 400–404. Springer, 1975.
- 413 Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*  
414 (*Adaptive Computation and Machine Learning*). The MIT Press, 2005.
- 415 Alex Graves. Practical variational inference for neural networks. *Advances in neural information*  
416 *processing systems*, 24:2348–2356, 2011.
- 417 Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in  
418 neural networks. *arXiv preprint arXiv:1505.05424*, 2015.
- 419 José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning  
420 of bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869,  
421 2015.
- 422 Leonard Hasenclever, Stefan Webb, Thibaut Lienart, Sebastian Vollmer, Balaji Lakshminarayanan,  
423 Charles Blundell, and Yee Whye Teh. Distributed bayesian learning with stochastic natural gradient  
424 expectation propagation and the posterior server. *The Journal of Machine Learning Research*, 18  
425 (1):3744–3780, 2017.
- 426 Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In  
427 *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688,  
428 2011.
- 429 Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, and Frank Hutter. Bayesian optimization  
430 with robust bayesian neural networks. *Advances in neural information processing systems*, 29:  
431 4134–4142, 2016.
- 432 Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- 433 Jeppe Thagaard, Søren Hauberg, Bert van der Vegt, Thomas Ebstrup, Johan D. Hansen, and Anders B.  
434 Dahl. Can you trust predictive uncertainty under real dataset shifts in digital pathology? In *Medical*  
435 *Image Computing and Computer-Assisted Intervention (MICCAI)*, Lima, Peru, October 2020.
- 436 Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua  
437 Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model’s uncertainty?  
438 evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing*  
439 *Systems*, pages 13991–14002, 2019.
- 440 Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- 441 Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov.  
442 Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine*  
443 *learning research*, 15(1):1929–1958, 2014.
- 444 Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution  
445 image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.
- 446 Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial  
447 examples. *arXiv preprint arXiv:1412.6572*, 2014a.
- 448 Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier  
449 exposure. *arXiv preprint arXiv:1812.04606*, 2018.

- 450 Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,  
451 Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information*  
452 *processing systems*, 27:2672–2680, 2014b.
- 453 Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Russ R Salakhutdinov. Good semi-  
454 supervised learning that requires a bad gan. In *Advances in neural information processing systems*,  
455 pages 6510–6520, 2017.
- 456 Andrew Stirn and David A Knowles. Variational variance: Simple and reliable predictive variance  
457 parameterization. *arXiv preprint arXiv:2006.04910*, 2020.
- 458 Pierre-Alexandre Mattei and Jes Frellsen. Leveraging the exact likelihood of deep latent variable  
459 models. In *Advances in Neural Information Processing Systems*, pages 3855–3866, 2018a.
- 460 Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin.  
461 *Bayesian data analysis*. CRC press, 2013.
- 462 Martin Jørgensen. *Stochastic Representations with Gaussian Processes and Geometry*. PhD thesis,  
463 Technical University of Denmark, 2020.
- 464 Elisa T Lee and John Wang. *Statistical methods for survival data analysis*, volume 476. John Wiley  
465 & Sons, 2003.
- 466 Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb.  
467 Learning from simulated and unsupervised images through adversarial training. In *Proceedings of*  
468 *the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.
- 469 Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International*  
470 *Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.
- 471 George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density  
472 estimation. *arXiv preprint arXiv:1705.07057*, 2017.
- 473 A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77  
474 (379):605–610, 1982.
- 475 David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians.  
476 *Journal of the American statistical Association*, 112(518):859–877, 2017.
- 477 Andrew YK Foong, Yingzhen Li, José Miguel Hernández-Lobato, and Richard E Turner. ‘in-  
478 between’ uncertainty in bayesian neural networks. *arXiv preprint arXiv:1906.11537*, 2019.
- 479 Jakob D. Havtorn, Jes Frellsen, Søren Hauberg, and Lars Maaløe. Hierarchical vaes know what they  
480 don’t know, 2021.
- 481 P.-A Mattei and J Frellsen. Refit your encoder when new data comes by. In *3rd NeurIPS workshop*  
482 *on Bayesian Deep Learning*, 2018b.
- 483 Norman L Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous univariate distri-*  
484 *butions, volume 1, 2nd Edition*. John wiley & sons, 1994.
- 485 Christian Bauckhage. Computing the kullback-leibler divergence between two generalized gamma  
486 distributions. *arXiv preprint arXiv:1401.6853*, 2014.
- 487 Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*  
488 *arXiv:1412.6980*, 2014.
- 489 Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by  
490 reducing internal covariate shift. In *International conference on machine learning*, pages 448–456.  
491 PMLR, 2015.

492 **Checklist**

- 493 1. For all authors...
- 494 (a) Do the main claims made in the abstract and introduction accurately reflect the paper's  
495 contributions and scope? [Yes]
- 496 (b) Did you describe the limitations of your work? [Yes] Yes, we have included a separate  
497 subsection on the limitations. See Section 4.
- 498 (c) Did you discuss any potential negative societal impacts of your work? [Yes] Covered  
499 specifically in Section 4.
- 500 (d) Have you read the ethics review guidelines and ensured that your paper conforms to  
501 them? [Yes] Yes.
- 502 2. If you are including theoretical results...
- 503 (a) Did you state the full set of assumptions of all theoretical results? [Yes] Our results are  
504 obtained through empirical evidence. We do, however, include relevant derivations in  
505 the supplementary material.
- 506 (b) Did you include complete proofs of all theoretical results? [N/A]
- 507 3. If you ran experiments...
- 508 (a) Did you include the code, data, and instructions needed to reproduce the main ex-  
509 perimental results (either in the supplemental material or as a URL)? [Yes] Code is  
510 included in the supplementary material.
- 511 (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they  
512 were chosen)? [Yes] Yes, every experiment is accompanied by a separate settings file  
513 in yaml format.
- 514 (c) Did you report error bars (e.g., with respect to the random seed after running experi-  
515 ments multiple times)? [Yes] Each experiment was run at least in triplicate.
- 516 (d) Did you include the total amount of compute and the type of resources used (e.g., type  
517 of GPUs, internal cluster, or cloud provider)? [Yes] Yes, in supplementary material.
- 518 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
- 519 (a) If your work uses existing assets, did you cite the creators? [Yes]
- 520 (b) Did you mention the license of the assets? [Yes] We ran our experiments on standard,  
521 well known datasets, reference the source, which itself includes licenses.
- 522 (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]  
523 Our code is available in the supplementary material
- 524 (d) Did you discuss whether and how consent was obtained from people whose data you're  
525 using/curating? [N/A] We did not run any experiments on personal data.
- 526 (e) Did you discuss whether the data you are using/curating contains personally identifiable  
527 information or offensive content? [N/A]
- 528 5. If you used crowdsourcing or conducted research with human subjects...
- 529 (a) Did you include the full text of instructions given to participants and screenshots, if  
530 applicable? [N/A]
- 531 (b) Did you describe any potential participant risks, with links to Institutional Review  
532 Board (IRB) approvals, if applicable? [N/A]
- 533 (c) Did you include the estimated hourly wage paid to participants and the total amount  
534 spent on participant compensation? [N/A]

535 **I. Student-t likelihood**

536 **I.1 Marginal distribution of a Gaussian likelihood with a Gamma precision**

537 In the case of a Gaussian likelihood with a latent Gamma distributed precision, the marginal distribu-  
538 tion follows:

$$\begin{aligned}
p_{\theta}(x) &= \int \mathcal{N}(x|\mu, \lambda)\Gamma(\lambda|\alpha, \beta)d\lambda \\
&= \int \frac{\lambda^{1/2}}{\sqrt{2\pi}}e^{-\frac{1}{2}\lambda(x-\mu)^2} \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}d\lambda \\
&= \frac{1}{\Gamma(\alpha)\sqrt{2\pi}} \frac{\beta^{\alpha}}{\left(\beta + \frac{(x-\mu)^2}{2}\right)^{\alpha-\frac{1}{2}}} \int \left[\left(\beta + \frac{1}{2}(x-\mu)^2\right)\lambda\right]^{(\alpha+\frac{1}{2})-1} e^{-\left(\beta+\frac{(x-\mu)^2}{2}\right)\lambda} d\lambda \\
&= \frac{\Gamma\left(\alpha + \frac{1}{2}\right)}{\Gamma(\alpha)\sqrt{2\pi}} \frac{\beta^{\alpha}}{\left(\beta + \frac{(x-\mu)^2}{2}\right)^{\alpha-\frac{1}{2}}} \\
&= \frac{\Gamma\left(\frac{2\alpha+1}{2}\right)}{\Gamma(\alpha)\sqrt{\pi}2\alpha\left(\frac{\beta}{\alpha}\right)^{1/2}} \left(1 + \frac{1}{2\alpha}\left(\frac{x-\mu}{\left(\frac{\beta}{\alpha}\right)^{1/2}}\right)^2\right)^{-\frac{2\alpha+1}{2}} \\
&= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}\hat{\sigma}} \left(1 + \frac{1}{\nu}\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\right)^{-\frac{\nu+1}{2}} \\
&= T\left(x|\nu = 2\alpha, \hat{\mu} = \mu, \hat{\sigma} = \sqrt{\beta/\alpha}\right).
\end{aligned} \tag{6}$$

539 The moments of the marginal distribution are, assuming  $\nu > 2$ ,

$$\begin{cases} \mathbb{E}[x] &= \hat{\mu} = \mu \\ \text{Var}[x] &= \hat{\sigma}^2 \frac{\nu}{\nu-2} = \frac{\beta}{\alpha} \frac{\alpha}{\alpha-1}. \end{cases} \tag{7}$$

540 **I.2 Decomposition of a Student-t's uncertainty**

541 The variance of a non standard Student-t distribution with  $2\alpha$  degrees of freedom and scaled by  
542  $\sqrt{\beta/\alpha}$  can be decomposed as  $\text{Var}[x] = \frac{\beta}{\alpha} \frac{\alpha}{\alpha-1}$ . The number of degrees of freedom scales with the  
543 number of observations that the distribution arises from. When the number of observations grows  
544 towards  $\infty$ , i.e towards perfect information, the term  $\frac{\alpha}{\alpha-1}$  converges towards 1, motivating its casting  
545 into an epistemic factor. The natural consequence is that  $\frac{\beta}{\alpha}$ , which accounts for the rest of the model's  
546 uncertainty, scales as the aleatoric uncertainty. For more details see Jørgensen [2020, p16].

547 **II. Regression experiments**

548 **II.1 Variational Variance's ELBO Closed Form**

549 For a Gaussian likelihood and a Gamma posterior, both terms of the ELBO have a closed form  
550 solution. Firstly the expected log-likelihood verifies:

$$\begin{aligned}
\mathbb{E}_{q(\lambda|x)}[\log p(y|x, \lambda)] &= \int \log \mathcal{N}(y|\mu(x), \lambda)\Gamma(\lambda|\alpha(x), \beta(x))d\lambda \\
&= \int -\frac{1}{2}(\log 2\pi - \log \lambda + \lambda(y - \mu(x))^2) \Gamma(\lambda|\alpha(x), \beta(x))d\lambda \tag{8} \\
&= -\frac{1}{2}\left(\log 2\pi - \mathbb{E}_{q(\lambda|x)}[\log \lambda] + (y - \mu(x))^2 \mathbb{E}_{q(\lambda|x)}[\lambda]\right).
\end{aligned}$$

551 The variational posterior being Gamma distributed, its expected value is defined as  $\mathbb{E}_{q(\lambda|x)}[\lambda] = \frac{\alpha(x)}{\beta(x)}$ .  
552 The logarithmic expectation of a Gamma distribution can be derived to yield [Johnson et al., 1994,  
553 337–349]  $\mathbb{E}_{q(\lambda|x)}[\log \lambda] = \psi(\alpha(x)) - \log \beta(x)$  where  $\psi$  is the digamma function. The closed-form  
554 expression of the expected likelihood is therefore:

$$\mathbb{E}_{q(\lambda|x)}[\log p(y|x, \lambda)] = -\frac{1}{2} \left( \log 2\pi - \psi(\alpha(x)) + \log \beta(x) + \frac{\alpha(x)}{\beta(x)} (y - \mu(x))^2 \right). \quad (9)$$

555 Secondly, the KL-divergence between the posterior  $\Gamma(\alpha(x), \beta(x))$  and the prior  $\Gamma(a, b)$  can be  
556 derived from Equation (28) in Bauckhage [2014, p6]. With Bauckhage’s notation, setting  $p_1 = p_2 =$   
557 1, to correspond to standard Gamma distributions, shape parameters  $d_1 = \alpha(x)$  and  $d_2 = a$ , and  
558 scale parameters  $a_1 = \frac{1}{\beta(x)}$  and  $a_2 = \frac{1}{b}$  the KL-divergence can be expressed as

$$\begin{aligned} D_{\text{KL}}(q(\lambda|x) || p(\lambda)) &= (\alpha(x) - a)\psi(\alpha(x)) \\ &\quad - \log \Gamma(\alpha(x)) + \log \Gamma(a) \\ &\quad + a(\log \beta(x) - \log b) \\ &\quad + \alpha(x) \frac{b - \beta(x)}{\beta(x)}. \end{aligned} \quad (10)$$

## 559 II.2 True posterior and heteroscedasticity

560 The true posterior for variational variance in a regression context can be written  $p(\lambda|y, x)$ . As first  
561 demonstrated in Sec. 8.2 of Stirn and Knowles [2020], it factorizes as:

$$p(\lambda|y, x) = \frac{p(y|x, \lambda)p(\lambda)}{\int p(y|x, \lambda)p(\lambda)d\lambda} \quad (11)$$

$$= \frac{\prod_{n=1}^N p(y_n|x_n, \lambda_n)p(\lambda_n)}{\int \prod_{n=1}^N p(y_n|x_n, \lambda_n)p(\lambda_n)d\lambda_n} \quad (12)$$

$$= \prod_{n=1}^N \frac{p(y_n|x_n, \lambda_n)p(\lambda_n)}{\int p(y_n|x_n, \lambda_n)p(\lambda_n)d\lambda_n} \quad (13)$$

$$= \prod_{n=1}^N p(\lambda_n|y_n, x_n). \quad (14)$$

562 As a result, the true posterior both depends on the inputs  $x_n$  and targets  $y_n$ . It means that a single  
563 input, could theoretically imply different latent precisions for different targets  $y_n \neq y_k$ , thus violating  
564 the x-surjectivity of the heteroscedastic definition.

## 565 II.3 Model architecture

566 We adopted a unified network architecture for the regression case. All neural-network parameter maps  
567 share the same underlying architecture, a single hidden layer with 50 hidden layers using *exponential*  
568 *linear unit (ELU)* activation functions. A final *softplus* layer is applied on the last layer of the  $\sigma$ ,  
569  $\alpha$  and  $\beta$  parameter maps. The  $\alpha$  parameter map is further shifted by +1 to ensure the definition of  
570 the marginal distribution’s variance. Regression models are trained with the *Adam* [Kingma and Ba,  
571 2014] optimiser, and both the inputs and targets are standardised prior to training and testing.

## 572 II.4 Pseudo-input generator

573 Tab. 3 presents the parameters used by the PIG in a regression setting. We remind that these  
parameters are parameters of a gradient descent, with learning rate  $\delta$ . For our experiments, we

Table 3: Parameters for the regression pseudo-input generator.

K	max_iterations	tolerance	$\delta$
N	5	0.005	4e-1

Table 4: UCI benchmarks

Name	Dimensions ( $N, D_x, D_y$ )	Link ( <a href="https://archive.ics.uci.edu/ml/">https://archive.ics.uci.edu/ml/</a> *)
Boston	(505,13,1)	machine-learning-databases/housing/
Carbon	(10721,5,3)	datasets/Carbon+Nanotubes
Concrete	(1030,8,1)	datasets/Concrete+Compressive+Strength
Energy	(768,8,2)	datasets/Energy+efficiency
Kin8nm	(8192,8,1)	<a href="https://www.openml.org/d/189">https://www.openml.org/d/189</a>
Naval	(11934,16,2)	datasets/Condition+Based+Maintenance+of+Naval+Propulsion+Plants
Power plant (CCPP)	(9568,4,1)	datasets/Combined+Cycle+Power+Plant
Protein	(45630, 9, 1)	datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure
Superconductivity	(21263,81,1)	datasets/Superconductivity+Data
Wine-red	(1599,11,1)	datasets/Wine+Quality
Wine-white	(4898,11,1)	datasets/Wine+Quality
Yacht	(308,6,1)	datasets/Yacht+Hydrodynamics

575 approximated the input density with a Bayesian Gaussian mixture model<sup>9</sup> with diagonal covariance  
576 matrices, and initialised with as many components as there are inputs in a batch.

## 577 **II.5 UCI experiments**

578 The UCI experiments consist of the datasets presented in Tab. 4. The results, for the different metrics,  
579 as presented in Tab. 5 to 14, were computed as the mean  $\pm$  the standard deviation over 10 trials with  
580 standardised inputs and targets. Due to a technical error, we were forced to re-run the experiments for  
581  $d$ - $VV$  and  $VV$  (*no PIG*) right before the submission deadline, and reduced the number of trials to 3 for  
582 these methods. The procedure for generating the colour table that summarizes the performance of the  
583 different models presented in Tab. 1 is detailed here. As shown in Fig. 11, each square represents  
584 the performance of a model, for a given metric on a single dataset. The intensity of the colouring  
585 is multiplied by a factor  $f$  ranging from 0, which results in a white square, to 1, which results in a  
586 completely coloured square. That factor is computed, for a method presenting a mean metric  $\mu$  and  $\sigma$   
587 deviation on a given dataset, as

$$f = \begin{cases} 2 F(\mu^*; \mu, \sigma) , & \text{if the metric is best when lower} \\ 2 (1 - F(\mu^*; \mu, \sigma)) , & \text{if the metric is best when higher} \end{cases} ,$$

588 where  $F(\cdot; \mu, \sigma)$  designates the cumulative density function of a Gaussian distribution with mean  $\mu$   
589 and standard deviation  $\sigma$ , and where  $\mu^*$  designates the mean metric of the best performing method.  
590 The factor amounts for the probability mass of the region where the method performs better than the  
591 best method. Consequently, the better a method performs compared to the best, the closer to 1 the  
592 factor will evaluate to, and the more pronounced the final colouring will be.

<sup>9</sup><https://scikit-learn.org/stable/modules/generated/sklearn.mixture.BayesianGaussianMixture.html>

Table 5: UCI benchmarks -  $\mathcal{L}$

	boston	carbon	ccpp	concrete	energy	kin8nm	naval	protein	superconduct	wine red	wine white	yacht
d-VV	-0.61±0.33	1.2±0.11	-0.14±0.015	-0.44±0.13	0.67±0.03	-0.25±0.024	0.52±0.16	-1.3±0.015	-0.54±0.033	-2.0±0.076	-1.8±0.056	-1.8±0.01±0.43
VV (no PIG)	-0.72±0.38	1.2±0.12	-0.16±0.035	-0.46±0.079	0.65±0.032	-0.28±0.038	0.12±0.4	-1.3±0.011	-0.56±0.0088	-2.4±0.19	-2.0±0.096	1.0±0.1
VV (no prior)	-6.4±0.1±2.9±0.1	-3.9±0.3±5.8±0.2	-1.1±0.01±1.1	-4.4±0.1±1.4±0.1	-1.2±0.2±8.1±0.1	-1.7±0.01±1.5	-9.6±4.6	-1.4±0.1±2.6	-2.7±0.02±5.9±0.1	-3.8±0.1±3.1±0.1	-8.7±0.02±1.7±0.3	-5.5±0.02±1.1±0.3
Mean-variance network	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
Deep ensemble	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
MC dropout	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
Skafte et al.	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan

Table 6: UCI benchmarks -  $\log p(y|x)$

	boston	carbon	ccpp	concrete	energy	kin8nm	naval	protein	superconduct	wine red	wine white	yacht
d-VV	-0.43±0.35	1.4±0.13	-0.067±0.013	-0.29±0.15	0.87±0.036	-0.17±0.022	0.71±0.19	-1.2±0.014	-0.38±0.021	-1.9±0.071	-1.7±0.062	0.9±0.018
VV (no PIG)	-0.42±0.39	1.5±0.11	-0.034±0.03	-0.25±0.096	0.89±0.031	-0.15±0.039	0.52±0.2	-1.1±0.014	-0.35±0.018	-2.1±0.21	-1.7±0.097	1.3±0.11
VV (no prior)	-3.3±1.4	0.98±3.1	0.047±0.06	-0.83±0.5	0.47±0.2	-0.36±0.073	-0.13±0.32	-1.4±0.38	-1.7±1.7	-7.8±6.4	-3.1±0.02±5.5±0.2	0.63±0.59
Mean-variance network	-0.76±0.07	-3.8±0.045	-0.58±0.14	-0.68±0.09	-1.2±0.11	-0.61±0.062	-2.3±0.079	-1.1±0.053	-0.66±0.045	-2.6±0.03±5.4±0.3	-2.8±0.1±4.9±0.1	-0.59±0.11
Deep ensemble	-0.68±0.042	-3.7±0.036	-0.61±0.046	-0.65±0.04	-1.2±0.04	-0.65±0.027	-2.3±0.063	-1.0±0.012	-0.68±0.028	-1.2±0.077	-1.2±0.077	-0.58±0.037
MC dropout	-0.81±0.51	0.29±1.1	-3.4±0.56	-0.9±0.33	0.36±0.26	-0.63±0.049	-0.2±0.74	-7.4±0.27	-1.7±0.25	-4.2±0.91	-5.9±1.1	0.33±0.69
Skafte et al.	-0.18±0.19	1.1±0.51	-0.18±0.12	-0.44±0.15	0.28±0.37	-0.61±0.12	-2.7±0.22	-1.5±0.74	-0.96±0.18	-1.1±0.036	-1.4±0.58	0.4±0.14

Table 7: UCI benchmarks - RMSE  $[y, \mu(x)]$

	boston	carbon	ccpp	concrete	energy	kin8nm	naval	protein	superconduct	wine red	wine white	yacht
d-VV	0.33±0.095	0.034±0.022	0.23±0.005	0.29±0.043	0.076±0.0067	0.26±0.0079	0.088±0.056	0.71±0.0076	0.35±0.011	0.89±0.0094	0.9±0.027	0.046±0.018
VV (no PIG)	0.33±0.084	0.035±0.021	0.23±0.011	0.29±0.044	0.078±0.0084	0.26±0.012	0.099±0.034	0.71±0.0047	0.35±0.0071	0.9±0.043	0.88±0.024	0.038±0.021
VV (no prior)	0.38±0.086	0.03±0.017	0.23±0.0085	0.33±0.037	0.3±0.03	0.28±0.0087	0.33±0.065	0.75±0.0091	0.4±0.018	0.77±0.071	0.82±0.058	0.82±0.12
Mean-variance network	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
Deep ensemble	0.33±0.05	0.078±0.005	0.24±0.0077	0.28±0.018	0.13±0.017	0.33±0.0072	0.2±0.074	0.7±0.0084	0.33±0.011	0.77±0.053	0.79±0.048	0.11±0.043
MC dropout	0.33±0.066	0.093±0.082	0.27±0.03	0.35±0.072	0.22±0.082	0.44±0.065	0.86±0.085	1.1±0.73	0.67±0.22	0.76±0.024	0.93±0.39	0.086±0.064
Skafte et al.												

Table 8: UCI benchmarks - RMSE  $\left[ \text{Var}[y|x], (y - \mu(x))^2 \right]$

	boston	carbon	cepp	concrete	energy	kin8nm	naval	protein	superconductor	wine red	wine white	yacht
d-VV	0.25±0.11	0.05±0.036	0.15±0.0084	0.24±0.15	0.026±0.00036	0.14±0.017	0.033±0.0013	0.8±0.036	0.39±0.043	1.4±0.085	1.5±0.011	0.028±0.0012
VV (no PIG)	0.37±0.26	0.05±0.036	0.15±0.012	0.22±0.13	0.025±0.0014	0.13±0.015	3.8e+03±6.6e+03	0.8±0.013	0.41±0.068	1.6±0.38	1.5±0.14	0.014±0.0039
VV (no prior)	inf±nan	0.026±0.029	0.24±0.3	1.3±3.5	0.15±0.038	6.3e+01±1.5e+02	inf±nan	inf±nan	5.4e+04±8.9e+04	3.5±4.1	3.8e+03±1.6e+04	2.3e+03±4.4e+03
Mean-variance network	0.61±0.14	1.7±0.057	0.5±0.18	0.55±0.15	0.55±0.058	0.48±0.087	2.0±0.2	0.88±0.087	0.53±0.079	2.7±0.78	1.5±0.33	0.59±0.13
Deep ensemble	0.53±0.062	1.6±0.03	0.47±0.044	0.47±0.049	0.49±0.024	0.47±0.036	1.9±0.16	0.77±0.03	0.53±0.053	3.0±2.4	1.1±0.2	0.52±0.052
MC dropout	0.29±0.17	0.088±0.013	0.14±0.039	0.18±0.035	0.081±0.013	0.2±0.013	0.12±0.045	0.88±0.03	0.33±0.061	1.2±0.21	1.2±0.16	0.089±0.035
Skafte et al.	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan

Table 9: UCI benchmarks - RMSE  $[y, \tilde{y}]$

	boston	carbon	cepp	concrete	energy	kin8nm	naval	protein	superconductor	wine red	wine white	yacht
d-VV	0.56±0.079	0.11±0.0069	0.42±0.0072	0.44±0.056	0.21±0.027	0.46±0.012	0.22±0.032	1.0±0.013	0.54±0.0039	1.1±0.1	1.2±0.045	0.18±0.03
VV (no PIG)	0.57±0.16	0.1±0.0028	0.4±0.012	0.43±0.053	0.17±0.011	0.43±0.0097	0.29±0.033	1.0±0.016	0.56±0.02	1.1±0.063	1.2±0.055	0.1±0.0093
VV (no prior)	0.58±0.28	0.043±0.016	0.33±0.0087	0.43±0.045	0.4±0.075	0.39±0.036	1.3±1.0	1.1±0.13	0.59±0.11	1.1±0.089	1.2±0.12	0.97±0.22
Mean-variance network	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
Deep ensemble	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
MC dropout	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
Skafte et al.	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan

Table 10: UCI benchmarks shifted -  $\mathcal{L}$

	boston	carbon	cepp	concrete	energy	kin8nm	naval	protein	superconduct	wine red	wine white	yacht
d-VV	-1.3±0.32	1.2±0.019	-0.25±0.076	-1.3±0.27	-0.69±1.1	-0.33±0.034	-4.7±4.6	-1.6±0.11	-1.4±0.32	-2.8±0.21	-2.0±0.11	-7.6±2.6
VV (no PIG)	-1.5±0.28	1.1±0.029	-0.33±0.14	-1.5±0.38	-0.56±0.9	-0.38±0.04	-5.3±6.6	-1.6±0.16	-1.5±0.34	-3.4±0.35	-2.5±0.29	0.35±0.19
Mean-variance network	-1.1e+02±1.1e+02	-4e+03±6.8e+02	-8.4±1.7	-7.6e+01±6e+01	-8.9e+02±1.6e+03	-2.4e+01±1.9	-2.6e+01±4.8e+01	-9.3±6.7	-2.2e+02±1.1e+02	-4.8e+01±5.6e+01	-3.5e+02±4.3e+02	-1.9e+02±1.5e+02
Deep ensemble	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
MC dropout	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
Skafte et al.	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan

Table 11: UCI benchmarks shifted -  $\log p(y|x)$

	boston	carbon	cepp	concrete	energy	kin8nm	naval	protein	superconduct	wine red	wine white	yacht
d-VV	-1.1±0.32	1.3±0.055	-0.18±0.085	-1.1±0.27	-0.47±1.1	-0.26±0.037	-4.5±4.6	-1.4±0.12	-1.3±0.32	-2.7±0.2	-2.0±0.12	0.71±0.76
VV (no PIG)	-1.2±0.27	1.4±0.019	-0.2±0.14	-1.2±0.36	-0.2±0.76	-0.22±0.039	-5.0±6.5	-1.4±0.15	-1.3±0.33	-3.1±0.32	-2.2±0.27	0.65±0.19
VV (no prior)	-1e+01±1.1e+01	-0.6±2.9	-0.1±0.079	-1.1e+01±1.5e+01	-9.7e+01±2.6e+02	-0.88±0.14	-1.5e+01±1.9e+01	-2.3±1.3	-8.3±1.2e+01	-1.2e+01±1.3e+01	-1.8e+02±3.2e+02	0.51±0.55
Mean-variance network	-0.84±0.088	-3.9±0.34	-0.54±0.037	-0.77±0.097	-1.4±0.26	-0.64±0.089	-3.6±0.89	-1.3±0.089	-0.98±0.11	-2.3e+02±7.5e+02	-1.7±0.23	-0.53±0.046
Deep ensemble	-0.79±0.071	-3.7±0.12	-0.65±0.021	-0.78±0.038	-1.4±0.23	-0.65±0.049	-3.6±0.89	-1.2±0.069	-0.94±0.061	-1.6±0.26	-1.3±0.04	-0.58±0.047
MC dropout	-2.5±1.3	0.55±0.11	-4.3±0.43	-2.4±0.52	-1.1±2.5	-0.86±0.21	-2.2e+01±1.2e+01	-9.9±1.3	-5.9±1.7	-4.4±0.5	-5.6±0.49	0.37±0.35
Skafte et al.	-0.16±0.089	1.1±0.25	-0.16±0.024	-0.38±0.061	0.2±0.25	-0.59±0.04	-2.8±0.15	-1.5±0.46	-1.1±0.16	-1.1±0.022	-1.3±0.13	0.43±0.065

Table 12: UCI benchmarks shifted - RMSE  $[y, \mu(x)]$

	boston	carbon	cepp	concrete	energy	kin8nm	naval	protein	superconduct	wine red	wine white	yacht
d-VV	0.52±0.078	0.032±0.0026	0.26±0.023	0.54±0.068	0.26±0.23	0.28±0.011	1.4±0.96	0.83±0.045	0.59±0.066	1.1±0.05	0.96±0.039	0.16±0.085
VV (no PIG)	0.52±0.078	0.032±0.0027	0.26±0.028	0.54±0.068	0.26±0.22	0.28±0.013	1.4±0.95	0.83±0.048	0.59±0.065	1.1±0.05	0.96±0.044	0.098±0.042
VV (no prior)	0.43±0.096	0.031±0.0025	0.25±0.013	0.44±0.054	0.39±0.14	0.29±0.012	1.6±0.71	0.84±0.038	0.52±0.057	0.79±0.031	0.87±0.025	0.74±0.22
Mean-variance network	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
Deep ensemble	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
MC dropout	0.41±0.055	0.076±0.0024	0.25±0.01	0.44±0.038	0.22±0.11	0.36±0.022	1.3±0.59	0.8±0.032	0.51±0.048	0.84±0.028	0.86±0.036	0.13±0.041
Skafte et al.	0.3±0.037	0.095±0.034	0.26±0.0056	0.35±0.03	0.23±0.048	0.43±0.022	0.9±0.064	1.0±0.22	0.6±0.084	0.76±0.014	0.84±0.052	0.066±0.022

Table 13: UCI benchmarks shifted - RMSE  $[\text{Var}[y|x], (y - \mu(x))^2]$

	boston	carbon	cepp	concrete	energy	kin8nm	naval	protein	superconductor	wine red	wine white	yacht
d-VV	0.74±0.37	0.044±0.0066	0.13±0.026	0.61±0.2	0.28±0.29	0.16±0.014	4.3±4.9	0.98±0.095	0.67±0.15	2.3±0.43	1.6±0.15	0.14±0.088
VV (no PIG)	0.74±0.35	0.042±0.0064	0.15±0.027	0.64±0.21	0.25±0.26	0.15±0.018	4.2±4.9	1.0±0.12	0.66±0.14	2.4±0.45	1.6±0.16	0.036±0.0074
VV (no prior)	inf±nan	0.039±0.0061	0.13±0.031	2.7±0.1±7.4e+01	1.2e+02±2.7e+02	3.4e+01±7.1e+01	inf±nan	inf±nan	inf±nan	4.0±5.2	2.8e+03±8.4e+03	7.1e+04±1.7e+05
Mean-variance network	0.72±0.19	1.3±0.51	0.44±0.058	0.57±0.16	0.55±0.092	0.49±0.1	2.9e+01±3.6e+01	1.0±0.08	0.73±0.11	4.9±1.4	1.8±0.32	0.52±0.074
Deep ensemble	0.71±0.16	1.5±0.22	0.49±0.015	0.54±0.086	0.59±0.2	0.46±0.058	3.1e+01±3.7e+01	0.88±0.085	0.7±0.097	7.7±6.9	1.3±0.15	0.51±0.046
MC dropout	0.46±0.21	0.089±0.0035	0.14±0.022	0.31±0.061	0.12±0.082	0.23±0.035	2.3±1.5	1.0±0.065	0.53±0.092	1.3±0.12	1.4±0.14	0.091±0.04
Skafte et al.	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan

Table 14: UCI benchmarks shifted - RMSE  $[y, \tilde{y}]$

	boston	carbon	cepp	concrete	energy	kin8nm	naval	protein	superconductor	wine red	wine white	yacht
d-VV	0.73±0.072	0.12±0.0081	0.44±0.013	0.73±0.071	0.44±0.16	0.48±0.025	1.7±0.8	1.2±0.035	0.79±0.052	1.3±0.055	1.2±0.066	0.22±0.093
VV (no PIG)	0.67±0.076	0.11±0.002	0.43±0.021	0.71±0.041	0.45±0.19	0.45±0.014	1.7±0.82	1.2±0.06	0.8±0.058	1.4±0.078	1.2±0.033	0.21±0.024
VV (no prior)	0.58±0.2	0.038±0.0057	0.35±0.018	0.52±0.062	0.69±0.49	0.44±0.023	2.3±1.3	1.2±0.13	0.93±0.42	1.0±0.048	1.2±0.094	0.97±0.28
Mean-variance network	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
Deep ensemble	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
MC dropout	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan
Skafte et al.	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan	nan±nan

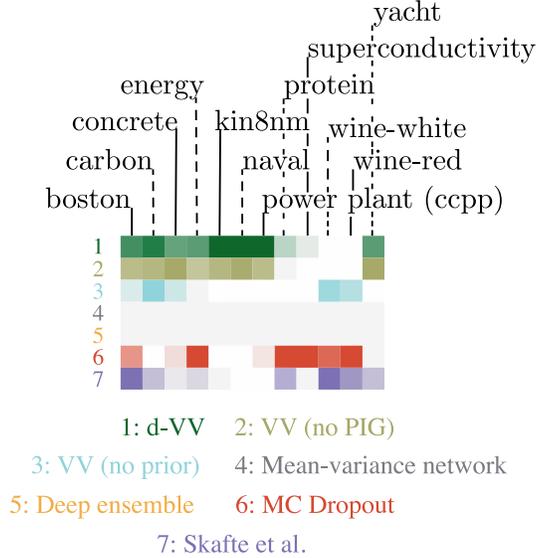


Figure 11: Decomposition of the comparison of models’ performance on a given metric for UCI benchmarks. Each row corresponds to a method, as indicated by the numbers, and each column corresponds to a UCI dataset. The colouring is a function of each model’s performance on each dataset wrt the given metric. Grey squares indicate impossible to compute values.

593 **II.6 Prior parameters**

594 For the toy experiments (Fig. 2, 4 and 5), an homoscedastic prior that matches the standard deviation  
 595 of the targets  $\bar{\sigma}$  is chosen. As shown in Fig. 12, for a Gamma prior, the rate  $\beta$  controls its informativity,  
 596 the closer  $\beta$  is to 0, the more spread out the prior is, and the less penalising it is for the posterior to  
 597 diverge from it. We thus deliberately choose a prior with low informativity,  $\beta = 1e-3$ , and infer the  
 shape as  $\alpha = 1 + \beta/\bar{\sigma}$ . For the UCI benchmarks, we aimed to adopt a prior that would match the

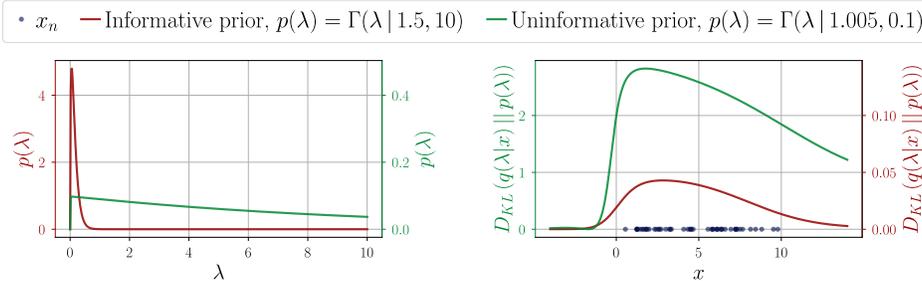


Figure 12: Effect of the informativity of the prior (as displayed on the left) on the KL divergence of the trained posterior on an artificial example (right). The scale of the respective KL divergences reveals that the heavy-tailed prior (green) allows the posterior to be significantly influenced by data, while its counterpart (red) is much more restrictive.

598 model’s empirical variance  $(y - \mu(x))^2$ . As such, we first ran a training run to determine the model’s  
 599 empirical variance on each dataset, and subsequently adopted  $\alpha = 1.5$  and  $\beta = (\alpha - 1) (y - \mu(x))^2$ ,  
 600 with the choice for  $\alpha$  being motivated by stability concerns, and obtained from an empirical study.  
 601 All prior parameters can be found in the configuration files present in the source code.  
 602

603 **III. Generative models experiments**

Table 15: Datasets for generative models

Name	Dimensions ( $N, C, D_x, D_y$ )	Link
MNIST	(60000, 1, 28, 28)	<a href="http://yann.lecun.com/exdb/mnist/">http://yann.lecun.com/exdb/mnist/</a>
FashionMNIST	(60000, 1, 28, 28)	<a href="https://github.com/zalandoresearch/fashion-mnist">https://github.com/zalandoresearch/fashion-mnist</a>
SVHN	(600000, 3, 32, 32)	<a href="http://ufldl.stanford.edu/housenumbers/">http://ufldl.stanford.edu/housenumbers/</a>
CIFAR	(60000, 3, 32, 32)	<a href="https://www.cs.toronto.edu/~kriz/cifar.html">https://www.cs.toronto.edu/~kriz/cifar.html</a>

Table 16: Parameters for the generative model pseudo-input generator

K	max_iterations	tolerance	$\delta$
N	10	0.007	$4e-1$

604 **III.1 Datasets**

605 Tab. 15 lists all the datasets used in the generative modelling experiments.

606 **III.2 Dissipative loss for generative models**

607 The full expression of the dissipative loss for the V3AE is given as:

$$\text{Loss}(q_\phi, \theta; \mathcal{D}_{\text{train}}) = - \left[ \left( \sum_{\mathbf{x} \in \mathcal{D}_{\text{train}}} \mathbb{E}_{q_\phi(z|\mathbf{x})} [\mathbb{E}_{q_\phi(\lambda|z)} [\log p_\theta(\mathbf{x}|z)] - D_{\text{KL}}(q_\phi(\lambda|z) || p(\lambda))] \right) - D_{\text{KL}}(q_\phi(z|\mathbf{x}) || p(z)) \right] + \mathbb{E}_{q_{\text{out}}(z)} [D_{\text{KL}}(q_\phi(\lambda|z) || p(\lambda))] \tag{15}$$

608 The expected likelihood w.r.t the posterior  $q_\phi(z|\mathbf{x})$  is intractable as it requires the integration of the  
 609 parameter maps  $\alpha_\phi(z)$  and  $\beta_\phi(z)$  and must be approximated through MC-integration, using multiple  
 610 sampled latent codes. We observed in practice that a low number of sampled codes, typically, 2 or 3,  
 611 is sufficient for ensuring convergence.

612 **III.3 Model architecture**

613 In our VAEs, the encoder and decoder networks’ architectures are mirrored, and all parameter maps of  
 614 each stage share on the same architecture. For the MNIST and FashionMNIST datasets, we relied on  
 615 fully connected encoder-decoders, with 2 hidden layers with respectively 512 and 256 neurons. Each  
 616 fully connected layer is followed by *batch normalisation* [Ioffe and Szegedy, 2015]. For the SVHN  
 617 and CIFAR datasets, we relied on a convolutional architecture, with hidden dimensions corresponding  
 618 to depths of 32, 64, 128, 256, and 512, for kernels of size 3, with a stride of 2 and a padding of 1.  
 619 Again, batch normalisation is applied after each layer. In both cases, we used *leaky rectified linear*  
 620 *units (Leaky ReLU)* for activations and here again, softplus and shifting might be applied on the last  
 621 layer of the different parameter maps to ensure the proper definition of the quantities they model, and  
 622 models are optimised with Adam.

623 **III.4 Pseudo-input generator**

624 Tab. 16 presents the parameters used by the PIG in a generative modelling setting. We remind  
 625 that these parameters are parameters of a gradient descent, with learning rate  $\delta$ . Because it is too  
 626 computationally expensive to use the complete aggregate posterior as the density estimate we base  
 627 the PIG on, we iteratively generated pseudo-inputs using the aggregate posterior established on one  
 628 batch at a time.

629 **III..5 Prior parameters**

630 As for the regression experiments, an homoscedastic Gamma prior, with the same parameters for  
 631 all image channels was chosen for model comparisons. The shape and rate parameters were tuned  
 632 with the same base intuition as for the UCI benchmarks; the prior uncertainty should be fairly close  
 633 to the empirical mean of the model. An empirical grid search was conducted to determine the best  
 634 combination of prior parameters wrt to the objective to optimise. In the case of out-of-distribution  
 635 detection, we adopted an heteroscedastic prior. Such prior adopts similar base parameters as the more  
 636 standard homoscedastic prior, but its rate parameter, and consequently its associated uncertainty,  
 637 increases linearly as a function of the distance to the closest of the  $C$  pre-determined K-means<sup>10</sup>  
 638 cluster center, where  $C$  is the number of classes in the dataset. Again, the prior parameters used for  
 639 running the experiments can be found in the configuration files provided with the source code.

640 **III..6 Pseudo-inputs training for VAE's with Bernoulli likelihood**

641 Motivated by the idea of not necessarily having to adopt a non trivial  $\Gamma(\lambda|a, b)$  prior, we explore  
 642 the use of pseudo-input training in simpler VAE's with Bernoulli likelihood. In this setting, the  
 643 combined epistemic and aleatoric uncertainty on the reconstructed  $\tilde{x}$  is approximated with a measure  
 644 of entropy. As uncertainty is high for distributions with high entropy, we reinterpret the decoded  
 645 Bernoulli distributed reconstruction  $\tilde{x}$  as normalized Categorical distribution and then proceed to  
 646 maximize its entropy for the pseudo inputs  $\hat{z}$ . The resulting loss function,

$$\text{Loss}(q_\phi, \theta; \mathcal{D}_{\text{train}}) = - \left[ \sum_{\mathbf{x} \in \mathcal{D}_{\text{train}}} \mathcal{L}(q_\phi, \theta; \mathbf{x}) - \sum_{\hat{z} \in \mathcal{D}_{\text{out}}} \text{H}[\tilde{x}|\hat{z}] \right], \quad (16)$$

647 balances the overall entropy of the reconstruction by promoting entropy increase,  $\text{H}[\tilde{x}|\hat{z}]$ . Figure 13  
 shows the effect of this method for a VAE trained on a subset of MNIST.

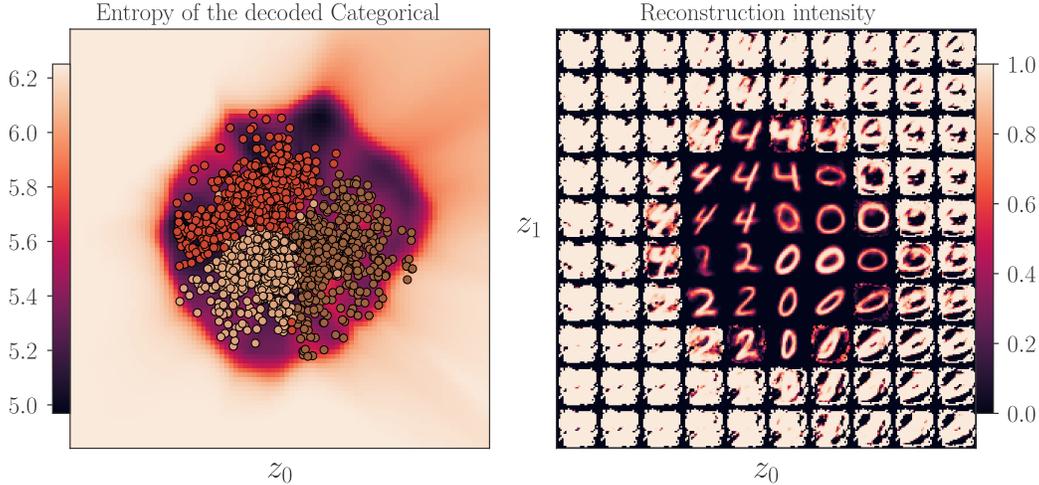


Figure 13: Pseudo inputs for Bernoulli VAE's

648

<sup>10</sup><https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

649 **IV. Implementation details**

650 **IV.1 Source code**

651 The source code is accessible in the GitHub repository [https://github.com/\\*\\*\\*\\*](https://github.com/****)<sup>11</sup>

652 **IV.2 Hardware**

653 Experiments were done on a 16-Core AMD Ryzen 5950X machine with a single Nvidia 3080 GPU.

654 **IV.3 Running times and carbon emissions**

Table 17: Regression models

Dataset	CO <sub>2</sub> (kg)	Time (s)
uci_boston	0.000064	15.969
uci_carbon	0.000695	138.548
uci_ccpp	0.001464	275.824
uci_concrete	0.000102	22.480
uci_energy	0.000095	21.140
uci_kin8nm	0.000869	168.273
uci_naval	0.001427	274.547
uci_protein	0.009920	1762.071
uci_superconduct	0.002010	360.634
uci_wine_red	0.000185	36.186
uci_wine_white	0.000482	89.373
uci_yacht	0.000046	11.404

Table 18: Generative models

Dataset	CO <sub>2</sub> (kg)	Time (s)
fashion_mnist	0.003713	504.190
cifar	0.037554	2573.566
svhn	0.036692	2585.144

Table 19: Generative models w/o pseudo inputs

Dataset	CO <sub>2</sub> (kg)	Time (s)
fashion_mnist	0.002750	407.991
cifar	0.025522	2022.803
svhn	0.025130	2023.293

655 Tab. 18 and 19 demonstrate that the generation of artificial pseudo-inputs incurs a limited additional  
656 computational burden ( $\sim +26\%$ ).

---

<sup>11</sup>Hidden for the review, please refer instead to the .zip folder attached.