Variational Learning Induces Adaptive Label Smoothing

Sin-Han Yang [1] **Zhedong Liu** [1] **Gian Maria Marconi** [1 2] **Mohammad Emtiyaz Khan** [1]

Abstract

We show that variational learning naturally induces an adaptive label smoothing where label noise is specialized for each example. Such label-smoothing is useful to handle examples with labeling errors and distribution shifts, but designing a good adaptivity strategy is not always easy. We propose to skip this step and simply use the natural adaptivity induced during the optimization of a variational objective. We show empirical results where a variational algorithm called IVON outperforms traditional label smoothing and yields adaptivity strategies similar to those of an existing approach. By connecting Bayesian methods to label smoothing, our work provides a new way to handle overconfident predictions.

1. Introduction

Adaptive strategies to Label Smoothing (LS) (Szegedy et al., 2016) aim to adapt the label noise according to the type of data example. Such adaptation can be more effective in practice than its traditional counterpart where the label noise is the same for all examples. Adaptation is useful to handle examples that may have labeling errors, distribution shift, or calibration issues. For such cases, the effectiveness of adaptation has been extensively studied, for example, see Ghoshal et al. (2021); Lee et al. (2022) for generalization improvements, Zhang et al. (2021); Ko et al. (2023) for mislabelled examples, Park et al. (2023) for miscalibration, and Xu et al. (2024) for out-of-distribution detection. Adaptivity is useful for label smoothing to handle all such cases.

One major problem with adaptive label smoothing is that it is not easy to design a good adaptivity strategy. For example, a simple approach is to adapt the label noise by using model's predictions but there are many ways to do this, for examples, Park et al. (2023) set the noise based on the logits, Zhang et al. (2021); Ko et al. (2023) use the predictive probabilities (obtained with softmax), while Lee et al. (2022) use their entropy. All of these are reasonable ideas but the choice of a good strategy for a given problem is not always straightforward. A strategy that reduces miscalibration may not be most effective for handling outliers or mislabeling. Focusing on one issue at a time has given rise to a lot of ad-hoc and heuristic strategies, and, despite their usefulness, designing an adaptive strategy for a task in hand remains tricky. Our goal here is to simplify the process by presenting and analyzing algorithms that naturally induce adaptivity.

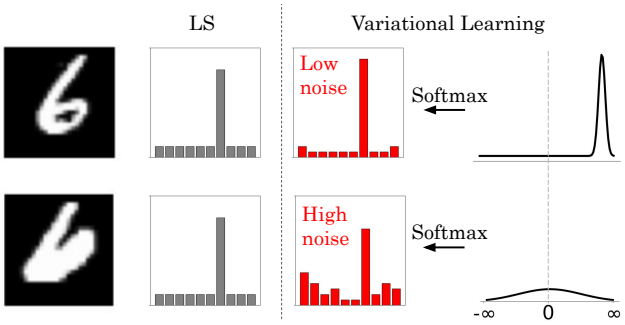

Figure 1. Given a regular 6 digit (top) and an atypical one (bottom), Label Smoothing (LS) assigns the same label noise to both (gray bars) while variational learning assigns higher noise to the atypical example (red bars). Adaptivity naturally arises due to the posterior.

We show that variational learning naturally induces an adaptive label smoothing. The smoothing arises due to the use of the expectation of the loss in the variational objective, taken with respect to the posterior distribution. The expectation gives rise to a label noise (among other types of noises) which is customized for each example through its features. Our key contribution is to derive the exact form of the label noise (Eq. 10) for many problems and study their behavior. We show extensive empirical results analyzing the label noise induced by Improved Variational Online Newton (IVON) (Shen et al., 2024). We show the following:

1. Variational learning assigns higher noise to atypical or ambiguous examples (Fig. 1 and Fig. 3).
2. IVON's adaptive label noise behaves similarly to the proposal of Zhang et al. (2021).
3. IVON consistently outperforms Label Smoothing in

[1]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan [2]Woven by Toyota. Correspondence to: Sin-Han Yang <sin-han.yang@riken.jp>, Mohammad Emtiyaz Khan <emtiyaz.khan@riken.jp>.

Workshop at the 7th Symposium on Advances in Approximate Bayesian Inference (non-archival), 2025.

presence of labeling errors, giving up to 9% accuracy boost for pair-flip noise (Fig. 5) and sometimes even around 50% for data-dependent noise (Fig. 6).

Our work connects label smoothing literature to Bayesian methods, thereby providing a new way to handle overconfident predictions in deep learning.

2. Label Smoothing and Adaptivity Strategies

Label Smoothing (LS) is a simple technique where the true label vector $\mathbf{y}_i$ (length K) are replaced by a smoothed version. In its simplest form, a convex combination is used where the smoothed labels are defined as

$$\mathbf{y}'_i = (1-\alpha)\mathbf{y}_i + \alpha\mathbf{u}, \tag{1}$$

for some scalar $\alpha \in (0,1)$ with $\mathbf{u}$ as a vector of $1/K$ with K being the number of classes. This simple technique is effective to penalize overconfident predictions because the noise $\alpha(\mathbf{u}-\mathbf{y}_i)$ reduces the importance of the label during training (Pereyra et al., 2017a). Multiple works have studied its effectiveness, for example, to improve calibration and representation (Müller et al., 2019), to favor flatter solutions (Damian et al., 2021), and improve robustness to mislabelled data (Lukasik et al., 2020; Liu, 2021) due to its connections to loss correction (Patrini et al., 2017). Despite its simplicity, LS has clear practical advantages.

Adaptive label smoothing aims to inject noise according to the type of data example, for example, during learning, we may want to inject a noise to get the smoothed label

$$\mathbf{y}_{i|t} = \mathbf{y}_i + \boldsymbol{\epsilon}_{i|t}. \tag{2}$$

The noise $\boldsymbol{\epsilon}_{i|t}$ depend on the model parameter $\boldsymbol{\theta}_t$ at iteration t, and can be varied according the model's opinion regarding the relevance of the examples. Adaptive label smoothing uses additive noise to reweigh examples during training. Many studies have shown the effectiveness of the adaptive noise, which ranges from improvements in generalization (Ghoshal et al., 2021; Lee et al., 2022), robustness to mislabeled data (Zhang et al., 2021; Ko et al., 2023), improving calibration (Park et al., 2023) and out-of-distribution (OOD) detection (Xu et al., 2024). By adapting label noise, such methods aim to down-weight the problematic examples.

While adaptivity is desirable, it also requires additional effort to design a good strategy to adapt. Each specific issues may require a different type of noise, for instance, what works to reduce miscalibration, may not be most effective for handling OOD detection or mislabling. Focusing on one issue or strategy at a time has given rise to a lot of ad-hoc and heuristic strategies, and, despite their usefulness, clarity of good ways to design adaptive strategy is lacking.

The simplest approach is to adapt by using the model predictions based on the logits $\mathbf{f}_i(\boldsymbol{\theta}_t)$, but there are many ways to use them. Zhang et al. (2021) use the following update for each example i in the epoch t

$$\mathbf{u} = \sum_i \mathcal{S}\left[\mathbf{f}_i(\boldsymbol{\theta}_t)\right], \tag{3}$$

where $\mathcal{S}[\mathbf{f}]$ vector (length K) with j'th entry defined as

$$\mathcal{S}_j\left[\mathbf{f}\right] = \frac{e^{f_j}}{\sum_{k=1}^{K} e^{f_k}}. \tag{4}$$

The noise injected by this method is

$$\boldsymbol{\epsilon}_{i|t} = \alpha(\bar{\mathbf{u}} - y_i), \tag{5}$$

where $\bar{\mathbf{u}}$ is the normalized $\mathbf{u}$. A similar rule is used by Ko et al. (2023). Instead of directly using the logits, Lee et al. (2022) use them to adjust α. They do so by using the entropy of the model-output distribution, assigning a smaller smoothing to high entropy samples and larger smoothing to low entropy samples. Another approach by Park et al. (2023) decrease the label noise linearly as the logit $\mathbf{f}_i(\boldsymbol{\theta}_t)$ increase. There are multiple ways to use predictions but the choice of a good strategy for a given problem is not always straightforward.

Intuitively, using model's predictions makes sense because predictions can tell us about the relevance of examples. Regions where model is inaccurate may also contain examples that need special attention but also those that are impossible to predict. Some works have explored this from the Bayesian viewpoint, although only using the posterior over the labels. For example, Pereyra et al. (2017b) motivate adaptive smoothing using Bayes error rate, which implies larger smoothing to example that lie near the decision boundary. Similarly, Ghoshal et al. (2021) use a PAC-Bayes bound to motivate adaptivity. However, there are no approaches investigating the effectiveness of posterior over $\boldsymbol{\theta}$.

In this paper, we show that directly learning the posterior using a variational method natural yields an adaptive label noise. Adaptivity introduced in this fashion directly takes various causes of uncertainty, some of which are then handled through the label noise. The uncertainty in parameter have other desired effect that are often missed when only focusing on the label noise. In our context, this can simplify the design of adaptive label smoothing or may even allow us to entirely skip the step. We will now discuss the adaptive label noise induced by variational learning.

3. Variational Learning Induces Adaptive LS

Variational learning aims to optimize for distribution over parameters $\boldsymbol{\theta}$ which is fundamentally different from traditional deep learning where we minimize empirical risk,

$$\bar{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) + \mathcal{R}_0(\boldsymbol{\theta}), \tag{6}$$

with loss $\ell_i(\boldsymbol{\theta})$ for the i'th example in the training dataset. The regularizer $\mathcal{R}_0(\theta)$ is often implicitly defined through various training choices, such as, weight-decay, initialization, and architecture design. In contrast, variational learning aims to find a distribution $q(\boldsymbol{\theta}) \in \mathcal{Q}$ which minimizes

$$\mathcal{L}(q) = \sum_{i=1}^{N} \mathbb{E}_q \left[\ell_i(\boldsymbol{\theta})\right] + \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta})]. \tag{7}$$

The second term is the Kullback-Leibler (KL) Divergence where the $p(\boldsymbol{\theta}) \propto \exp(-\mathcal{R}_0(\boldsymbol{\theta}))$ can be defined implicitly similar to deep learning. Throughout, we will set $q(\boldsymbol{\theta})$ to take Gaussian forms and show that, despite their differences, variational learning can be implicitly seen as minimizing a noisy version of Eq. 6. Existing works have studied the weight-noise (Zhang et al., 2018; Khan et al., 2018) but our goal here is to specifically study its effect on label noise.

3.1. A Simple Example: Logistic Regression

We start with logistic regression where we can write a closed-form expression for the adaptive label noise. The result extends to all loss functions using generalized linear model. We will consider all such extensions (including neural networks) afterwards. For now, we consider a loss function for binary labels $y_i \in \{0, 1\}$ with model output $f_i(\boldsymbol{\theta})$,

$$\ell_i(\boldsymbol{\theta}) = -y_i f_i(\boldsymbol{\theta}) + \log\left(1 + e^{f_i(\boldsymbol{\theta})}\right). \tag{8}$$

In logistic regression, we have $f_i(\boldsymbol{\theta}) = \boldsymbol{\phi}_i^\top \boldsymbol{\theta}$ where $\boldsymbol{\phi}_i \in \mathbb{R}^P$ is the feature vector. For simplicity, let us assume $\mathcal{R}_0(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|^2$ to be a quadratic regularizer. For such a model, we can solve Eq. 6 with gradient descent (GD),

$$\boldsymbol{\theta}_{t+1} = (1 - \rho_t)\boldsymbol{\theta}_t - \rho_t \sum_{i=1}^{N} \boldsymbol{\phi}_i \left[\sigma(f_i(\boldsymbol{\theta}_t)) - y_i\right] \tag{9}$$

The result is obtained by simply taking the derivative of Eq. 8 which gives rise to $\sigma(f) = 1/(1 + e^{-f})$, a binary version of the softmax function from Eq. 4. We will now show that, by choosing the family $\mathcal{Q}$ appropriately, variational learning can be seen as GD with label noise.

We choose the distribution $q_t(\boldsymbol{\theta})$ at iteration t to take a Gaussian form with mean $\boldsymbol{\theta}_t$ and covariance set to the identity,

$$q_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} | \boldsymbol{\theta}_t, \mathbf{I}),$$

and perform GD to minimize the variational objective in Eq. 7, now denoted as $\mathcal{L}(\boldsymbol{\theta}_t)$, with respect to $\boldsymbol{\theta}_t$. Below is a formal statement of the result.

Theorem 1 *A gradient update $\boldsymbol{\theta}_{t+t} = \boldsymbol{\theta}_t - \rho_t \nabla_{\boldsymbol{\theta}_t} \mathcal{L}(\boldsymbol{\theta}_t)$ is equivalent to the gradient update in Eq. 9 where the label y_i are replaced by $y_i + \epsilon_{i|t}$ with noise defined as*

$$\epsilon_{i|t} = \sigma(f_i(\boldsymbol{\theta}_t)) - \mathbb{E}_{q_t}[\sigma(f_i(\boldsymbol{\theta}))]. \tag{10}$$

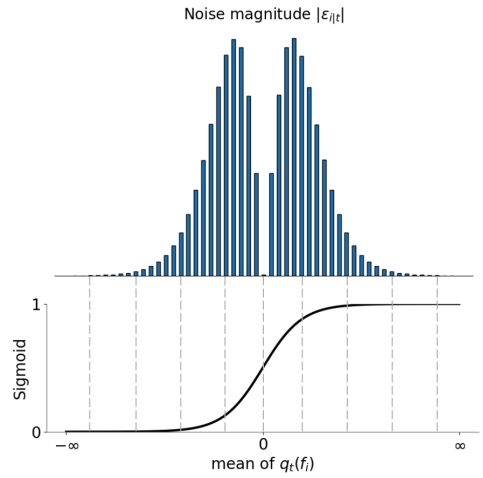

Figure 2. We plot label noise magnitude $\epsilon_{i|t}$ from Eq. 12 by varying the mean $f_{i|t}$ of $q_t(f_i)$ while fixing its variance to 1. The noise is large around 0 (but not at 0) with large peaks on both sides.

The proof is in the App. B.1.

The result shows that the GD steps to optimize Eq. 7 are equivalent to those to optimize Eq. 6 but with a noisy label. The noise is adaptive and depends on where the Gaussian distribution is located. To show this, we derive the distribution over $f_i = \boldsymbol{\phi}_i^\top \boldsymbol{\theta}$, which takes a Gaussian form:

$$q_t(f_i) = \mathcal{N}(f_i | f_{i|t}, \boldsymbol{\phi}_i^\top \boldsymbol{\phi}_i), \tag{11}$$

where we denote $f_{i|t} = \boldsymbol{\phi}_i^\top \boldsymbol{\theta}_t$. The label noise then is simply the difference between the sigmoid $\sigma(f_{i|t})$ of the mean $f_{i|t}$ and mean of $\sigma(f_i)$ with respect to $q_t(f_i)$, that is

$$\epsilon_{i|t} = \sigma(f_{i|t}) - \mathbb{E}_{q_t}[\sigma(f_i)], \tag{12}$$

where the form is similar to the noise of Zhang et al. (2021) in Eq. 5. Fig. 2 plots the magnitude of this quantity as a function of the mean $f_{i|t}$ but fixing the variance $\boldsymbol{\phi}_i^\top \boldsymbol{\phi}_i = 1$. We see the noise to be large whenever $f_{i|t}$ around 0, with the maximum in areas slightly away from it. The $\sigma(f)$ is flat far away from 0 and uncertainty in q_t is amplified around 0, which makes the difference large around 0 (but not at 0).

The other factor that affects the noise is the feature $\boldsymbol{\phi}_i$. Inputs with larger features induce larger variance. When the features are normalized, this is unlikely to have an effect, but this is important for the neural networks case where features are learned.

The two factors explain why we would expect high label noise for atypical or ambiguous examples. This is because the predictive distribution $q(f_{i|t})$ is close to 0 and may also have a higher variance. An alternate way to understand the impact of the two factors is to use a Taylor's approximation

at a sample $e \sim \mathcal{N}(0, 1)$,

$$\epsilon_{i|t} \approx \sigma(f_{i|t}) - \sigma(f_{i|t} + \|\boldsymbol{\phi}_i\|_2 e) \approx \sigma'(f_{i|t})(\boldsymbol{\phi}_i^\top \boldsymbol{\phi}_i)^{1/2} e. \tag{13}$$

We again see the two factors: one is $\sigma'(f)$ (which peaks around 0) and the other is the feature norm. Note that this approximation does not get better for larger number of samples, but it roughly captures the behavior away from 0.

3.2. Generalized Linear Model (GLM)

The result generalizes to any loss function derived using exponential-family distribution, for instance, the following generalization of Eq. 8

$$\ell_i(\boldsymbol{\theta}) = -\mathbf{y}_i^\top \mathbf{f}_i(\boldsymbol{\theta}) + A(\mathbf{f}_i(\boldsymbol{\theta})), \tag{14}$$

where $A(\mathbf{f})$ is a convex function called the log-partition function. The regularizer can also be a general convex function. For such models, we can derive the label noise following almost the same procedure as in the previous section. Due to its similarity, we omit the derivation and only give the final form of the noise,

$$\boldsymbol{\epsilon}_{i|t} = A'(\mathbf{f}_i(\boldsymbol{\theta}_t)) - \mathbb{E}_{q_t}[A'(\mathbf{f}_i(\boldsymbol{\theta}))]. \tag{15}$$

Essentially, we replace the $\sigma(f)$ by the derivative $A'(f)$. For logistic regression, $A(f) = \log(1 + e^f)$, derivative of which is $\sigma(f)$ and we recover the result in Eq. 10. We can extend this result to multiclass classification by considering $A(\mathbf{f}) = \log \sum_{k=1}^K e^{f_k}$, derivative of which is the softmax function defined in Eq. 4. Similarly to the binary case, we expect uncertainty in q_t to be amplified near the boundary. The label noise is therefore low for examples where softmax yields probabilities close to 0 or 1.

In App. A, we show a variational-learning algorithm called Variational Online Newton (VON) algorithm (Khan et al., 2018) that can be seen as a noisy-label version of Newton's method. It induces more adaptive noises because it learns the covariance $\boldsymbol{\Sigma}_t$ for a Gaussian form q_t.

3.3. Neural Network training with IVON

We will now show that the label noise expression have similar form for the neural network case, but to derive them we need to use Taylor's approximation. Essentially, the form of the expression then is similar to Eq. 13 there the adaptive nature should roughly stay the same. We validate these findings later through numerical experiments.

We will illustrate the derivation for the binary case which can then be extended to other case as we did in previous section. Taylor's approximations is required because the gradient of ℓ_i, shown below,

$$\nabla \ell_i(\boldsymbol{\theta}) = \nabla f_i(\boldsymbol{\theta}_t) \left[\sigma(f_i(\boldsymbol{\theta}_t)) - y_i\right],$$

replaces the $\boldsymbol{\phi}_i$ term in Eq. 9 by $\nabla f_i(\boldsymbol{\theta}_t)$. As a result, we cannot simply move the expectation over q_t to derive the label noise as we did in Eq. 20. However, we can simplify these by using Taylor's approximation.

We show this by using a single-sample $\boldsymbol{\theta}_t^{(1)} \sim q_t$ Monte-Carlo approximation (multiple samples can also be used),

$$\mathbb{E}_{q_t}\left[\nabla \ell_i(\boldsymbol{\theta})\right] \approx \nabla f_i(\boldsymbol{\theta}_t^{(1)}) \left[\sigma(f_i(\boldsymbol{\theta}_t^{(1)})) - y_i\right].$$

Then, we do the following two approximations where we use Taylor's expansion but ignore the second-order terms,

$$\sigma(f_i(\boldsymbol{\theta}_t^{(1)})) \approx \sigma(f_i(\boldsymbol{\theta}_t)) + \sigma'(f_i(\boldsymbol{\theta}_t))\nabla f_i(\boldsymbol{\theta}_t)(\boldsymbol{\theta}_t^{(1)} - \boldsymbol{\theta}_t)$$
$$\nabla f_i(\boldsymbol{\theta}_t^{(1)}) \approx \nabla f_i(\boldsymbol{\theta}_t)$$

With these approximations, we can write,

$$\mathbb{E}_{q_t}\left[\nabla \ell_i(\boldsymbol{\theta})\right] \approx \nabla f_i(\boldsymbol{\theta}_t) \left[\sigma(f_i(\boldsymbol{\theta}_t)) - (y_i + \epsilon_{i|t})\right],$$

where the noise takes a very similar form to Eq. 13,

$$\epsilon_{i|t} \approx \sigma'(f_i(\boldsymbol{\theta}_t))\nabla f_i(\boldsymbol{\theta}_t)\boldsymbol{\Sigma}_t^{1/2}\mathbf{e} \tag{16}$$

where $\mathbf{e}$ is a sample from a standard normal distribution. The derivation generalizes to all GLM losses by replacing $\sigma(\cdot)$ by $A'(\cdot)$. It also extends to variational GD and VON.

In practice, neural networks are trained with Adam-style algorithm. In our experiments, we will use an Adam-like version of VON, called IVON, which is recently proposed by Shen et al. (2024). The key different to VON is that it estimates a diagonal covariance by using an Adam-like preconditioning update; a pseudo-code is added in Alg. 1. The diagonal covariance is estimated through the scale vector. We will use the label noise expression given in Eq. 16 where $\boldsymbol{\Sigma}_t$ is replaced by the diagonal covariance estimated by IVON. Note that variational learning for neural neworks with IVON introduces many other noise other than label noise, for instance, the noise is introduced in the features $\nabla f(\boldsymbol{\theta})$, as shown above. We will analyze only the label noise but the performance is affected by other noises too.

In our experiments, we also compare to Sharpness-Aware Minimization (SAM) (Foret et al., 2021) which has a variational interpretation (Möllenhoff & Khan, 2023) and has been shown to perform well with mislabelled data. Using our techniques, it is possible to derive the label noise of SAM but the expression would be similar to the one derived here. The difficulty with SAM is that we need to tune the 'size' of adversarial perturbation, often denoted by a scalar ρ, while IVON can automatically estimate it using the posterior variance. In our experiments, we show that IVON performs comparably to SAM with a highly tuned ρ, and it does not need to set any such hyperparameters.

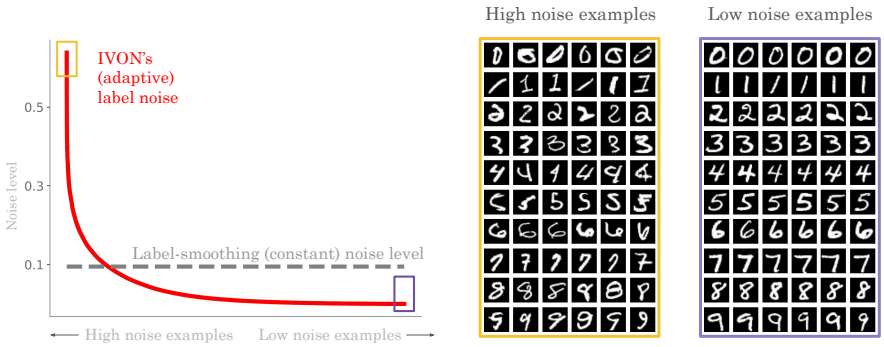

Figure 3. Label noise assigned by IVON and LS in MNIST dataset. Examples are ordered according to IVON's noise, and highest and lowest noise examples are visualized. We see that high noise is assigned to atypical examples while low noise is assigned to regular ones.

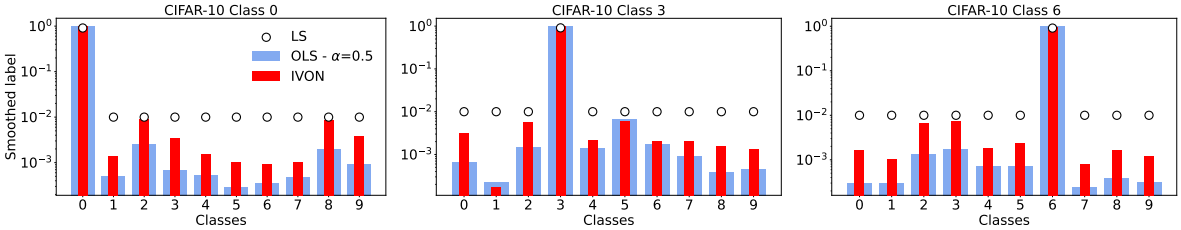

Figure 4. Smoothed label comparison among IVON, LS and Online Label Smoothing (OLS) in CIFAR-10. IVON has a similar adaptive label smoothing effect as OLS. α is the smoothing rate defined in Eq. 1. Y-axis is in the log scale.

4. Experiments

We do extensive experiments to show adaptive label noise via variational learning and its benefits. In Sec. 4.1, we show that IVON adapts the label noise for each example, and generally assigns higher noise magnitude to ambiguous ones. In Sec. 4.2, we show that IVON's smoothed labels are similar to an existing adaptive smoothing method (Zhang et al., 2021). In Sec. 4.3, we show that IVON consistently outperforms LS when datasets have labeling errors in various settings. Additional experiments are reported in App. C, and experiment details are reported in App. D.

4.1. IVON's Adaptive Label Noise

We demonstrate IVON label noise's adaptivity on MNIST dataset (LeCun & Cortes, 2010). We plot IVON's label noise distribution in Fig. 3, which shows that IVON adds different label noise on each example whereas traditional Label Smoothing defines a uniformly distributed noise for all. By further visualizing the data, we see that IVON induces stronger noise to unclear examples, which prevent models from being overconfident in these datapoints.

4.2. Comparisons to Existing Adaptive LS Strategies

In this section, we show that IVON's label smoothing is similar to an adaptive method called Online Label Smoothing (OLS) (Zhang et al., 2021). In the CIFAR-10 and CIFAR-100 dataset (Krizhevsky & Hinton, 2009), we compare the smoothed labels of IVON with traditional LS (Szegedy et al., 2016) and Online Label Smoothing (OLS) (Zhang et al., 2021). OLS adjusts the label noise according to the model's predictions, as described in Sec. 2. Fig. 4 shows the CIFAR-10 results, and CIFAR-100 results are in App. C.1. IVON has surprisingly similar smoothed label distributions as the OLS in both datasets, while IVON tends to induce stronger label noises. Variational learning's adaptive label smoothing is similar to existing work's, without needing any additional effort to design or estimate the adaptive label noise.

4.3. Comparisons on Datasets with Labeling Errors

We compare IVON to Label Smoothing (LS) (Szegedy et al., 2016) and SAM (Foret et al., 2021) in presence of labeling errors, and the results show that IVON consistently outperforms LS in various settings. To find the best performance of the baselines, we tune several LS's smoothing rates α (defined in Eq. 1), and various SAM's adversarial perturbation size ρ (discussed in Sec. 3.3). We conduct studies on benchmark datasets with synthetic noise, where the noise level can be adjusted. We also designed a specific scenario where labeling errors are dependent on the classes in Sec. 4.3.2. In App. C.3, we evaluate on datasets with natural noise, where the noise level is fixed and unknown. For synthetic noise

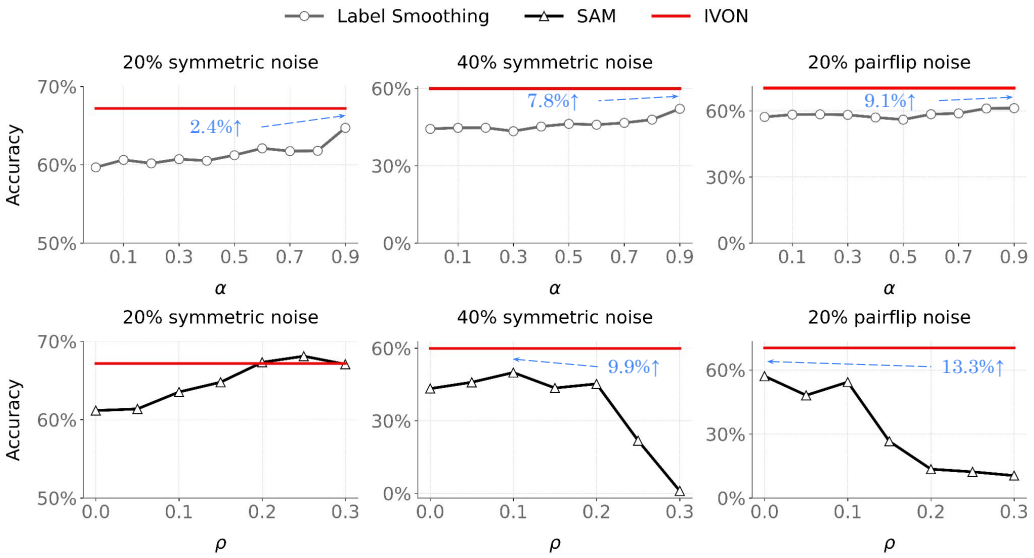

Figure 5. Results on CIFAR-100 with symmetric noisy labels. Top: IVON outperforms Label Smoothing (LS) with different smoothing rates α. Down: IVON has comparable results with SAM peak performances, while SAM is sensitive to the choice of perturbation ρ. Accuracy improvements are shown in blue. Results are reported over 5 random seeds.

experiments, we use the CIFAR-10 and CIFAR-100 datasets (Yu et al., 2019). For natural noise experiments in App. C.3, we use the benchmark Clothing1M (Xiao et al., 2015). All datasets include a clean test set.

4.3.1. Synthetic Noisy Datasets

We consider two commonly used corruptions (Patrini et al., 2017; Li et al., 2019; Yu et al., 2019): Symmetric flipping and Pair flipping. In symmetric flipping, a true label is replaced by a randomly generated class with a probability. In pair flipping, it tries to mimic real world mistakes for similar classes, where a true label is replaced by the next class with a probability. For training dataset, we use previous work's (Yu et al., 2019) code to generate noisy labels. More experiment details are in App. D.2.

In CIFAR-100, Fig. 5 shows that IVON outperforms Label Smoothing and SAM in different scenarios. We also observe that SAM is sensitive to the choice of ρ, while IVON does not need to tune any hyperparameters to perform well. More experiment results in CIFAR-10 and the comparison with fixed diagonal variance are in App. C.2.

4.3.2. Data Dependent Labeling Errors

In this section, we try to understand the adaptivity of these methods in the data-dependent noisy dataset. When each class has different noise levels, we expect LS will fail since it adds uniform noises to all classes, while IVON's adaptivity makes it stand. The experiment setting is in App. D.3.

The experiment results are in Fig. 6. IVON outperforms LS and SAM in all noise levels. Meanwhile, IVON can learn in very noisy scenarios $\kappa = \{0.4, 0.5\}$ while baselines can only reach around 10% accuracy. The results support our claim that adaptive label noise induced by variational learning is more effective than traditional label smoothing.

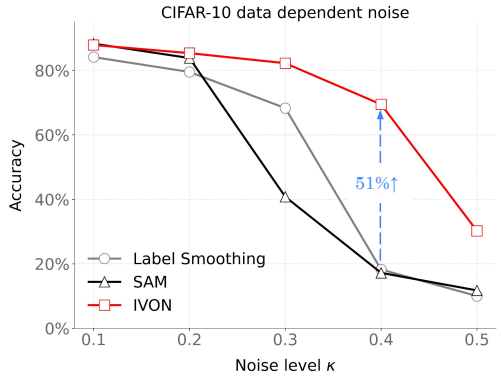

Figure 6. Results for CIFAR-10 with data dependent noise. IVON outperforms LS and SAM in all noise levels. Furthermore, IVON can learn extremely noisy scenario, while LS and SAM cannot.

5. Conclusion

In this paper, we show that variational learning induces an adaptive label smoothing similar to an existing adaptive approach (Zhang et al., 2021) but does not require any additional effort to design. We derive the exact form for simple models and extend them to neural networks. We empirically confirm the effectiveness of noise, showing that the IVON

method consistently performs better than LS, and comparably to SAM, without requiring hyperparameters to achieve desired smoothing. Our work suggests that Bayesian frameworks are naturally suitable for label noise. Specifically, we believe that variational learning algorithms, such as IVON, provide a flexible framework to further add noise to handle both the abnormalities and typicalities in the data.

Acknowledgements

Mohammad Emtiyaz Khan was supported by the Bayes duality project, JST CREST Grant Number JPMJCR2112. Some of the experiments were carried out using the TSUBAME4.0 supercomputer at Institute of Science Tokyo.

We also would like to thank our former post-doc Lu Xu who started this project and contributed heavily to it. For reasons beyond our control, Lu was unable to join the author list but this work would not have been possible without her work. We also would like to thank Thomas Möllenhoff (RIKEN) for his help in deriving some of the results.

Author Contributions Statement

Author list: Sin-Han Yang (SHY), Zhedong Liu (ZL), Gian Maria Marconi (GMM) and Mohammad Emtiyaz Khan (MEK).

MEK proposed the label-noise idea which was developed by Lu Xu and GMM who also performed the first set of experiments to validate the initial hypothesis. Subsequently, ZL extended these experiments to the Clothing 1M settings. Building up on all these experiments, SHY did most of the experiments involving IVON which are currently in the final version of the paper. MEK wrote the theory in Sec 3 with help from SHY. Both SHY and MEK contributed to the final writing with feedback from ZL and GMM.

References

Damian, A., Ma, T., and Lee, J. D. Label noise SGD provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021. 2

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 4, 5, 14

Ghoshal, A., Chen, X., Gupta, S., Zettlemoyer, L., and Mehdad, Y. Learning better structured representations using low-rank adaptive label smoothing. In *International Conference on Learning Representations*, 2021. 1, 2

Khan, M., Nielsen, D., Tangkaratt, V., Lin, W., Gal, Y., and Srivastava, A. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. In *International conference on machine learning*, pp. 2611–2620. PMLR, 2018. 3, 4, 9

Khan, M. E. and Rue, H. The Bayesian learning rule. *Journal of Machine Learning Research*, 24(281):1–46, 2023. 9

Khan, M. E., Immer, A., Abedi, E., and Korzepa, M. Approximate inference turns deep networks into Gaussian processes. *Advances in Neural Information Processing Systems*, 2019. 10

Ko, J., Yi, B., and Yun, S.-Y. A gift from label smoothing: robust training with adaptive label smoothing via auxiliary classifier under label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 1, 2

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 5

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/. 5

Lee, D., Cheung, K. C., and Zhang, N. L. Adaptive label smoothing with self-knowledge in natural language generation. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pp. 9781–9792. Association for Computational Linguistics, 2022. 1, 2

Li, J., Wong, Y., Zhao, Q., and Kankanhalli, M. S. Learning to learn from noisy labeled data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pp. 5051–5059. Computer Vision Foundation / IEEE, 2019. 6

Liu, Y. Understanding instance-level label noise: Disparate impacts and treatments. In *International Conference on Machine Learning*, pp. 6725–6735. PMLR, 2021. 2

Lukasik, M., Bhojanapalli, S., Menon, A. K., and Kumar, S. Does label smoothing mitigate label noise? In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6448–6458. PMLR, 2020. 2

Möllenhoff, T. and Khan, M. E. SAM as an optimal relaxation of bayes. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023. 4

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019. 2

Nickl, P., Xu, L., Tailor, D., Möllenhoff, T., and Khan, M. E. The memory-perturbation equation: Understanding model's sensitivity to data. *Advances in Neural Information Processing Systems*, 36, 2024. 10

Park, H., Noh, J., Oh, Y., Baek, D., and Ham, B. ACLS: Adaptive and conditional label smoothing for network calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3936–3945, 2023. 1, 2

Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 2233–2241. IEEE Computer Society, 2017. 2, 6

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., and Hinton, G. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017a. 2

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., and Hinton, G. E. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017b. 2

Shen, Y., Daheim, N., Cong, B., Nickl, P., Marconi, G. M., Bazan, C., Yokota, R., Gurevych, I., Cremers, D., Khan, M. E., and Möllenhoff, T. Variational learning is effective for large deep networks. In *Forty-first International Conference on Machine Learning, ICML 2024*. OpenReview.net, 2024. 1, 4, 11, 14

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016. 1, 5, 11, 14

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 6, 13, 14

Xu, M., Lee, J., Yoon, S., and Park, D. S. Adaptive label smoothing for out-of-distribution detection. *arXiv preprint arXiv:2410.06134*, 2024. 1, 2

Yu, X., Han, B., Yao, J., Niu, G., Tsang, I. W., and Sugiyama, M. How does disagreement help generalization against label corruption? In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97, pp. 7164–7173. PMLR, 2019. 6

Zhang, C.-B., Jiang, P.-T., Hou, Q., Wei, Y., Han, Q., Li, Z., and Cheng, M.-M. Delving deep into label smoothing. *IEEE Transactions on Image Processing*, 30:5984–5996, 2021. 1, 2, 3, 5, 6, 11, 13

Zhang, G., Sun, S., Duvenaud, D., and Grosse, R. Noisy natural gradient as variational inference. In *International conference on machine learning*, pp. 5852–5861. PMLR, 2018. 3

A. Generalized Linear Model with Newton's Method

We now discuss the Newton's method and show that a specific variational-learning algorithm can be seen as a noisy-label version of Newton's method. Here, we find that the form of the noise has exactly same form as Eq. 15 but the distribution q_t has a flexible covariance which improves the adaptivity of the label noise.

We consider the following Newton's update,

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \left[\nabla^2 \bar{\ell}(\boldsymbol{\theta}_t)\right]^{-1} \nabla \bar{\ell}(\boldsymbol{\theta}_t) \tag{17}$$

which is commonly used for generalized linear models. As shown by Khan & Rue (2023), the update can be seen as a special case of a Variational Online Newton (VON) algorithm (Khan et al., 2018) to learn a full Gaussian with covariance $\boldsymbol{\Sigma}_t$,

$$q_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\theta}_t, \boldsymbol{\Sigma}_t)$$

The VON updates are given as follows,

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \rho_t \boldsymbol{\Sigma}_{t+1} \mathbb{E}_{q_t}[\nabla \bar{\ell}(\boldsymbol{\theta})] \\ \boldsymbol{\Sigma}_{t+1}^{-1} &= (1-\rho_t)\boldsymbol{\Sigma}_t^{-1} + \rho_t \mathbb{E}_{q_t}[\nabla^2 \bar{\ell}(\boldsymbol{\theta})]. \end{aligned} \tag{18}$$

Setting $\rho_t = 1$ yields a Newton-like update where gradients $\nabla \bar{\ell}$ and Hessian $\nabla^2 \bar{\ell}$ are replaced by their terms where expectations are taken, namely, $\mathbb{E}_{q_t}[\nabla \bar{\ell}]$ and $\mathbb{E}_{q_t}[\nabla^2 \bar{\ell}]$. Similarly to the previous cases, the label noise in VON arises due to the expectation of the gradient, while expectation of the Hessian gives rise to other types of noise.

As shown in in App. B.2, the VON updates in Eq. 18 are equivalent to Newton's update in Eq. 17 where labels are replaced by the noisy ones with noise shown in Eq. 15. The proof technique relies on comparing the form of the surrogates for the two algorithms. Even though the noise has the same form, there is an important difference here. Essentially, the Gaussian q_t now is more flexible because its covariance $\boldsymbol{\Sigma}_t$ is not fixed but learned using the Hessian. As a result the distribution over f_i now has adaptive variances,

$$q_t(f_i) = \mathcal{N}(f_i | f_{i|t}, \boldsymbol{\phi}_i^\top \boldsymbol{\Sigma}_t \boldsymbol{\phi}_i). \tag{19}$$

Therefore, now both the location and spread of the Gaussians are changed for each example, and they both contribute to the adaptivity. The result shows that second-order methods yield more adaptive label noise than first order methods, and are expected to perform better in practice. We will later present experiments that support this finding.

B. Derivations

B.1. Proof of the Logistic Regression's Noise

Proof: The gradient of the expected loss in Eq. 7 can be simplified to take a form very similar to the one in Eq. 9,

$$\begin{aligned} \nabla_{\boldsymbol{\theta}_t} \mathbb{E}_{q_t}[\ell_i(\boldsymbol{\theta})] &= \nabla_{\boldsymbol{\theta}_t} \mathbb{E}_{\mathcal{N}(\mathbf{e}|0,\mathbf{I})}[\ell_i(\boldsymbol{\theta}_t + \mathbf{e})] \\ &= \mathbb{E}_{\mathcal{N}(\mathbf{e}|0,\mathbf{I})}\left[\nabla_{\boldsymbol{\theta}_t} \ell_i(\boldsymbol{\theta}_t + \mathbf{e})\right] \\ &= \boldsymbol{\phi}_i \left[\mathbb{E}_{q_t}[\sigma(f_i(\boldsymbol{\theta}))] - y_i\right] \end{aligned} \tag{20}$$

The gradient of KL also simplifies to

$$\mathbb{D}_{\text{KL}}[q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta})] = \mathbb{E}_q\left[\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}\right] = \tfrac{1}{2}\|\boldsymbol{\theta}\|^2 + \text{const.}$$

Using these, we can write the GD to minimize Eq. 7 as

$$\boldsymbol{\theta}_{t+1} = (1-\rho_t)\boldsymbol{\theta}_t - \rho_t \sum_{i=1}^{N} \boldsymbol{\phi}_i \left[\mathbb{E}_{q_t}[\sigma(f_i(\boldsymbol{\theta}))] - y_i\right],$$

which has a similar form as Eq. 9 but with one difference: $\sigma(f_i(\boldsymbol{\theta}))$ are replaced by their expectation over q (highlighted in red). By adding and subtracting $\sigma(f_i(\boldsymbol{\theta}_t))$, we can rewrite the update as Eq. 9 which has the label noise defined in Eq. 10. ■

B.2. Derivation of GLM with Newton's method

Newton's update shown in Eq. 17 is equivalent to the following surrogate minimization,

$$\boldsymbol{\theta}_{t+1} = \arg\min_{\boldsymbol{\theta}} \; \boldsymbol{\theta}^\top \nabla \bar{\ell}(\boldsymbol{\theta}_t) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top \nabla^2 \bar{\ell}(\boldsymbol{\theta}_t)(\boldsymbol{\theta} - \boldsymbol{\theta}_t). \tag{21}$$

This can be verified by simply taking the derivative of the above objective and setting it to zero. We will now show that the VON update give rise to a similar surrogate but where the gradient and Hessian are replaced by their expected values.

To do so, we use the result of Khan et al. (2019) who show that each step of VON algorithm can be seen as inference on a linear model. Essentially, the VON update can be expressed as follows (see Nickl et al. (2024, App. C.3) for a derivation):

$$\begin{aligned} q_{t+1}(\boldsymbol{\theta}) &\propto q_t(\boldsymbol{\theta})^{1-\rho_t} \prod_{i=1}^N e^{\rho_t \left(\boldsymbol{\theta}^\top \mathbb{E}_{q_t}[-\nabla \ell_i(\boldsymbol{\theta}) + \nabla^2 \ell_i(\boldsymbol{\theta})\boldsymbol{\theta}_t] - \frac{1}{2}\boldsymbol{\theta}^\top \mathbb{E}_{q_t}[\nabla^2 \ell_i(\boldsymbol{\theta})]\boldsymbol{\theta} \right)} \\ &\propto q_t(\boldsymbol{\theta})^{1-\rho_t} \prod_{i=1}^N e^{-\rho_t \left(\boldsymbol{\theta}^\top \mathbb{E}_{q_t}[\nabla \ell_i(\boldsymbol{\theta})] + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top \mathbb{E}_{q_t}[\nabla^2 \ell_i(\boldsymbol{\theta})](\boldsymbol{\theta} - \boldsymbol{\theta}_t) \right)}, \end{aligned} \tag{22}$$

where we subtracted $\boldsymbol{\theta}_t^\top \mathbb{E}_{q_t}[\nabla^2 \ell_i(\boldsymbol{\theta})]\boldsymbol{\theta}_t$ and completed the square. This is a constant which is absorbed in the normalizing constant of q_{t+1}. From here, we can simply match the mode $\boldsymbol{\theta}_{t+1}$ of q_{t+1} to the mode of the right hand side. For $\rho_t = 1$, this gives us the following minimization problem to recover $\boldsymbol{\theta}_{t+1}$:

$$\boldsymbol{\theta}_{t+1} = \arg\min_{\boldsymbol{\theta}} \; \boldsymbol{\theta}^\top \mathbb{E}_{q_t}[\nabla \bar{\ell}(\boldsymbol{\theta})] + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top \mathbb{E}_{q_t}[\nabla^2 \bar{\ell}(\boldsymbol{\theta})](\boldsymbol{\theta} - \boldsymbol{\theta}_t). \tag{23}$$

This shows that, for $\rho_t = 1$, VON updates can be seen as Newton step where gradient and Hessian are replaced by their expected values. The proof is identical to the one shown in the main tex, therefore we omit it.

Theorem 2 *For the loss function of Eq. 8, the VON update in Eq. 18 with $\rho_t = 1$ is equivalent to Newton's update in Eq. 17 but where the label y_i are replaced by $y_i + \epsilon_{i|t}$ with noise defined as*

$$\epsilon_{i|t} = \sigma\left(f_{i|t}\right) - \mathbb{E}_{\mathcal{N}(e|0,1)}\left[\sigma\left(f_{i|t} + e\sqrt{\boldsymbol{\phi}_i^\top \boldsymbol{\Sigma}_t \boldsymbol{\phi}_i}\right)\right], \tag{24}$$

and the Hessian $\nabla^2 \bar{\ell}(\boldsymbol{\theta}_t)$ is replaced by its noisy version $\mathbb{E}_{q_t}[\nabla^2 \bar{\ell}(\boldsymbol{\theta})]$.

B.3. IVON pseudo code

The pseudo-code is given in Alg. 1.

Algorithm 1 Improved Variational Online Newton (IVON) (Shen et al., 2024).

Require: Learning rates $\{\alpha_t\}$, weight-decay $\delta > 0$.
Require: Momentum parameters $\beta_1, \beta_2 \in [0, 1)$.
Require: Hessian init $h_0 > 0$.
Init: $\mathbf{m} \leftarrow$ (NN-weights), $\mathbf{h} \leftarrow h_0$, $\mathbf{g} \leftarrow 0$, $\lambda \leftarrow N$.
Init: $\boldsymbol{\sigma} \leftarrow 1/\sqrt{\lambda(\mathbf{h} + \delta)}$.
Optional: $\alpha_t \leftarrow (h_0 + \delta)\alpha_t$ for all t.

for $t = 1, 2, \ldots$ **do**
  $\widehat{\mathbf{g}} \leftarrow \widehat{\nabla}\bar{\ell}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} \sim q$
  $\widehat{\mathbf{h}} \leftarrow \widehat{\mathbf{g}} \cdot (\boldsymbol{\theta} - \mathbf{m})/\boldsymbol{\sigma}^2$
  $\mathbf{g} \leftarrow \beta_1 \mathbf{g} + (1 - \beta_1)\widehat{\mathbf{g}}$
  $\mathbf{h} \leftarrow \beta_2 \mathbf{h} + (1 - \beta_2)\widehat{\mathbf{h}} + \frac{1}{2}(1 - \beta_2)^2(\mathbf{h} - \widehat{\mathbf{h}})^2/(\mathbf{h} + \delta)$
  $\bar{\mathbf{g}} \leftarrow \mathbf{g}/(1 - \beta_1^t)$
  $\mathbf{m} \leftarrow \mathbf{m} - \alpha_t(\bar{\mathbf{g}} + \delta\mathbf{m})/(\mathbf{h} + \delta)$
  $\boldsymbol{\sigma} \leftarrow 1/\sqrt{\lambda(\mathbf{h} + \delta)}$
end for
return $\mathbf{m}, \boldsymbol{\sigma}$

C. Additional Experiments

C.1. Comparisons on CIFAR-100 Smoothed Label Distributions

In Fig. 7, we visualize the label noise distributions of IVON (Shen et al., 2024), traditional LS (Szegedy et al., 2016) and Online Label Smoothing (OLS) (Zhang et al., 2021).

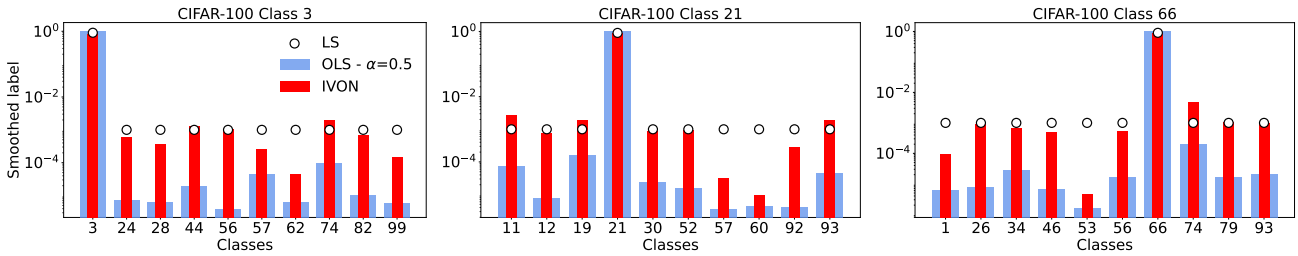

Figure 7. Smoothed label comparison among IVON, LS and Online Label Smoothing (OLS) in CIFAR-100. Results are similar to CIFAR-10 results in Fig. 4. Y-axis is in the log scale. We randomly pick 10 classes for CIFAR-100 due to image size limit.

C.2. Synthetic Noisy Datasets: More Experiments

In CIFAR-10, Fig. 8 shows similar trends as CIFAR-100 results in Fig. 5. Meanwhile, we test the effectiveness of flexible $\mathbf{\Sigma}_t$ by comparing it with the fixed diagonal variance. Fig. 9 shows that learned $\mathbf{\Sigma}_t$ consistently outperforms fixed $\mathbf{\Sigma}_t$ in three noisy datasets. The experiment results demonstrate the importance of flexible $\mathbf{\Sigma}_t$ as stated in App. A.

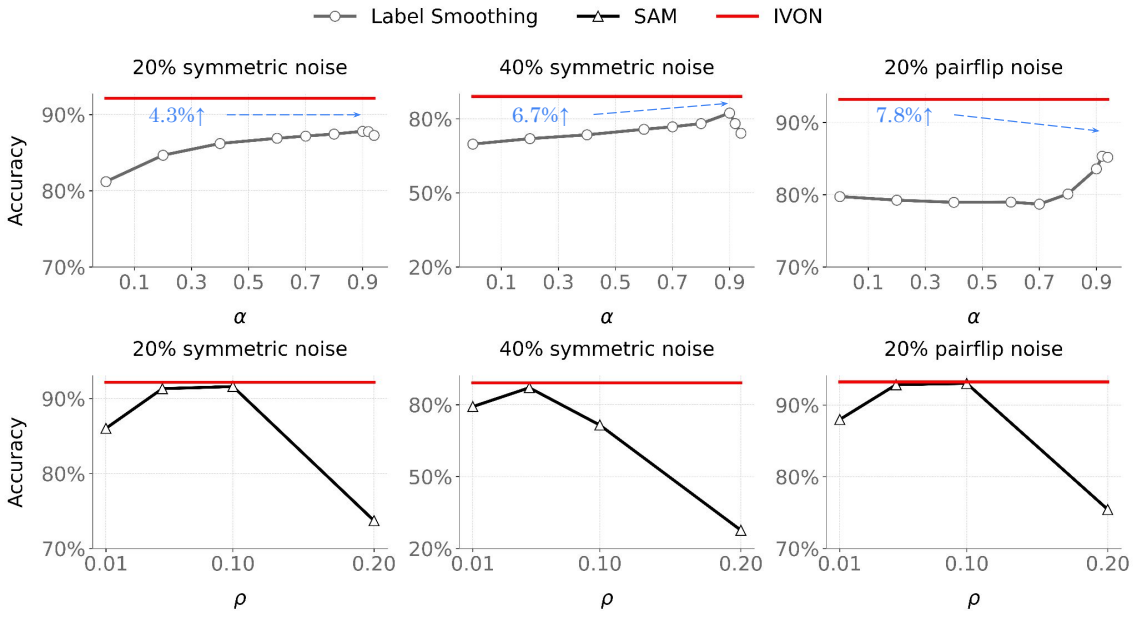

Figure 8. Results on CIFAR-10 with symmetric noisy labels over 5 random seeds, which is similar to CIFAR-100 results in Fig. 5.

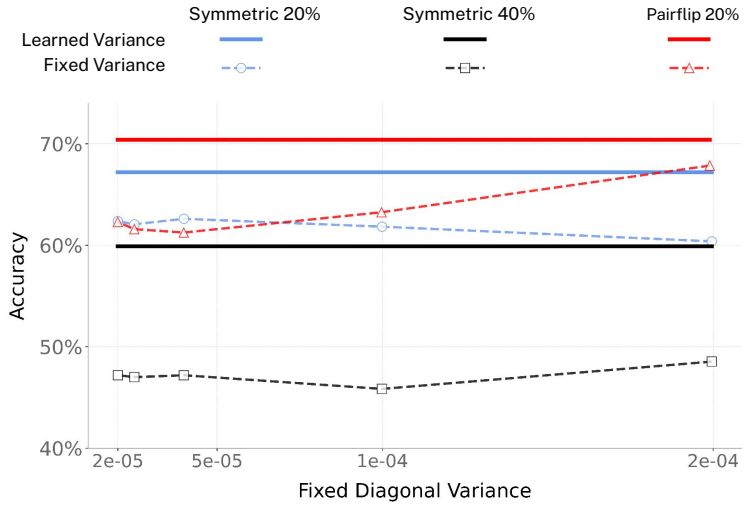

Figure 9. In synthetic noisy datasets of CIFAR-100, we test IVON with multiple fix diagonal variance $\mathbf{\Sigma}_t$. The fixed diagonal $\mathbf{\Sigma}_t$ is worse than learned diagonal variance in all datasets.

C.3. Uncontrolled Noisy Datasets

We now report results on Clothing1M (Xiao et al., 2015), a large-scale dataset that features natural label noise from the web and consists of 1 million images across 14 categories. We conduct experiments by using ResNet-50 as the model.

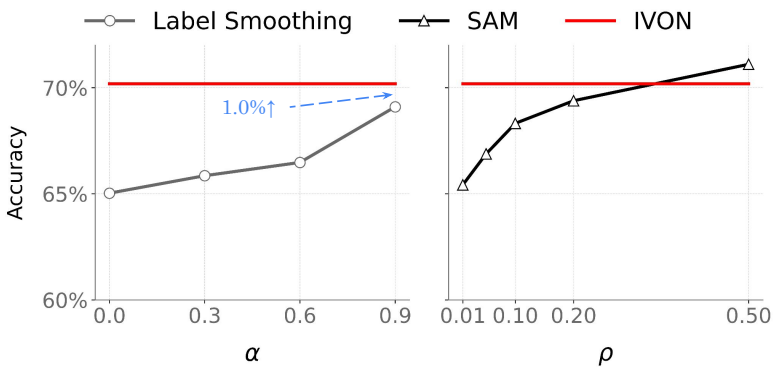

Figure 10. Clothing 1M experiment result is similar to synthetic noisy datasets reported in Sec. 4.3.1. Results are reported over 5 seeds.

The results on Clothing1M, illustrated in Fig. 10, demonstrate that IVON outperforms Label Smoothing and is comparable to SAM. This experiment shows that IVON's performance is consistent in the large scale dataset.

C.4. Hessian Initialization

We analyze how Hessian initialization h_0 of IVON affects the accuracy. The results are in Fig. 11. IVON's accuracy can only vary by up to 10% when the Hessian is bigger than 0.05, and this variation is less sensitive compared to SAM's sensitivity to ρ, as shown in Fig. 8, Fig. 5 and Fig. 10.

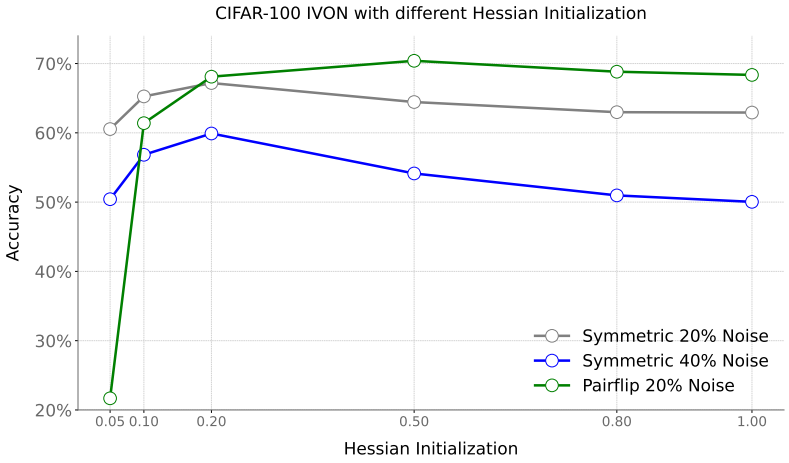

Figure 11. Results for IVON on CIFAR-100 with multiple Hessian initializations. IVON's accuracy is consistent when having different Hessian initializations.

D. Experiment details

D.1. Experiment Details of Sec. 4.1 and Sec. 4.2

In Fig. 3, we test IVON on a 3-layers convolutional neural networks. In Fig. 4, we do experiments on ResNet-34 model. We uses the PyTorch implementation verison[1] of Online Label Smoothing (Zhang et al., 2021).

[1] https://github.com/ankandrew/online-label-smoothing-pt

D.2. Experiments on Synthetic Noisy Datasets

For pairflip setting in CIFAR-10, the classes flipping order is: AIRPLANE → AUTOMOBILE → BIRD → CAT → DEER → DOG → FROG → HORSE → SHIP → TRUCK → AIRPLANE. In CIFAR-10 experiments, we train a ResNet 34 for 200 epochs with batch size set to 50 and weight decay set to 0.001. For SAM and LS, we set initial learning rate as 0.05 and reduce it by 0.1 at 100 epoch and 150 epoch, following hyper-parameters from previous papers. For IVON (Shen et al., 2024), we follow the original paper to set initial learning rate as 0.2 and anneal the learning rate to zero with a cosine schedule after a linear warmup phase over 5 epochs. We set momentum to 0.9 for all methods, and hessian momentum β_2 to $1 - e^{-5}$, hessian initial h_0 to 0.9, scaling parameter λ to the number of training data for IVON. For SAM, we follow the original paper (Foret et al., 2021) and choose best neighborhood size ρ from $[0.01, 0.05, 0.1, 0.2, 0.5]$. In CIFAR-100 experiments, we tune the hyperparamters to the best for each method. The hyperparameters are specified in Table 1.

In Fig. 9, we fix the Hessian of IVON by setting $\beta_2 = 1$ in Line 5 of Alg. 1. Therefore, standard deviation $\boldsymbol{\sigma}$ defined in Line 8 is fixed since Hessian $\mathbf{h}$ is fixed.

Table 1. Hyperparamters of each method for CIFAR-100. We denote learning rate as lr, Hessian Initialization as Hessian init.

	Symmetric 20%	Symmetric 40%	Pairflip 20%
Weight decay	2e-4	2e-4	5e-4
LS (Szegedy et al., 2016) lr	0.1	0.1	0.1
SAM (Foret et al., 2021) lr	0.05	0.1	0.1
IVON (Shen et al., 2024) lr	0.8	0.8	0.5
IVON (Shen et al., 2024) Hessian init	0.2	0.2	0.5

D.3. Data Dependent Labeling Errors Setting

First, we create a new transition matrix P of noisy label $\mathbf{y}' = P\mathbf{y}$, where $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^K, P \in \mathbb{R}^{K \times K}$. We inject difference noise level to each class, so the noise level of each class is different:

$$P_{i,i} = 1 - (\kappa + \beta i), i \in [1, K]. \tag{25}$$

where κ is the starting noise level and β is the increase factor. Afterwards, we give the same transition probability to the rest of the wrong classes:

$$P_{i,j} = \frac{\kappa + \beta i}{K - 1}, i, j \in [1, K], i \neq j. \tag{26}$$

In experiments, we follow the hyperparameters in CIFAR-10 synthetic noise experiment from Sec. 4.3.1. For LS, we run smoothing rate $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ and report the best accuracy. For SAM, we run ρ for $\{0, 0.05, 0.1, 0.15, 0.2, 0.5\}$ and report the best accuracy. The experiment results are reported for $\kappa = \{0.1 \sim 0.5\}$ and $\beta = 0.05$.

D.4. Clothing 1M Details

The noisy labels in Clothing1M (Xiao et al., 2015) are derived from the text surrounding the images on the web. In constructing the dataset, noisy labels are assigned to images based on this contextual text.