

Improved Dropout for Shallow and Deep Learning

Zhe Li¹, Boqing Gong², Tianbao Yang¹

¹The University of Iowa, Iowa city, IA 52245

²University of Central Florida, Orlando, FL 32816

{zhe-li-1, tianbao-yang}@uiowa.edu

bgong@crcv.ucf.edu

Abstract

Dropout has been witnessed with great success in training deep neural networks by independently zeroing out the outputs of neurons at random. It has also received a surge of interest for shallow learning, e.g., logistic regression. However, the independent sampling for dropout could be suboptimal for the sake of convergence. In this paper, we propose to use multinomial sampling for dropout, i.e., sampling features or neurons according to a multinomial distribution with different probabilities for different features/neurons. To exhibit the optimal dropout probabilities, we analyze the shallow learning with multinomial dropout and establish the risk bound for stochastic optimization. By minimizing a sampling dependent factor in the risk bound, we obtain a distribution-dependent dropout with sampling probabilities dependent on the second order statistics of the data distribution. To tackle the issue of evolving distribution of neurons in deep learning, we propose an efficient adaptive dropout (named **evolutional dropout**) that computes the sampling probabilities on-the-fly from a mini-batch of examples. Empirical studies on several benchmark datasets demonstrate that the proposed dropouts achieve not only much faster convergence and but also a smaller testing error than the standard dropout. For example, on the CIFAR-100 data, the evolutional dropout achieves relative improvements over 10% on the prediction performance and over 50% on the convergence speed compared to the standard dropout.

1 Introduction

Dropout has been widely used to avoid overfitting of deep neural networks with a large number of parameters [9, 16], which usually identically and independently at random samples neurons and sets their outputs to be zeros. Extensive experiments [4] have shown that dropout can help obtain the state-of-the-art performance on a range of benchmark data sets. Recently, dropout has also been found to improve the performance of logistic regression and other single-layer models for natural language tasks such as document classification and named entity recognition [21].

In this paper, instead of identically and independently at random zeroing out features or neurons, we propose to use multinomial sampling for dropout, i.e., sampling features or neurons according to a multinomial distribution with different probabilities for different features/neurons. Intuitively, it makes more sense to use non-uniform multinomial sampling than identical and independent sampling for different features/neurons. For example, in shallow learning if input features are centered, we can drop out features with small variance more frequently or completely allowing the training to focus on more important features and consequentially enabling faster convergence. To justify the multinomial sampling for dropout and reveal the optimal sampling probabilities, we conduct a rigorous analysis on the risk bound of shallow learning by stochastic optimization with multinomial dropout, and demonstrate that a distribution-dependent dropout leads to a smaller expected risk (i.e., faster convergence and smaller generalization error).

Inspired by the distribution-dependent dropout, we propose a data-dependent dropout for shallow learning, and an evolutionary dropout for deep learning. For shallow learning, the sampling probabilities are computed from the second order statistics of features of the training data. For deep learning, the sampling probabilities of dropout for a layer are computed on-the-fly from the second-order statistics of the layer’s outputs based on a mini-batch of examples. This is particularly suited for deep learning because (i) the distribution of each layer’s outputs is evolving over time, which is known as internal covariate shift [5]; (ii) passing through all the training data in deep neural networks (in particular deep convolutional neural networks) is much more expensive than through a mini-batch of examples. For a mini-batch of examples, we can leverage parallel computing architectures to accelerate the computation of sampling probabilities.

We note that the proposed evolutionary dropout achieves similar effect to the batch normalization technique (Z-normalization based on a mini-batch of examples) [5] but with different flavors. Both approaches can be considered to tackle the issue of internal covariate shift for accelerating the convergence. Batch normalization tackles the issue by normalizing the output of neurons to zero mean and unit variance and then performing dropout independently¹. In contrast, our proposed evolutionary dropout tackles this issue from another perspective by exploiting a distribution-dependent dropout, which adapts the sampling probabilities to the evolving distribution of a layer’s outputs. In other words, it uses normalized sampling probabilities based on the second order statistics of internal distributions. Indeed, we notice that for shallow learning with Z-normalization (normalizing each feature to zero mean and unit variance) the proposed data-dependent dropout reduces to uniform dropout that acts similarly to the standard dropout. Because of this connection, the presented theoretical analysis also sheds some lights on the power of batch normalization from the angle of theory. Compared to batch normalization, the proposed distribution-dependent dropout is still attractive because (i) it is rooted in theoretical analysis of the risk bound; (ii) it introduces no additional parameters and layers without complicating the back-propagation and the inference; (iii) it facilitates further research because it shares the same mathematical foundation as standard dropout (e.g., equivalent to a form of data-dependent regularizer) [18].

We summarize the main contributions of the paper below.

- We propose a multinomial dropout and demonstrate that a distribution-dependent dropout leads to a faster convergence and a smaller generalization error through the risk bound analysis for shallow learning.
- We propose an efficient evolutionary dropout for deep learning based on the distribution-dependent dropout.
- We justify the proposed dropouts for both shallow learning and deep learning by experimental results on several benchmark datasets.

In the remainder, we first review some related work and preliminaries. We present the main results in Section 4 and experimental results in Section 5.

2 Related Work

In this section, we review some related work on dropout and optimization algorithms for deep learning.

Dropout is a simple yet effective technique to prevent overfitting in training deep neural networks [16]. It has received much attention recently from researchers to study its practical and theoretical properties. Notably, Wager et al. [18], Baldi and Sadowski [2] have analyzed the dropout from a theoretical viewpoint and found that dropout is equivalent to a data-dependent regularizer.

The most simple form of dropout is to multiply hidden units by i.i.d Bernoulli noise. Several recent works also found that using other types of noise works as well as Bernoulli noise (e.g., Gaussian noise), which could lead to a better approximation of the marginalized loss [20, 7]. Some works tried to optimize the hyper-parameters that define the noise level in a Bayesian framework [23, 7]. Graham et al. [3] used the same noise across a batch of examples in order to speed up the computation. The adaptive dropout proposed in [1] overlays a binary belief network over a neural network, incurring more computational overhead to dropout because one has to train the additional binary

¹The author also reported that in some cases dropout is even not necessary

belief network. In contrast, the present work proposes a new dropout with noise sampled according to distribution-dependent sampling probabilities. To the best of our knowledge, this is the first work that rigorously studies this type of dropout with theoretical analysis of the risk bound. It is demonstrated that the new dropout can improve the speed of convergence.

Stochastic gradient descent with back-propagation has been used a lot in optimizing deep neural networks. However, it is notorious for its slow convergence especially for deep learning. Recently, there emerge a battery of studies trying to accelerate the optimization of deep learning [17, 12, 22, 5, 6], which tackle the problem from different perspectives. Among them, we notice that the developed evolutionary dropout for deep learning achieves similar effect as batch normalization [5] addressing the internal covariate shift issue (i.e., evolving distributions of internal hidden units).

3 Preliminaries

In this section, we present some preliminaries, including the framework of risk minimization in machine learning and learning with dropout noise. We also introduce the multinomial dropout, which allows us to construct a distribution-dependent dropout as revealed in the next section.

Let (\mathbf{x}, y) denote a feature vector and a label, where $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathcal{Y}$. Denote by \mathcal{P} the joint distribution of (\mathbf{x}, y) and denote by \mathcal{D} the marginal distribution of \mathbf{x} . The goal of risk minimization is to learn a prediction function $f(\mathbf{x})$ that minimizes the expected loss, i.e., $\min_{f \in \mathcal{H}} \mathbb{E}_{\mathcal{P}}[\ell(f(\mathbf{x}), y)]$, where $\ell(z, y)$ is a loss function (e.g., the logistic loss) that measures the inconsistency between z and y and \mathcal{H} is a class of prediction functions. In deep learning, the prediction function $f(\mathbf{x})$ is determined by a deep neural network. In shallow learning, one might be interested in learning a linear model $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. In the following presentation, the analysis will focus on the risk minimization of a linear model, i.e.,

$$\min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}(\mathbf{w}) \triangleq \mathbb{E}_{\mathcal{P}}[\ell(\mathbf{w}^\top \mathbf{x}, y)] \quad (1)$$

In this paper, we are interested in learning with dropout, i.e., the feature vector \mathbf{x} is corrupted by a dropout noise. In particular, let $\epsilon \sim \mathcal{M}$ denote a dropout noise vector of dimension d , and the corrupted feature vector is given by $\hat{\mathbf{x}} = \mathbf{x} \circ \epsilon$, where the operator \circ represents the element-wise multiplication. Let $\hat{\mathcal{P}}$ denote the joint distribution of the new data $(\hat{\mathbf{x}}, y)$ and $\hat{\mathcal{D}}$ denote the marginal distribution of $\hat{\mathbf{x}}$. With the corrupted data, the risk minimization becomes

$$\min_{\mathbf{w} \in \mathbb{R}^d} \hat{\mathcal{L}}(\mathbf{w}) \triangleq \mathbb{E}_{\hat{\mathcal{P}}}[\ell(\mathbf{w}^\top (\mathbf{x} \circ \epsilon), y)] \quad (2)$$

In standard dropout [18, 4], the entries of the noise vector ϵ are sampled independently according to $\Pr(\epsilon_j = 0) = \delta$ and $\Pr(\epsilon_j = \frac{1}{1-\delta}) = 1 - \delta$, i.e., features are dropped with a probability δ and scaled by $\frac{1}{1-\delta}$ with a probability $1 - \delta$. We can also write $\epsilon_j = \frac{b_j}{1-\delta}$, where $b_j \in \{0, 1\}$, $j \in [d]$ are i.i.d Bernoulli random variables with $\Pr(b_j = 1) = 1 - \delta$. The scaling factor $\frac{1}{1-\delta}$ is added to ensure that $\mathbb{E}_{\epsilon}[\hat{\mathbf{x}}] = \mathbf{x}$. It is obvious that using the standard dropout different features will have equal probabilities to be dropped out or to be selected independently. However, in practice some features could be more informative than the others for learning purpose. Therefore, it makes more sense to assign different sampling probabilities for different features and make the features compete with each other.

To this end, we introduce the following multinomial dropout.

Definition 1. (Multinomial Dropout) A multinomial dropout is defined as $\hat{\mathbf{x}} = \mathbf{x} \circ \epsilon$, where $\epsilon_i = \frac{m_i}{kp_i}$, $i \in [d]$ and $\{m_1, \dots, m_d\}$ follow a multinomial distribution $\text{Mult}(p_1, \dots, p_d; k)$ with $\sum_{i=1}^d p_i = 1$ and $p_i \geq 0$.

Remark: The multinomial dropout allows us to use non-uniform sampling probabilities p_1, \dots, p_d for different features. The value of m_i is the number of times that the i -th feature is selected in k independent trials of selection. In each trial, the probability that the i -th feature is selected is given by p_i . As in the standard dropout, the normalization by kp_i is to ensure that $\mathbb{E}_{\epsilon}[\hat{\mathbf{x}}] = \mathbf{x}$. The parameter k plays the same role as the parameter $1 - \delta$ in standard dropout, which controls the number of features to be dropped. In particular, the expected total number of the kept features using multinomial dropout is k and that using standard dropout is $d(1 - \delta)$. In the sequel, to make

fair comparison between the two dropouts, we let $k = d(1 - \delta)$. In this case, when a uniform distribution $p_i = 1/d$ is used in multinomial dropout to which we refer as *uniform dropout*, then $\epsilon_i = \frac{m_i}{1-\delta}$, which acts similarly to the standard dropout using i.i.d Bernoulli random variables. Note that another choice to make the sampling probabilities different is still using i.i.d Bernoulli random variables but with different probabilities for different features. However, multinomial dropout is more suitable because (i) it is easy to control the level of dropout by varying the value of k ; (ii) it gives rise to natural competition among features because of the constraint $\sum_i p_i = 1$; (iii) it allows us to minimize the sampling dependent risk bound for obtaining a better distribution than uniform sampling.

Dropout is a data-dependent regularizer Dropout as a regularizer has been studied in [18, 2] for logistic regression, which is stated in the following proposition for ease of discussion later.

Proposition 1. If $\ell(z, y) = \log(1 + \exp(-yz))$, then

$$\mathbb{E}_{\hat{\mathcal{P}}}[\ell(\mathbf{w}^\top \hat{\mathbf{x}}, y)] = \mathbb{E}_{\mathcal{P}}[\ell(\mathbf{w}^\top \mathbf{x}, y)] + R_{\mathcal{D}, \mathcal{M}}(\mathbf{w}) \quad (3)$$

where \mathcal{M} denotes the distribution of ϵ and $R_{\mathcal{D}, \mathcal{M}}(\mathbf{w}) = \mathbb{E}_{\mathcal{D}, \mathcal{M}} \left[\log \frac{\exp(\mathbf{w}^\top \frac{\mathbf{x} \circ \epsilon}{2}) + \exp(-\mathbf{w}^\top \frac{\mathbf{x} \circ \epsilon}{2})}{\exp(\mathbf{w}^\top \mathbf{x}/2) + \exp(-\mathbf{w}^\top \mathbf{x}/2)} \right]$.

Remark: It is notable that $R_{\mathcal{D}, \mathcal{M}} \geq 0$ due to the Jensen inequality. Using the second order Taylor expansion, [18] showed that the following approximation of $R_{\mathcal{D}, \mathcal{M}}(\mathbf{w})$ is easy to manipulate and understand:

$$\hat{R}_{\mathcal{D}, \mathcal{M}}(\mathbf{w}) = \frac{\mathbb{E}_{\mathcal{D}}[q(\mathbf{w}^\top \mathbf{x})(1 - q(\mathbf{w}^\top \mathbf{x}))\mathbf{w}^\top C_{\mathcal{M}}(\mathbf{x} \circ \epsilon)\mathbf{w}]}{2} \quad (4)$$

where $q(\mathbf{w}^\top \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x}/2)}$, and $C_{\mathcal{M}}$ denotes the covariance matrix in terms of ϵ . In particular, if ϵ is the standard dropout noise, then $C_{\mathcal{M}}[\mathbf{x} \circ \epsilon] = \text{diag}(x_1^2 \delta / (1 - \delta), \dots, x_d^2 \delta / (1 - \delta))$, where $\text{diag}(s_1, \dots, s_n)$ denotes a $d \times d$ diagonal matrix with the i -th entry equal to s_i . If ϵ is the multinomial dropout noise in Definition 1, we have

$$C_{\mathcal{M}}[\mathbf{x} \circ \epsilon] = \frac{1}{k} \text{diag}(x_i^2 / p_i) - \frac{1}{k} \mathbf{x} \mathbf{x}^\top \quad (5)$$

4 Learning with Multinomial Dropout

In this section, we analyze a stochastic optimization approach for minimizing the dropout loss in (2). Assume the sampling probabilities are known. We first obtain a risk bound of learning with multinomial dropout for stochastic optimization. Then we try to minimize the factors in the risk bound that depend on the sampling probabilities. We would like to emphasize that our goal here is not to show that using dropout would render a smaller risk than without using dropout, but rather focus on the impact of different sampling probabilities on the risk. Let the initial solution be \mathbf{w}_1 . At the iteration t , we sample $(\mathbf{x}_t, y_t) \sim \mathcal{P}$ and $\epsilon_t \sim \mathcal{M}$ as in Definition 1 and then update the model by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t^\top (\mathbf{x}_t \circ \epsilon_t), y_t) \quad (6)$$

where $\nabla \ell$ denotes the (sub)gradient in terms of \mathbf{w}_t and η_t is a step size. Suppose we run the stochastic optimization by n steps (i.e., using n examples) and compute the final solution as $\hat{\mathbf{w}}_n = \frac{1}{n} \sum_{t=1}^n \mathbf{w}_t$.

We note that another approach of learning with dropout is to minimize the empirical risk by marginalizing out the dropout noise, i.e., replacing the true expectations $\mathbb{E}_{\mathcal{P}}$ and $\mathbb{E}_{\mathcal{D}}$ in (3) with empirical expectations over a set of samples $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ denoted by $\mathbb{E}_{\mathcal{P}_n}$ and $\mathbb{E}_{\mathcal{D}_n}$. Since the data dependent regularizer $R_{\mathcal{D}_n, \mathcal{M}}(\mathbf{w})$ is difficult to compute, one usually uses an approximation $\hat{R}_{\mathcal{D}_n, \mathcal{M}}(\mathbf{w})$ (e.g., as in (4)) in place of $R_{\mathcal{D}_n, \mathcal{M}}(\mathbf{w})$. However, the resulting problem is a non-convex optimization, which together with the approximation error would make the risk analysis much more involved. In contrast, the update in (6) can be considered as a stochastic gradient descent update for solving the convex optimization problem in (2), allowing us to establish the risk bound based on previous results of stochastic gradient descent for risk minimization [14, 15]. Nonetheless, this restriction does not lose the generality. Indeed, stochastic optimization is usually employed for solving empirical loss minimization in big data and deep learning.

The following theorem establishes a risk bound of $\hat{\mathbf{w}}_n$ in expectation.

Theorem 1. Let $\mathcal{L}(\mathbf{w})$ be the expected risk of \mathbf{w} defined in (1). Assume $\mathbb{E}_{\hat{\mathcal{D}}}[\|\mathbf{x} \circ \epsilon\|_2^2] \leq B^2$ and $\ell(z, y)$ is G -Lipschitz continuous. For any $\|\mathbf{w}_*\|_2 \leq r$, by appropriately choosing η , we can have

$$\mathbb{E}[\mathcal{L}(\hat{\mathbf{w}}_n) + R_{\mathcal{D}, \mathcal{M}}(\hat{\mathbf{w}}_n)] \leq \mathcal{L}(\mathbf{w}_*) + R_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*) + \frac{GBr}{\sqrt{n}}$$

where $\mathbb{E}[\cdot]$ is taking expectation over the randomness in $(\mathbf{x}_t, y_t, \epsilon_t), t = 1, \dots, n$.

Remark: In the above theorem, we can choose \mathbf{w}_* to be the best model that minimizes the expected risk in (1). Since $R_{\mathcal{D}, \mathcal{M}}(\mathbf{w}) \geq 0$, the upper bound in the theorem above is also the upper bound of the risk of $\hat{\mathbf{w}}_n$, i.e., $\mathcal{L}(\hat{\mathbf{w}}_n)$, in expectation. The proof of the above theorem follows the standard analysis of stochastic gradient descent. The detailed proof of theorem is included in the appendix.

4.1 Distribution Dependent Dropout

Next, we consider the sampling dependent factors in the risk bounds. From Theorem 1, we can see that there are two terms that depend on the sampling probabilities, i.e., B^2 - the upper bound of $\mathbb{E}_{\hat{\mathcal{D}}}[\|\mathbf{x} \circ \epsilon\|_2^2]$, and $R_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*) - R_{\mathcal{D}, \mathcal{M}}(\hat{\mathbf{w}}_n) \leq R_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*)$. We note that the second term also depends on \mathbf{w}_* and $\hat{\mathbf{w}}_n$, which is more difficult to optimize. We first try to minimize $\mathbb{E}_{\hat{\mathcal{D}}}[\|\mathbf{x} \circ \epsilon\|_2^2]$ and present the discussion on minimizing $R_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*)$ later. From Theorem 1, we can see that minimizing $\mathbb{E}_{\hat{\mathcal{D}}}[\|\mathbf{x} \circ \epsilon\|_2^2]$ would lead to not only a smaller risk (given the same number of total examples, smaller $\mathbb{E}_{\hat{\mathcal{D}}}[\|\mathbf{x} \circ \epsilon\|_2^2]$ gives a smaller risk bound) but also a faster convergence (with the same number of iterations, smaller $\mathbb{E}_{\hat{\mathcal{D}}}[\|\mathbf{x} \circ \epsilon\|_2^2]$ gives a smaller optimization error).

Due to the limited space, the proofs of Proposition 2, 3, 4 are included in supplement. The following proposition simplifies the expectation $\mathbb{E}_{\hat{\mathcal{D}}}[\|\mathbf{x} \circ \epsilon\|_2^2]$.

Proposition 2. Let ϵ follow the distribution \mathcal{M} defined in Definition 1. Then

$$\mathbb{E}_{\hat{\mathcal{D}}}[\|\mathbf{x} \circ \epsilon\|_2^2] = \frac{1}{k} \sum_{i=1}^d \frac{1}{p_i} \mathbb{E}_{\mathcal{D}}[x_i^2] + \frac{k-1}{k} \sum_{i=1}^d \mathbb{E}_{\mathcal{D}}[x_i^2] \quad (7)$$

Given the expression of $\mathbb{E}_{\hat{\mathcal{D}}}[\|\mathbf{x} \circ \epsilon\|_2^2]$ in Proposition 2, we can minimize it over \mathbf{p} , leading to the following result.

Proposition 3. The solution to $\mathbf{p}_* = \arg \min_{\mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1} \mathbb{E}_{\hat{\mathcal{D}}}[\|\mathbf{x} \circ \epsilon\|_2^2]$ is given by

$$p_i^* = \frac{\sqrt{\mathbb{E}_{\mathcal{D}}[x_i^2]}}{\sum_{j=1}^d \sqrt{\mathbb{E}_{\mathcal{D}}[x_j^2]}}, i = 1, \dots, d \quad (8)$$

Next, we examine $R_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*)$. Since direct manipulation on $R_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*)$ is difficult, we try to minimize the second order Taylor expansion $\hat{R}_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*)$ for logistic loss. The following theorem establishes an upper bound of $\hat{R}_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*)$.

Proposition 4. Let ϵ follow the distribution \mathcal{M} defined in Definition 1. We have $\hat{R}_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*) \leq \frac{1}{8k} \|\mathbf{w}_*\|_2^2 \left(\sum_{i=1}^d \frac{\mathbb{E}_{\mathcal{D}}[x_i^2]}{p_i} - \mathbb{E}_{\mathcal{D}}[\|\mathbf{x}\|_2^2] \right)$

Remark: By minimizing the relaxed upper bound in Proposition 4, we obtain the same sampling probabilities as in (8). We note that a tighter upper bound can be established, however, which will yield sampling probabilities dependent on the unknown \mathbf{w}_* .

In summary, using the probabilities in (8), we can reduce both $\mathbb{E}_{\hat{\mathcal{D}}}[\|\mathbf{x} \circ \epsilon\|_2^2]$ and $R_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*)$ in the risk bound, leading to a faster convergence and a smaller generalization error. In practice, we can use empirical second-order statistics to compute the probabilities, i.e.,

$$p_i = \frac{\sqrt{\frac{1}{n} \sum_{j=1}^n [\mathbf{x}_j]_i^2}}{\sum_{i'=1}^d \sqrt{\frac{1}{n} \sum_{j=1}^n [\mathbf{x}_j]_{i'}^2}} \quad (9)$$

where $[\mathbf{x}_j]_i$ denotes the i -th feature of the j -th example, which gives us a data-dependent dropout. We state it formally in the following definition.

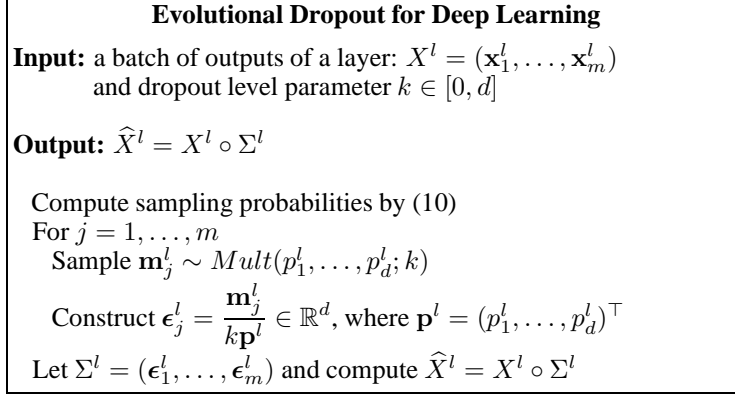


Figure 1: Evolutional Dropout applied to a layer over a mini-batch

Definition 2. (Data-dependent Dropout) Given a set of training examples $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$. A data-dependent dropout is defined as $\hat{\mathbf{x}} = \mathbf{x} \circ \epsilon$, where $\epsilon_i = \frac{m_i}{k p_i}$, $i \in [d]$ and $\{m_1, \dots, m_d\}$ follow a multinomial distribution $\text{Mult}(p_1, \dots, p_d; k)$ with p_i given by (9).

Remark: Note that if the data is normalized such that each feature has zero mean and unit variance (i.e., according to Z-normalization), the data-dependent dropout reduces to uniform dropout. It implies that the data-dependent dropout achieves similar effect as Z-normalization plus uniform dropout. In this sense, our theoretical analysis also explains why Z-normalization usually speeds up the training [13].

4.2 Evolutional Dropout for Deep Learning

Next, we discuss how to implement the distribution-dependent dropout for deep learning. In training deep neural networks, the dropout is usually added to the intermediate layers (e.g., fully connected layers and convolutional layers). Let $\mathbf{x}^l = (x_1^l, \dots, x_d^l)$ denote the outputs of the l -th layer (with the index of data omitted). Adding dropout to this layer is equivalent to multiplying \mathbf{x}^l by a dropout noise vector ϵ^l , i.e., feeding $\hat{\mathbf{x}}^l = \mathbf{x}^l \circ \epsilon^l$ as the input to the next layer. Inspired by the data-dependent dropout, we can generate ϵ^l according to a distribution given in Definition 1 with sampling probabilities p_i^l computed from $\{\mathbf{x}_1^l, \dots, \mathbf{x}_m^l\}$ similar to that (9). However, deep learning is usually trained with big data and a deep neural network is optimized by mini-batch stochastic gradient descent. Therefore, at each iteration it would be too expensive to afford the computation to pass through all examples. To address this issue, we propose to use a mini-batch of examples to calculate the second-order statistics similar to what was done in batch normalization. Let $X^l = (\mathbf{x}_1^l, \dots, \mathbf{x}_m^l)$ denote the outputs of the l -th layer for a mini-batch of m examples. Then we can calculate the probabilities for dropout by

$$p_i^l = \frac{\sqrt{\frac{1}{m} \sum_{j=1}^m [\|\mathbf{x}_j^l\|_i^2]}}{\sum_{i'=1}^d \sqrt{\frac{1}{m} \sum_{j=1}^m [\|\mathbf{x}_j^l\|_{i'}^2]}}, i = 1, \dots, d \quad (10)$$

which define the evolutional dropout named as such because the probabilities p_i^l will also evolve as the the distribution of the layer's outputs evolve. We describe the evolutional dropout as applied to a layer of a deep neural network in Figure 1.

Finally, we would like to compare the evolutional dropout with batch normalization. Similar to batch normalization, evolutional dropout can also address the internal covariate shift issue by adapting the sampling probabilities to the evolving distribution of layers' outputs. However, different from batch normalization, evolutional dropout is a randomized technique, which enjoys many benefits as standard dropout including (i) the back-propagation is simple to implement (just multiplying the gradient of \hat{X}^l by the dropout mask to get the gradient of X^l); (ii) the inference (i.e., testing) remains the same ²; (iii) it is equivalent to a data-dependent regularizer with a clear mathematical explanation;

²Different from some implementations for standard dropout which do not scale by $1/(1 - \delta)$ in training but scale by $1 - \delta$ in testing, here we do scale in training and thus do not need any scaling in testing.

(iv) it prevents units from co-adapting of neurons, which facilitate generalization. Moreover, the evolutionary dropout has its root in distribution-dependent dropout, which has theoretical guarantee to accelerate the convergence and improve the generalization for shallow learning.

5 Experimental Results

In the section, we present some experimental results to justify the proposed dropouts. In all experiments, we set $\delta = 0.5$ in the standard dropout and $k = 0.5d$ in the proposed dropouts for fair comparison, where d represents the number of features or neurons of the layer that dropout is applied to. For the sake of clarity, we divided the experiments into three parts. In the first part, we compare the performance of the data-dependent dropout (**d-dropout**) to the standard dropout (**s-dropout**) for logistic regression. In the second part, we compare the performance of evolutionary dropout (**e-dropout**) to the standard dropout for training deep convolutional neural networks. Finally, we compare e-dropout with batch normalization.

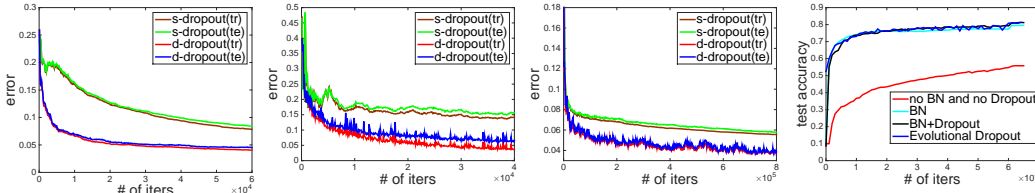


Figure 2: Left three: data-dependent dropout vs. standard dropout on three data sets (real-sim, news20, RCV1) for logistic regression; Right: Evolutional dropout vs BN on CIFAR-10. (best seen in color).

5.1 Shallow Learning

We implement the presented stochastic optimization algorithm. To evaluate the performance of data-dependent dropout for shallow learning, we use the three data sets: **real-sim, news20 and RCV1**³. In this experiment, we use a fixed step size and tune the step size in $[0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001]$ and report the best results in terms of convergence speed on the training data for both standard dropout and data-dependent dropout. The left three panels in Figure 2 show the obtained results on these three data sets. In each figure, we plot both the training error and the testing error. We can see that both the training and testing errors using the proposed data-dependent dropout decrease much faster than using the standard dropout and also a smaller testing error is achieved by using the data-dependent dropout.

5.2 Evolutional Dropout for Deep Learning

We would like to emphasize that we are not aiming to obtain better prediction performance by trying different network structures and different engineering tricks such as data augmentation, whitening, etc., but rather focus on the comparison of the proposed dropout to the standard dropout using Bernoulli noise on the same network structure. In our experiments, we use the default splitting of training and testing data in all data sets. We directly optimize the neural networks using all training images without further splitting it into a validation data to be added into the training in later stages, which explains some marginal gaps from the literature results that we observed (e.g., on CIFAR-10 compared with [19]).

We conduct experiments on **four benchmark data sets for comparing e-dropout and s-dropout: MNIST [10], SVHN [11], CIFAR-10 and CIFAR-100 [8]**. We use the same or similar network structure as in the literatures for the four data sets. In general, the networks consist of convolution layers, pooling layers, locally connected layers, fully connected layers, softmax layers and a cost layer. For the detailed neural network structures and their parameters, please refer to the supplementary materials. The dropout is added to some fully connected layers or locally connected layers. The **rectified linear activation function is used for all neurons**. All the experiments are conducted using the cuda-convnet library⁴. The training procedure is similar to [9] using mini-batch SGD with

³<https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>

⁴<https://code.google.com/archive/p/cuda-convnet/>

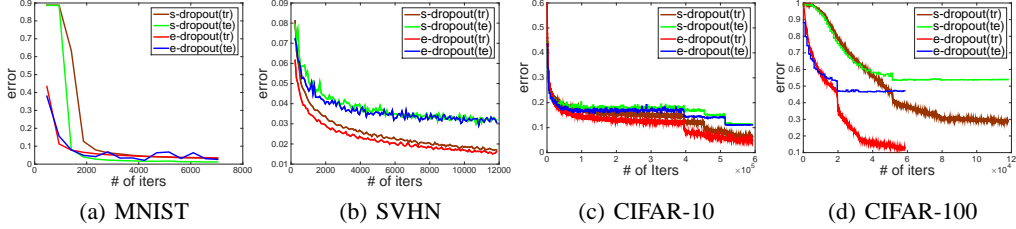


Figure 3: Evolutional dropout vs. standard dropout on four benchmark datasets for deep learning (best seen in color).

momentum (0.9). The size of mini-batch is fixed to 128. The weights are initialized based on the Gaussian distribution with mean zero and standard deviation 0.01. The learning rate (i.e., step size) is decreased after a number of epochs similar to what was done in previous works [9]. We tune the initial learning rates for s-dropout and e-dropout separately from 0.001, 0.005, 0.01, 0.1 and report the best result on each data set that yields the fastest convergence.

Figure 3 shows the training and testing error curves in the optimization process on the four data sets using the standard dropout and the evolutionary dropout. For SVHN data, we only report the first 12000 iterations, after which the error curves of the two methods almost overlap. We can see that using the evolutionary dropout generally converges faster than using the standard dropout. On CIFAR-100 data, we have observed significant speed-up. In particular, the evolutionary dropout achieves relative improvements over 10% on the testing performance and over 50% on the convergence speed compared to the standard dropout.

5.3 Comparison with the Batch Normalization (BN)

Finally, we make a comparison between the evolutionary dropout and the batch normalization. For batch normalization, we use the implementation in Caffe⁵. We compare the evolutionary dropout with the batch normalization on CIFAR-10 data set. The network structure is from the Caffe package and can be found in the supplement, which is different from the one used in the previous experiment. It contains three convolutional layers and one fully connected layer. Each convolutional layer is followed by a pooling layer. We compare four methods: (1) **No BN and No dropout** - without using batch normalization and dropout; (2) **BN**; (3) **BN with standard dropout**; (4) **Evolutional Dropout**. The rectified linear activation is used in all methods. We also tried BN with the sigmoid activation function, which gives worse results. For the methods with BN, three batch normalization layers are inserted before or after each pooling layer following the architecture given in Caffe package (see supplement). For the evolutionary dropout training, only one layer of dropout is added to the last convolutional layer. The mini-batch size is set to 100, the default value in Caffe. The initial learning rates for the four methods are set to the same value (0.001), and they are decreased once by ten times. The testing accuracy versus the number of iterations is plotted in the right panel of Figure 2, from which we can see that the evolutionary dropout training achieves comparable performance with BN + standard dropout, which justifies our claim that evolutionary dropout also addresses the internal covariate shift issue.

6 Conclusion

In this paper, we have proposed a distribution-dependent dropout for both shallow learning and deep learning. Theoretically, we proved that the new dropout achieves a smaller risk and faster convergence. Based on the distribution-dependent dropout, we developed an efficient evolutionary dropout for training deep neural networks that adapts the sampling probabilities to the evolving distributions of layers' outputs. Experimental results on various data sets verified that the proposed dropouts can dramatically improve the convergence and also reduce the testing error.

Acknowledgments

We thank anonymous reviewers for their comments. Z. Li and T. Yang are partially supported by National Science Foundation (IIS-1463988, IIS-1545995). B. Gong is supported in part by NSF (IIS-1566511) and a gift from Adobe.

⁵<https://github.com/BVLC/caffe/>

References

- [1] Jimmy Ba and Brendan Frey. Adaptive dropout for training deep neural networks. In *Advances in Neural Information Processing Systems*, pages 3084–3092, 2013.
- [2] Pierre Baldi and Peter J Sadowski. Understanding dropout. In *Advances in Neural Information Processing Systems*, pages 2814–2822, 2013.
- [3] Benjamin Graham, Jeremy Reizenstein, and Leigh Robinson. Efficient batchwise dropout training using submatrices. *CoRR*, abs/1502.02478, 2015.
- [4] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.
- [5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- [7] Diederik P. Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *CoRR*, abs/1506.02557, 2015.
- [8] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images, 2009.
- [9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [10] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [11] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, Spain, 2011.
- [12] Behnam Neyshabur, Ruslan R Salakhutdinov, and Nati Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2413–2421, 2015.
- [13] Marc’Aurelio Ranzato, Alex Krizhevsky, and Geoffrey E. Hinton. Factored 3-way restricted boltzmann machines for modeling natural images. In *AISTATS*, pages 621–628, 2010.
- [14] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *The 22nd Conference on Learning Theory (COLT)*, 2009.
- [15] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pages 2199–2207, 2010.
- [16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [17] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th international conference on machine learning (ICML-13)*, pages 1139–1147, 2013.
- [18] Stefan Wager, Sida Wang, and Percy S Liang. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems*, pages 351–359, 2013.
- [19] Li Wan, Matthew Zeiler, Sixin Zhang, Yann L Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1058–1066, 2013.
- [20] Sida Wang and Christopher Manning. Fast dropout training. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 118–126, 2013.
- [21] Sida I Wang, Mengqiu Wang, Stefan Wager, Percy Liang, and Christopher D Manning. Feature noising for log-linear structured prediction. In *EMNLP*, pages 1170–1179, 2013.
- [22] Sixin Zhang, Anna Choromanska, and Yann LeCun. Deep learning with elastic averaging sgd. *arXiv preprint arXiv:1412.6651*, 2014.
- [23] Jingwei Zhuo, Jun Zhu, and Bo Zhang. Adaptive dropout rates for learning with corrupted features. In *IJCAI*, pages 4126–4133, 2015.
- [24] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. 2003.

7 Supplement

7.1 Proof of Theorem 1

The update given by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \ell(\mathbf{w}_t^\top (\mathbf{x}_t \circ \boldsymbol{\epsilon}_t), y_t)$ can be considered as the stochastic gradient descent (SGD) update of the following problem

$$\min_{\mathbf{w}} \{ \widehat{\mathcal{L}}(\mathbf{w}) \triangleq \mathbb{E}_{\widehat{\mathcal{D}}} [\ell(\mathbf{w}^\top (\mathbf{x} \circ \boldsymbol{\epsilon}), y)] \}$$

Define \mathbf{g}_t as $\mathbf{g}_t = \nabla \ell(\mathbf{w}_t^\top (\mathbf{x}_t \circ \boldsymbol{\epsilon}_t), y_t) = \ell'(\mathbf{w}_t^\top (\mathbf{x}_t \circ \boldsymbol{\epsilon}_t), y_t) \mathbf{x}_t \circ \boldsymbol{\epsilon}_t$, where $\ell'(z, y)$ denotes the derivative in terms of z . Since the loss function is G -Lipschitz continuous, therefore $\|\mathbf{g}_t\|_2 \leq G \|\mathbf{x}_t \circ \boldsymbol{\epsilon}_t\|_2$. According to the analysis of SGD [24], we have the following lemma.

Lemma 1. *Let $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_t$ and $\mathbf{w}_1 = 0$. Then for any $\|\mathbf{w}_*\|_2 \leq r$ we have*

$$\sum_{t=1}^n \mathbf{g}_t^\top (\mathbf{w}_t - \mathbf{w}_*) \leq \frac{r^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^n \|\mathbf{g}_t\|_2^2 \quad (11)$$

By taking expectation on both sides over the randomness in $(\mathbf{x}_t, y_t, \boldsymbol{\epsilon}_t)$ and noting the bound on $\|\mathbf{g}_t\|_2$, we have

$$\mathbb{E}_{[n]} \left[\sum_{t=1}^n \mathbf{g}_t^\top (\mathbf{w}_t - \mathbf{w}_*) \right] \leq \frac{r^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^n G^2 \mathbb{E}_{[n]} [\|\mathbf{x}_t \circ \boldsymbol{\epsilon}_t\|_2^2]$$

where $\mathbb{E}_{[t]}$ denote the expectation over $(\mathbf{x}_i, y_i, \boldsymbol{\epsilon}_i), i = 1, \dots, t$. Let $\mathbb{E}_t[\cdot]$ denote the expectation over $(\mathbf{x}_t, y_t, \boldsymbol{\epsilon}_t)$ with $(\mathbf{x}_i, y_i, \boldsymbol{\epsilon}_i), i = 1, \dots, t-1$ given. Then we have

$$\sum_{t=1}^n \mathbb{E}_{[t]} [\mathbf{g}_t^\top (\mathbf{w}_t - \mathbf{w}_*)] \leq \frac{r^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^n G^2 \mathbb{E}_t [\|\mathbf{x}_t \circ \boldsymbol{\epsilon}_t\|_2^2]$$

Since

$$\mathbb{E}_{[t]} [\mathbf{g}_t^\top (\mathbf{w}_t - \mathbf{w}_*)] = \mathbb{E}_{[t-1]} [\mathbb{E}_t [\mathbf{g}_t]^\top (\mathbf{w}_t - \mathbf{w}_*)] = \mathbb{E}_{[t-1]} [\nabla \widehat{\mathcal{L}}(\mathbf{w}_t)^\top (\mathbf{w}_t - \mathbf{w}_*)] \geq \mathbb{E}_{[t-1]} [\widehat{\mathcal{L}}(\mathbf{w}_t) - \widehat{\mathcal{L}}(\mathbf{w}_*)]$$

As a result

$$\mathbb{E}_{[n]} \left[\sum_{t=1}^n (\widehat{\mathcal{L}}(\mathbf{w}_t) - \widehat{\mathcal{L}}(\mathbf{w}_*)) \right] \leq \frac{r^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^n G^2 \mathbb{E}_{\widehat{\mathcal{D}}} [\|\mathbf{x}_t \circ \boldsymbol{\epsilon}_t\|_2^2] \leq \frac{r^2}{2\eta} + \frac{\eta}{2} G^2 B^2 n \quad (12)$$

where the last inequality follows the assumed upper bound of $\mathbb{E}_{\widehat{\mathcal{D}}} [\|\mathbf{x}_t \circ \boldsymbol{\epsilon}_t\|_2^2]$. Following the definition of $\widehat{\mathbf{w}}_n$ and the convexity of $\mathcal{L}(\mathbf{w})$ we have

$$\mathbb{E}_{[n]} [\widehat{\mathcal{L}}(\widehat{\mathbf{w}}_n) - \widehat{\mathcal{L}}(\mathbf{w}_*)] \leq \mathbb{E}_{[n]} \left[\frac{1}{n} \sum_{t=1}^n (\widehat{\mathcal{L}}(\mathbf{w}_t) - \widehat{\mathcal{L}}(\mathbf{w}_*)) \right] \leq \frac{r^2}{2\eta n} + \frac{\eta}{2} G^2 B^2$$

By minimizing the upper bound in terms of η , we have $\mathbb{E}_{[n]} [\widehat{\mathcal{L}}(\widehat{\mathbf{w}}_n) - \widehat{\mathcal{L}}(\mathbf{w}_*)] \leq \frac{GBr}{\sqrt{n}}$. According to Proposition 1 in the paper $\widehat{\mathcal{L}}(\mathbf{w}) = \mathcal{L}(\mathbf{w}) + R_{\mathcal{D}, \mathcal{M}}(\mathbf{w})$, therefore

$$\mathbb{E}_{[n]} [\mathcal{L}(\widehat{\mathbf{w}}_n) + R_{\mathcal{D}, \mathcal{M}}(\widehat{\mathbf{w}}_n)] \leq \mathcal{L}(\mathbf{w}_*) + R_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*) + \frac{GBr}{\sqrt{n}}$$

7.2 Proof of Lemma 1

We have the following:

$$\frac{1}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2 = \frac{1}{2} \|\mathbf{w}_t - \eta \mathbf{g}_t - \mathbf{w}_*\|_2^2 = \frac{1}{2} \|\mathbf{w}_t - \mathbf{w}_*\|_2^2 + \frac{\eta^2}{2} \|\mathbf{g}_t\|_2^2 - \eta (\mathbf{w}_t - \mathbf{w}_*)^\top \mathbf{g}_t$$

Then

$$(\mathbf{w}_t - \mathbf{w}_*)^\top \mathbf{g}_t \leq \frac{1}{2\eta} \|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - \frac{1}{2\eta} \|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2 + \frac{\eta}{2} \|\mathbf{g}_t\|_2^2$$

By summing the above inequality over $t = 1, \dots, n$, we obtain

$$\sum_{t=1}^n \mathbf{g}_t^\top (\mathbf{w}_t - \mathbf{w}_*) \leq \frac{\|\mathbf{w}_* - \mathbf{w}_1\|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^n \|\mathbf{g}_t\|_2^2$$

By noting that $\mathbf{w}_1 = 0$ and $\|\mathbf{w}_*\|_2 \leq r$, we obtain the inequality in Lemma 1.

7.3 Proof of Proposition 2

We have

$$\mathbb{E}_{\mathcal{D}}[\|\mathbf{x} \circ \boldsymbol{\epsilon}\|_2^2] = \mathbb{E}_{\mathcal{D}} \left[\sum_{i=1}^d \frac{x_i^2}{k^2 p_i^2} \mathbb{E}[m_i^2] \right]$$

Since $\{m_1, \dots, m_d\}$ follows a multinomial distribution $Mult(p_1, \dots, p_d; k)$, we have

$$\mathbb{E}[m_i^2] = \text{var}(m_i) + (\mathbb{E}[m_i])^2 = kp_i(1 - p_i) + k^2 p_i^2$$

The result in the Proposition follows by combining the above two equations.

7.4 Proof of Proposition 3

Note that only the first term in the R.H.S of Eqn. (7) depends on p_i . Thus,

$$\mathbf{p}_* = \arg \min_{\mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1} \sum_{i=1}^d \frac{\mathbb{E}_{\mathcal{D}}[x_i^2]}{p_i}$$

The result then follows the KKT conditions.

7.5 Proof of Proposition 4

We prove the first upper bound first. From Eqn. (4) in the paper, we have

$$\hat{R}_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*) \leq \frac{1}{8} \mathbb{E}_{\mathcal{D}}[\mathbf{w}_*^\top C_{\mathcal{M}}(\mathbf{x} \circ \boldsymbol{\epsilon}) \mathbf{w}_*]$$

where we use the fact $\sqrt{ab} \leq \frac{a+b}{2}$ for $a, b \geq 0$. Using Eqn. (5) in the paper, we have

$$\mathbb{E}_{\mathcal{D}}[\mathbf{w}_*^\top C_{\mathcal{M}}(\mathbf{x} \circ \boldsymbol{\epsilon}) \mathbf{w}_*] = \mathbb{E}_{\mathcal{D}} \left[\mathbf{w}_*^\top \left(\frac{1}{k} \text{diag}(x_i^2/p_i) - \frac{1}{k} \mathbf{x} \mathbf{x}^\top \right) \mathbf{w}_* \right] = \frac{1}{k} \mathbb{E}_{\mathcal{D}} \left[\sum_{i=1}^d \frac{w_{*i}^2 x_i^2}{p_i} - (\mathbf{w}_*^\top \mathbf{x})^2 \right]$$

This gives a tight bound of $\hat{R}_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*)$, i.e.,

$$\hat{R}_{\mathcal{D}, \mathcal{M}}(\mathbf{w}_*) \leq \frac{1}{8k} \left\{ \sum_{i=1}^d \frac{w_{*i}^2 \mathbb{E}_{\mathcal{D}}[x_i^2]}{p_i} - \mathbb{E}_{\mathcal{D}}(\mathbf{w}_*^\top \mathbf{x})^2 \right\}$$

By minimizing the above upper bound over p_i , we obtain following probabilities

$$p_i^* = \frac{\sqrt{w_{*i}^2 \mathbb{E}_{\mathcal{D}}[x_i^2]}}{\sum_{j=1}^d \sqrt{w_{*j}^2 \mathbb{E}_{\mathcal{D}}[x_j^2]}} \quad (13)$$

which depend on unknown \mathbf{w}_* . We address this issue, we derive a relaxed upper bound. We note that

$$\begin{aligned} C_{\mathcal{M}}(\mathbf{x} \circ \boldsymbol{\epsilon}) &= \mathbb{E}_{\mathcal{M}}[(\mathbf{x} \circ \boldsymbol{\epsilon} - \mathbf{x})(\mathbf{x} \circ \boldsymbol{\epsilon} - \mathbf{x})^\top] \\ &\leq (\mathbb{E}_{\mathcal{M}}[\|\mathbf{x} \circ \boldsymbol{\epsilon} - \mathbf{x}\|_2^2]) \cdot I_d = (\mathbb{E}_{\mathcal{M}}[\|\mathbf{x} \circ \boldsymbol{\epsilon}\|_2^2] - \|\mathbf{x}\|_2^2) I_d \end{aligned}$$

where I_d denotes the identity matrix of dimension d . Thus

$$\mathbb{E}_{\mathcal{D}}[\mathbf{w}_*^\top C_{\mathcal{M}}(\mathbf{x} \circ \boldsymbol{\epsilon}) \mathbf{w}_*] \leq \|\mathbf{w}_*\|_2^2 (\mathbb{E}_{\mathcal{D}}[\|\mathbf{x} \circ \boldsymbol{\epsilon}\|_2^2] - \mathbb{E}_{\mathcal{D}}[\|\mathbf{x}\|_2^2])$$

By noting the result in Proposition 2 in the paper, we have

$$\mathbb{E}_{\mathcal{D}}[\mathbf{w}_*^\top C_{\mathcal{M}}(\mathbf{x} \circ \boldsymbol{\epsilon}) \mathbf{w}_*] \leq \frac{1}{k} \|\mathbf{w}_*\|_2^2 \left(\sum_{i=1}^d \frac{\mathbb{E}_{\mathcal{D}}[x_i^2]}{p_i} - \mathbb{E}_{\mathcal{D}}[\|\mathbf{x}\|_2^2] \right)$$

which proves the upper bound in Proposition 4.

7.6 Neural Network Structures

In this section we present the neural network structures and the number of filters, filter size, padding and stride parameters for MNIST, SVHN, CIFAR-10 and CIFAR-100, respectively. Note that in Table 2, Table 3 and Table 4, the rnorm layer is the local response normalization layer and the local layer is the locally-connected layer with unshared weights.

7.6.1 MNIST

We used the similar neural network structure to [19]: two convolution layers, two fully connected layers, a softmax layer and a cost layer at the end. The dropout is added to the first fully connected layer. Tables 1 presents the neural network structures and the number of filters, filter size, padding and stride parameters for MNIST.

Table 1: The Neural Network Structure for MNIST

Layer Type	Input Size	#Filters	Filter size	Padding/Stride	Output Size
conv1	$28 \times 28 \times 1$	32	4×4	0/1	$21 \times 21 \times 32$
pool1(max)	$21 \times 21 \times 32$		2×2	0/2	$11 \times 11 \times 32$
conv2	$11 \times 11 \times 32$	64	5×5	0/1	$7 \times 7 \times 64$
pool2(max)	$7 \times 7 \times 64$		3×3	0/3	$3 \times 3 \times 64$
fc1	$3 \times 3 \times 64$				150
dropout	150				150
fc2	150				10
softmax	10				10
cost	10				1

7.6.2 SVHN

The neural network structure used for this data set is from [19], including 2 convolutional layers, 2 max pooling layers, 2 local response layers, 2 fully connected layers, a softmax layer and a cost layer with one dropout layer. Tables 2 presents the neural network structures and the number of filters, filter size, padding and stride parameters used for SVHN data set.

Table 2: The Neural Network Structure for SVHN

Layer Type	Input Size	#Filters	Filter Size	Padding/Stride	Output Size
conv1	$28 \times 28 \times 3$	64	5×5	0/1	$24 \times 24 \times 64$
pool1(max)	$24 \times 24 \times 64$		3×3	0/2	$12 \times 12 \times 64$
rnorm1	$12 \times 12 \times 64$				$12 \times 12 \times 64$
conv2	$12 \times 12 \times 64$	64	5×5	2/1	$12 \times 12 \times 64$
rnorm2	$12 \times 12 \times 64$				$12 \times 12 \times 64$
pool2(max)	$12 \times 12 \times 64$		3×3	0/2	$6 \times 6 \times 64$
local3	$6 \times 6 \times 64$	64	3×3	1/1	$6 \times 6 \times 64$
local4	$6 \times 6 \times 64$	32	3×3	1/1	$6 \times 6 \times 32$
dropout	1152				1152
fc1	1152				512
fc10	512				10
softmax	10				10
cost	10				1

7.6.3 CIFAR-10

The neural network structure is adopted from [19], which consists two convolutional layer, two pooling layers, two local normalization response layers, 2 locally connected layers, two fully connected

layers and a softmax and a cost layer. Table 3 presents the detail neural network structure and the number of filters, filter size, padding and stride parameters used.

Table 3: The Neural Network Structure for CIFAR-10

Layer Type	Input Size	#Filters	Filter Size	Padding/Stride	Output Size
conv1	$24 \times 24 \times 3$	64	5×5	2/1	$24 \times 24 \times 64$
pool1(max)	$24 \times 24 \times 64$		3×3	0/2	$12 \times 12 \times 64$
rnorm1	$12 \times 12 \times 64$				$12 \times 12 \times 64$
conv2	$12 \times 12 \times 64$	64	5×5	2/1	$12 \times 12 \times 64$
rnorm2	$12 \times 12 \times 64$				$12 \times 12 \times 64$
pool2(max)	$12 \times 12 \times 64$		3×3	0/2	$6 \times 6 \times 64$
local3	$6 \times 6 \times 64$	64	3×3	1/1	$6 \times 6 \times 64$
local4	$6 \times 6 \times 64$	32	3×3	1/1	$6 \times 6 \times 32$
dropout	1152				1152
fc1	1152				128
fc10	128				10
softmax	10				10
cost	10				1

7.6.4 CIFAR-100

The network structure for this data set is similar to the neural network structure in [8], which consists of 2 convolution layers, 2 max pooling layers, 2 local response normalization layers, 2 locally connected layers, 3 fully connected layers, and a softmax and a cost layer. Table 4 presents the neural network structures and the number of filters, filter size, padding and stride parameters used for CIFAR-100 data set.

7.6.5 The Neural Network Structure used for BN

Tables 5 and 6 present the network structures of different methods in subsection 5.3 in the paper. The layer pool(ave) in Table 5 and Table 6 represents the average pooling layer.

Table 4: The Neural Network Structure for CIFAR-100

Layer Type	Input Size	#Filters	Filter Size	Padding/Stride	Output Size
conv1	$32 \times 32 \times 3$	64	5×5	2/1	$32 \times 32 \times 64$
pool1(max)	$32 \times 32 \times 64$		3×3	0/2	$16 \times 16 \times 64$
rnorm1	$16 \times 16 \times 64$				$16 \times 16 \times 64$
conv2	$16 \times 16 \times 64$	64	5×5	2/1	$16 \times 16 \times 64$
rnorm2	$16 \times 16 \times 64$				$16 \times 16 \times 64$
pool2(max)	$16 \times 16 \times 64$		3×3	0/2	$8 \times 8 \times 64$
local3	$8 \times 8 \times 64$	64	3×3	1/1	$8 \times 8 \times 64$
local4	$8 \times 8 \times 64$	32	3×3	1/1	$8 \times 8 \times 32$
fc1	2048				128
dropout	128				128
fc2	128				128
fc100	128				100
softmax	100				100
cost	100				1

Table 5: Layers of networks for the experiment comparing with BN on CIFAR-10

Layer Type	noBN-noDropout	BN	e-dropout
Layer 1	conv1	conv1	conv1
Layer 2	pool1(max)	pool(max)	pool1(max)
Layer 3	N/A	bn1	N/A
Layer 4	conv2	conv2	conv2
Layer 5	N/A	bn2	N/A
Layer 6	pool2(ave)	pool2(ave)	pool2(ave)
Layer 7	conv3	conv3	conv3
Layer 8	N/A	bn3	e-dropout
Layer 9	pool3(ave)	pool3(ave)	pool3(ave)
Layer 10	fc1	fc1	fc1
Layer 11	softmax	softmax	softmax

Table 6: Sizes in networks for the experiment comparing with BN on CIFAR-10

Layer Type	Input size	#Filters	Filter size	Padding/Stride	Output size
conv1	$32 \times 32 \times 3$	32	5×5	2/1	$32 \times 32 \times 32$
pool1(max)	$32 \times 32 \times 32$		3×3	0/2	$16 \times 16 \times 32$
conv2	$16 \times 16 \times 32$	32	5×5	2/1	$16 \times 16 \times 32$
pool2(ave)	$16 \times 16 \times 32$		3×3	0/2	$8 \times 8 \times 32$
conv3	$8 \times 8 \times 32$	64	5×5	2/1	$8 \times 8 \times 64$
pool3(ave)	$8 \times 8 \times 64$		3×3	0/2	$4 \times 4 \times 64$
fc1	$4 \times 4 \times 64$				10
softmax	10				10
cost	10				1