Variational Classification

Anonymous authors
Paper under double-blind review

Abstract

We present a latent variable generalisation of neural network softmax classification trained with cross-entropy loss, referred to as *variational classification* (VC). Our approach offers a novel probabilistic perspective on the highly familiar softmax classification model, to which it relates similarly to how variational and traditional autoencoders relate. We derive a training objective based on the evidence lower bound (ELBO) that is non-trivial to optimize, and therefore propose an adversarial approach to maximise it. We show that VC addresses an inherent inconsistency within softmax classification, whilst also allowing more flexible choices of prior distributions in the latent space in place of implicit assumptions revealed within off-the-shelf softmax classifiers. Empirical evaluation on image and text classification datasets demonstrates that variational classification maintains prediction accuracy while improving other desirable properties such as calibration and adversarial robustness, particularly under distribution shift and low data settings.

1 Introduction

Classification is a core task in machine learning, from categorising objects (Klasson et al., 2019) and providing medical diagnoses (Adem et al., 2019; Mirbabaie et al., 2021), to identifying potentially life-supporting planets (Tiensuu et al., 2019). Classification also arises within other learning paradigms, such as to select actions in reinforcement learning or distinguish positive and negative samples in contrastive learning, and relates to the *attention* mechanism in transformer models. Classification tasks are commonly tackled by training domain-specific neural networks with a *sigmoid* or *softmax* output layer.[1] Data samples x (in a domain $\mathcal{X}$) are mapped deterministically by a network f_ω (with weights ω) to a real vector $z = f_\omega(x)$, which is transformed in the softmax layer to a point on the simplex $\Delta^{|\mathcal{Y}|}$, that parameterises $p_\theta(y|x)$, a discrete distribution over class labels $y \in \mathcal{Y}$:

$$p_\theta(y|x) = \frac{\exp\{z^\top w_y + b_y\}}{\sum_{y' \in \mathcal{Y}} \exp\{z^\top w_{y'} + b_{y'}\}} \ . \tag{1}$$

Although softmax classifiers often outperform alternatives, how they achieve that is not well understood theoretically and they are not without issue. For example: (i) the full mapping from $\mathcal{X}$ to $\Delta^{|\mathcal{Y}|}$ is learned by numerically minimising a loss function over finite training samples and is, in many respects, a "black box", giving predictions $p_\theta(y|x)$ that are *hard to explain*; (ii) highly flexible neural networks, or function classes $\{f_\omega\}_{\omega \in \Omega}$, are typically used, so that $p_\theta(y|x)$ can approximate the true class distributions $p(y|x)$, which may, for various settings of ω, give similar accuracy over training samples but different yet *confident* predictions elsewhere, which are thus *uncertain*; (iii) predictions can vary significantly and erroneously for imperceptible changes in the data, as shown by *adversarial examples*; and (iv) predictions may correctly identify the most likely class by their mode, but poorly reflect the full class distribution $p(y|x)$, known as *miscalibration.*

To better understand softmax classification and potentially mitigate some of its shortcomings, we take a latent variable perspective, treating z of Equation 1 as the realisation of a latent variable z in a generative (Markov) model $\mathrm{y} \rightarrow \mathrm{z} \rightarrow \mathrm{x}$. In a typical image classification case, this generative model might be interpreted as first selecting a sample's class (i.e. what it *is*), then choosing its descriptive parameters (e.g. size, colour),

---

[1] Since the softmax function generalises sigmoid to multiple classes, we refer to softmax throughout, but arguments also apply to the sigmoid case.

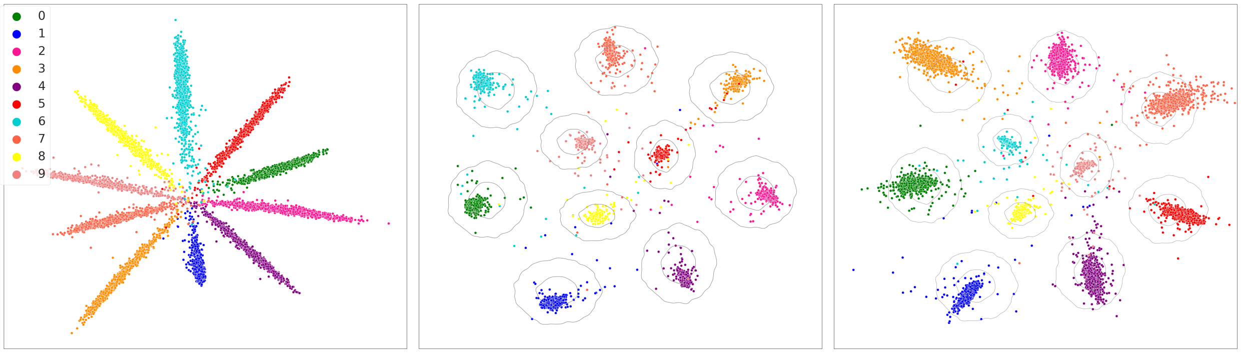

Figure 1: Empirical distributions of inputs to the output layer $q_\phi(z|y)$ for classifiers trained under incremental components of the VC objective (Eq. 9) on MNIST (*cf* the central $\mathcal{Z}$-plane in figure 2). (l) “MLE” objective = softmax cross-entropy; (c) “MAP” objective = MLE + Gaussian class priors $p_\theta(\mathrm{z}|y)$ (in contour); (r) VC objective = MAP + entropy of $p_\theta(\mathrm{z}|y)$. Colour indicates class y; $\mathcal{Z}=\mathbb{R}^2$ for visualisation purposes.

from which x is determined subject to stochasticity (e.g. noise, background variation). Under this model, labels can be inferred by reversing the generative process, first predicting z from x, then y from z, and integrating over z: $p_{\theta,\phi}(y|x)=\int_z p_\theta(y|z)q_\phi(z|x)$.[2] In general, it is intractable to learn the parameters of this model by maximising the conditional log likelihood, $\int p(x,y)\log p_{\theta,\phi}(y|x)$. Instead, a lower bound can be maximised, comparable to the evidence lower bound (ELBO), as used to train a variational auto-encoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014).

We first show that the standard softmax cross-entropy (SCE) objective *is* such a lower bound for specific choices of $q_\phi(\mathrm{z}|x)$ and $p_\theta(\mathrm{y}|z)$, and therefore the considered latent variable model generalises softmax classification. For this correspondence, $q_\phi(z|x)$ is parameterised by a neural network, f_ω (up to, but excluding, the softmax layer, analogous to a VAE *encoder*); and $p_\theta(y|z)=\frac{p_\theta(z|y)p_\theta(y)}{\sum_{y'}p_\theta(z|y')p_\theta(y')}$ is defined by Bayes' rule with exponential family class-conditional distributions $p_\theta(\mathrm{z}|y)$, which a softmax layer can be seen to encode.

We then show that *two* versions of the class-conditional latent distributions $p(\mathrm{z}|y)$ can be described: (i) $p_\theta(\mathrm{z}|y)$ encoded in the softmax layer (as above); and (ii) $q_\phi(\mathrm{z}|y)=\int_x q_\phi(\mathrm{z}|x)p(x|y)$, sampled as $z\sim q_\phi(\mathrm{z}|x)$, parameterised by the network f_ω, given class instances $x\sim p(\mathrm{x}|y)$. These distributions can be interpreted, respectively, as that *anticipated*, or required, by the output layer for accurate label distributions to be output; and that *empirically* received. Notably, the SCE objective does not encourage consistency between them.

We show theoretically and empirically that in practical settings the anticipated and empirical latent distributions, $p_\theta(\mathrm{z}|y)$ and $q_\phi(\mathrm{z}|y)$, can differ materially. In particular, the SCE objective can be optimal if empirical class-conditional latent distributions “*collapse*” to distinct points rather than *fit* the anticipated distributions. Such collapse equates to a loss of information required to estimate the confidence of a prediction and possibly other downstream tasks, reducing the use of z as a *representation* of x. In particular, **a softmax classifier may give inaccurate label distributions**, i.e. be *miscalibrated*, despite giving *accurate* class predictions (the latter require only *modes* of predicted distributions $p_\theta(\mathrm{y}|x)$ to match those of the ground truth $p(\mathrm{y}|x)$).

To address this, we introduce a regularisation term to align empirical latent distributions $q_\phi(\mathrm{z}|y)$ to the anticipated $p_\theta(\mathrm{z}|y)$. Specifically, we minimise a Kullback-Leibler (KL) divergence, which is non-trivial since $q_\phi(\mathrm{z}|y)$ can only be sampled not evaluated. Similarly to previous works (Gutmann & Hyvärinen, 2010; Makhzani et al., 2015; Mescheder et al., 2017), we take an *adversarial* approach to implicitly learn the required log probability ratios as an auxiliary task. The resulting *Variational Classification* (**VC**) objective generalises typical softmax cross-entropy classification from a latent perspective and fits empirical latent distributions $q_\phi(\mathrm{z}|y)$ to chosen *class priors* $p_\theta(\mathrm{z}|y)$. Considered within this more general VC framework, latent variables learned by a typical softmax classifier can be viewed as *maximium likelihood* point estimates that maximise $p_\theta(y|z)$. The two KL components introduced in VC can be interpreted sequentially as giving

---

[2] We use the notation q_ϕ to distinguish distributions, as will be made clear.

(i) *maximum a posteriori* point estimates of z; and (ii) a *Bayesian* approach where $q_\phi(\mathrm{z}|y)$ approximates the full class prior $p_\theta(\mathrm{z}|y)$ (see Figure 1). (We note that terms of the ELBO can be interpreted similarly.)

Since VC reduces the difference between empirical and anticipated latent distributions, which is expected to diminish as the sample size increases, VC is expected to offer greatest benefit over softmax classification in lower data regimes or where accurate estimates of the full label distribution $p(\mathrm{y}|x)$ are required. Through a series of experiments on vision and text datasets, we demonstrate that VC achieves comparable accuracy to regular softmax classification while the imposed latent structure improves calibration, robustness to adversarial perturbations, performance in low data regimes and generalisation under domain shift. We note that a number of proposed methods target any *one* of these aspects of softmax classification, often requiring additional hyperparameters to be tuned on held-out validation sets, whereas *VC simultaneously addresses them all* without tailoring to any one or requiring issue-specific hyperparameters or validation sets.

The VC framework also gives clearer mathematical insight into softmax classification, interpreting the neural network component as mapping a mixture of unknown distributions in the data space to a mixture of chosen class priors in the latent space, which the softmax (or more generally *output*) layer "flips" by Bayes' rule. This may enable principled improvement of classification and integration with other latent variable paradigms, such as VAEs and self-supervised learning.

2 Background (Variational Auto-Encoder)

The proposed generalisation of softmax to variational classification (section 3) is analogous to how a deterministic auto-encoder relates to a *variational auto-encoder* (VAE). We briefly summarise the VAE.

Estimating parameters of a latent variable model $p_\theta(x) = \int_z p_\theta(x|\mathrm{z})p_\theta(\mathrm{z})$ by maximising the data likelihood $\int_x p(x)\log p_\theta(x)$ is intractable in general. Instead, one can maximise the *evidence lower bound* (ELBO):

$$\int_x p(x)\log p_\theta(x) = \int_x p(x)\int_z q_\phi(z|x)\Big\{\log p_\theta(x|z) - \log\tfrac{q_\phi(z|x)}{p_\theta(z)} + \log\tfrac{q_\phi(z|x)}{p_\theta(z|x)}\Big\}, \tag{2}$$

$$\geq \int_x p(x)\int_z q_\phi(z|x)\Big\{\log p_\theta(x|z) - \log\tfrac{q_\phi(z|x)}{p_\theta(z)}\Big\} \;\doteq\; \mathbf{ELBO}, \tag{3}$$

where $q_\phi(z|x)$ is termed the *approximate posterior* and the term dropped to give the inequality is of the form $\int_z q(z)\log\frac{q(z)}{p(z)} \doteq D_{\mathrm{KL}}[\,q(z)\|\,p(z)] \geq 0$, a Kullback-Leibler (KL) divergence. By rearranging terms, it can be seen that maximising the ELBO is equivalent to minimising

$$D_{\mathrm{KL}}[\,p(x)\|\,p_\theta(x)] \;+\; \mathbb{E}_x\big[D_{\mathrm{KL}}[\,q_\phi(z|x)\|\,p_\theta(z|x)]\big], \tag{4}$$

and so fits the model $p_\theta(x)$ to the data distribution $p(x)$ while also fitting the approximate posterior $q_\phi(z|x)$ to the (implicit) model posterior $p_\theta(z|x) = \frac{p_\theta(x|z)p_\theta(z)}{p_\theta(x)}$. An alternate interpretation is that the two explicitly modelled distributions, $q_\phi(z|x)$ and $p_\theta(x|z)$, become *consistent under Bayes' rule.*

The variational auto-encoder (Kingma & Welling, 2014; Rezende et al., 2014) is an implementation of the ELBO in which all distributions are assumed Gaussian, with $p_\theta(x|z)$, $q_\phi(z|x)$ parameterised by neural networks. As its variance tends to zero, $q_\phi(\mathrm{z}|x)$ tends to a delta distribution and the first ("reconstruction") term of Equation 3 tends to the loss function of a deterministic auto-encoder. Hence, the VAE can be seen to probabilistically generalise a deterministic auto-encoder allowing for uncertainty or stochasticity in the latent $\mathrm{z}|x$, whose entropy is promoted whilst also being constrained to a prior in the second ("regularisation") term.

3 Variational Classification

3.1 A Latent Variable Model for Classification

Data $x \in \mathcal{X}$ and labels $y \in \mathcal{Y}$ are treated as samples of random variables x, y jointly distributed by $p(\mathrm{x}, \mathrm{y})$. A softmax classifier is a deterministic function that maps each data point x, via a sequence of intermediate representations, to a point on the simplex $\Delta^{|\mathcal{Y}|}$ that parameterises a categorical label distribution $p_\theta(\mathrm{y}|x)$.

Any intermediate representation $z = g(x)$ can be considered a realisation of a *latent* random variable z sampled from a conditional (delta) distribution $z \sim p(\mathrm{z}|x) = \delta_{z-g(x)}$. More generally, under a (Markov) generative latent variable model $\mathrm{y} \to \mathrm{z} \to \mathrm{x}$:

$$p(x) = \int_{y,z} p(x|z)p(z|y)p(y) \ , \tag{5}$$

class labels can be predicted in reverse (figure 2):

$$p_\theta(y|x) = \int_{z} p_\theta(y|z)p_\theta(z|x) \ . \tag{6}$$

A softmax classifier is a special case of Equation 6 where: (i) f_ω, the neural network up to the softmax layer, parameterises $p_\theta(z|x) = \delta_{z-f_\omega(x)}$, a delta distribution; (ii) the softmax layer input is considered a sample from $p_\theta(z|x)$; and (iii) $p_\theta(\mathrm{y}|z)$ is defined by the softmax layer (see RHS of Equation 1).

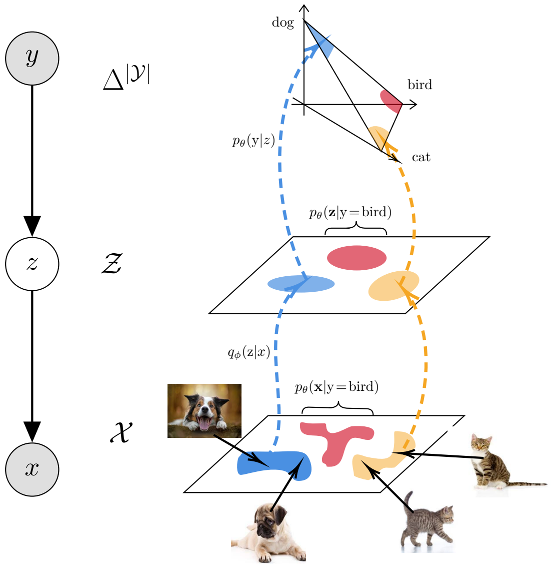

Figure 2: Variational Classification, reversing the generative process: $q_\phi(\mathrm{z}|x)$ stochastically maps data $x \in \mathcal{X}$ to the latent space $\mathcal{Z}$, where *empirical* latent distributions $q_\phi(\mathrm{z}|y) \doteq \int q_\phi(\mathrm{z}|y)p(x|y)$ are fitted to *class priors* $p_\theta(\mathrm{z}|y)$; the output layer computes $p_\theta(y|z)$ by Bayes' rule to give the class prediction $p(y|x)$.

3.1.1 Training a Classification LVM

Similarly to the latent variable model for $p_\theta(\mathrm{x})$ (section 2), parameters of equation 6 cannot generally be learned by directly maximising the likelihood, but rather a lower bound on it (*cf* equation 3):

$$\begin{aligned} \int_{x,y} p(x,y) \log p_\theta(y|x) \ &= \int_{x,y} p(x,y) \int_z q_\phi(z|x) \Big\{ \log p_\theta(y|z,\cancel{x}) - \cancel{\log \tfrac{q_\phi(z|x)}{p_\theta(z|x)}} + \log \tfrac{q_\phi(z|x)}{p_\theta(z|x,y)} \Big\} \\ &\geq \int_{x,y} p(x,y) \int_z q_\phi(z|x) \log p_\theta(y|z) \ \doteq \ \mathbf{ELBO_{VC}} \end{aligned} \tag{7}$$

Here, $p_\theta(y|z,x) = p_\theta(y|z)$ by the Markov model, and the (freely chosen) variational posterior q_ϕ is assumed to depend only on x and, furthermore, equal to $p_\theta(z|x)$ (eliminating the second term).[3] This derivation is analogous to equation 2 conditioned on x; a more direct derivation also follows from Jensen's inequality.

Unlike the derivation of the standard ELBO, the "dropped" KL term $D_{\mathrm{KL}}[\,q_\phi(z|x) \| \, p_\theta(z|x,y)]$ (minimised implicitly as $\mathrm{ELBO_{VC}}$ is maximised) may not minimise to zero – except in the limiting case $p_\theta(y|x,z) = p_\theta(y|z)$, i.e. when z is a *sufficient statistic* for y given x, intuitively meaning that z contains all information in x about y.[4] Hence, **maximising $\mathbf{ELBO_{VC}}$ implicitly encourages $\boldsymbol{z}$ to learn a sufficient statistic for $\boldsymbol{y|x}$.**

3.2 $\mathrm{ELBO_{VC}}$ Generalises Softmax Cross Entropy

Here, we show that softmax cross-entropy (SCE) loss is a special case of $\mathrm{ELBO_{VC}}$. Set (i) $q_\phi(z|x) = \delta_{z-f_\omega(x)}$ and (ii) $p_\theta(z|y) \propto \exp\{z^\top w_y + b'_y\}, \forall y \in \mathcal{Y}$ in equation 7, for constants w_y, b'_y, and let $b_y = b'_y - \log p_\theta(y)$:

$$\begin{aligned} \int_{x,y} p(x,y) \int_z q_\phi(z|x) \log p_\theta(y|z) \overset{(i)}{=} \int_{x,y} p(x,y) \log p_\theta(y|z = f_\omega(x)) \overset{(Bayes)}{=} \int_{x,y} p(x,y) \log \tfrac{p_\theta(z=f_\omega(x)|y)p_\theta(y)}{\sum_{y'} p_\theta(z=f_\omega(x)|y')p_\theta(y')} \\ \overset{(ii)}{=} \int_{x,y} p(x,y) \log \tfrac{\exp\{f_\omega(x)^\top w_y + b_y\}}{\sum_{y'} \exp\{f_\omega(x)^\top w_{y'} + b_{y'}\}} \doteq \mathbf{SCE}. \end{aligned} \tag{8}$$

This shows that for softmax classification to output correct label distributions $p(y|x)$, latents z are **anticipated to follow *(equi-scale/variate) exponential family* class-conditional distributions**, $p_\theta(z|y)$.

---

[3] We use the notation "q_ϕ" by analogy to the VAE and to later distinguish $q_\phi(z|y)$, derived from $q_\phi(z|x)$, from $p_\theta(z|y)$.

[4] Simple proof: from $p(z|x,y)p(y|x) = p(y|x,z)p(z|x)$ and Markovianity, we see that $D_{\mathrm{KL}}[\,q_\phi(z|x) \| \, p_\theta(z|x,y)] = 0 \Leftrightarrow p_\theta(z|x,y) = q_\phi(z|x) \Leftrightarrow p_\theta(y|x) = p_\theta(y|x,z) = p_\theta(y|z) \Leftrightarrow$ z a sufficient statistic for $\mathrm{y}|x$.

3.3 Two versions of Class-conditional Latent Distributions

Training a classifier under the $\mathrm{ELBO}_{\mathrm{VC}}$ (equation 7) requires specifying $p_\theta(y|z)$, e.g. from class-conditionals $p_\theta(z|y)$ using Bayes' rule. SCE loss has been seen to be a special case that implicitly assumes $p_\theta(z|y)$ to be of (equi-scale) exponential family form (section 3.2). More generally, the choice of $p_\theta(z|y)$, encoded in the output layer (a generalised term for the *softmax* layer), defines the *anticipated* form that class-conditional latent distributions must follow in order for accurate label predictions $p(y|x)$ to be output. A natural question then is: *do latent variables for each class* empirically *follow the* anticipated *distributions* $p_\theta(z|y)$*?*

Empirical latent distributions are not fixed, but rather described by $q_\phi(z|y) \doteq \int_x q_\phi(z|x)p(x|y)$, i.e. by sampling $q_\phi(z|x)$ (parameterised by the neural network f_ω), given class samples $x \sim p(\mathrm{x}|y)$. Since $\mathrm{ELBO}_{\mathrm{VC}}$ is optimised w.r.t. parameters ϕ, denoting optimal parameters ϕ^*, the question becomes: *does* $q_{\phi^*}(z|y) = p_\theta(z|y)$*?*

It can be seen that $\mathrm{ELBO}_{\mathrm{VC}}$ is optimised w.r.t ϕ if $q_{\phi^*}(z|x) = \delta_{z-z_x}$, for $z_x = \arg\max_z \mathbb{E}_{y|x}[\log p_\theta(y|z)]$ (see appendix A.1).[5] In practice, *true* label distributions $p(y|x)$ are unknown and we have only finite samples from them. For a continuous data domain $\mathcal{X}$, e.g. images or sounds, any empirically observed x is *re*-sampled with probability zero and so *is observed once with a single label* $y(x)$. A similar situation occurs for *any* $\mathcal{X}$ where each x has only one ground truth label $y(x)$ *as a property of the data*, i.e. labels are mutually exclusive and *partition* the data (as in popular image datasets, e.g. MNIST, CIFAR, ImageNet, where *samples belong to one class or another*). In either case, the expectation over labels falls away and, for a given class y, $z_x = \arg\max_z p_\theta(y|z)$ is identical for all samples x labelled y, subject to uniqueness of $\arg\max_z p_\theta(y|z)$. This uniqueness is not guaranteed in general, but is assumed for suitable choices of $p_\theta(z|y)$, such as in the softmax case of central interest.[6] Letting z_y denote the optimal latent variable for all x of class y, optimal class-level distributions are simply $q_{\phi^*}(z|y) = \delta_{z-z_y}$, and **$\mathrm{ELBO}_{\mathrm{VC}}$ is maximised if all latent representations of a class, and hence $q_\phi(z|y)$, "collapse" to a point**, irrespective of the variance of $p_\theta(z|y)$.

Since softmax classification is a special case, this reveals the potential, depending on the data distribution and model flexibility, for softmax classifiers to learn over-concentrated, or *over-confident*, latent distributions relative to anticipated distributions. In practical terms, the softmax cross entropy loss may be minimised when all samples of a given class are mapped (by f_ω) to the same latent variable/representation, regardless of differences in the samples' probabilities or semantics, thus disregarding information that may, for example, be useful for calibration or downstream tasks. We note that this "loss of information" accords with the Information Bottleneck Theory (Tishby et al., 2000; Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017), but as we see below, losing such information is not *necessary* for classification and may be undesirable.

3.3.1 Aligning the Anticipated and Empirical Latent Distributions

We have shown that the $\mathrm{ELBO}_{\mathrm{VC}}$ objective, a generalisation of SCE loss, effectively involves two versions of the latent class conditional distributions, $p_\theta(y|z)$ and $q_\phi(z|y)$, and that a mismatch between them may have undesirable consequences in terms of information loss. We therefore propose to align $p_\theta(z|y)$ and $q_\phi(z|y)$, or, equivalently, for $p_\theta(y|z)$ and $q_\phi(z|y)$ to be made *consistent under Bayes' rule* (analogous to $p_\theta(x|z)$ and $q_\phi(z|x)$ in the ELBO, section 2). Specifically, we minimise $D_{\mathrm{KL}}[\,q_\phi(\mathrm{z}|y)\|\,p_\theta(\mathrm{z}|y)]$, $\forall y \in \mathcal{Y}$. Including this constraint (weighted by $\beta > 0$) and learning required class distribution $p_\pi(\mathrm{y})$ gives the full **VC objective**:

$$\max_{\theta,\phi,\pi} \int_{x,y} p(x,y) \Big\{ \int_z q_\phi(z|x) \log \frac{p_\theta(z|y)p_\pi(y)}{\sum_{y'} p_\theta(z|y')p_\pi(y')} - \beta \int_z q_\phi(z|y) \log \frac{q_\phi(z|y)}{p_\theta(z|y)} + \log p_\pi(y) \Big\} \,. \tag{9}$$

Considered incrementally, q_ϕ–terms of the VC objective can be interpreted as treating latent the variable z from a *maximum likelihood* (MLE), *maximum a posteriori* (MAP) and *Bayesian* perspective:

(i) maximising $\int_z q_\phi(z|x) \log p_\theta(y|z)$ may overfit $q_\phi(z|y)$ to δ_{z-z_y} for finite samples (as above); [MLE]

(ii) adding *class priors* $\int_z q_\phi(z|y) \log p_\theta(z|y)$ constrains the MLE point estimates z_y [MAP]

(iii) adding *entropy* $= -\int_z q_\phi(z|y) \log q_\phi(z|y)$ encourages $q_\phi(\mathrm{z}|y)$ to "fill out" $p_\theta(\mathrm{z}|y)$. [Bayesian]

---

[5] We assume the parametric family q_ϕ is sufficiently flexible to approximate the analytic maximiser of $\mathrm{ELBO}_{\mathrm{VC}}$.

[6] If all x have a (single) label $y(x)$, $p(y|x) = \mathbf{1}_{y=y(x)}$, a "one-hot" vector. Letting $\|w_y\| = \|z\| = \alpha$, for some $\alpha > 0$, equation 1 best approximates $p(y|x)$ when w_y are evenly dispersed (i.e. unit vectors $\hat{w}_y$ form a regular polytope on the unit sphere) and $z = f_\omega(x) = w_{y(x)}$, i.e. $\arg\max_z p_\theta(y|z) = w_{y(x)}$ (unique). Without constraint, $\alpha \to \infty$ during optimisation.

Figure 1 shows samples from empirical latent distributions $q_\phi(\mathrm{z}|y)$ for classifiers trained under incremental terms of the VC objective. This confirms empirically that softmax cross entropy loss does not impose the anticipated latent distribution encoded in the output layer (*left*). Adding class priors $p_\theta(\mathrm{z}|y)$ observably induces latent structure (*centre*), and including entropy encourages class priors to be "filled out" (*right*). As noted above, if each x is observed with a single label (e.g. MNIST), the MLE/MAP training objectives are theoretically optimised when class distributions $q_\phi(\mathrm{z}|y)$ collapse to a point. Although latent values grow large, collapse is not observed in practice (Figure 1, *left*, *centre*). We conjecture that this is due to typical constraints on f_ω, in particular continuity and ℓ_2 regularisation, and possibly heuristics such as batch norm.

Compared to the KL form of the ELBO (equation 4), maximising equation 9 is equivalent to minimising:

$$\underline{\mathbb{E}_x\big[D_{\text{KL}}[\,p(y|x)\|\,p_\theta(y|x)]} + \mathbb{E}_{x,y}\big[D_{\text{KL}}[\,q_\phi(z|x)\|\,p_\theta(z|x,y)]\big] + \mathbb{E}_y\big[D_{\text{KL}}[\,q_\phi(z|y)\|\,p_\theta(z|y)]\big] + D_{\text{KL}}[\,p(y)\|\,p_\pi(y)]\big] \quad (10)$$

showing the extra constraints over the initial objective of modelling $p(y|x)$ by $p_\theta(y|x)$ (underlined). Variational Classification abstracts a typical neural network classifier, giving interpretability to its components:

- the neural network up to the last layer (f_ω) transforms a mixture of analytically unknown class-conditional data distributions $p(\mathrm{x}|y)$ to a mixture of analytically defined latent distributions $p_\theta(\mathrm{z}|y)$;
- assuming latent variables follow the anticipated class distributions $p_\theta(\mathrm{z}|\mathrm{y})$, the output layer applies Bayes' rule to give $p_\theta(\mathrm{y}|\mathrm{z})$ (see figure 2) and thus the class prediction $p(\mathrm{y}|x)$ (by equation 6).

3.4 Summary: Softmax Classification vs the full VC Objective

From section 3.2, we have seen that ELBO_{VC} extends softmax cross-entropy, treating the input to the softmax layer as a latent variable and identifying the anticipated class-conditionals $p_\theta(z|y)$ implicitly encoded within the softmax layer. The full VC objective then encourages the empirical latent distributions $q_\phi(z|y)$ to fit, rather than maximise, $p_\theta(z|y)$. Conversely, SCE loss is recovered from the VC objective by setting (i) $q_\phi(\mathrm{z}|x) = \delta_{z-f_\omega(x)}$; (ii) class priors $p(\mathrm{z}|y)$ to (equal-scale) exponential family distributions, e.g. equivariate Gaussians; and (iii) $\beta = 0$. This is highly analogous to how a deterministic auto-encoder relates to a VAE.

Understanding how softmax classification fits within it, **the VC framework elucidates assumptions that are made implicitly in softmax classification**. By generalising the softmax case, the VC objective allows such assumptions, e.g. the form of $p_\theta(\mathrm{z}|y)$, to be considered and revised on a task/data-specific basis.

3.5 Optimising the VC Objective

The VC objective (equation 9) is a lower bound that can be maximised by gradient methods, e.g. SGD:

- the first term can be calculated by sampling $q_\phi(\mathrm{z}|x)$ (using the "reparameterisation trick" as necessary (Kingma & Welling, 2014)) and computing $p_\theta(\mathrm{y}|\mathrm{z})$ by Bayes' rule;
- the third term is standard multinomial cross-entropy;
- the second term, however, is not readily computable since $q_\phi(\mathrm{z}|y)$ is implicit and cannot easily be evaluated, only sampled, as $z \sim q_\phi(\mathrm{z}|x)$ (parameterised by f_ω) for class samples $x \sim p(\mathrm{x}|y)$.

Fortunately, we require log ratios $\log\frac{q_\phi(z|y)}{p_\theta(z|y)}$ for each class y, which can be approximated by training a binary classifiers to distinguish between samples of $q_\phi(\mathrm{z}|y)$ and $p_\theta(\mathrm{z}|y)$. This *contrastive* "trick" has been employed elsewhere, underpinning learning methods such as Noise Contrastive Estimation (Gutmann & Hyvärinen, 2010) and contrastive self-supervised learning (e.g. Oord et al., 2018; Chen et al., 2020) and, comparably, to train variants of the VAE (Makhzani et al., 2015; Mescheder et al., 2017).

Specifically, we maximise the following *auxiliary objective* w.r.t. parameters ψ of a set of binary classifiers:

$$\int_y p(y)\big\{\int_z q_\phi(z|y)\log\sigma(T^y_\psi(z)) + \int_z p_\theta(z|y)\log(1-\sigma(T^y_\psi(z))\big\} \quad (11)$$

where σ is the logistic sigmoid function $\sigma(x) = (1+e^{-x})^{-1}$, $T^y_\psi(z) = w_y^\top z + b_y$ and $\psi = \{w_y, b_y\}_{y\in\mathcal{Y}}$.

Algorithm 1 Variational Classification (VC)

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

It is easy to show that Eq. 11 is optimised if $T^y_\psi(z) = \log \frac{q_\phi(z|y)}{p_\theta(z|y)}$, $\forall y \in \mathcal{Y}$. Hence, when all binary classifiers are trained, $T^y_\psi(z)$ approximates the log ratio for class y required in the VC objective (Eq. 9). Optimising the VC objective might, in principle, also require gradients of these approximated log ratios w.r.t parameters θ and ϕ. However, the gradient w.r.t. the ϕ within the log ratio is always zero (Mescheder et al., 2017), and the gradient w.r.t. θ can be computed from Eq. 9. See Algorithm 1 for a summary.

This approach is *adversarial* since (a) the VC objective is maximised when log ratios give a *minimal* KL divergence (0), i.e. $q_\phi(z|y) = p_\theta(z|y)$ and latents sampled $q_\phi(z|y)$ are indistinguishable from those of $p_\theta(z|y)$; whereas (b) the auxiliary objective is maximised if the ratio is *maximal* and the two distributions are optimally discriminated. Relating to a Generative Adversarial Network (GAN) (Goodfellow et al., 2014a), the network f_ω can be considered a *generator* and each binary classifier a *discriminator*. Unlike a GAN, VC requires a discriminator *per class* that each distinguish generated samples from a learned, not static, reference/noise distribution $p_\theta(\mathrm{z}|y)$. However, whereas a GAN discriminator distinguishes between complex distributions in the data domain, a VC discriminator compares a Gaussian to an approximate Gaussian in the lower dimensional latent domain, a far simpler task. The auxiliary objective can be parallelised across classes and adds marginal computational overhead per class.

3.6 Optimum of the VC Objective

In section 3.3, the empirical distribution $q_\phi(z|x)$ that opitimises the $\mathrm{ELBO}_{\mathrm{VC}}$ was considered, showing that it need not match anticipated $p_\theta(z|y)$. Here, we perform similar analysis to identify $q_\phi(z|x)$ that maximises the VC objective, which is expected to better fit the anticipated distribution.

Letting $\beta = 1$ to simplify (see appendix A.2 for general case), the VC objective is maximised w.r.t. $q_\phi(z|x)$ if:

$$\mathbb{E}_{p(y|x)}[\log q_\phi(z|y)] = \mathbb{E}_{p(y|x)}[\log p_\theta(y|z) p_\theta(z|y)] + c \ , \tag{12}$$

for a constant c. This is satisfied if, for each class y,

$$q_\phi(z|y) = p_\theta(z|y) \frac{p_\theta(y|z)}{\mathbb{E}_{p_\theta(z'|y)}[p_\theta(y|z')]} \ , \tag{13}$$

given a unique solution if each x has a single label y (see section 3.3; see appendix A.2 for proof). We see that each $q_\phi(z|y)$ fits $p_\theta(z|y)$ scaled by a ratio of $p_\theta(y|z)$ to its weighted average. Where $p_\theta(y|z)$ is *above average*, $q_\phi(z|y) > p_\theta(z|y)$, and vice versa. In simple terms, $q_\phi(z|y)$ reflects $p_\theta(z|y)$ but is "peakier" (fitting observation in Figure 1). We have thus shown empirically (Figure 1) and theoretically that the VC objective aligns the empirical and anticipated latent distributions. However, these distributions are not identical and we leave to future work the derivation of an objective that achieves both $p_\theta(y|x) = p(y|x)$ and $q_\phi(z|y) = p_\theta(z|y)$.

4 Related Work

Despite notable differences, the *energy-based* interpretation of softmax classification of Grathwohl et al. (2019) is perhaps most comparable to our own in taking an abstract view of softmax classification to improve aspects of it. However, their achieved benefits, e.g. to calibration and adversarial robustness, come at a significant cost to the main goal of classification accuracy. Further, the MCMC normalisation required reportedly slows and destabilises training, whereas we use tractable probability distributions and retain the order of complexity.

Several previous works adapt the standard ELBO, used to learn a model of $p(x)$, to a conditional analog for learning $p(y|x)$ (Tang & Salakhutdinov, 2013; Sohn et al., 2015). However, such works focus on generative scenarios rather than discriminative classification, e.g. x being a face image and $y|x$ being the same face in a different pose determined by latent z; or x being part of an image and $y|x$ its completion given latent content z. The *Gaussian stochastic neural network* (GSNN) model (Sohn et al., 2015) is closer to our own by conditioning $q(z|x, y)$ only on x, however neither model generalises softmax classification or considers class-level latent priors $q(z|y)$ as in variational classification.

Variational classification subsumes many works that add a regularisation term to a softmax cross-entropy loss function, which can be interpreted as a prior over latent variables in the "MAP" case (section 3.3.1). For example, several semi-supervised learning models can be interpreted as treating the softmax *outputs* as latent variables and using the latent prior to guide the predictions of unlabelled data (Allen et al., 2020). Closer to variational classification, several works can be interpreted as treating softmax *inputs* as latent variables with a regularisation term that encourages prior beliefs, such as *deterministic* label predictions (i.e. all probability mass on a single class), which can be encouraged by imposing a *large margin* between class-conditional latent distributions (Liu et al., 2016; Wen et al., 2016; Wan et al., 2018; 2022; Scott et al., 2021).

Variational classification also sits amongst works spanning several learning paradigms in which a Gaussian mixture prior is imposed in the latent space, e.g. for representation learning (Xie et al., 2016; Caron et al., 2018), in auto-encoders (Song et al., 2013; Ghosh et al., 2019) and in variational auto-encoders (Jiang et al., 2016; Yang et al., 2017; Prasad et al., 2020; Manduchi et al., 2021).

5 Empirical Validation

Our goal is to empirically demonstrate that the latent structure induced by the VC objective is beneficial relative to the standard softmax classifier. A variational classifier can be substituted wherever a softmax classifier is used, by making distributional choices appropriate for the data. In particular, variational classification does not set out to address any one drawback of a softmax classifier, rather it aims to better reverse the generative process and so capture the data distribution, providing multiple benefits. We illustrate the effectiveness of a VC through a variety of tasks on familiar datasets from the visual and text domains.

Specifically, we set out to validate the following hypotheses:

H1: The VC objective improves uncertainty estimation, leading to a more calibrated model.

H2: The VC objective increases model robustness to changes in the data distribution.

H3: The VC objective enhances resistance to adversarial perturbations.

H4: The VC objective aids learning from fewer samples.

For fair comparison, we make minimal changes to adapt a standard softmax classifier to a variational classifier. As described in section 3.4, we train with the VC objective (equation 9) under the following assumptions: $q_\phi(\mathrm{z}|x)$ is a delta distribution parameterised by a neural network $f_\omega : \mathcal{X} \to \mathcal{Z}$; class-conditional priors $p_\theta(\mathrm{z}|y)$ are multi-variate Gaussians with parameters learned from the data (we use diagonal covariance for simplicity). To provide an ablation across the components of the VC objective, we compare classifiers trained to maximise three objective functions (see section 3):

	CIFAR-10				CIFAR-100				TINY-IMAGENET		
	CE	GM$^\diamond$	VC	vMF*	CE	GM $^\diamond$	VC	vMF*	CE	GM$^\diamond$	VC
Acc. (%, ↑)											
WRN	96.2 ± 0.1	95.0 ± 0.2	96.3 ± 0.2	-	80.3 ± 0.1	79.8 ± 0.2	80.3 ± 0.1	-	-	-	-
RNET	93.7 ± 0.1	93.0 ± 0.1	93.2 ± 0.1	94.0 ± 0.1	73.2 ± 0.1	74.2 ± 0.1	73.4 ± 0.1	69.94 ± 0.2	59.7 ± 0.2	59.3 ± 0.1	59.3 ± 0.1
ECE (%, ↓)											
WRN	3.1 ± 0.2	3.5 ± 0.3	**2.1** ± 0.2	-	11.1 ± 0.7	19.6 ± 0.4	**4.8** ± 0.3	-	-	-	-
RNET	3.8 ± 0.3	4.1 ± 0.2	**3.2** ± 0.2	5.9 ± 0.2	8.7 ± 0.2	10.5 ± 0.2	**5.1** ± 0.2	7.9 ± 0.3	12.3 ± 0.4	8.75 ± 0.2	**7.4** ± 0.5

Table 1: Classification Accuracy and Expected Calibration Error (mean, std.dev. over 5 runs). Accuracy is comparable across all VC-based models and data sets; calibration notably improves.
$\star$ from (Scott et al., 2021), $\diamond$ our implementation of (Wan et al., 2018)

CE: equivalent to standard softmax cross-entropy under the above assumptions and corresponds to the MLE form of the VC objective (section 3.3.1, (i)).

$$J_{\mathbf{CE}} = \int_{x,y} p(x,y) \left(\int_z q_\phi(z|x) \log p_\theta(y|z) \; + \; \log p_\pi(y) \right)$$

GM: includes class priors and corresponds to the MAP form of the VC objective (section 3.3.1, (ii)). This is equivalent to Wan et al. (2018) with just the Gaussian Prior.

$$J_{\mathbf{GM}} = J_{\mathbf{CE}} + \int_{x,y} p(x,y) \int_z q_\phi(z|y) \log p_\theta(z|y)$$

VC: includes entropy of the empirical latent distributions and corresponds to the Bayesian form of the VC objective (section 3.3.1, (iii)).

$$J_{\mathbf{VC}} = J_{\mathbf{GM}} - \int_{x,y} p(x,y) \int_z q_\phi(z|y) \log q_\phi(z|y)$$

5.1 Accuracy and Calibration

We first compare the classification accuracy and calibration of each model on three standard benchmarks (CIFAR-10, CIFAR-100, and TINY-IMAGENET), across two standard ResNet model architectures (*WideResNet-28-10* (WRN) and *ResNet-50* (RNET)) (He et al., 2016; Zagoruyko & Komodakis, 2016). Calibration is evaluated in terms of the *Expected Calibration Error* (ECE) (see Appendix C).

Table 1 shows that the VC and GM models achieve comparable accuracy to softmax cross entropy (CE), but that the VC model is consistently, significantly more calibrated (**H1**). Unlike approaches such as Platt's scaling (Platt et al., 1999) and temperature scaling (Guo et al., 2017), no *post hoc* calibration is performed, no external calibration set is needed or calibration-specific hyperparameters tuned.

5.2 Generalization under distribution shift

When used in real-world settings, machine learning models may encounter *distribution shift* relative to the training data. It can be important to know when a model's output is reliable and can be trusted, requiring the model to be **calibrated on out-of-distribution (OOD) data** and *know when they do not know.* To test performance under distribution shift, we use the robustness benchmarks, CIFAR-10-C, CIFAR-100-C and TINY-IMAGENET-C, proposed by Hendrycks & Dietterich (2019), which *simulate* distribution shift by adding various *synthetic* corruptions of varying intensities to a dataset. We compare the CE model, with and without temperature scaling, to the VC model. Temperature scaling was performed as in Guo et al. (2017) with the temperature tuned on an in-distribution validation set.

Both models are found to perform comparably in terms of classification accuracy (Figure 8), according to previous results (section 5.1). However, Figure 3 shows that the VC model has a consistently lower calibration error as the corruption intensity increases (left to right) (**H2**). We note that the improvement in calibration between the CE and VC models increases as the complexity of the dataset increases.

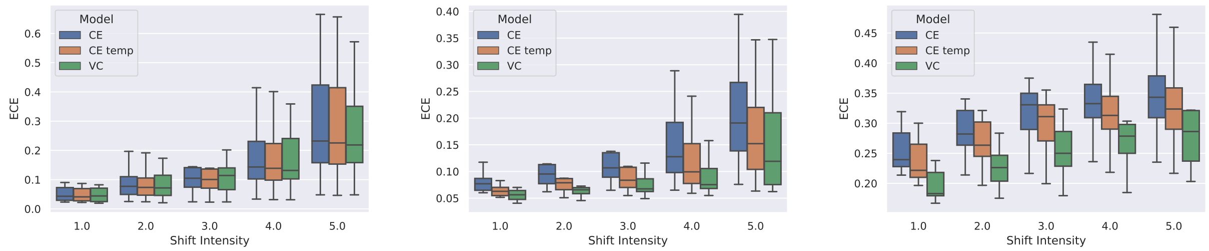

Figure 3: Calibration under distribution shift: boxes indicate quartiles for results across 16 shift types, whiskers indicate min/max across shift types. *(left)* CIFAR-10-C, *(middle)* CIFAR-100-C, *(right)* TINY-IMAGENET-C

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 $p(z)$ 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.

5.3 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., 2014b). 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.

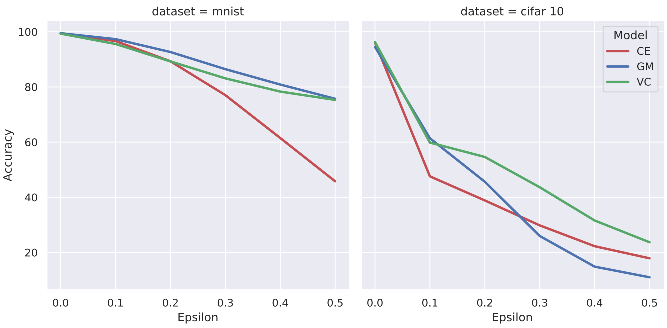

Figure 4: Prediction accuracy as FGSM adversarial attacks increase *(l)* MNIST; *(r)* CIFAR-10

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**).

5.4 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 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.

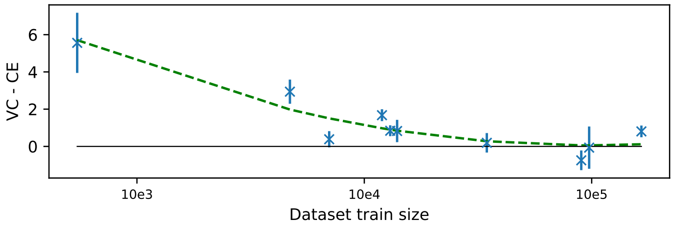

Figure 5: Accuracy increase of VC over CE on MedMNIST datasets of varying training set size (mean, std.err., 3 runs)

Together with the results in Table 4, this suggests that the VC model offers most significant benefit for small, complex datasets.

6 Conclusion

We have presented Variational Classification (VC), a latent generalisation of standard softmax classification trained under cross-entropy loss, mirroring the relationship between the variational auto-encoder and the deterministic auto-encoder (section 3). We show that softmax classification is a special case of VC under specific assumptions that are effectively taken for granted when using the softmax output layer. We present a training objective analogous to the ELBO, together with an adversarial optimisation regime. A series of experiments on image and text datasets show that, with marginal computational overhead and without tuning hyper-parameters other than for the original classification task, variational classification achieves comparable prediction accuracy to standard softmax classification while significantly improving performance in terms of calibration, adversarial robustness, under distribution shift or in a low data regime.

In terms of limitations, we intentionally focus on the *output* layer of a classifier, treating the function beneath, f_ω, as a "black-box". This leaves open the question of how the underlying network achieves its role of transforming a mixture of unknown data distributions to a mixture of specified latent distributions, or how that might be improved. We also prove that learned *empirical* latent distributions $q_\phi(z|y)$ are "peaky" approximations to the *anticipated* $p_\theta(z|y)$, leaving open the possibility of further improving the VC objective.

The VC framework gives new theoretical insight into the highly familiar softmax classifier, opening up several interesting future directions. For example, $q(z|x)$ might be modelled by a stochastic distribution, rather than a delta distribution, to reflect uncertainty in the latent variables, similarly to a VAE. VC may also be extended to semi-supervised learning and related to approaches that impose structure in the latent space.

References

Kemal Adem, Serhat Kiliçarslan, and Onur Cömert. Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. *Expert Systems with Applications*, 115:557–564, 2019.

Carl Allen, Ivana Balažević, and Timothy Hospedales. A probabilistic model for discriminative and neuro-symbolic semi-supervised learning. *arXiv preprint arXiv:2006.05896*, 2020.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. In *IEEE Transactions on Medical Imaging*, 2018.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

Partha Ghosh, Mehdi SM Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*, 2019.

Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*, 2013.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014a.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics*, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*, 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Marcus Klasson, Cheng Zhang, and Hedvig Kjellström. A hierarchical grocery store image dataset with visual and semantic labels. In *IEEE Winter Conference on Applications of Computer Vision*, 2019.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Learning*, 2016.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, 2015.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Laura Manduchi, Kieran Chin-Cheong, Holger Michel, Sven Wellmann, and Julia Vogt. Deep conditional gaussian mixture model for constrained clustering. In *Advances in Neural Information Processing Systems*, 2021.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, 2017.

Milad Mirbabaie, Stefan Stieglitz, and Nicholas RJ Frick. Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. *Health and Technology*, 11(4):693–731, 2021.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A simple baseline. *arXiv e-prints*, pp. arXiv–2102, 2021.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Vignesh Prasad, Dipanjan Das, and Brojeshwar Bhowmick. Variational clustering: Leveraging variational autoencoders for image clustering. In *International Joint Conference on Neural Networks*, 2020.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Tyler R Scott, Andrew C Gallagher, and Michael C Mozer. von mises-fisher loss: An exploration of embedding geometries for supervised learning. In *International Conference on Computer Vision*, 2021.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, 2015.

Chunfeng Song, Feng Liu, Yongzhen Huang, Liang Wang, and Tieniu Tan. Auto-encoder based data clustering. In *Iberoamerican Congress on Pattern Recognition*, 2013.

Charlie Tang and Russ R Salakhutdinov. Learning stochastic feedforward neural networks. In *Advances in Neural Information Processing Systems*, 2013.

Jacob Tiensuu, Maja Linderholm, Sofia Dreborg, and Fredrik Örn. Detecting exoplanets with machine learning: A comparative study between convolutional neural networks and support vector machines, 2019.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop*, 2015.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Weitao Wan, Yuanyi Zhong, Tianpeng Li, and Jiansheng Chen. Rethinking feature distribution for loss functions in image classification. In *Conference on Computer Vision and Pattern Recognition*, 2018.

Weitao Wan, Jiansheng Chen, Cheng Yu, Tong Wu, Yuanyi Zhong, and Ming-Hsuan Yang. Shaping deep feature space towards gaussian mixture for visual classification. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, 2016.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, 2016.

Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *International Conference on Machine Learning*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

A Proofs

A.1 Optimising the ELBO_{VC} w.r.t q

Rearranging Equation 7, the ELBO_{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 9 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

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

We note, it is straightforward to include β to show

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

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: $\mathrm{y} \rightarrow \mathrm{z} \rightarrow \mathrm{x}$, 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 $h : \mathcal{Z} \rightarrow \mathcal{X}$ (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 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.

$$\mathbb{E}_{P(\hat{y}|x)}\Big[\mathbb{P}(\hat{y} = y | P(\hat{y}|x) = p) - p\Big] \tag{14}$$

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.

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| acc(B_m) - conf(B_m) \right| \tag{15}$$

D OOD Detection

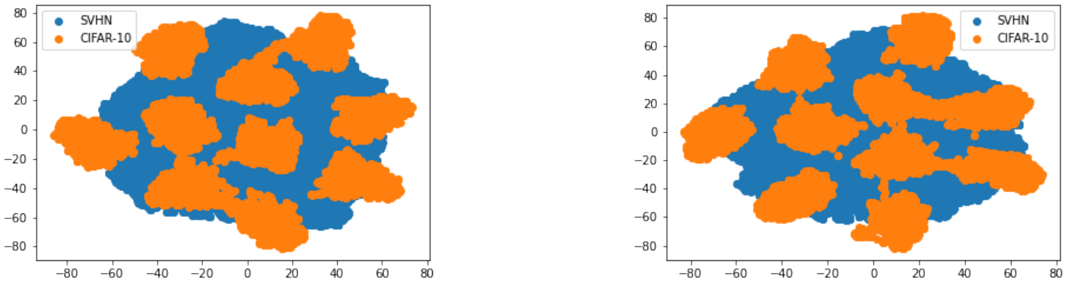

Figure 6: t-SNE plots of the feature space for a classifier trained on CIFAR-10. *(l)* Trained using CE. *(r)* Trained using VC. We posit that similar to CE, VC model is unable to meaningfully represent data from an entirely different distribution.

E 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.

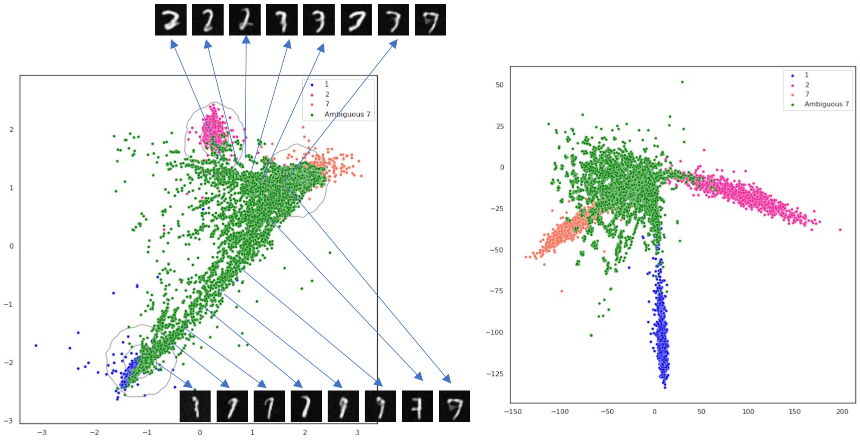

Figure 7: Interpolating in the latent space: Ambiguous MNIST when mapped on the latent space. *(l) VC, (r) CE*

F 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.

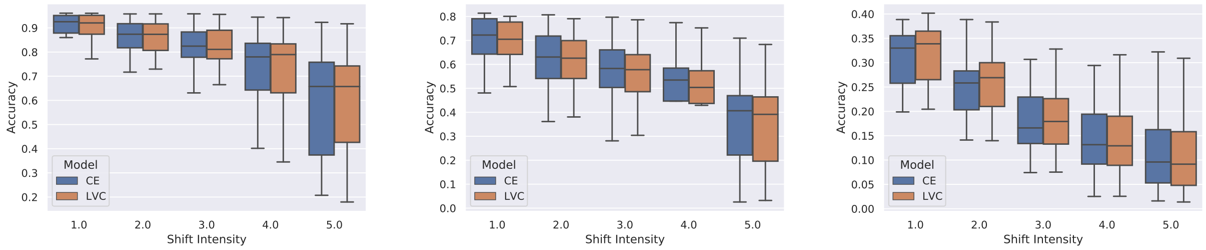

Figure 8: Classification accuracy under distributional shift: *(left)* CIFAR-10-C *(middle) CIFAR-100-C (right)* *TINY-IMAGENET-C*