Supplementary material: Improving Generalization by Controlling Label-Noise Information in Neural Network Weights

Algorithm 1 LIMIT: limiting label information memorization in training.
Our implementation is available at https://github.com/hrayrhar/limit-label-memorization.

```
Input: Training dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$.
Input: Gradient norm regularization coefficient $\beta$. $\{\lambda$ is set to $1\}$
Initialize classifier $f(y \mid x, w)$ and gradient predictor $q_\phi(\cdot \mid \mathbf{x}, g_{<t})$.
for $t = 1..T$ do
    Fetch the next batch $(x_t, y_t)$ and compute the predicted logits $a_t$.
    Compute the cross-entropy gradient, $g_t^{\mathcal{L}} \leftarrow s(a_t) - y_t$.
    if sampling of gradients is enabled then
        $g_t \sim q_\phi(\cdot \mid \mathbf{x}, g_{<t})$.
    else
        $g_t \leftarrow \mu_t$ {the mean of predicted gradient}
    end if
    Starting with $g_t$, backpropagate to compute the gradient with respect to $w$.
    Update $w_{t-1}$ to $w_t$.
    Update $\phi$ using the gradient of the following loss: $-\log q_\phi(\tilde{g}_t^{\mathcal{L}} \mid \mathbf{x}, g_{<t}) + \beta \|\mu_t\|_2^2$.
end for
```

A. Proofs

This section presents the proofs and some remarks that were not included in the main text due to space constraints.

A.1. Proof of Thm. 2.1

Theorem A.1. **(Thm. 2.1 restated)** Consider a dataset $S = (\mathbf{x}, \mathbf{y})$ of n i.i.d. samples, $\mathbf{x} = \{x^{(i)}\}_{i=1}^n$ and $\mathbf{y} = \{y^{(i)}\}_{i=1}^n$, where the domain of labels is a finite set, $\mathcal{Y}$, with $|\mathcal{Y}| > 2$. Let $\mathcal{A}(w \mid S)$ be any training algorithm, producing weights for possibly stochastic classifier $f(y \mid x, w)$. Let $\widehat{y}^{(i)}$ denote the prediction of the classifier on i-th example and $e^{(i)} = \mathbb{1}\{\widehat{y}^{(i)} \neq y^{(i)}\}$ be a random variable corresponding to predicting $y^{(i)}$ incorrectly. Then, the following holds

$$\mathbb{E}\left[\sum_{i=1}^n e^{(i)}\right] \geq \frac{H(\mathbf{y} \mid \mathbf{x}) - I(w; \mathbf{y} \mid \mathbf{x}) - \sum_{i=1}^n H(e^{(i)})}{\log\left(|\mathcal{Y}| - 1\right)}.$$

Proof. For each example we consider the following Markov chain:

$$y^{(i)} \rightarrow \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \rightarrow \begin{bmatrix} x^{(i)} \\ w \end{bmatrix} \rightarrow \widehat{y}^{(i)}.$$

In this setup Fano's inequality gives a lower bound for the error probability:

$$H(e^{(i)}) + \mathbb{P}(e^{(i)} = 1) \log\left(|\mathcal{Y}| - 1\right) \geq H(y^{(i)} \mid x^{(i)}, w), \tag{13}$$

which can be written as:

$$\mathbb{P}(e^{(i)} = 1) \geq \frac{H(y^{(i)} \mid x^{(i)}, w) - H(e^{(i)})}{\log\left(|\mathcal{Y}| - 1\right)}.$$

Summing this inequality for $i = 1, \dots, n$ we get

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{P}(e^{(i)} = 1) &\geq \frac{\sum_{i=1}^{n} \left(H(y^{(i)} \mid x^{(i)}, w) - H(e^{(i)})\right)}{\log\left(|\mathcal{Y}| - 1\right)} \\ &\geq \frac{\sum_{i=1}^{n} \left(H(y^{(i)} \mid \mathbf{x}, w) - H(e^{(i)})\right)}{\log\left(|\mathcal{Y}| - 1\right)} \\ &\geq \frac{H(\mathbf{y} \mid \mathbf{x}, w) - \sum_{i=1}^{n} H(e^{(i)})}{\log\left(|\mathcal{Y}| - 1\right)}.\end{aligned}$$

The correctness of the last step follows from the fact that total correlation is always non-negative (Cover & Thomas, 2006):

$$\sum_{i=1}^{n} H(y^{(i)} \mid \mathbf{x}, w) - H(\mathbf{y} \mid \mathbf{x}, w) = \text{TC}(\mathbf{y} \mid \mathbf{x}, w) \geq 0.$$

Finally, using the fact that $H(\mathbf{y} \mid \mathbf{x}, w) = H(\mathbf{y} \mid \mathbf{x}) - I(w; \mathbf{y} \mid \mathbf{x})$, we get that the desired result:

$$\mathbb{E}\left[\sum_{i=1}^{n} e^{(i)}\right] \geq \frac{H(\mathbf{y} \mid \mathbf{x}) - I(w; \mathbf{y} \mid \mathbf{x}) - \sum_{i=1}^{n} H(e^{(i)})}{\log\left(|\mathcal{Y}| - 1\right)}. \tag{14}$$

□

A.2. Proof of Prop. 3.1

Proposition A.1. (Prop. 3.1 restated) If $g_t = \mu_t + \epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, \sigma_q^2 I_d)$ is an independent noise and $\mathbb{E}\left[\mu_t^T \mu_t\right] \leq L^2$, then the following inequality holds:

$$I(g_t; \mathbf{y} \mid \mathbf{x}, g_{<t})) \leq \frac{d}{2} \log\left(1 + \frac{L^2}{d\sigma_q^2}\right).$$

Proof. Given that ϵ_t and μ_t are independent, let us bound the expected L2 norm of g_t:

$$\begin{aligned}\mathbb{E}\left[g_t^T g_t\right] &= \mathbb{E}\left[(\epsilon_t + \mu_t)^T (\epsilon_t + \mu_t)\right] \\ &= \mathbb{E}\left[\epsilon_t^T \epsilon_t\right] + \mathbb{E}\left[\mu_t^T \mu_t\right] \\ &\leq d\sigma_q^2 + L^2.\end{aligned}$$

Among all random variables Z with $\mathbb{E}[Z^T Z] \leq C$ the Gaussian distribution $Y \sim \mathcal{N}\left(0, \frac{C}{d} I_d\right)$ has the largest entropy, given by $H(Y) = \frac{d}{2} \log\left(\frac{2\pi e C}{d}\right)$. Therefore,

$$H(g_t) \leq \frac{d}{2} \log\left(\frac{2\pi e(d\sigma_q^2 + L^2)}{d}\right).$$

With this we can upper bound the $I(g_t; \mathbf{y} \mid \mathbf{x}, g_{<t})$ as follows:

$$\begin{aligned}I(g_t; \mathbf{y} \mid \mathbf{x}, g_{<t}) &= H(g_t \mid \mathbf{x}, g_{<t}) - H(g_t \mid \mathbf{x}, \mathbf{y}, g_{<t}) \\ &= H(g_t \mid \mathbf{x}, g_{<t}) - H(\epsilon_t) \\ &\leq \frac{d}{2} \log\left(\frac{2\pi e(d\sigma_q^2 + L^2)}{d}\right) - \frac{d}{2} \log\left(2\pi e \sigma_q^2\right) \qquad (15) \\ &= \frac{d}{2} \log\left(1 + \frac{L^2}{d\sigma_q^2}\right).\end{aligned}$$

□

Layer type	Parameters
Conv	32 filters, 4×4 kernels, stride 2, padding 1, batch normalization, ReLU
Conv	32 filters, 4×4 kernels, stride 2, padding 1, batch normalization, ReLU
Conv	64 filters, 3×3 kernels, stride 2, padding 0, batch normalization, ReLU
Conv	256 filters, 3×3 kernels, stride 1, padding 0, batch normalization, ReLU
FC	128 units, ReLU
FC	10 units, linear activation

Table 3: The architecture of MNIST classifiers.

Note that the proof will work for arbitrary ϵ_t that has zero mean and independent components, where the L2 norm of each component is bounded by σ_q^2. This holds because in such cases $H(\epsilon_t) \le \frac{d}{2} \log(2\pi e \sigma_q^2)$ (as Gaussians have highest entropy for fixed L2 norm) and the transition of (15) remains correct. Therefore, the same result holds when ϵ_t is sampled from a product of univariate zero-mean Laplace distributions with scale parameter $\sigma_q/\sqrt{2}$ (which makes the second moment equal to σ_q^2).

A similar result has been derived by Pensia et al. (2018) (lemma 5) to bound $I(w_t; (x_t, y_t) \mid w_{t-1})$.

B. Experimental Details

In this section we describe the details of experiments and implementations.

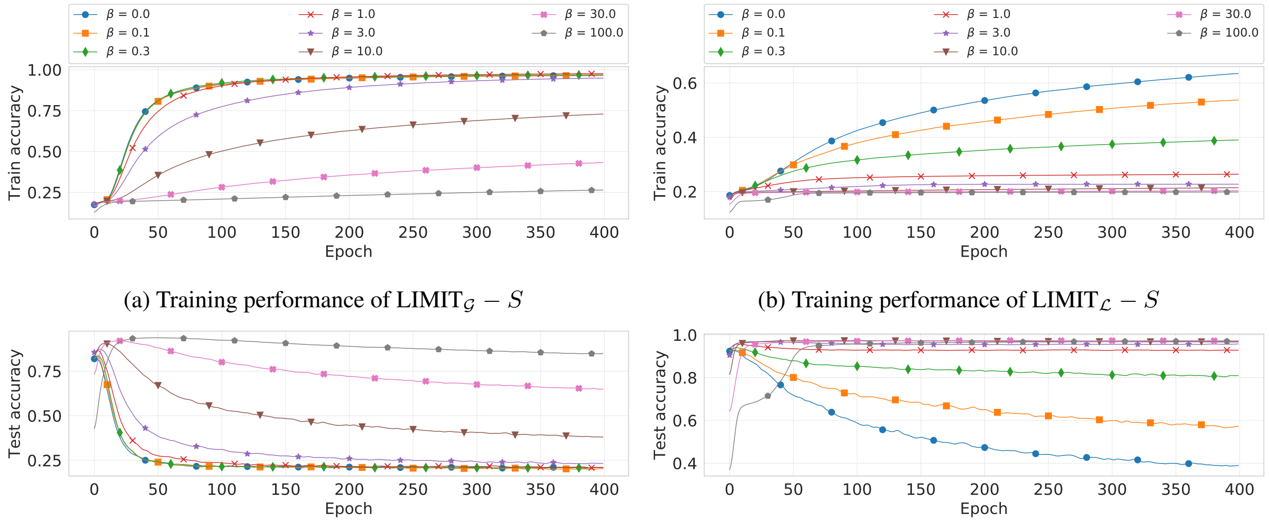
Classifier architectures. The architecture of classifiers used in MNIST experiments is presented in Table 3. The ResNet-34 used in CIFAR-10 and CIFAR-100 experiments differs from the standard ResNet-34 architecture (which is used for 224×224 images) in two ways: (a) the first convolutional layer has 3x3 kernels and stride 1 and (b) the max pooling layer after it is skipped. The architecture of ResNet-50 used in the Clothing1M experiment follows the original (He et al., 2016).

Hyperparameter search. The CE, MAE, and FW baselines have no hyperparameters. For the DMI, we tuned the learning rate by setting the best value from the following list: $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. The soft regularization approach of (11) has two hyperparameters: λ and β. We select λ from $[0.001, 0.01, 0.03, 0.1]$ and β from $[0.0, 0.01, 0.1, 1.0, 10.0]$. The objective of LIMIT instances has two terms: $\lambda H_{p,q}$ and $\beta \|\mu_t\|_2^2$. Consequently, we need only one hyperparameter instead of two. We choose to set $\lambda = 1$ and select β from $[0.0, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0, 100.0]$. When sampling is enabled, we select σ_q from $[0.01, 0.03, 0.1, 0.3]$. In MNIST and CIFAR experiments, we trained all models for 400 epochs and terminated the training early when the best validation accuracy was not improved in the last 100 epochs. All models for Clothing1M were trained for 30 epochs.

C. Additional Results

Effectiveness of gradient norm penalty. In the main text we discussed that the proposed approach may overfit if the gradient predictor $q_\phi(\cdot \mid \mathbf{x}, g_{<t})$ overfits and proposed to penalize the L2 norm of predicted gradients as a simply remedy for this issue. To demonstrate the effectiveness of this regularization, we present the training and testing accuracy curves of LIMIT with varying values of β in Fig. A0. We see that increasing β decreases overfitting on the training set and usually results in better generalization.

Detecting incorrect samples. In the proposed approach, the auxiliary network q should not be able to distinguish correct and incorrect samples, unless it overfits. In fact, Fig. A1 shows that if we look at the norm of predicted gradients, examples with correct and incorrect labels are indistinguishable in easy cases (MNIST with 80% uniform noise and CIFAR-10 with 40% uniform noise) and have large overlap in harder cases (CIFAR-10 with 40% pair noise and CIFAR-100 with 40% uniform noise). Therefore, we hypothesize that the auxiliary network learns to utilize incorrect samples effectively by predicting "correct" gradients. This also hints that the distance between the predicted and cross-entropy gradients might be useful for detecting samples with incorrect or confusing labels. Fig. A2 confirms this intuition, demonstrating that this distance separates correct and incorrect samples perfectly in easy cases (MNIST with 80% uniform noise and CIFAR-10 with 40% uniform noise) and separates them well in harder cases (CIFAR-10 with 40% pair noise and CIFAR-100 with 40% uniform noise). If we interpret this distance as a score for classifying correctness of a label, we get 91.1% ROC AUC score



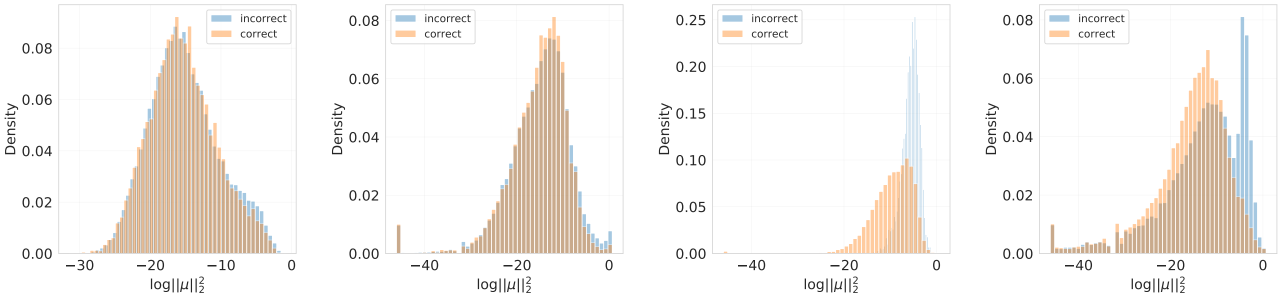
(a) Training performance of LIMIT$_{\mathcal{G}} - S$

(b) Training performance of LIMIT$_{\mathcal{L}} - S$

(c) Testing performance of LIMIT$_{\mathcal{G}} - S$

(d) Testing performance LIMIT$_{\mathcal{L}} - S$

Figure A0: Training and testing accuracies of "LIMIT$_{\mathcal{G}} - S$" and "LIMIT$_{\mathcal{L}} - S$" instances with varying values of β on MNIST with 80% uniform label noise. The curves are smoothed for better presentation.



(a) MNIST
80% uniform noise

(b) CIFAR-10
40% uniform noise

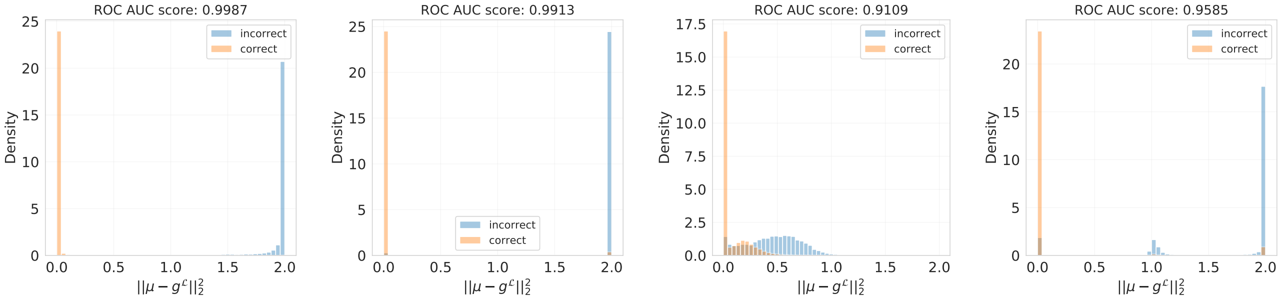
(c) CIFAR-10
40% pair noise

(d) CIFAR-100
40% uniform noise

Figure A1: Histograms of the norm of predicted gradients for examples with correct and incorrect labels. The gradient predictions are done using the best instances of LIMIT.

in the hardest case: CIFAR-10 with 40% pair noise, and more than 99% score in the easier cases.Motivated by this results, we use this analysis to detect samples with incorrect or confusing labels in the original MNIST, CIFAR-10, and Clothing1M datasets. We present a few incorrect/confusing labels for each class in Figures A3 and A4.

Quantitative results. Tables 4, 5, 6, and 7 present test accuracy comparisons on multiple corrupted versions of MNIST and CIFAR-10. The presented error bars are standard deviations. In case of MNIST, we compute them over 5 training/validation splits. In the case of CIFAR-10, due to high computational cost, we have only one run for each model and dataset pair. The standard deviations are computed by resampling the corresponding test sets 1000 times with replacement.



(a) MNIST
80% uniform noise

(b) CIFAR-10
40% uniform noise

(c) CIFAR-10
40% pair noise

(d) CIFAR-100
40% uniform noise

Figure A2: Histograms of the distance between predicted and actual gradient for examples with correct and incorrect labels. The gradient predictions are done using the best instances of LIMIT.

Method	$p = 0.0$			$p = 0.5$		
	$n = 10^3$	$n = 10^4$	All	$n = 10^3$	$n = 10^4$	All
CE	94.3 ± 0.5	98.4 ± 0.2	**99.2 ± 0.0**	71.8 ± 4.3	93.1 ± 0.6	97.2 ± 0.2
CE + GN	89.5 ± 0.8	95.4 ± 0.5	97.1 ± 0.5	70.5 ± 3.5	92.3 ± 0.7	97.4 ± 0.5
CE + LN	90.0 ± 0.5	95.3 ± 0.6	96.7 ± 0.7	66.8 ± 1.3	92.0 ± 1.5	97.6 ± 0.1
MAE	94.6 ± 0.5	98.3 ± 0.2	**99.1 ± 0.1**	75.6 ± 5.0	95.7 ± 0.5	98.1 ± 0.1
FW	93.6 ± 0.6	98.4 ± 0.1	**99.2 ± 0.1**	64.3 ± 9.1	91.6 ± 2.0	97.3 ± 0.3
DMI	94.5 ± 0.5	98.5 ± 0.1	**99.2 ± 0.0**	79.8 ± 2.9	95.7 ± 0.3	98.3 ± 0.1
Soft reg. (11)	**95.7 ± 0.2**	98.4 ± 0.1	**99.2 ± 0.0**	76.4 ± 2.4	95.7 ± 0.0	98.2 ± 0.1
LIMIT$_\mathcal{G}$ + S	**95.6 ± 0.3**	**98.6 ± 0.1**	**99.3 ± 0.0**	82.8 ± 4.6	97.0 ± 0.1	98.7 ± 0.1
LIMIT$_\mathcal{L}$ + S	94.8 ± 0.3	**98.6 ± 0.2**	**99.3 ± 0.0**	**88.7 ± 3.8**	**97.6 ± 0.1**	**98.9 ± 0.0**
LIMIT$_\mathcal{G}$ - S	**95.7 ± 0.2**	**98.7 ± 0.1**	**99.3 ± 0.1**	83.3 ± 2.3	97.1 ± 0.2	98.6 ± 0.1
LIMIT$_\mathcal{L}$ - S	95.0 ± 0.2	**98.7 ± 0.1**	**99.3 ± 0.1**	**88.2 ± 2.9**	**97.7 ± 0.1**	**99.0 ± 0.1**

Table 4: Test accuracy comparison on multiple versions of MNIST corrupted with uniform label noise.

Method	$p = 0.8$			$p = 0.89$		
	$n = 10^3$	$n = 10^4$	All	$n = 10^3$	$n = 10^4$	All
CE	27.0 ± 3.8	69.9 ± 2.6	87.2 ± 1.0	10.3 ± 1.6	13.4 ± 3.3	13.2 ± 1.8
CE + GN	25.9 ± 4.6	51.9 ± 10.5	85.3 ± 8.3	10.4 ± 4.5	10.2 ± 3.3	11.1 ± 0.4
CE + LN	30.2 ± 4.8	53.1 ± 6.4	74.5 ± 19.1	11.9 ± 3.9	8.8 ± 5.4	14.1 ± 4.3
MAE	25.1 ± 3.3	74.6 ± 2.7	93.2 ± 1.1	10.9 ± 1.4	12.1 ± 3.9	17.6 ± 8.1
FW	19.0 ± 4.1	61.2 ± 5.0	89.1 ± 2.1	8.7 ± 2.8	11.4 ± 1.4	12.3 ± 1.8
DMI	30.3 ± 5.1	79.0 ± 1.5	88.8 ± 0.9	10.5 ± 1.2	14.1 ± 5.1	12.5 ± 1.5
Soft reg. (11)	28.8 ± 2.2	67.0 ± 1.9	89.3 ± 0.6	10.3 ± 1.6	10.5 ± 0.8	12.7 ± 2.6
LIMIT$_\mathcal{G}$ + S	**35.9 ± 6.3**	80.6 ± 2.8	93.4 ± 0.5	10.0 ± 1.0	14.3 ± 5.4	13.1 ± 4.3
LIMIT$_\mathcal{L}$ + S	**35.6 ± 3.2**	**93.3 ± 0.3**	**97.6 ± 0.3**	10.1 ± 0.7	12.5 ± 2.1	**28.3 ± 8.1**
LIMIT$_\mathcal{G}$ - S	**37.1 ± 5.4**	82.0 ± 1.5	94.7 ± 0.6	9.9 ± 1.0	12.6 ± 0.3	16.0 ± 5.9
LIMIT$_\mathcal{L}$ - S	**35.9 ± 4.3**	**93.9 ± 0.8**	**97.7 ± 0.2**	11.1 ± 0.7	11.8 ± 1.0	**28.6 ± 4.0**

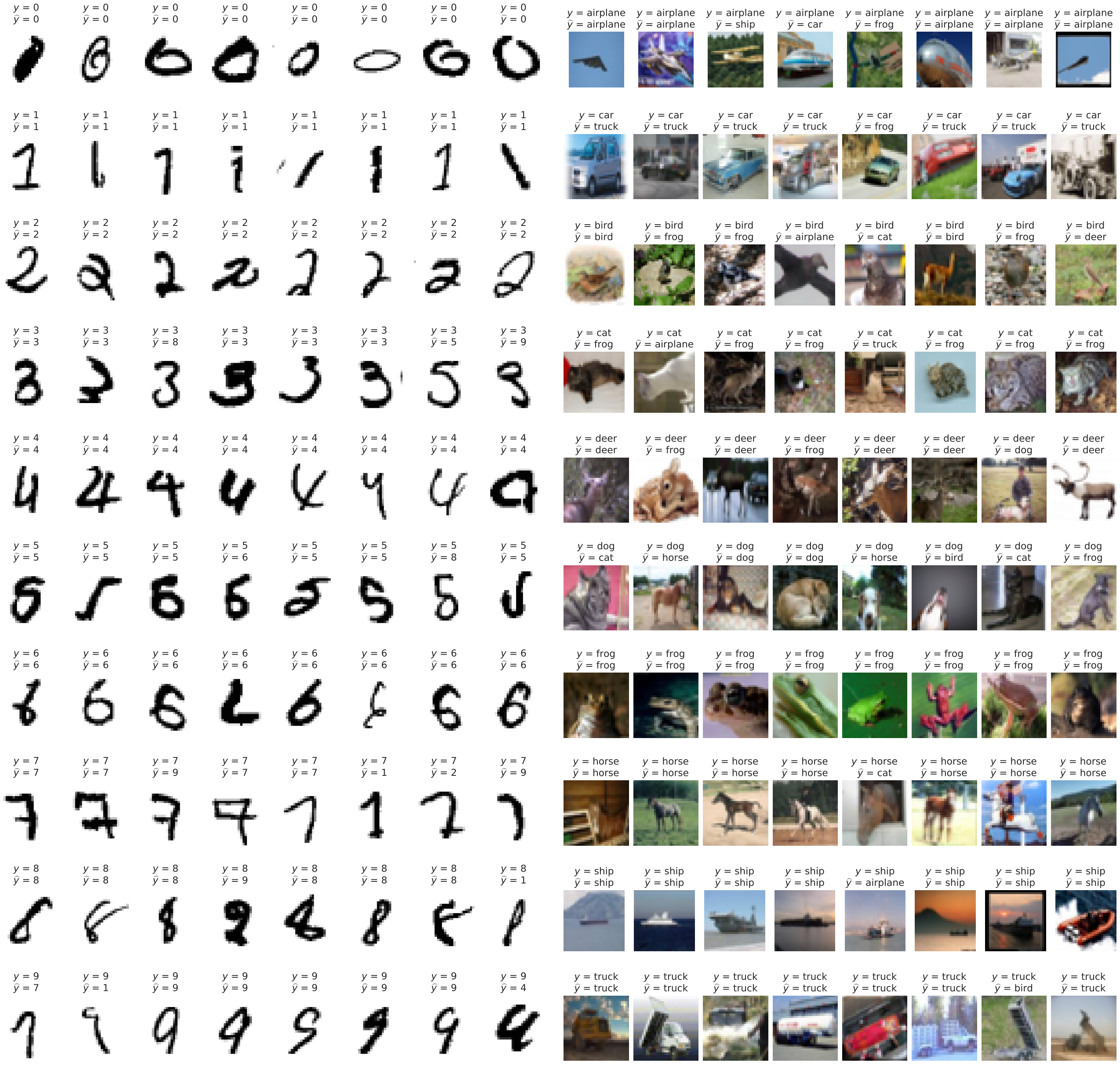
Table 5: Test accuracy comparison on multiple versions of MNIST corrupted with uniform label noise.

Method	$p = 0.0$	$p = 0.2$	$p = 0.4$	$p = 0.6$	$p = 0.8$
CE	92.7 ± 0.3	85.2 ± 0.4	81.0 ± 0.4	69.0 ± 0.5	38.8 ± 0.5
MAE	84.4 ± 0.4	85.4 ± 0.4	64.6 ± 0.5	15.4 ± 0.4	12.0 ± 0.3
FW	92.9 ± 0.3	86.2 ± 0.3	81.4 ± 0.4	69.7 ± 0.5	34.4 ± 0.5
DMI	93.0 ± 0.3	88.3 ± 0.3	85.0 ± 0.3	72.5 ± 0.4	38.9 ± 0.5
LIMIT$_\mathcal{G}$	**93.5 ± 0.2**	90.7 ± 0.3	86.6 ± 0.3	73.7 ± 0.4	38.7 ± 0.5
LIMIT$_\mathcal{L}$	93.1 ± 0.3	91.5 ± 0.3	88.2 ± 0.3	75.7 ± 0.4	35.8 ± 0.5
LIMIT$_\mathcal{G}$ + init.	**93.3 ± 0.3**	**92.4 ± 0.3**	**90.3 ± 0.3**	81.9 ± 0.4	**44.1 ± 0.5**
LIMIT$_\mathcal{L}$ + init.	**93.3 ± 0.2**	**92.2 ± 0.3**	**90.2 ± 0.3**	**82.9 ± 0.4**	**44.3 ± 0.5**

Table 6: Test accuracy comparison on CIFAR-10, corrupted with uniform label noise. The error bars are computed by bootstrapping the test set 1000 times.

Method	$p = 0.0$	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$
CE	92.7 ± 0.3	90.0 ± 0.3	88.1 ± 0.3	87.2 ± 0.3	81.8 ± 0.4
MAE	84.4 ± 0.4	88.6 ± 0.3	83.2 ± 0.4	72.1 ± 0.4	61.1 ± 0.5
FW	92.9 ± 0.3	90.1 ± 0.3	88.0 ± 0.3	86.8 ± 0.3	84.6 ± 0.3
DMI	93.0 ± 0.3	91.4 ± 0.3	90.6 ± 0.3	90.4 ± 0.3	**89.6 ± 0.3**
LIMIT$_\mathcal{G}$	**93.5 ± 0.2**	92.8 ± 0.3	91.3 ± 0.3	89.2 ± 0.3	86.0 ± 0.3
LIMIT$_\mathcal{L}$	93.1 ± 0.3	91.9 ± 0.3	91.1 ± 0.3	88.8 ± 0.3	84.2 ± 0.4
LIMIT$_\mathcal{G}$ + init.	**93.3 ± 0.3**	**93.3 ± 0.3**	**92.9 ± 0.3**	**90.8 ± 0.3**	88.3 ± 0.3
LIMIT$_\mathcal{L}$ + init.	**93.3 ± 0.2**	**93.0 ± 0.2**	92.3 ± 0.3	**91.1 ± 0.3**	**90.0 ± 0.3**

Table 7: Test accuracy comparison on CIFAR-10, corrupted with pair noise, described in Sec. 4.2. The error bars are computed by bootstrapping the test set 1000 times.



(a) MNIST

(b) CIFAR-10

Figure A3: Most confusing 8 labels per class in the MNIST (on the left) and CIFAR-10 (on the right) datasets, according to the distance between predicted and cross-entropy gradients. The gradient predictions are done using the best instances of LIMIT.

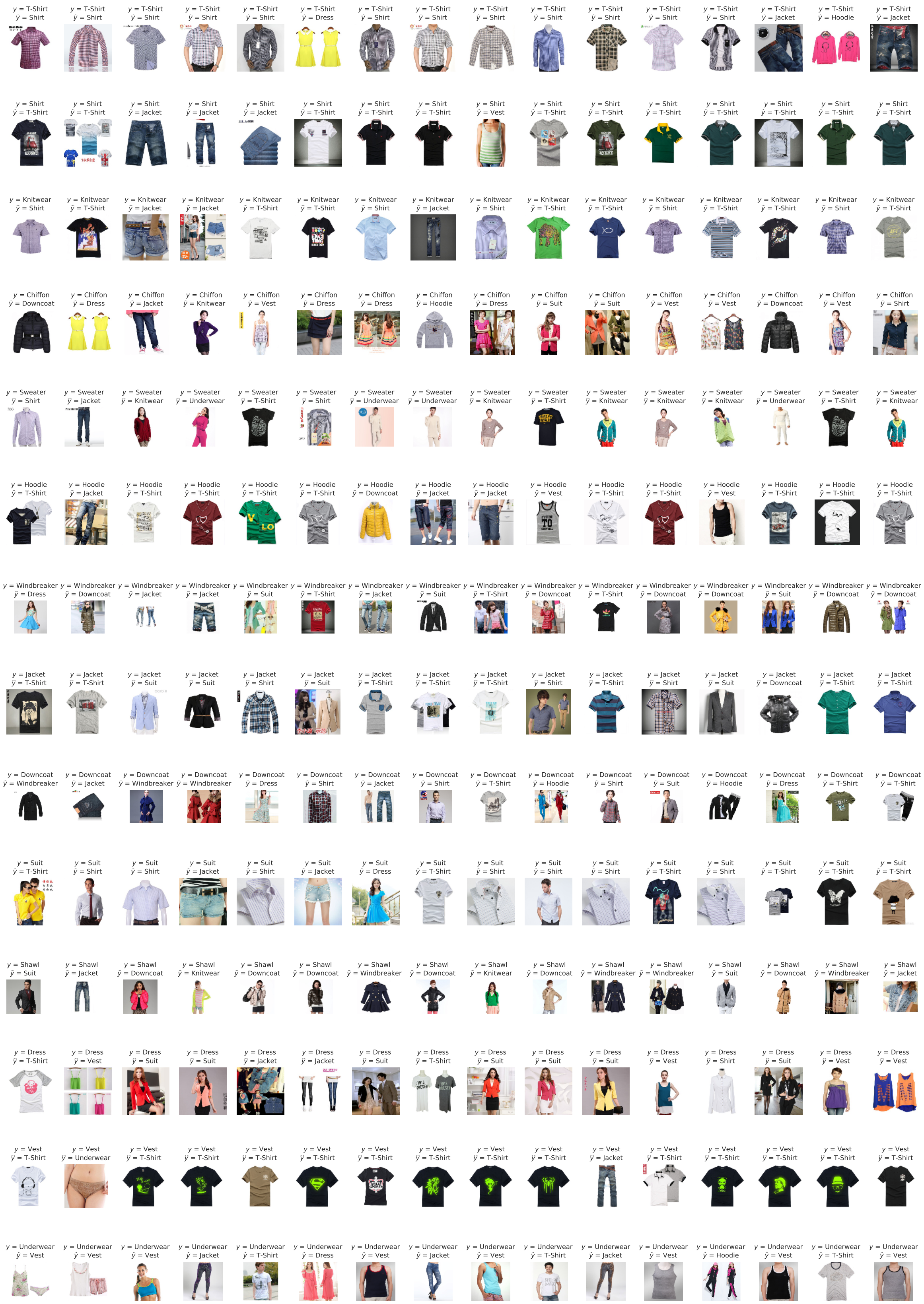

y = T-Shirt $\tilde{y}$ = Shirt
y = T-Shirt $\tilde{y}$ = Shirt
y = T-Shirt $\tilde{y}$ = Shirt
y = T-Shirt $\tilde{y}$ = Shirt
y = T-Shirt $\tilde{y}$ = Shirt
y = T-Shirt $\tilde{y}$ = Dress
y = T-Shirt $\tilde{y}$ = Shirt
y = T-Shirt $\tilde{y}$ = Shirt
y = T-Shirt $\tilde{y}$ = Shirt
y = T-Shirt $\tilde{y}$ = Shirt
y = T-Shirt $\tilde{y}$ = Shirt
y = T-Shirt $\tilde{y}$ = Shirt
y = T-Shirt $\tilde{y}$ = Shirt
y = T-Shirt $\tilde{y}$ = Jacket
y = T-Shirt $\tilde{y}$ = Hoodie
y = T-Shirt $\tilde{y}$ = Jacket
y = Shirt $\tilde{y}$ = T-Shirt
y = Shirt $\tilde{y}$ = T-Shirt
y = Shirt $\tilde{y}$ = Jacket
y = Shirt $\tilde{y}$ = Jacket
y = Shirt $\tilde{y}$ = Jacket
y = Shirt $\tilde{y}$ = T-Shirt
y = Shirt $\tilde{y}$ = T-Shirt
y = Shirt $\tilde{y}$ = T-Shirt
y = Shirt $\tilde{y}$ = Vest
y = Shirt $\tilde{y}$ = T-Shirt
y = Shirt $\tilde{y}$ = T-Shirt
y = Shirt $\tilde{y}$ = T-Shirt
y = Shirt $\tilde{y}$ = T-Shirt
y = Shirt $\tilde{y}$ = T-Shirt
y = Shirt $\tilde{y}$ = T-Shirt
y = Shirt $\tilde{y}$ = T-Shirt
y = Knitwear $\tilde{y}$ = Shirt
y = Knitwear $\tilde{y}$ = T-Shirt
y = Knitwear $\tilde{y}$ = Jacket
y = Knitwear $\tilde{y}$ = Jacket
y = Knitwear $\tilde{y}$ = T-Shirt
y = Knitwear $\tilde{y}$ = T-Shirt
y = Knitwear $\tilde{y}$ = Shirt
y = Knitwear $\tilde{y}$ = Jacket
y = Knitwear $\tilde{y}$ = Shirt
y = Knitwear $\tilde{y}$ = T-Shirt
y = Knitwear $\tilde{y}$ = T-Shirt
y = Knitwear $\tilde{y}$ = Shirt
y = Knitwear $\tilde{y}$ = T-Shirt
y = Knitwear $\tilde{y}$ = T-Shirt
y = Knitwear $\tilde{y}$ = Shirt
y = Knitwear $\tilde{y}$ = T-Shirt
y = Chiffon $\tilde{y}$ = Downcoat
y = Chiffon $\tilde{y}$ = Dress
y = Chiffon $\tilde{y}$ = Jacket
y = Chiffon $\tilde{y}$ = Knitwear
y = Chiffon $\tilde{y}$ = Vest
y = Chiffon $\tilde{y}$ = Dress
y = Chiffon $\tilde{y}$ = Dress
y = Chiffon $\tilde{y}$ = Hoodie
y = Chiffon $\tilde{y}$ = Dress
y = Chiffon $\tilde{y}$ = Suit
y = Chiffon $\tilde{y}$ = Suit
y = Chiffon $\tilde{y}$ = Vest
y = Chiffon $\tilde{y}$ = Vest
y = Chiffon $\tilde{y}$ = Downcoat
y = Chiffon $\tilde{y}$ = Vest
y = Chiffon $\tilde{y}$ = Shirt
y = Sweater $\tilde{y}$ = Shirt
y = Sweater $\tilde{y}$ = Jacket
y = Sweater $\tilde{y}$ = Knitwear
y = Sweater $\tilde{y}$ = Underwear
y = Sweater $\tilde{y}$ = T-Shirt
y = Sweater $\tilde{y}$ = Shirt
y = Sweater $\tilde{y}$ = Underwear
y = Sweater $\tilde{y}$ = Underwear
y = Sweater $\tilde{y}$ = Knitwear
y = Sweater $\tilde{y}$ = T-Shirt
y = Sweater $\tilde{y}$ = Knitwear
y = Sweater $\tilde{y}$ = Knitwear
y = Sweater $\tilde{y}$ = Knitwear
y = Sweater $\tilde{y}$ = Underwear
y = Sweater $\tilde{y}$ = T-Shirt
y = Sweater $\tilde{y}$ = Knitwear
y = Hoodie $\tilde{y}$ = T-Shirt
y = Hoodie $\tilde{y}$ = Jacket
y = Hoodie $\tilde{y}$ = T-Shirt
y = Hoodie $\tilde{y}$ = T-Shirt
y = Hoodie $\tilde{y}$ = T-Shirt
y = Hoodie $\tilde{y}$ = T-Shirt
y = Hoodie $\tilde{y}$ = Downcoat
y = Hoodie $\tilde{y}$ = Jacket
y = Hoodie $\tilde{y}$ = Jacket
y = Hoodie $\tilde{y}$ = Vest
y = Hoodie $\tilde{y}$ = T-Shirt
y = Hoodie $\tilde{y}$ = T-Shirt
y = Hoodie $\tilde{y}$ = Vest
y = Hoodie $\tilde{y}$ = T-Shirt
y = Hoodie $\tilde{y}$ = T-Shirt
y = Hoodie $\tilde{y}$ = T-Shirt
y = Windbreaker $\tilde{y}$ = Dress
y = Windbreaker $\tilde{y}$ = Downcoat
y = Windbreaker $\tilde{y}$ = Jacket
y = Windbreaker $\tilde{y}$ = Jacket
y = Windbreaker $\tilde{y}$ = Suit
y = Windbreaker $\tilde{y}$ = T-Shirt
y = Windbreaker $\tilde{y}$ = Jacket
y = Windbreaker $\tilde{y}$ = Suit
y = Windbreaker $\tilde{y}$ = T-Shirt
y = Windbreaker $\tilde{y}$ = Downcoat
y = Windbreaker $\tilde{y}$ = T-Shirt
y = Windbreaker $\tilde{y}$ = Downcoat
y = Windbreaker $\tilde{y}$ = Downcoat
y = Windbreaker $\tilde{y}$ = Suit
y = Windbreaker $\tilde{y}$ = Downcoat
y = Windbreaker $\tilde{y}$ = Downcoat
y = Jacket $\tilde{y}$ = T-Shirt
y = Jacket $\tilde{y}$ = T-Shirt
y = Jacket $\tilde{y}$ = Suit
y = Jacket $\tilde{y}$ = Suit
y = Jacket $\tilde{y}$ = Shirt
y = Jacket $\tilde{y}$ = Suit
y = Jacket $\tilde{y}$ = T-Shirt
y = Jacket $\tilde{y}$ = T-Shirt
y = Jacket $\tilde{y}$ = T-Shirt
y = Jacket $\tilde{y}$ = Shirt
y = Jacket $\tilde{y}$ = T-Shirt
y = Jacket $\tilde{y}$ = Shirt
y = Jacket $\tilde{y}$ = Suit
y = Jacket $\tilde{y}$ = Downcoat
y = Jacket $\tilde{y}$ = T-Shirt
y = Jacket $\tilde{y}$ = T-Shirt
y = Downcoat $\tilde{y}$ = Windbreaker
y = Downcoat $\tilde{y}$ = Jacket
y = Downcoat $\tilde{y}$ = Windbreaker
y = Downcoat $\tilde{y}$ = Windbreaker
y = Downcoat $\tilde{y}$ = Dress
y = Downcoat $\tilde{y}$ = Shirt
y = Downcoat $\tilde{y}$ = Jacket
y = Downcoat $\tilde{y}$ = Shirt
y = Downcoat $\tilde{y}$ = T-Shirt
y = Downcoat $\tilde{y}$ = Hoodie
y = Downcoat $\tilde{y}$ = Shirt
y = Downcoat $\tilde{y}$ = Suit
y = Downcoat $\tilde{y}$ = Hoodie
y = Downcoat $\tilde{y}$ = Dress
y = Downcoat $\tilde{y}$ = T-Shirt
y = Downcoat $\tilde{y}$ = T-Shirt
y = Suit $\tilde{y}$ = T-Shirt
y = Suit $\tilde{y}$ = Shirt
y = Suit $\tilde{y}$ = Shirt
y = Suit $\tilde{y}$ = Jacket
y = Suit $\tilde{y}$ = Shirt
y = Suit $\tilde{y}$ = Jacket
y = Suit $\tilde{y}$ = Dress
y = Suit $\tilde{y}$ = T-Shirt
y = Suit $\tilde{y}$ = Shirt
y = Suit $\tilde{y}$ = Shirt
y = Suit $\tilde{y}$ = Shirt
y = Suit $\tilde{y}$ = T-Shirt
y = Suit $\tilde{y}$ = Shirt
y = Suit $\tilde{y}$ = T-Shirt
y = Suit $\tilde{y}$ = Shirt
y = Suit $\tilde{y}$ = T-Shirt
y = Shawl $\tilde{y}$ = Suit
y = Shawl $\tilde{y}$ = Jacket
y = Shawl $\tilde{y}$ = Downcoat
y = Shawl $\tilde{y}$ = Knitwear
y = Shawl $\tilde{y}$ = Downcoat
y = Shawl $\tilde{y}$ = Downcoat
y = Shawl $\tilde{y}$ = Windbreaker
y = Shawl $\tilde{y}$ = Downcoat
y = Shawl $\tilde{y}$ = Knitwear
y = Shawl $\tilde{y}$ = Downcoat
y = Shawl $\tilde{y}$ = Windbreaker
y = Shawl $\tilde{y}$ = Windbreaker
y = Shawl $\tilde{y}$ = Suit
y = Shawl $\tilde{y}$ = Downcoat
y = Shawl $\tilde{y}$ = Windbreaker
y = Shawl $\tilde{y}$ = Jacket
y = Dress $\tilde{y}$ = T-Shirt
y = Dress $\tilde{y}$ = Vest
y = Dress $\tilde{y}$ = Suit
y = Dress $\tilde{y}$ = Suit
y = Dress $\tilde{y}$ = Jacket
y = Dress $\tilde{y}$ = Jacket
y = Dress $\tilde{y}$ = Suit
y = Dress $\tilde{y}$ = T-Shirt
y = Dress $\tilde{y}$ = Suit
y = Dress $\tilde{y}$ = Suit
y = Dress $\tilde{y}$ = Suit
y = Dress $\tilde{y}$ = Vest
y = Dress $\tilde{y}$ = Shirt
y = Dress $\tilde{y}$ = Suit
y = Dress $\tilde{y}$ = Vest
y = Dress $\tilde{y}$ = Vest
y = Vest $\tilde{y}$ = T-Shirt
y = Vest $\tilde{y}$ = Underwear
y = Vest $\tilde{y}$ = T-Shirt
y = Vest $\tilde{y}$ = T-Shirt
y = Vest $\tilde{y}$ = T-Shirt
y = Vest $\tilde{y}$ = T-Shirt
y = Vest $\tilde{y}$ = T-Shirt
y = Vest $\tilde{y}$ = T-Shirt
y = Vest $\tilde{y}$ = T-Shirt
y = Vest $\tilde{y}$ = T-Shirt
y = Vest $\tilde{y}$ = Jacket
y = Vest $\tilde{y}$ = T-Shirt
y = Vest $\tilde{y}$ = T-Shirt
y = Vest $\tilde{y}$ = T-Shirt
y = Vest $\tilde{y}$ = T-Shirt
y = Vest $\tilde{y}$ = T-Shirt
y = Underwear $\tilde{y}$ = Vest
y = Underwear $\tilde{y}$ = Vest
y = Underwear $\tilde{y}$ = Vest
y = Underwear $\tilde{y}$ = Jacket
y = Underwear $\tilde{y}$ = T-Shirt
y = Underwear $\tilde{y}$ = Dress
y = Underwear $\tilde{y}$ = Vest
y = Underwear $\tilde{y}$ = Jacket
y = Underwear $\tilde{y}$ = Vest
y = Underwear $\tilde{y}$ = T-Shirt
y = Underwear $\tilde{y}$ = Jacket
y = Underwear $\tilde{y}$ = Vest
y = Underwear $\tilde{y}$ = Hoodie
y = Underwear $\tilde{y}$ = Vest
y = Underwear $\tilde{y}$ = T-Shirt
y = Underwear $\tilde{y}$ = Vest

Figure A4: Most confusing 16 labels per class in the Clothing1M dataset, according to the distance between predicted and cross-entropy gradients. The gradient predictions are done using the best instance of LIMIT.