Main contribution

- Proposed a multinomial dropout for shallow learning.
- Demonstrated that this proposed distribution-dependent dropout leads to a faster convergence and a smaller generalization error through the risk bound analysis.
- Proposed an efficient evolutionary dropout for deep learning.
- Justified the proposed dropouts for both shallow and deep learning empirically.

Problem Setup

- Let \( \mathbf{x} \) denote a feature vector and a label, where \( \mathbf{x} \in \mathbb{R}^d \) and \( y \in \mathcal{Y} \).
- Denote by \( \mathcal{D} \) the joint distribution of \( (\mathbf{x}, y) \) and by \( \mathcal{D} \) the marginal distribution of \( \mathbf{x} \).
- The goal is to learn a linear prediction function \( f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} \) that minimizes the expected risk (considering loss function \( \ell(\cdot, \cdot) \)):

\[
\min_{\mathbf{w} \in \mathbb{R}^d} \mathbb{E}_{\mathcal{D}}[\ell(\mathbf{w}^T \mathbf{x}, y)]
\]  

(1)

- By \( \epsilon \sim \mathcal{M} \) a dropout noise vector of dimension \( d \).
- The corrupted feature vector is given by \( \mathbf{x} = \mathbf{x} \circ \epsilon \), where the operator \( \circ \) represents the element-wise multiplication.
- Denote by \( \mathcal{D} \) the joint distribution of the new data \( (\mathbf{x}, y) \) and by \( \mathcal{D} \) the marginal distribution of \( \mathbf{x} \).
- With the corrupted data, the risk minimization becomes

\[
\min_{\mathbf{w} \in \mathbb{R}^d} \tilde{\mathbb{E}}_{\mathcal{D}}[\ell(\tilde{\mathbf{w}}^T \mathbf{x}, y)]
\]  

(2)

Learning with Multinomial Dropout

Definition 1. A multinomial dropout is defined as \( \mathbf{x} = \mathbf{x} \circ \epsilon \), where \( \epsilon_i = \frac{1}{\mathcal{M}} \mathcal{M}(i), i \in [d] \) and \( \{m_1, \ldots, m_d\} \) follow a multinomial distribution \( \text{Mult}(p_1, \ldots, p_d) \) with \( \epsilon \mathbb{1} = 1 \) and \( p_i \geq 0 \).

Proposition 1. If \( \ell(\cdot, \cdot) = \log(1 + \exp\{-y_\mathbf{x}\}) \), then

\[
\mathbb{E}_{\mathcal{D}}[\ell(\mathbf{w}^T \mathbf{x}, y)] = \mathbb{E}_{\mathcal{D}}[\ell(\mathbf{w}^T \mathbf{x}, y)] + R_{D,M}(\mathbf{w})
\]

where \( \mathcal{M} \) denotes the distribution of \( \epsilon \) and \( R_{D,M}(\mathbf{w}) = \mathbb{E}_{\mathcal{M}}[\log(\frac{\exp(\mathbf{w}^T \mathbf{x})}{\exp(\mathbf{w}^T \mathbf{x})})] \). Dropout is a data-dependent regularizer.

Improved Dropout for Shallow Learning

Theorem 1. Let \( \mathcal{L}(\mathbf{w}) \) be the expected risk of \( \mathbf{w} \) defined in (1). Assume \( \mathbb{E}_{\mathcal{D}}[\ell(\mathbf{x} \circ \epsilon, y)] \leq B^e \) and \( \ell(\cdot, \cdot) \) is convex and \( G \)-Lipschitz continuous. For any \( \mathbf{w}, \mathbf{w}' \leq r \), by appropriately choosing \( y \), we can have

\[
\mathbb{E}[\mathcal{L}(\tilde{\mathbf{w}})] + R_{D,M}(\tilde{\mathbf{w}}) \leq \mathcal{L}(\mathbf{w}) + R_{D,M}(\mathbf{w}) + GB y \sqrt{r}
\]

Theoretically, minimizing the term \( \mathbb{E}_{\mathcal{D}}[\ell(\mathbf{x} \circ \epsilon, y)] \) and the relaxed upper bound of term \( R_{D,M}(\mathbf{w}) \) yields the optimal sampling probabilities:

\[
p_i^* = \sqrt{\frac{\mathbb{E}_{\mathcal{D}}[\ell(\mathbf{x} \circ \epsilon, y)]}{\sum_{j=1}^{d} \sqrt{\mathbb{E}_{\mathcal{D}}[\ell(\mathbf{x} \circ \epsilon, y)]}}, i = 1, \ldots, d
\]  

(3)

Practically, we use the empirical second-order statistics to compute the probabilities:

\[
p_i = \sqrt{\frac{\mathbb{E}_{\mathcal{D}}[\ell(\mathbf{x} \circ \epsilon, y)]}{\sum_{j=1}^{d} \sqrt{\mathbb{E}_{\mathcal{D}}[\ell(\mathbf{x} \circ \epsilon, y)]}}, i = 1, \ldots, d
\]  

(4)

Improved Dropout for Deep Learning

Let \( X^{l} = (x_1^{l}, \ldots, x_n^{l}) \) denote the outputs of the \( l \)th layer for a minibatch of \( m \) examples, calculate the probabilities for dropout by

\[
p_i^{l} = \sqrt{\frac{\mathbb{E}_{\mathcal{D}}[\ell(\mathbf{x} \circ \epsilon, y)]}{\sum_{j=1}^{d} \sqrt{\mathbb{E}_{\mathcal{D}}[\ell(\mathbf{x} \circ \epsilon, y)]}}, i = 1, \ldots, d
\]  

(5)

Experimental Results

Figure 2: data-dependent dropout vs. standard dropout on three datasets (real-sim, news20 and RCV1) for logistic regression; Lower Right Corner: Evolutional dropout vs. BN on CIFAR-10.

Figure 3: Evolutional dropout vs. standard dropout on four benchmark datasets (MNIST, SVHN, CIFAR-10 and CIFAR-100) for deep learning.