A Proofs

A.1 Optimising the ELBO$_{\text{VC}}$ w.r.t q

Rearranging Equation 5, the ELBO$_{\text{VC}}$ is optimised by

$$\begin{aligned}&\arg\max_{q_\phi(z|x)} \int_x \sum_y p(x,y) \int_z q_\phi(z|x) \log p_\theta(y|z)\\ =&\arg\max_{q_\phi(z|x)} \int_x p(x) \int_z q_\phi(z|x) \sum_y p(y|x) \log p_\theta(y|z)\end{aligned}$$

The integral over z is a $q_\phi(z|x)$-weighted sum of $\sum_y p(y|x) \log p_\theta(y|z)$ terms. Since $q_\phi(z|x)$ is a probability distribution, the integral is upper bounded by $\max_z \sum_y p(y|x) \log p_\theta(y|z)$. This maximum is attained *iff* support of $q_\phi(z|x)$ is restricted to $z^* = \arg\max_z \sum_y p(y|x) \log p_\theta(y|z)$ (which may not be unique). $\square$

A.2 Optimising the VC objective w.r.t. q

Setting $\beta = 1$ in Equation 6 to simplify and adding a lagrangian term to constrain $q_\phi(z|x)$ to a probability distribution, we aim to find

$$\begin{aligned}\arg\max_{q_\phi(z|x)} \int_x \sum_y p(x,y) \Big\{ &\int_z q_\phi(z|x) \log p_\theta(y|z)\\ -&\int_z q_\phi(z|y) \log \tfrac{q_\phi(z|y)}{p_\theta(z|y)} \; + \; \log p_\pi(y) \Big\} + \lambda(1 - \int_z q_\phi(z|x)) \,.\end{aligned}$$

Recalling that $q_\phi(z|y) = \int_x q_\phi(z|x) p(x|y)$ and using calculus of variations, we set the derivative of this functional w.r.t. $q_\phi(z|x)$ to zero

$$\sum_y p(x,y) \Big\{ \log p_\theta(y|z) - (\log \tfrac{q_\phi(z|y)}{p_\theta(z|y)} + 1) \Big\} - \lambda = 0$$

Rearranging and diving through by $p(x)$ gives

$$\mathbb{E}_{p(y|x)}[\log q_\phi(z|y)] = \mathbb{E}_{p(y|x)}[\log p_\theta(y|z) p_\theta(z|y)] + c \,,$$

where $c = -(1 + \frac{\lambda}{p(x)})$. Further, if each label y occurs once with each x, due to sampling or otherwise, then this simplifies to

$$q_\phi(z|y^*) e^c = p_\theta(y^*|z) p_\theta(z|y^*) \,,$$

which holds for all classes $y \in \mathcal{Y}$. Integrating over z shows $e^c = \int_z p_\theta(y|z) p_\theta(z|y)$ to give

$$q_\phi(z|y) = \tfrac{p_\theta(y|z) p_\theta(z|y)}{\int_z p_\theta(y|z) p_\theta(z|y)} = p_\theta(z|y) \tfrac{p_\theta(y|z)}{\mathbb{E}_{p_\theta(z|y)}[p_\theta(y|z)]} \,. \quad \square$$

We note, it is straightforward to include β to show

$$q_\phi(z|y) = p_\theta(z|y) \tfrac{p_\theta(y|z)^{1/\beta}}{\mathbb{E}_{p_\theta(z|y)}[p_\theta(y|z)^{1/\beta}]} \,.$$

B Justifying the Latent Prior in Variational Classification

Choosing Gaussian class priors in Variational classification can be interpreted in two ways:

Well-specified generative model: Assume data $x \in \mathcal{X}$ is generated from the hierarchical model: $\mathrm{y} \to \mathrm{z} \to \mathrm{x}$, where $p(\mathrm{y})$ is categorical; $p(\mathrm{z}|y)$ are analytically known distributions, e.g. $\mathcal{N}(z; \mu_y, \Sigma_y)$; the dimensionality of z is not large; and $x = h(z)$ for an arbitrary invertible function $h : \mathcal{Z} \to \mathcal{X}$ (if $\mathcal{X}$ is of higher dimension than $\mathcal{Z}$, assume h maps one-to-one to a manifold in $\mathcal{X}$). Accordingly, $p(\mathrm{x})$ is a mixture of unknown distributions. If $\{p_\theta(\mathrm{z}|y)\}_\theta$ includes the true distribution $p(\mathrm{z}|y)$, variational classification effectively aims to invert h and learn the parameters of the true generative model. In practice, the model parameters and h^{-1} may only be identifiable up to some equivalence, but by reflecting the true latent variables, the learned latent variables should be semantically meaningful.

Miss-specified model: Assume data is generated as above, but with z having a large, potentially uncountable, dimension with complex dependencies, e.g. details of every blade of grass or strand of hair in an image. In general, it is impossible to learn all such latent variables with a lower dimensional model. The latent variables of a VC might learn a complex function of multiple true latent variables.

The first scenario is ideal since the model might learn disentangled, semantically meaningful features of the data. However, it requires distributions to be well-specified and a low number of true latent variables. For natural data with many latent variables, the second case seems more plausible but choosing $p_\theta(\mathrm{z}|y)$ to be Gaussian may nevertheless be justifiable by the Central Limit Theorem.

C Variational Classification Algorithm

Algorithm 1 Variational Classification (VC)

Input $\quad p_\theta(\mathrm{z}|y)$, $q_\phi(\mathrm{z}|x)$, $p_\pi(\mathrm{y})$, $T_\psi(z)$; learning rate schedule $\{\eta^t_\theta, \eta^t_\phi, \eta^t_\pi, \eta^t_\psi\}_t$
Initialise $\ \theta, \phi, \pi, \psi;\ t \leftarrow 0$
while not converged **do**
$\quad \{x_i, y_i\}_{i=1}^m \sim \mathcal{D}$ [sample batch from data distribution $p(\mathrm{x}, \mathrm{y})$]
$\quad$ **for** z = {1 ... m} **do**
$\qquad z_i \sim q_\phi(\mathrm{z}|x_i),\ z'_i \sim p_\theta(\mathrm{z}|y_i)$ [e.g. $q_\phi(z|x_i) \doteq \delta_{z - f_\omega(x_i)}, \phi \doteq \omega \Rightarrow\ z_i = f_\omega(x_i)$]
$\qquad p_\theta(y_i|z_i) = \frac{p_\theta(z_i|y_i)p_\pi(y_i)}{\sum_y p_\theta(z_i|y)p_\pi(y)}$
$\quad$ **end for**
$\quad g_\theta \leftarrow \frac{1}{m}\sum_{i=1}^m \nabla_\theta \left[\log p_\theta(y_i|z_i) + p_\theta(z_i|y_i)\right]$
$\quad g_\phi \leftarrow \frac{1}{m}\sum_{i=1}^m \nabla_\phi \left[\log p_\theta(y_i|z_i) - T_\psi(z_i)\right]$ [e.g. using "reparameterisation trick"]
$\quad g_\pi \leftarrow \frac{1}{m}\sum_{i=1}^m \nabla_\pi \log p_\pi(y_i)$
$\quad g_\psi \leftarrow \frac{1}{m}\sum_{i=1}^m \nabla_\psi \left[\log \sigma(T_\psi(z_i)) + \log(1 - \sigma(T_\psi(z'_i)))\right]$
$\quad \theta \leftarrow \theta + \eta^t_\theta\, g_\theta, \quad \phi \leftarrow \phi + \eta^t_\phi\, g_\phi, \quad \pi \leftarrow \pi + \eta^t_\pi\, g_\pi, \quad \psi \leftarrow \psi + \eta^t_\psi\, g_\psi, \qquad t \leftarrow t + 1$
end while

D Calibration Metrics

One way to measure if a model is calibrated is to compute the expected difference between the confidence and expected accuracy of a model.

$$\mathbb{E}_{P(\hat{y}|x)}\Big[\mathbb{P}(\hat{y} = y | P(\hat{y}|x) = p) - p\Big] \tag{8}$$

This is known as expected calibration error (ECE) (Naeini et al., 2015). Practically, ECE is estimated by sorting the predictions by their confidence scores, partitioning the predictions in *M* equally spaced bins $(B_1 \dots B_M)$ and taking the weighted average of the difference between the average accuracy and average confidence of the bins. In our experiments we use 20 equally spaced bins.

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| acc(B_m) - conf(B_m) \right| \tag{9}$$

E Further Results

E.1 Distribution Shift (continued)

When deployed in the wild, *natural* distributional shifts may occur in the data due to subtle changes in the data generation process, e.g. a change of camera. We test resilience to *natural* distributional shifts on two tasks: Natural Language Inference (NLI) and detecting whether cells are cancerous from microscopic images. NLI requires verifying if a hypothesis logically follows from a premise. Models are trained on the SNLI dataset (Bowman et al., 2015) and tested on the MNLI dataset (Williams et al., 2018) taken from more diverse sources. Cancer detection uses the CAMELYON17 dataset (Bandi et al., 2018) from the WILDs datasets (Koh et al., 2021), where the `train` and `eval` sets contain images from different hospitals.

Table 2 shows that the VC model achieves better calibration under these natural distributional shifts (**H2**). The CAMELYON17 (CAM) dataset has a relatively small number (1000) of training samples (hence wide error bars are expected), which combines distribution shift with a low data setting (**H4**) and shows that the VC model achieves higher (average) accuracy in this more challenging real-world setting.

	Accuracy (↑)		Calibration (↓)	
	CE	VC	CE	VC
NLI	**71.2** ± 0.1	**71.2** ± 0.1	7.3 ± 0.2	**3.4** ± 0.2
CAM	79.2 ± 2.8	**84.5** ± 4.0	8.4 ± 2.5	**1.8** ± 1.3

Table 2: Accuracy and Calibration (ECE) under distributional shift (mean, std. err., 5 runs)

We also test the ability to **detect OOD examples**. We compute the AUROC when a model is trained on CIFAR-10 and evaluated on the CIFAR-10 validation set mixed (in turn) with SVHN, CIFAR-100, and CELEBA (Goodfellow et al., 2013; Liu et al., 2015). We compare the VC and CE models using the probability of the predicted class $\arg\max_y p_\theta(y|x)$ as a means of identifying OOD samples.

Table 3 shows that the VC model performs comparably to the CE model. We also consider $p(\mathrm{z})$ as a metric to detect OOD samples and achieve comparable results, which is broadly consistent with the findings of (Grathwohl et al., 2019). Although the VC model learns to map the data to a more structured latent space and, from the results above, makes more calibrated predictions for OOD data, it does not appear to be better able to distinguish OOD data than a standard softmax classifier (CE) using the metrics tested (we note that "OOD" is a loosely defined term).

Model	SVHN	C-100	CelebA
$P_{\text{CE}}(y\|x)$	0.92	0.88	0.90
$P_{\text{VC}}(y\|z)$	0.93	0.86	0.89

Table 3: AUROC for the OOD detection task. Models are trained on CIFAR-10 and evaluated on in and out-of-distribution samples.

E.2 Adversarial Robustness

We test model robustness by measuring performance on adversarially generated images using the common *Fast Gradient Sign Method* (FGSM) of adversarial attack (Goodfellow et al., 2014). Perturbations are generated as $P = \epsilon \times sign\,(\mathcal{L}(x,y))$, where $\mathcal{L}(x,y)$ is the model loss for data sample x and correct class y; and ϵ is the *magnitude* of the attack. We compare all models trained on MNIST and CIFAR-10 against FGSM attacks of different magnitudes.

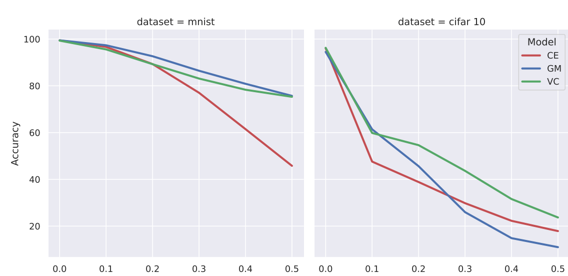

Figure 4: Prediction accuracy as FGSM adversarial attacks increase (l) MNIST; (r) CIFAR-10

Results in Figure 4 show that the VC model is consistently more adversarially robust relative to the standard CE model, across attack magnitudes on both datasets (**H3**).

E.3 Low Data Regime

In many real-world settings, datasets may have relatively few data samples and it may be prohibitive or impossible to acquire more, e.g. historic data or rare medical cases. We investigate model performance when data is scarce on the hypothesis that a prior over the latent space enables the model to better generalise from fewer samples. Models are trained on 500 samples from MNIST, 1000 samples from CIFAR-10 and 50 samples from AGNEWS.

	CE	GM	VC
MNIST	93.1 ± 0.2	**94.4** ± 0.1	**94.2** ± 0.2
CIFAR-10	52.7 ± 0.5	54.2 ± 0.6	**56.3** ± 0.6
AGNEWS	56.3 ± 5.3	61.5± 2.9	**66.3** ± 4.6

Table 4: Accuracy in low data regime (mean, std.err., 5 runs)

Results in Table 4 show that introducing the prior (GM) improves performance in a low data regime and that the additional entropy term in the VC model maintains or further improves accuracy (**H4**), particularly on the more complex datasets.

We further probe the relative benefit of the VC model over the CE baseline as the training sample size varies (**H4**) on MedMNIST, a collection of real-world medical datasets of varying sizes.

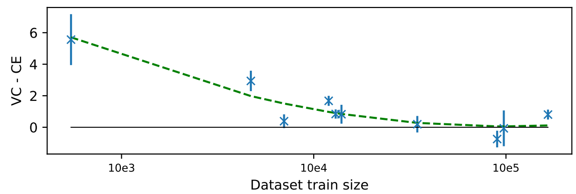

Figure 5: Accuracy increase of VC over CE on MedMNIST datasets of varying training set size (mean, std.err., 3 runs)

Figure 5 shows the increase in classification accuracy for the VC model relative to the CE model against number of training samples (log scale). The results show a clear trend that the benefit of the additional latent structure imposed in the VC model increases exponentially as the number of training samples decreases. Together with the results in Table 4, this suggests that the VC model offers most significant benefit for small, complex datasets.

E.4 Classification under Domain Shift

A comparison of accuracy between the VC and CE models under 16 different synthetic domain shifts. We find that VC performs comparably well as CE.

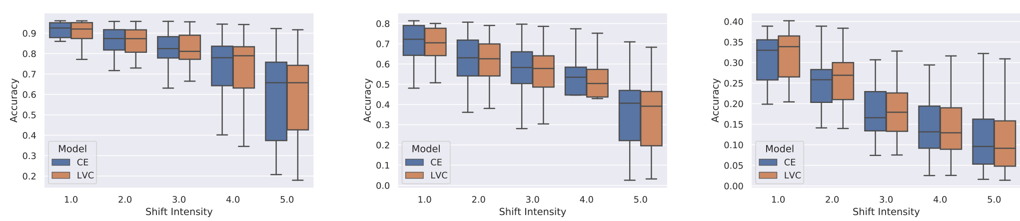

Figure 6: Classification accuracy under distributional shift: *(left)* CIFAR-10-C *(middle)* CIFAR-100-C *(right)* TINY-IMAGENET-C

E.5 OOD Detection

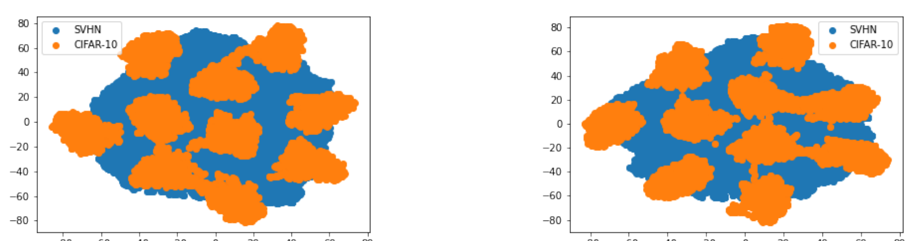

Figure 7: t-SNE plots of the feature space for a classifier trained on CIFAR-10. *(l)* Trained using CE. *(r)* Trained using VC. We posit that similar to CE, VC model is unable to meaningfully represent data from an entirely different distribution.

F Semantics of the latent space

To try to understand the semantics captured in the latent space, we use a pre-trained MNIST model on the *Ambiguous* MNIST dataset (Mukhoti et al., 2021). We interpolate between ambiguous 7's that are mapped close to the Gaussian clusters of classes of "1" and "2". It can be observed that traversing from the mean of the "7" Gaussian to that on the "1" class, the ambiguous 7's begin to look more like "1"s.

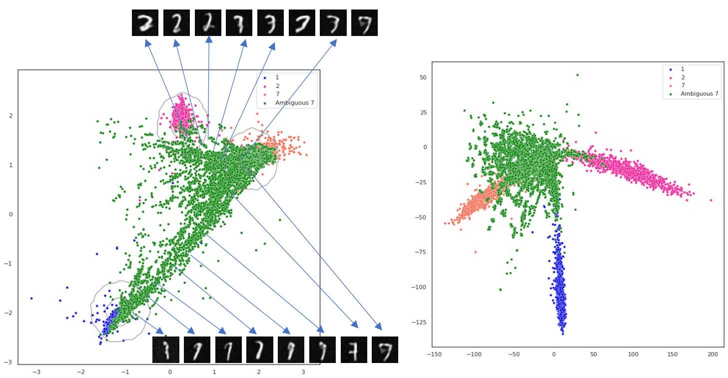

Figure 8: Interpolating in the latent space: Ambiguous MNIST when mapped on the latent space. *(l)* VC, *(r)* CE