SEMI-SUPERVISED LEARNING OBJECTIVES AS LOG-LIKELIHOODS IN A GENERATIVE MODEL OF DATA CURATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We currently do not have an understanding of semi-supervised learning (SSL) objectives such as pseudo-labelling and entropy minimization as log-likelihoods, which precludes the development of e.g. Bayesian SSL. Here, we note that benchmark image datasets such as CIFAR-10 are carefully curated, and we formulate SSL objectives as a log-likelihood in a generative model of data curation that was initially developed to explain the cold-posterior effect (Aitchison 2020). SSL methods, from entropy minimization and pseudo-labelling, to state-of-the-art techniques similar to FixMatch can be understood as lower-bounds on our principled log-likelihood. We are thus able to give a proof-of-principle for Bayesian SSL on toy data. Finally, our theory suggests that SSL is effective in part due to the statistical patterns induced by data curation. This provides an explanation of past results which show SSL performs better on clean datasets without any "out of distribution" examples. Confirming these results we find that SSL gave much larger performance improvements on curated than on uncurated data, using matched curated and uncurated datasets based on Galaxy Zoo 2.[1]

1 INTRODUCTION

To build high-performing deep learning models for industrial and medical applications, it is necessary to train on large human-labelled datasets. For instance, Imagenet (Deng et al., 2009), a classic benchmark dataset for object recognition, contains over 1 million labelled examples. Unfortunately, human labelling is often prohibitively expensive. In contrast obtaining unlabelled data is usually very straightforward. For instance, unlabelled image data can be obtained in almost unlimited volumes from the internet. Semi-supervised learning (SSL) attempts to leverage this unlabelled data to reduce the required number of human labels (Seeger, 2000; Zhu, 2005; Chapelle et al., 2006; Zhu & Goldberg, 2009; Van Engelen & Hoos, 2020). One family of SSL methods — those based on low-density separation — assume that decision boundaries lie in regions of low probability density, far from all labelled and unlabelled points. To achieve this, pre deep learning (DL) low-density separation SSL methods such as entropy minimization and pseudo-labelling (Grandvalet & Bengio, 2005; Lee, 2013) use objectives that repel decision boundaries away from unlabelled points by encouraging the network to make more certain predictions on those points. Entropy minimization (as the name suggests) minimizes the predictive entropy, whereas pseudo-labelling treats the currently most-probable label as a pseudo-label, and minimizes the cross entropy to that pseudo-label. More modern work uses the notion of consistency regularisation, which augments the unlabelled data (e.g. using translations and rotations), then encourages the neural network to produce similar outputs for different augmentations of the same underlying image (Sajjadi et al., 2016; Xie et al., 2019; Berthelot et al., 2019b; Sohn et al., 2020). Further developments of this line of work have resulted in many variants/combinations of these algorithms, from directly encouraging the smoothness of the classifier outputs around unlabelled datapoints (Miyato et al., 2018) to the "FixMatch" family of algorithms (Berthelot et al., 2019b;a; Sohn et al., 2020), which combine pseudo-labelling and consistency regularisation by augmenting each image twice, and using one of the augmented images to provide a pseudo-label for the other augmentation.

[1] Our code: https://anonymous.4open.science/r/GZ_SSL-B6CC; MIT Licensed

$Y_1 = Y_2 = Y_3 = \text{train}$

$Y = \text{train}$

$Y_1 = Y_2 = Y_3 = \text{bus}$

$Y = \text{bus}$

$Y_1 = Y_2 = \text{train}; \quad Y_3 = \text{bus}$

$Y = \texttt{None}$

Figure 1: A depiction of the generative model of data curation, with $S = 3$. Annotators are instructed to classify images as trains or buses. The left-hand image is clearly a train, so the annotators agree and consensus is reached. The middle image is clearly a bus, so annotators agree and consensus is reached. The right-hand image, however, is ambiguous or even has an ill-defined class label. So annotators disagree, consensus is not reached and the image is excluded from the dataset.

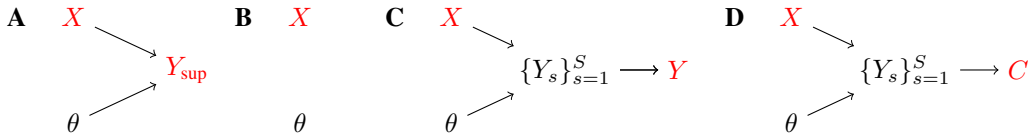

Figure 2: Graphical models under consideration. The observed variables are highlighted in red. **A** The generative model for standard supervised learning with no data curation. Note that the label, $Y_{\text{sup}} \in \mathcal{Y}$, only takes on values in the label set, so it differs from our label, Y, which could also be `None`. **B** The generative model for standard supervised learning, omitting the label. Under this model $\mathrm{P}(\theta|X)$ reduces to $\mathrm{P}(\theta)$ (see Sec. 2.3). **C** The generative model with data curation for labelled points, note that if there is consensus, $Y \in \mathcal{Y}$ and if there is no consensus, $Y = \texttt{None}$. **D** The generative model with data curation for unlabelled points. C is a random variable representing whether or not consensus was reached, so we have $C = 1$ if consensus was reached (i.e. $Y \in \mathcal{Y}$) and $C = 0$ if consensus was not reached (i.e. $Y = \texttt{None}$). Critically, C acts as a "label", so the posterior over parameters, $\mathrm{P}(\theta|X, C)$, does not reduce to $\mathrm{P}(\theta)$, unlike in the case of standard supervised learning when we omit the label (panel **B**).

However, some of the biggest successes of deep learning, from supervised learning to many generative models, have been built on a principled statistical framework as maximum (marginal) likelihood inference (e.g. the cross-entropy objective in supervised learning can be understood as the log-likelihood for a Categorical-softmax model of the class-label MacKay, 2003). Low-density separation SSL methods such as pseudo-labelling and entropy minimization are designed primarily to encourage the class-boundary to lie in low-density regions. Therefore they cannot be understood as log-likelihoods and cannot be combined with principled statistical methods such as Bayesian inference.

Here, we give a formal account of SSL methods based on low-density separation (Chapelle et al., 2006) as lower bounds on a principled log-likelihood. In particular, we consider pseudo-labelling (Lee, 2013), entropy minimization (Grandvalet & Bengio, 2005), and modern methods similar to FixMatch (Sohn et al., 2020). This log-likelihood arises from a generative model of data curation that was initially developed to explain the cold-posterior effect (Aitchison, 2021). Critically, this approach gives an explanation for previous findings that SSL is most effective when unlabelled data is obtained by throwing away labels from the carefully curated training set, and is less effective when unlabelled data is taken from uncurated images, especially those that do not depict one of the classes of interest (Cozman et al., 2003; Oliver et al., 2018; Chen et al., 2020; Guo et al., 2020). We confirmed the importance of data curation for SSL on toy data generated from a known model and on real data from Galaxy Zoo 2 (Willett et al., 2013).

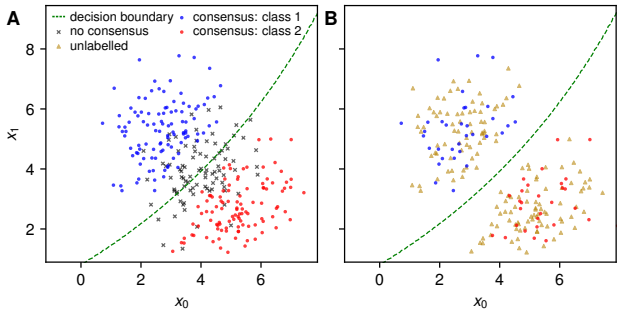

Figure 3: Our generative model of data curation applied to a simple 2D dataset. Data from each class was sampled from a different Gaussian, and the true decision boundary (green dashed line) was given by the posterior probability of class given (x_0, x_1). **A** Datapoints far from the decision boundary are unambiguous, so annotators agree and consensus is reached (red and blue points). Datapoints close to the decision boundary are ambiguous, so consensus is not reached (grey crosses). The consensus datapoints thus exhibit artificially induced low-density separation. **B** When using benchmark datasets such as CIFAR-10, the unlabelled points (yellow triangles) are selected from the consensus points (red or blue points) as the noconensus points are not available. The unlabelled points therefore also exhibit artificially induced low-density separation.

2 Background

Our work brings together many disparate areas. Here, we give an introduction to a generative model of data curation (Aitchison, 2021) initially developed to explain the cold posterior effect (Wenzel et al., 2020), pseudo-labelling and entropy minimization (Grandvalet & Bengio, 2005; Lee, 2013), and the treatment of unlabelled points in the standard supervised learning setup.

2.1 A generative model of data curation

To develop a model of data curation, remember that image datasets including CIFAR-10 and ImageNet are curated to ensure they only contain images whose class-labels are unambiguous. For instance, in CIFAR-10, annotators were instructed that "It's worse to include one that shouldn't be included than to exclude one.", and Krizhevsky (2009) "personally verified every label submitted by the annotators". In creating ImageNet, Deng et al. (2009) made sure that a number of Amazon Mechanical Turk annotators agreed upon the class before including an image in the dataset.

Thus, these datasets have two odd properties. First, consensus labels exist only for a subset of images, e.g. for a white-noise image, consensus cannot be reached and the image cannot be labelled. Second, inclusion of an image in a dataset like CIFAR-10 is informative in and of itself, as it indicates that the image shows an unambiguous example of one of the ten classes. To understand these odd properties of curated datasets, consider a simplified generative model of consensus-formation: draw a random image, X, from the distribution over images, $\mathrm{P}\left(X\right)$, and ask S human annotators, indexed s, to give a label, $\{Y_s\}_{s=1}^S$ (e.g. using Mechanical Turk). Importantly, every annotator is forced to label every image and if the image is ambiguous they should give a random label. If all the annotators agree, $Y_1 = Y_2 = \cdots = Y_S$, they have consensus and the datapoint is included in the dataset. However, in the case of any disagreement, consensus is not reached and the datapoint is excluded (Fig. 1), Concretely, the final label, Y is Y_1 (which is the same as all the other labels) if consensus was reached and `None` otherwise (Fig. 2C),

$$Y|\{Y_s\}_{s=1}^S = \begin{cases} Y_1 & \text{if } Y_1 = Y_2 = \cdots = Y_S \\ \texttt{None} & \text{otherwise} \end{cases} \tag{1}$$

Taking $\mathcal{Y}$ to be the label set, we have $Y_s \in \mathcal{Y}$, and the final label, Y, could be any of the underlying labels in $\mathcal{Y}$, or `None` if consensus is not reached, so $Y \in \mathcal{Y} \cup \{\texttt{None}\}$. When consensus was reached, the likelihood is,

$$\mathrm{P}\left(Y = y|X, \theta\right) = \mathrm{P}\left(\{Y_s = y\}_{s=1}^S|X, \theta\right) = \prod_{s=1}^S \mathrm{P}\left(Y_s = y|X, \theta\right) = \mathrm{P}\left(Y_s = y|X, \theta\right)^S = \left(p_y(X)\right)^S \tag{2}$$

where we have assumed annotators are IID, and $p_y(X) = \mathrm{P}\left(Y_s{=}y|X,\theta\right)$ is the single-annotator probability. From here, it is possible to see how this model might be taken to give an account of tempering, as we have taken the underlying single-annotator likelihood, $p_y(X)$ to the power S (for further details see Aitchison, 2021).

2.2 Low-density separation semi-supervised learning objectives

The intuition behind low-density separation objectives for semi-supervised learning is that decision boundaries should be in low-density regions away from both labelled and unlabelled data. As such, it is sensible to "repel" decision boundaries away from labelled and unlabelled datapoints and this can be achieved by making the classifier as certain as possible on those points. This happens automatically for labelled points as the standard supervised objective encourages the classifier to be as certain as possible about the true class label. But for unlabelled points we need a new objective that encourages certainty, and we focus on two approaches. First, and perhaps most direct is entropy minimization (Grandvalet & Bengio, 2005)

$$\mathcal{L}_{\text{entropy}}(X) = \sum_{y\in\mathcal{Y}} p_y(X) \log p_y(X) \tag{3}$$

where, following the typical probabilistic approach, we write the negative entropy as an objective to be maximized. Alternatively, we could use pseudo-labelling, which takes the current classification, y^*, to be the true label, and maximizes the log-probability of that label (Lee, 2013),

$$\mathcal{L}_{\text{pseudo}}(X) = \log p_{y^*}(X) \qquad\qquad y^* = \arg\max_{y\in\mathcal{Y}} \log p_y(X). \tag{4}$$

Lee (2013) regarded pseudo-labelling as closely related to entropy miminization as the optimal value of both objectives is reached when all the probability mass is assigned to one class. However, they are not formulated as a principled log-likelihood, which gives rise to at least three problems. First, these methods cannot be combined with other principled statistical methods such as Bayesian inference. Second, it is unclear how to combine these objectives with standard supervised objectives, except by taking a weighted sum and doing hyperparameter optimization over the weight. Third, these objectives risk reinforcing any initial poor classifications and it is unclear whether this is desirable.

2.3 In standard supervised learning, unlabelled points should be uninformative

It is important to note that under the standard supervised-learning generative model (Fig. 2A), unlabelled points should not give any information about the weights. Omitting the label, Y_{sup}, we obtain the graphical model in Fig. 2B. This model emphasises that the images, X, and the model parameters, θ, are marginally independent, so we cannot obtain any information about θ from X alone (Fig. 2B). Formally, the posterior over θ conditioned on X is equal to the prior,

$$\begin{aligned}\mathrm{P}\left(\theta|X\right) &= \frac{\mathrm{P}\left(\theta, X\right)}{\mathrm{P}\left(X\right)} = \frac{\sum_{y\in\mathcal{Y}} \mathrm{P}\left(\theta, X, Y_{\text{sup}}{=}y\right)}{\mathrm{P}\left(X\right)} \\ &= \frac{\mathrm{P}\left(\theta\right)\mathrm{P}\left(X\right)\sum_{y\in\mathcal{Y}} \mathrm{P}\left(Y_{\text{sup}}{=}y|\theta, X\right)}{\mathrm{P}\left(X\right)} = \mathrm{P}\left(\theta\right).\end{aligned} \tag{5}$$

as $1 = \sum_{y\in\mathcal{Y}} \mathrm{P}\left(Y_{\text{sup}}{=}y|\theta, X\right)$. To confirm this result is intuitively sensible, note that are many situations where encouraging the decision boundary to lie in low density regions would be very detrimental to performance. Consider a classifier with two input features: x_0 and x_1 (Fig. 4A). The class boundary lies in the high-density region crossing both clusters, so to obtain a reasonable result, the classifier should ignore the low-density region lying between the clusters. However, strong low-density separation SSL terms in the objective may align the cluster boundaries with the class boundaries, leading the classifier to wrongly believe that one cluster is entirely one class and the other cluster is entirely the other class. In contrast, supervised learning without SSL will ignore clustering and obtain a reasonable answer close to the grey dashed line. Importantly, this is just an illustrative example to demonstrate that without further assumptions, the standard supervised approach of ignoring unlabelled data is sensible; semi-supervised learning without loss of performance in such settings has been studied and is known as Safe SSL (Li & Zhou, 2014; Krijthe & Loog, 2014; Kawakita & Takeuchi, 2014; Loog, 2015; Krijthe & Loog, 2016).

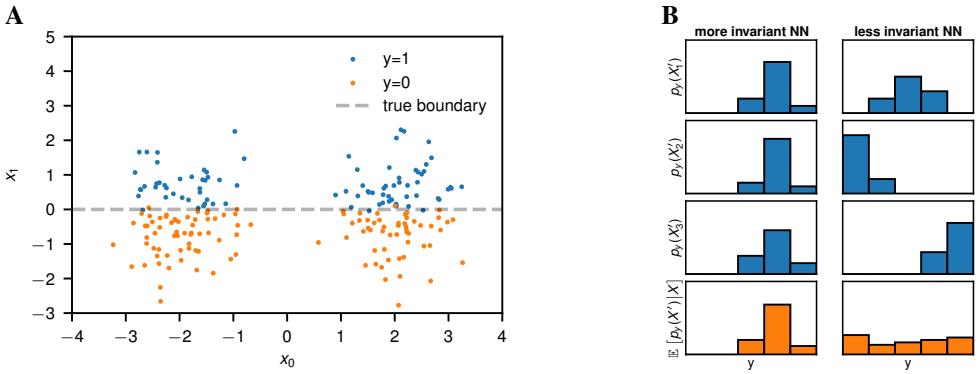

Figure 4: **A**. A toy dataset generated to illustrate the dangers of using the clustering of the input points to inform classification boundaries. The input features, x_0 and x_1 are plotted on the x and y-axes and the class is represented by colour. **B**. A schematic diagram demonstrating the effect of our principled likelihood incorporating data-augmentation on the certainty of predictions for different degrees of invariance. More invariant NNs (left) give similar predictive distributions for different augmentations (blue), and hence a certain averaged predictive distribution (bottom; orange). Less invariant NNs (right) give different predictive distributions for different augmentations (blue), and hence highly uncertain averaged predictive distributions (bottom; orange).

3 THEORY

SSL methods are usually applied to benchmark datasets such as CIFAR-10 or ImageNet. These datasets were first carefully curated during the labelling process: (Fig. 3A), implying that ambiguous images close to the decision boundary were excluded. Critically, unlabelled points for these benchmark datasets are obtained by taking labelled points (which have reached consensus) and throwing away their labels (Fig. 3B). The likelihood for consensus ($Y{\neq}\texttt{None}$) is

$$\mathrm{P}\left(Y{\neq}\texttt{None}|X,\theta\right) = \sum_{y\in\mathcal{Y}} \left(p_y(X)\right)^S. \tag{6}$$

This probability is close to 1 (for $S > 1$) if the underlying distribution, $\left(p_y(X)\right)^S$ puts most of its mass onto one class, and the probability is smaller if the mass is spread out over classes. As such, the likelihood "repels" decision boundaries away from unlabelled points, which is the common intuition behind low-density separation SSL methods, and which should be beneficial if class boundaries indeed lie in regions of low probability density away from both labelled and unlabelled points.

If noconsensus images are observed (Fig. 2C), we can include a likelihood term for those images,

$$\mathrm{P}\left(Y{=}\texttt{None}|X,\theta\right) = 1 - \mathrm{P}\left(Y \neq \texttt{None}|X,\theta\right) = 1 - \sum_{y\in\mathcal{Y}} \left(p_y(X)\right)^S. \tag{7}$$

If noconsensus images are not observed, we could in principle integrate over the underlying distribution over images, $\mathrm{P}\left(X{=}x\right)$. However, we do not even have samples from the underlying distributions over images (and if we did, we would have the noconsensus images so we could use Eq. 7). As such this term is usually omitted (e.g. Aitchison, 2021), but the use of out-of-distribution (OOD) datasets as surrogate noconsensus points is an important direction for future work.

3.1 ENTROPY MINIMIZATION AND PSEUDO-LABELS ARE LOWER BOUNDS ON OUR PRINCIPLED LOG-LIKELIHOOD

To prove that entropy minimization forms a lower-bound on our log-likelihood (Eq. 6), we begin by writing the log-likelihood of consensus in terms of an expectation over labels, y,

$$\log \mathrm{P}\left(Y{\neq}\texttt{None}|X,\theta\right) = \log \sum_{y\in\mathcal{Y}} p_y(X)\left(p_y(X)\right)^{S-1} = \log \mathbb{E}_{p_y(X)}\left[\left(p_y(X)\right)^{S-1}\right]. \tag{8}$$

Applying Jensen's inequality, the negative entropy gives a lower-bound on our log-likelihood,

$$\begin{aligned}\log \mathrm{P}\left(Y{\neq}\texttt{None}|X,\theta\right) &\geq \mathbb{E}_{p_y(X)}\left[\log\left(p_y(X)\right)^{S-1}\right] \\ &= (S-1)\textstyle\sum_{y\in\mathcal{Y}} p_y(X)\log p_y(X) = (S-1)\mathcal{L}_{\text{entropy}}(X)\end{aligned} \tag{9}$$

This bound is tight for a uniform predictive distribution,

$$\log \mathrm{P}\left(Y{\neq}\texttt{None}|X,\theta\right) = \log \textstyle\sum_{y\in\mathcal{Y}} \left(p_y(X)\right)^S = \log S \left(\frac{1}{S}\right)^S = (S-1)\log S \tag{10}$$

$$(S-1)\mathcal{L}_{\text{entropy}}(X) = -(S-1)\log \textstyle\sum_{y\in\mathcal{Y}} p_y(X)\log p_y(X) = (S-1)\log S. \tag{11}$$

Pseudo-labelling forms an alternative lower bound on the log-likelihood which is obtained by noting that all $\left(p_y(X)\right)^S$ are positive, so selecting any subset of terms in the sum gives a lower bound,

$$\log \mathrm{P}\left(Y{\neq}\texttt{None}|X,\theta\right) = \log \textstyle\sum_{y\in\mathcal{Y}} \left(p_y(X)\right)^S \geq \log \left(p_{y^*}(X)\right)^S = S\log p_{y^*}(X) = S\mathcal{L}_{\text{pseudo}}(X). \tag{12}$$

The inequality holds if we choose y^* to be any class, but will be tightest if we choose the highest probability class. This bound is tight for a predictive distribution that puts all its mass on y^*, so $p_{y^*}(X) = 1$ and $p_{y\neq y^*} = 0$

$$\log \mathrm{P}\left(Y{\neq}\texttt{None}|X,\theta\right) = \log \textstyle\sum_{y\in\mathcal{Y}} \left(p_y(X)\right)^S = \log \left(p_{y^*}(X)\right)^S = \log 1 = 0 \tag{13}$$

$$S\mathcal{L}_{\text{pseudo}}(X) = S\log p^*_y(X) = S\log 1 = 0. \tag{14}$$

As such, entropy minimization and pseudo-labelling optimize different lower-bounds on our principled log-likelihood, $\log \mathrm{P}\left(Y{\neq}\texttt{None}|X,\theta\right)$, which gives a potential explanation for the effectiveness of pseudo-labelling and entropy minimization. Additionally, low-density separation SSL objectives encourages class-labels to be more certain. We can therefore expect pseudo-labelling to be the more relevant bound, as that bound is tight when the predictive distribution puts all its mass onto one class. In contrast, the entropy maximization bound is tight when the predictive distribution is uniform, which is discouraged by all low-density separation SSL objectives. This provides a potential explanation for the use of psuedo-labelling rather than entropy regularisation in modern SSL approaches such as (Sohn et al., 2020).

3.2 Data augmentation priors and FixMatch family methods

FixMatch family methods combine data augmentation and pseudo-labelling. To understand FixMatch as a bound on a principled log-likelihood, we therefore need a principled account of data augmentation as a likelihood. Inspired by Wenzel et al. (2020) (their Appendix K), we consider a distribution, $\mathrm{P}\left(X'|X\right)$, over augmented images, X', given the underlying unaugmented image, X. We choose the single-annotator predictive distribution as the average over predictive distributions for many different augmented images,

$$\mathrm{P}\left(Y_s{=}y|X,\theta\right) = \mathbb{E}\left[p_y(X')|X\right] \tag{15}$$

where $p_y(X')$ is the predictive probabilities resulting from applying the neural network to the augmented image, and remember $s \in \{1,\ldots,S\}$ indexes the annotator. This is a sensible prior because we expect the neural network to be invariant under data-augmentation, and if the predictions are approximately invariant, then averaging the predictive distributions has little impact (Fig. 4B left). However, if the predictions do vary dramatically with different data augmentations then we should not trust the network's classifications (i.e. we should have an uncertain predictive distribution), and averaging over very different predictive distributions for different augmentations indeed gives rise to broader, more uncertain predictions (Fig. 4B right).

To obtain a tractable objective in the supervised setting, we use a multi-sample version of Jensen's inequality, with K augmented images denoted X'_k,

$$\log \mathrm{P}\left(Y_s{=}y|X,\theta\right) \geq \mathbb{E}\left[\log \tfrac{1}{K}\textstyle\sum_k p_y(X'_k)\big|X\right]. \tag{16}$$

Combining this single-annotator probability with our generative model of curation, we obtain,

$$\begin{aligned}\log \mathrm{P}\left(Y{=}y|X,\theta\right) &= S\log \mathrm{P}\left(Y_s{=}y|X,\theta\right)\\ &= S\log \mathbb{E}\left[p_y(X')|X\right] \geq S\,\mathbb{E}\left[\log \tfrac{1}{K}\textstyle\sum_k p_y(X'_k)\big|X\right],\end{aligned} \tag{17}$$

The resulting objective for unlabelled points is,

$$\begin{aligned}\log \mathrm{P}\left(Y{\neq}\texttt{None}|X,\theta\right) &= \log \textstyle\sum_{y\in\mathcal{Y}} \mathrm{P}\left(Y{=}y|X,\theta\right)\\ &= \log \textstyle\sum_{y\in\mathcal{Y}} \mathbb{E}\left[p_y(X')|X\right]^S \approx \log \textstyle\sum_{y\in\mathcal{Y}} \left(\frac{1}{K}\sum_k p_y(X'_k)\right)^S,\end{aligned} \tag{18}$$

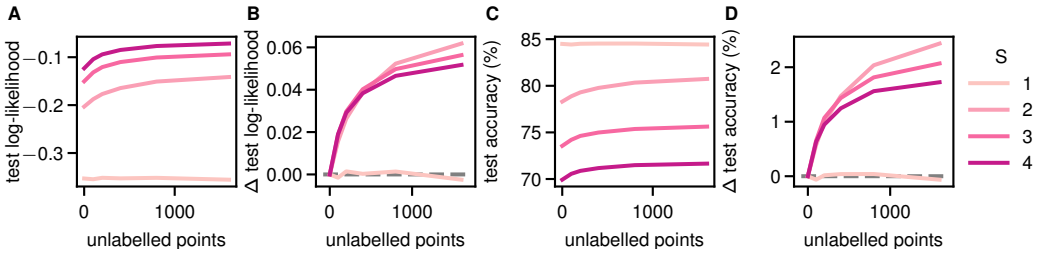

Figure 5: Test log-likelihood and accuracy for Langevin sampling for Bayesian SSL on toy datasets sampled from the model as a function of the number of unlabelled points.

where we approximate the expectation with K different samples of X', denoted X'_k. Unfortunately, this approach does not immediately form a bound on the log-likelihood due to the convex nonlinearity in taking the power of S. Nonetheless, one key problem with approximating machine learning losses is that the optimizer learns to exploit approximation errors to find a pathological solution that makes the objective unboundedly large. We appear to be safe from that pathology here, as we are simply forming predictions by averaging over K augmentations of the underlying image. Nonetheless, to form a lower bound, we can follow FixMatch family algorithms by pseudo-labelling, i.e. by taking only one term in the sum for class y^*. FixMatch chooses y^* by using the highest-probability class for a weakly-augmented image. An alternative approach is to choose the y^* giving the tightest bound, i.e. $\arg\max_y \frac{1}{K}\sum_k p_y(X'_k)$. In either case,

$$\log \mathrm{P}\left(Y \neq \texttt{None} | X, \theta\right) \geq \log \mathbb{E}\left[p_{y^*}(X')|X\right]^S \geq S\, \mathbb{E}\left[\log \tfrac{1}{K}\textstyle\sum_k p_{y^*}(X'_k)\big|X\right], \tag{19}$$

If $K = 1$ and y^* is chosen using a separate "weak" augmentation, then this is exactly equal to the FixMatch objective for unlabelled points.

Note that both of these objectives (Eq. 18 and 19) promote reduced predictive uncertainty. Importantly, this does not just increase confidence in the single-augmentation predictive distributions, $p_y(X'_k)$, but also increases alignment between the predictive distributions for different augmentations (Fig. 4B). In particular, if the single-augmentation predictives are all highly confident, but place that high-confidence on different classes, then the multi-augmentation predictive formed by averaging will have low-confidence (Fig. 4B right). The only way for the multi-augmentation predictive to have high confidence is if the underlying single-augmentation predictive distributions have high confidence in the same class (Fig. 4B left), which encourages the underlying network to become more invariant. This makes sense: if data-augmentation changes the class predicted by the neural network, then any predictions *should* be low confidence. And it implies that combining principled data augmentation with a generative model of data curation automatically gives rise to an objective encouraging invariance.

4 Results

We begin by giving a proof-of-principle for Bayesian SSL on a toy dataset generated from a known model. Next, we tested our theoretical results (rather than trying to achieve SOTA performance) on real-world datasets. In particular, our theory gives one explanation for why SSL is typically more effective when unlabelled data is taken from the original, curated training set. To confirm these results, we used Galaxy Zoo 2 as this was a real-world dataset which allowed us to generate matched curated and uncurated datasets.

4.1 Bayesian SSL on a generated dataset

Our formulation of SSL as a likelihood implies that it should be possible to take entirely novel approaches, such as using low-density separation SSL in a Bayesian neural network (BNN).

We considered a toy dataset generated from a "true" neural network model with one hidden layer and 30 hidden units, 5 dimensional inputs and 2 output classes. We generated inputs IID from a Gaussian, then passed them through the "true" neural network, then sampled multiple categorical

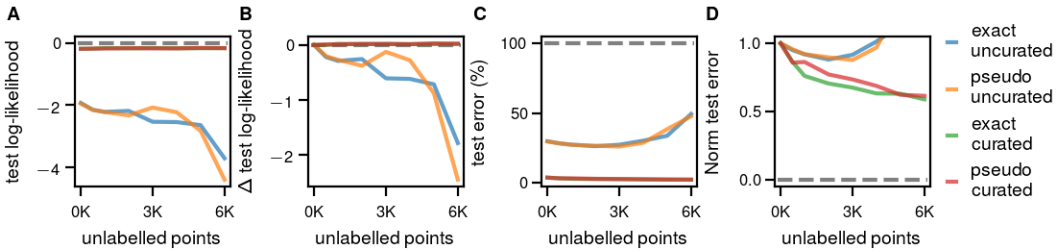

Figure 6: Test log-likelihood and error for curated and uncurated GZ2 datasets as a function of the number of unlabelled points. Exact corresponds to Eq. (18) (which is exact in the limit as $K \to \infty$) and pseudo corresponds to the pseudo labelling version of the augmented objective (Eq. 19).

class-labels corresponding to different annotators. If all the simulated annotators agreed, consensus was reached and if any simulated annotators disagreed, consensus was not reached. We used 100 labelled datapoints, though not all of them will have reached consensus, and we used up to 1600 unlabelled points, though again not all of them will have reached consensus. Note that as the consensus/noconsensus status of a point arises from the generative model, we cannot independently specify the number of consensus/noconsensus points. We used Eq. (2) as the likelihood for labelled points, Eq. (6) as the likelihood for unlabelled points and Eq. (7) as the likelihood for noconsensus points. We sampled (and trained networks on) 500 datasets in parallel. We trained using Langevin dynamics with all data simultaneously (no minibatching) with no momentum and no rejection.

For a generative model with $S = 1$, consensus is always reached and the problem is equivalent to standard supervised learning. As such, we found no benefits from including unlabelled points for $S = 1$. In contrast, for any setting of $S > 1$ we found that increasing the number of unlabelled points improved the test log-likelihood (Fig. 5A–B) and the test accuracy (Fig. 5C–D).

4.2 Galaxy Zoo 2

Our data curation based theory predicts that low-density separation based SSL should be much more effective on curated than uncurated data. To test this prediction on real-world data, we turned to Galaxy Zoo 2[2] (GZ2) (Willett et al., 2013) which uses images from the Sloan Digital Sky Survey. This dataset is particularly useful for us as it has received only very minimal filtering based on criteria such as object brightness and spatial extent. We defined 9 labels by truncating the complex decision tree followed by the annotators (for further details see Aitchison, 2021). Further, as each GZ2 image has received ~ 50 labels, we can define a consensus coefficient by taking the fraction of annotators that agreed upon the highest probability class. We can then define a curated dataset by taking the images with consensus coefficient above some threshold within each class. Note that we needed to select images on a per-class basis, because annotators tend to be more confident on some classes than others, so taking the highest consensus coefficients overall would dramatically change the class balance. In particular, we used the top 8.2% of images, which gave a full curated dataset of just over 20,000 images. Of those, we randomly selected 2000 as labelled examples, 10000 as test examples, and 0 – 6000 as unlabelled examples. The images were preprocessed by center-cropping to 212×212 and then scaled to 32×32. We applied a FixMatch-inspired semi-supervised learning algorithm, with a standard supervised objective, with unlabelled objective given by Eq. (18) with $K = 2$. Data augmentation was given by vertical and horizontal flips, rotations from $-180°$ to $180°$, translations by up to 40% on both axes and scaling from 20% to 180%. Note that as we were trying to mirror the standard SSL setup, we did not include noconsensus points in the objective. We trained a ResNet18 with our maximum likelihood objective using SGD with a batch size of 500, a learning rate of 0.01 and 1500 epochs. We used an internal cluster of nVidia 1080 and 2080 GPUs, and the experiments took roughly 300 GPU hours.

We found that the test-log-likelihood for curated data improved slightly as more unlabelled points were included, whereas the test-log-likelihood for uncurated dramatically declined as unlabelled points were added (Fig. 6A–B). We saw strong improvements in test accuracy with the number of

[2]https://data.galaxyzoo.org; www.sdss.org/collaboration/image-use-policy/

unlabelled points for curated datasets (Fig. 6C–D). Note that in Fig. 6C the error rate for curated datasets is already very small, so to see any effect we needed to plot the test error, normalized to the initial test error (Fig. 6D). For uncurated data, the inclusion of large numbers of unlabelled points dramatically worsened performance, though the inclusion of a small number of unlabelled points gave very small performance improvements (Fig. 6C–D). Thus, this experiment is consistent with the idea that the effectiveness of SSL arises at least in part from curation of the underlying dataset.

5 Related work

There are at least three main approaches to semi-supervised learning (Seeger, 2000; Zhu, 2005; Chapelle et al., 2006; Zhu & Goldberg, 2009). First there is low-density separation, where we assume that the class boundary lies in a region of low probability density away from both labelled and unlabelled points. This approach dates back at least to transductive support vector machines (SVMs) where the model is to be tested on a finite number of known test locations (Vapnik, 1998; Chapelle et al., 1999). Those known test locations are treated as unlabelled points, and we find the decision boundary that perfectly classifies the limited number of labelled points, while at the same time being as far as possible from labelled and unlabelled data. Alternative approaches include pseudo-labelling and entropy minimization (Grandvalet & Bengio, 2005; Lee, 2013). Second, there are graph-based methods such as (Zhu & Ghahramani, 2002) which are very different from the methods considered here. Third, there are approaches that use unlabelled points to build a generative model of the *inputs* and leverage that model to improve classification (e.g. Kingma et al., 2014; Odena, 2016; Gordon & Hernández-Lobato, 2017). This approach was originally explored in a considerable body of classical work (e.g. McLachlan, 1975; Castelli & Cover, 1995; Druck et al., 2007) for a review, see Seeger (2000) and references therein. These approaches are fundamentally different from the SSL approaches considered here, as they require a generative model of inputs, while low-density separation methods do not. Generative modelling can be problematic as training a generative model can be more involved than training a discriminative model and because the even when the model can produce excellent samples, the high-level representation may be "entangled" (Higgins et al., 2017) in which case it may not offer benefits for classification.

6 Discussion

Our theory provides a theoretical understanding of past results showing that SSL is more effective when unlabelled data is drawn from the original, curated training set (Cozman et al., 2003; Oliver et al., 2018; Chen et al., 2020; Guo et al., 2020). In the extreme, our theory might be taken to imply that if data has not been curated, then SSL cannot work, and therefore that low-density separation SSL methods will not be effective in messy, uncurated real-world datasets. However, this is not the complete picture. Low-density separation SSL methods, including our log-likelihood, fundamentally exploit class-boundaries lying in low-density regions. As such, low-density separation could equally come from the real underlying data or could be artificially induced by data curation (Fig. 3). None of these methods are able to distinguish between these different underlying sources of low-density separation and as such any of them may work on uncurated data where the underlying distribution displays low-density separation. However, the possibility for curation to artificially induce low-density separation does imply that we should be cautious about overinterpreting spectacular results obtained on very carefully curated benchmark datasets such as CIFAR-10.

Surprisingly, the generative model of data curation used here also explains the cold-posterior effect in Bayesian neural networks (Wenzel et al., 2020; Aitchison, 2021), revealing a profound and previously unsuspected connection.

In conclusion, we showed that low-density separation SSL objectives can be understood as a lower-bound on a log-probability which arises from a principled generative model of data curation. This gives a theoretical understanding of recent results showing that SSL is more effective on curated data, which we confirmed by developing a Bayesian SSL model applied to toy data, using GZ2, which allowed us to consider a completely uncurated dataset.

REFERENCES

Laurence Aitchison. A statistical theory of cold posteriors in deep neural networks. *ICLR*, 2021.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019a.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 5050–5060, 2019b.

Vittorio Castelli and Thomas M Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16(1):105–111, 1995.

Olivier Chapelle, Vladimir Vapnik, and Jason Weston. Transductive inference for estimating values of functions. In *NIPS*, pp. 421–427. Citeseer, 1999.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. 2006.

Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *AAAI*, pp. 3569–3576. AAAI Press, 2020.

Fabio Gagliardi Cozman, Ira Cohen, Marcelo Cesar Cirelo, et al. Semi-supervised learning of mixture models. In *ICML*, volume 4, pp. 24, 2003.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Gregory Druck, Chris Pal, Andrew McCallum, and Xiaojin Zhu. Semi-supervised classification with hybrid generative/discriminative methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 280–289, 2007.

Jonathan Gordon and José Miguel Hernández-Lobato. Bayesian semisupervised learning with deep generative models. *arXiv preprint arXiv:1706.09751*, 2017.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pp. 529–536, 2005.

Lan-Zhe Guo, Zhenyu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3897–3906. PMLR, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2017.

Masanori Kawakita and Jun'ichi Takeuchi. Safe semi-supervised learning based on weighted likelihood. *Neural Networks*, 53:146–164, 2014.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27: 3581–3589, 2014.

Jesse H Krijthe and Marco Loog. Implicitly constrained semi-supervised linear discriminant analysis. In *2014 22nd International Conference on Pattern Recognition*, pp. 3762–3767. IEEE, 2014.

Jesse H Krijthe and Marco Loog. The pessimistic limits and possibilities of margin-based losses in semi-supervised learning. *arXiv preprint arXiv:1612.08875*, 2016.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Tech. report*, 2009.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 2, 2013.

Yu-Feng Li and Zhi-Hua Zhou. Towards making unlabeled data never hurt. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):175–188, 2014.

Marco Loog. Contrastive pessimistic likelihood estimation for semi-supervised classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):462–475, 2015.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Geoffrey J McLachlan. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, 70(350):365–369, 1975.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

Avital Oliver, Augustus Odena, Colin Raffel, Ekin Dogus Cubuk, and Ian J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, pp. 3239–3250, 2018.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in neural information processing systems*, pp. 1163–1171, 2016.

Matthias Seeger. Learning with labeled and unlabeled data. 2000.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

V Vapnik. Statistical learning theory. *NY: Wiley*, 1998.

Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Świątkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

Kyle W Willett, Chris J Lintott, Steven P Bamford, Karen L Masters, Brooke D Simmons, Kevin RV Casteels, Edward M Edmondson, Lucy F Fortson, Sugata Kaviraj, William C Keel, et al. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 2013.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv:1904.12848*, 2019.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002.

Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005.