Journal of Machine Learning Research 16 (2015) 3403-3454 Submitted 12/14; Revised 4/15; Published 12/15

On the Inductive Bias of Dropout

David P. Helmbold DPH@SOE.UCSC.EDU
Department of Computer Science
University of California, Santa Cruz
Santa Cruz, CA 95064, USA

Philip M. Long PLONG@MICROSOFT.COM
Microsoft
1020 Enterprise Way
Sunnyvale, CA 94089, USA

Editor: Samy Bengio

Abstract

Dropout is a simple but effective technique for learning in neural networks and other settings. A sound theoretical understanding of dropout is needed to determine when dropout should be applied and how to use it most effectively. In this paper we continue the exploration of dropout as a regularizer pioneered by Wager et al. We focus on linear classification where a convex proxy to the misclassification loss (i.e. the logistic loss used in logistic regression) is minimized. We show:

- when the dropout-regularized criterion has a unique minimizer,
- when the dropout-regularization penalty goes to infinity with the weights, and when it remains bounded,
- that the dropout regularization can be non-monotonic as individual weights increase from 0, and
- that the dropout regularization penalty may *not* be convex.

This last point is particularly surprising because the combination of dropout regularization with any convex loss proxy is always a convex function.

In order to contrast dropout regularization with L_2 regularization, we formalize the notion of when different random sources of data are more compatible with different regularizers. We then exhibit distributions that are provably more compatible with dropout regularization than L_2 regularization, and vice versa. These sources provide additional insight into how the inductive biases of dropout and L_2 regularization differ. We provide some similar results for L_1 regularization.

Keywords: dropout, inductive bias, learning theory, regularization, feature noising

1. Introduction

Since its prominent role in a win of the ImageNet Large Scale Visual Recognition Challenge (Hinton, 2012; Hinton et al., 2012; Srivastava et al., 2014), there has been intense interest in dropout (see the work by Dahl, 2012; Deng et al., 2013; Dahl et al., 2013; Wan et al., 2013; Wager et al., 2013; Baldi and Sadowski, 2014; Van Erven et al., 2014). Dropout is a modification of stochastic gradient descent where each update is performed on a reduced network created by temporarily removing a random subset of the nodes. This paper studies the inductive bias of dropout: when one chooses to train with dropout, what prior preference over models results? We show that dropout training shapes the learner's search space in a much different way than L_1 or L_2 regularization. Our results shed new insight into why dropout prefers rare features, how the dropout probability affects the strength of regularization, and how dropout restricts the co-adaptation of weights.

Our theoretical study will concern learning a linear classifier via convex optimization. The learner wishes to find a parameter vector $\mathbf{w}$ so that, for a random feature-label pair $(\mathbf{x}, y) \in \mathbf{R}^n \times \{-1, 1\}$ drawn from some joint distribution P, the probability that $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y$ is small. It does this by using training data to try to minimize $\mathbf{E}(\ell(y\mathbf{w} \cdot \mathbf{x}))$, where $\ell(z) = \ln(1 + \exp(-z))$ is the loss function associated with logistic regression.

We have chosen to focus on this problem for several reasons. First, the inductive bias of dropout is not well understood even in this simple setting. Second, linear classifiers remain a popular choice for practical problems, especially in the case of very high-dimensional data. Third, we view a thorough understanding of dropout in this setting as a mandatory prerequisite to understanding the inductive bias of dropout when applied in a deep learning architecture. This is especially true when the preference over deep learning models is decomposed into preferences at each node. In any case, the setting that we are studying faithfully describes the inductive bias of a deep learning system at its output nodes.

We will borrow the following clean and illuminating description of dropout as artificial noise due to Wager et al. (2013). An algorithm for linear classification using loss ℓ and dropout updates its parameter vector $\mathbf{w}$ online, using stochastic gradient descent. Given an example $(\mathbf{x}, y)$, the dropout algorithm independently perturbs each feature i of $\mathbf{x}$: with probability q, x_i is replaced with 0, and, with probability $p = 1 - q$, x_i is replaced with x_i/p. Equivalently, $\mathbf{x}$ is replaced by $\mathbf{x} + \boldsymbol{\nu}$, where

$$\nu_i = \begin{cases} -x_i & \text{with probability } q \\ (1/p - 1)x_i & \text{with probability } p = 1 - q \end{cases}$$

before performing the stochastic gradient update step. (Note that, while $\boldsymbol{\nu}$ obviously depends on $\mathbf{x}$, if we sample the components of $\mathbf{b} \in \{-1, 1/p-1\}^n$ independently of one another and $\mathbf{x}$, by choosing $b_i = -1$ with the dropout probability q, then we may write $\nu_i = b_i x_i$.)

Stochastic gradient descent is known to converge under a broad variety of conditions (Kushner and Yin, 1997). Thus, if we abstract away sampling issues as done by Breiman (2004); Zhang (2004); Bartlett et al. (2006); Long and Servedio (2010), we are led to consider

$$\mathbf{w}^* \stackrel{\text{def}}{=} \mathrm{argmin}_{\mathbf{w}} \mathbf{E}_{(\mathbf{x},y)\sim P, \boldsymbol{\nu}}(\ell(y\mathbf{w} \cdot (\mathbf{x} + \boldsymbol{\nu})))$$

as dropout can be viewed as a stochastic gradient update of this global objective function. We call this objective the *dropout criterion*, and it can be viewed as a risk on the dropout-induced distribution. (Abstracting away sampling issues is consistent with our goal of concentrating on the inductive bias of the algorithm. From the point of view of a bias-variance decomposition, we do not intend to focus on the large-sample-size case, where the variance is small, but rather to focus on the contribution from the bias where P could be an empirical sample distribution.)

We start with the observation of Wager et al. (2013) that the dropout criterion may be decomposed as

$$\mathbf{E}_{(\mathbf{x},y)\sim P,\boldsymbol{\nu}}(\ell(y\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu}))) = \mathbf{E}_{(\mathbf{x},y)\sim P}(\ell(y\mathbf{w}\cdot\mathbf{x})) + \mathbf{reg}_{D,q}(\mathbf{w}), \tag{1}$$

where $\mathbf{reg}_{D,q}(\mathbf{w})$ is non-negative, and depends only on the marginal distribution D over the feature vectors $\mathbf{x}$ (along with the dropout probability q), and not on the labels. This leads naturally to a view of dropout as a regularizer.

A popular style of learning algorithm minimizes an objective function like the RHS of (1), but where $\mathbf{reg}_{D,q}(\mathbf{w})$ is replaced by a norm of $\mathbf{w}$. One motivation for algorithms in this family is to first replace the training error with a convex proxy to make optimization tractable, and then to regularize using a convex penalty such as a norm, so that the objective function remains convex.

We show that $\mathbf{reg}_{D,q}(\mathbf{w})$ formalizes a preference for classifiers that assign a very large weight to a single feature. This preference is stronger than what one gets from a penalty proportional to $||\mathbf{w}||_1$. In fact, despite the convexity of the dropout risk, we show that $\mathbf{reg}_{D,q}(\mathbf{w})$ is *not* convex. Therefore that dropout provides a way to realize the inductive bias arising from a non-convex penalty while still enjoying the benefit of convexity in the overall objective function (see the plots in Figures 1, 2 and 3). Figure 1 shows the even more surprising result that the dropout regularization penalty is not even monotonic in the absolute values of the individual weights.

It is not hard to see that $\mathbf{reg}_{D,q}(\mathbf{0}) = 0$. Thus, if $\mathbf{reg}_{D,q}(\mathbf{w})$ is greater than the expected loss incurred by $\mathbf{0}$ (which is $\ln 2$), then it might as well be infinity, because dropout will prefer $\mathbf{0}$ to $\mathbf{w}$. However, in some cases, dropout never reaches this extreme—it remains willing to use a models with arbitrarily large parameters, unlike methods that use a convex penalty. In particular,

$$\mathbf{reg}_{D,q}(w_1, 0, 0, 0, ..., 0) < \ln 2$$

for all D, no matter how large w_1 gets. On the other hand, except for some special cases (which are detailed in the body of the paper),

$$\mathbf{reg}_{D,q}(cw_1, cw_2, 0, 0, ..., 0)$$

goes to infinity with c. It follows that $\mathbf{reg}_{D,q}(\mathbf{w})$ cannot be approximated to within any factor, constant or otherwise, by a convex function of $\mathbf{w}$.

To get a sense of which sources dropout can be successfully applied to, we compare dropout with an algorithm that regularizes using L_2, by minimizing the L_2 *criterion*:

$$\mathbf{E}_{(\mathbf{x},y)\sim P}(\ell(y\mathbf{w}\cdot\mathbf{x})) + \frac{\lambda}{2}||\mathbf{w}||_2^2. \tag{2}$$

Will will use "L_2" as a shorthand to refer to an algorithm that minimizes (2). Note that q, the probability of dropping out an input feature, plays a role in dropout analogous to λ. In particular, as q goes to zero the examples remain unperturbed and the dropout regularization has no effect.

Informally, we say that joint probability distributions P and Q *separate* dropout from L_2 if, when the same parameters λ and q are used for both P and Q, then using dropout leads to a much more accurate hypothesis for P, and using L_2 leads to a much more accurate hypothesis for Q. This enables us to illustrate the inductive biases of the algorithms through contrasting sources that either align or are incompatible with the algorithms' inductive bias. Comparing with another regularizer helps to restrict these illustrative examples to "reasonable" sources, which can be handled using the other regularizer. Ensuring that the same values of the regularization parameter are used for both P and Q controls for the amount of regularization, and ensures that the difference is due to the model preferences of the respective regularizers. This style of analysis is new, as far as we know, and may be a useful tool for studying the inductive biases of other algorithms and in other settings.

Related previous work. Our research builds on the work of Wager et al. (2013), who analyzed dropout for random (x, y) pairs where the distribution of y given x comes from a member of the exponential family, and the quality of a model is evaluated using the log-loss. They pointed out that, in these cases, the dropout criterion can be decomposed into the original loss and a term that does not depend on y, which therefore can be viewed as a regularizer. They then proposed an approximation to this dropout regularizer, discussed its relationship with other regularizers and training algorithms, and evaluated it experimentally. Baldi and Sadowski (2014) exposed properties of dropout when viewed as an ensemble method (see also Bachman et al., 2014). Van Erven et al. (2014) showed that applying dropout for online learning in the experts setting leads to algorithms that adapt to important properties of the input without requiring doubling or other parameter-tuning techniques, and Abernethy et al. (2014) analyzed a class of methods including dropout by viewing these methods as smoothers. The impact of dropout on generalization (roughly, how much dropout restricts the search space of the learner, or, from a bias-variance point of view, its impact on variance) was studied by Wan et al. (2013) and Wager et al. (2014). The latter paper considers a variant of dropout compatible with a Poisson source, and shows that under some assumptions this dropout variant converges more quickly to its infinite sample limit than non-dropout training, and that the Bayes-optimal predictions are preserved under the modified dropout distribution. Our results complement theirs by focusing on the effect of the original dropout on the algorithm's bias.

Section 2 defines our notation and characterizes when the dropout criterion has a unique minimizer. Section 3 presents many additional properties of the dropout regularizer. Section 4 formally defines when two distributions separate two algorithms or regularizers. Sections 5 and 6 give sources over $\mathbf{R}^2$ that separate dropout and L_2; these exploit the preference of dropout for hypotheses that concentrate weight on a single feature. Section 7 provides plots demonstrating that the same distributions separate dropout from L_1 regularization. Section 8 gives a definition of co-adaptation and shows (using plots) that distributions exploiting dropout's bias against co-adapted weights can also be used to separate dropout from L_2 and L_1 regularization. Sections 9 and 10 give additional separation results using distributions with many features.

2. Preliminaries

We use $\mathbf{w}^*$ for the optimizer of the dropout criterion, q for the probability that a feature is dropped out, and $p = 1 - q$ for the probability that a feature is kept throughout the paper.

As in the introduction, if $X \subseteq \mathbf{R}^n$ and P is a joint distribution over $X \times \{-1, 1\}$, define

$$\mathbf{w}^*(P, q) \stackrel{\text{def}}{=} \text{argmin}_{\mathbf{w}} \mathbf{E}_{(\mathbf{x},y)\sim P,\boldsymbol{\nu}}(\ell(y\mathbf{w} \cdot (\mathbf{x} + \boldsymbol{\nu}))) \tag{3}$$

where $\nu_i = b_i x_i$ for $b_1, ..., b_n$ sampled independently at random from $\{-1, 1/p - 1\}$ with $\mathbf{Pr}(b_i = 1/p - 1) = p = 1 - q$, and $\ell(z)$ is the logistic loss function:

$$\ell(z) = \ln(1 + \exp(-z)).$$

For some analyses, an alternative representation of $\mathbf{w}^*(P, q)$ will be easier to work with. Let $r_1, ..., r_n$ be sampled randomly from $\{0, 1\}$, independently of $(\mathbf{x}, y)$ and one another, with $\mathbf{Pr}(r_i = 1) = p$. Defining $\mathbf{r} \odot \mathbf{x} = (x_1 r_1, ..., x_n r_n)$, we have the equivalent definition

$$\mathbf{w}^*(P, q) = p \text{ argmin}_{\mathbf{w}} \mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y\mathbf{w} \cdot (\mathbf{r} \odot \mathbf{x}))). \tag{4}$$

To see that they are equivalent, note that

$$\begin{aligned}\mathbf{E}(\ell(y\mathbf{w} \cdot (\mathbf{x} + \boldsymbol{\nu}))) &= \mathbf{E}\left(\ell\left(y\mathbf{w} \cdot \left(\frac{\mathbf{r} \odot \mathbf{x}}{p}\right)\right)\right) \\ &= \mathbf{E}(\ell(y(\mathbf{w}/p) \cdot (\mathbf{r} \odot \mathbf{x}))).\end{aligned}$$

Although this paper focuses on the logistic loss, the above definitions can be used for any loss function $\ell()$. Since the dropout criterion is an expectation of $\ell()$, we have the following obvious consequence.

Proposition 1 *If loss $\ell(\cdot)$ is convex, then the dropout criterion is also a convex function of* $\mathbf{w}$.

The remainder of the paper focuses on the logistic loss, $\ell(y\mathbf{w} \cdot \mathbf{x}) = \ln(1 + \exp(-y\mathbf{w} \cdot \mathbf{x}))$. We use $\mathbf{v}$ for the optimizer of the L_2 regularized criterion:

$$\mathbf{v}(P, \lambda) \stackrel{\text{def}}{=} \text{argmin}_{\mathbf{w}} \mathbf{E}_{(\mathbf{x},y)\sim P}(\ell(y\mathbf{w} \cdot \mathbf{x})) + \frac{\lambda}{2}||\mathbf{w}||^2. \tag{5}$$

It is not hard to see that the $\frac{\lambda}{2}||\mathbf{w}||^2$ term implies that $\mathbf{v}(P, \lambda)$ is always well-defined. On the other hand, $\mathbf{w}^*(P, q)$ is *not* always well-defined, as can be seen by considering any distribution concentrated on a single example. This motivates the following definition.

Definition 2 *Let P be a joint distribution with support contained in $\mathbf{R}^n \times \{-1, 1\}$. A feature i is perfect modulo ties for P if either $yx_i \geq 0$ for all $\mathbf{x}$ in the support of P, or $yx_i \leq 0$ for all $\mathbf{x}$ in the support of P.*

Put another way, i is perfect modulo ties if there is a linear classifier that only pays attention to feature i and is perfect on the part of P where x_i is nonzero.

Proposition 3 *For all finite domains $X \subseteq \mathbf{R}^n$, all distributions P with support in X, and all $q \in (0, 1)$, we have that $\mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y\mathbf{w} \cdot (\mathbf{r} \odot \mathbf{x})))$ has a unique minimum in $\mathbf{R}^n$ if and only if no feature is perfect modulo ties for P.*

$\mathbf{x} = (x_1, \ldots, x_n)$	feature vector in $\mathbf{R}^n$
y	label in $\{-1, +1\}$
$\mathbf{w} = (w_1, \ldots, w_n)$	weight vector in $\mathbf{R}^n$
$\ell(y\mathbf{w} \cdot \mathbf{x})$	loss function, generally the logistic loss: $\ln(1 + \exp(-y\mathbf{w} \cdot \mathbf{x}))$
P, Q	source distributions over $(\mathbf{x}, y)$ pairs, varies by section
D	marginal distribution over $\mathbf{x}$
q	feature dropout probability in $(0, 1)$
$p = 1 - q$	probability of keeping a feature
λ	L_2 regularization parameter
$\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$	additive dropout noise, $\nu_i \in \{-x_i, x_i/p - x_i\}$
$\mathbf{r} = (r_1, \ldots, r_n)$	multiplicative dropout noise, $r_i \in \{0, 1\}$
$\odot$	component-wise product: $\mathbf{r} \odot \mathbf{x} = (r_1 x_1, \ldots, r_n x_n)$
$\mathbf{w}^*(P, q)$ and $\mathbf{w}^*$	minimizer of dropout criterion: $\mathbf{E}(\ell(y\ \mathbf{w} \cdot (\mathbf{x} + \boldsymbol{\nu})))$
$\mathbf{w}^{\circledast} = \mathbf{w}^*/p$	minimizer of expected loss $\mathbf{E}(\ell(y\ \mathbf{w} \cdot (\mathbf{r} \odot \mathbf{x})))$
$\mathbf{v}(P, \lambda)$ and $\mathbf{v}$	minimizer of L_2-regularized loss
$\mathbf{reg}_{D,q}(\mathbf{w})$	regularization due to dropout
J, K	criteria to be optimized, varies by sub-section
$g(\mathbf{w})$, $\mathbf{g}$	gradients of the current criterion
$\mathrm{er}_P(\mathbf{w})$	0-1 classification generalization error of $\mathrm{sign}(\mathbf{w} \cdot x)$

Table 1: Summary of notation used throughout the paper.

Proof: Assume for contradiction that feature i is perfect modulo ties for P and some $\mathbf{w}^{\circledast}$ is the unique minimizer of $\mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y\mathbf{w} \cdot (\mathbf{r} \odot \mathbf{x})))$. Assume w.l.o.g. that $yx_i \geq 0$ for all $\mathbf{x}$ in the support of P (the case where $yx_i \leq 0$ is analogous). Increasing $w_i^{\circledast}$ keeps the loss unchanged on examples where $x_i = 0$ and decreases the loss on the other examples in the support of P, contradicting the assumption that $\mathbf{w}^{\circledast}$ was a unique minimizer of the expected loss.

Now, suppose then each feature i has both examples where $yx_i > 0$ and examples where $yx_i < 0$ in the support of P. Since the support of P is finite, there is a positive lower bound on the probability of any example in the support. With probability $p(1-p)^{n-1}$, component r_i of random vector $\mathbf{r}$ is non-zero and the remaining $n-1$ components are all zero. Therefore as w_i increases without bound in the positive or negative direction, $\mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y\mathbf{w} \cdot (\mathbf{r} \odot \mathbf{x})))$ also increases without bound. Since $\mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y\mathbf{0} \cdot (\mathbf{r} \odot \mathbf{x}))) = \ln 2$, there is a value M depending only on distribution P and the dropout probability such that minimizing $\mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y\mathbf{w} \cdot (\mathbf{r} \odot \mathbf{x})))$ over $\mathbf{w} \in [-M, M]^n$ is equivalent to minimizing $\mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y\mathbf{w} \cdot (\mathbf{r} \odot \mathbf{x})))$ over $\mathbf{R}^n$. Since $\mathbf{Pr}_{(\mathbf{x},y)}(x_i = 0) \neq 1$ for all i, $\{\mathbf{r} \odot \mathbf{x} : \mathbf{r} \in \{0, 1\}^n, \mathbf{x} \in X\}$ has full rank and therefore $\mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y\mathbf{w} \cdot (\mathbf{r} \odot \mathbf{x})))$ is strictly convex. Since a strictly convex function defined on a compact set has a unique minimum, $\mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y\mathbf{w} \cdot (\mathbf{r} \odot \mathbf{x})))$ has a unique minimum on $[-M, M]^n$, and therefore on $\mathbf{R}^n$. ∎

See Table 1 for a summary of the notation used in the paper.

3. Properties of the Dropout Regularizer

We start by rederiving the regularization function corresponding to dropout training previously presented by Wager et al. (2013), specialized to our context and using our notation. The first step is to write $\ell(y\mathbf{w}\cdot\mathbf{x})$ in an alternative way that exposes some symmetries:

$$\begin{aligned}\ell(y\mathbf{w}\cdot\mathbf{x}) &= \ln(1+\exp(-y\mathbf{w}\cdot\mathbf{x}))\\ &= \ln\left(\frac{\exp(y(\mathbf{w}\cdot\mathbf{x})/2)+\exp(-y(\mathbf{w}\cdot\mathbf{x})/2)}{\exp(y(\mathbf{w}\cdot\mathbf{x})/2)}\right)\\ &= \ln\left(\frac{\exp((\mathbf{w}\cdot\mathbf{x})/2)+\exp(-(\mathbf{w}\cdot\mathbf{x})/2)}{\exp(y(\mathbf{w}\cdot\mathbf{x})/2)}\right). \end{aligned}\tag{6}$$

This then implies

$$\begin{aligned}&\mathbf{reg}_{D,q}(\mathbf{w})\\ &= \mathbf{E}(\ell(y\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu}))) - \mathbf{E}(\ell(y\mathbf{w}\cdot\mathbf{x}))\\ &= \mathbf{E}\left(\ln\left(\frac{\exp((\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu}))/2)+\exp(-(\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu}))/2)}{\exp(y(\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu}))/2)}\times\frac{\exp(y(\mathbf{w}\cdot\mathbf{x})/2)}{\exp((\mathbf{w}\cdot\mathbf{x})/2)+\exp(-(\mathbf{w}\cdot\mathbf{x})/2)}\right)\right)\\ &= \mathbf{E}\Big(\ln\left(\frac{\exp((\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu}))/2)+\exp(-(\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu}))/2)}{\exp((\mathbf{w}\cdot\mathbf{x})/2)+\exp(-(\mathbf{w}\cdot\mathbf{x})/2)}\right) - y(\mathbf{w}\cdot\boldsymbol{\nu})/2\Big).\end{aligned}$$

Since $\mathbf{E}(\boldsymbol{\nu}) = \mathbf{0}$, we get the following.

Proposition 4 *(Wager et al., 2013)*

$$\mathbf{reg}_{D,q}(\mathbf{w}) = \mathbf{E}\left(\ln\left(\frac{\exp(\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu})/2)+\exp(-\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu})/2)}{\exp((\mathbf{w}\cdot\mathbf{x})/2)+\exp(-(\mathbf{w}\cdot\mathbf{x})/2)}\right)\right). \tag{7}$$

Using a Taylor expansion, Wager et al. (2013) arrived at the following approximation:

$$\frac{q}{2(1-q)}\sum_i w_i^2 \mathbf{E}_{\mathbf{x}}\left(\frac{x_i^2}{(1+\exp(-\frac{\mathbf{w}\cdot\mathbf{x}}{2}))(1+\exp(\frac{\mathbf{w}\cdot\mathbf{x}}{2}))}\right). \tag{8}$$

This approximation suggests two properties: the strength of the regularization penalty decreases exponentially in the prediction confidence $|\mathbf{w}\cdot\mathbf{x}|$, and that the regularization penalty goes to infinity as the dropout probability q goes to 1. However, $\mathbf{w}\cdot\boldsymbol{\nu}$ can be quite large, making a second-order Taylor expansion inaccurate.[1] In fact, the analysis in this section suggests that the regularization penalty does not decrease with the confidence and that the regularization penalty increases linearly with $q = 1-p$ (Figure 1, Theorem 8, Proposition 9).

The following propositions show that $\mathbf{reg}_{D,q}(\mathbf{w})$ satisfies at least some of the intuitive properties of a regularizer.

Proposition 5 $\mathbf{reg}_{D,q}(\mathbf{0}) = 0$.

Proposition 6 *(Wager et al., 2013) The contribution of each* $\mathbf{x}$ *to the dropout regularization penalty (7) is non-negative: for all* $\mathbf{x}$*,*

$$\mathbf{E}_{\boldsymbol{\nu}}\left(\ln\left(\frac{\exp((\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu}))/2)+\exp(-(\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu}))/2)}{\exp((\mathbf{w}\cdot\mathbf{x})/2)+\exp(-(\mathbf{w}\cdot\mathbf{x})/2)}\right)\right) \geq 0.$$

[1] Wager et al. (2013) experimentally evaluated the accuracy of a related approximation in the case that, instead of using dropout, $\boldsymbol{\nu}$ was distributed according to a zero-mean Gaussian.

Proof: The proposition follows from Jensen's Inequality. ∎

The $\mathbf{w}^*(P,q)$ vector learned by dropout training minimizes

$$\mathbf{E}_{(\mathbf{x},y)\sim P}(\ell(y\mathbf{w}\cdot\mathbf{x})) + \mathbf{reg}_{D,q}(\mathbf{w}).$$

However, the $\mathbf{0}$ vector has $\ell(y\mathbf{0}\cdot\mathbf{x}) = \ln(2)$ and $\mathbf{reg}_{D,q}(\mathbf{0}) = 0$, implying:

Proposition 7 $\mathbf{reg}_{D,q}(\mathbf{w}^*) \le \ln(2)$.

Thus any regularization penalty greater than $\ln(2)$ is effectively equivalent to a regularization penalty of ∞.

We now present new results based on analyzing the exact $\mathbf{reg}_{D,q}(\mathbf{w})$. The next properties show that the dropout regularizer is emphatically *not* like other convex or norm-based regularization penalties in that the dropout regularization penalty always remains bounded when a single component of the weight vector goes to infinity (see also Figure 1).

Theorem 8 *For all dropout probabilities* $1-p \in (0,1)$*, all* n*, all marginal distributions* D *over* n*-feature vectors, and all indices* $1 \le i \le n$*,*

$$\sup_{w_i} \mathbf{reg}_{D,q}(\underbrace{0,\ldots,0}_{i-1}, w_i, \underbrace{0,\ldots,0}_{n-i}) \le \mathbf{Pr}_D(x_i \ne 0)(1-p)\ln(2) \;<\; \ln 2.$$

Proof: Fix arbitrary n, p, i, and D. We have

$$\begin{aligned}&\mathbf{reg}_{D,q}(\underbrace{0,\ldots,0}_{i-1}, w_i, \underbrace{0,\ldots,0}_{n-i})\\ &= \mathbf{E}_{\mathbf{x},\boldsymbol{\nu}}\left(\ln\left(\frac{\exp(-w_i(x_i+\nu_i)/2)+\exp(w_i(x_i+\nu_i)/2)}{\exp(-w_ix_i/2)+\exp(w_ix_i/2)}\right)\right).\end{aligned}$$

Fix an arbitrary $\mathbf{x}$ in the support of D and examine the expectation over $\boldsymbol{\nu}$ for that $\mathbf{x}$. Recall that $x_i+\nu_i$ is 0 with probability $1-p$ and is x_i/p with probability p, and we will use the substitution $z = |w_ix_i|/2$.

$$\mathbf{E}_{\boldsymbol{\nu}}\left(\ln\left(\frac{\exp(\frac{-w_i(x_i+\nu_i)}{2}) + \exp(\frac{w_i(x_i+\nu_i)}{2})}{\exp(\frac{-w_ix_i}{2}) + \exp(\frac{w_ix_i}{2})}\right)\right) \tag{9}$$

$$= (1-p)\ln(2) + p\ln\left(\exp(\frac{z}{p}) + \exp(\frac{-z}{p})\right) - \ln\left(\exp(z)+\exp(-z)\right). \tag{10}$$

We now consider cases based on whether or not z is 0. When $z=0$ (so either w_i or x_i is 0) then (10) is also 0.

If $z \ne 0$ then consider the derivative of (10) w.r.t. z, which is

$$\frac{\exp(z/p)-\exp(-z/p)}{\exp(z/p)+\exp(-z/p)} - \frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)}.$$

This derivative is positive since $z>0$ and $0<p<1$. Therefore (10) is bounded by its limit as $z\to\infty$, which is $(1-p)\ln(2)$, in this case.

Since (9) is 0 when $x_i = 0$ and is bounded by $(1-p)\ln(2)$ otherwise, the expectation over $\mathbf{x}$ of (9) is bounded $\mathbf{Pr}_D(x_i \ne 0)(1-p)\ln(2)$, completing the proof. ∎

Since line (10) is derived using a chain of equalities, the same proof ideas can be used to show that Theorem 8 is tight.

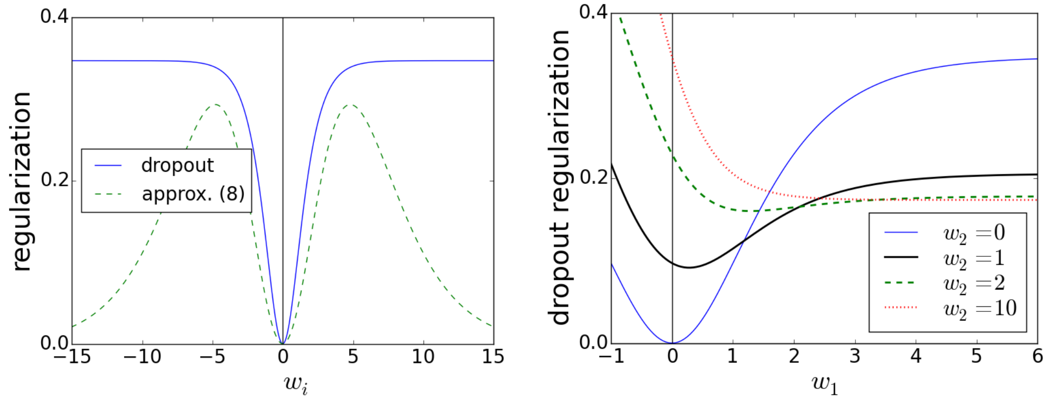


Figure 1: The $p = 1/2$ dropout regularization for $\mathbf{x} = (1, 1)$ as a function of w_i when the other weights are 0 together with its approximation (8) (left) and as a function of w_1 for different values of the second weight (right).

Proposition 9 *Under the conditions of Theorem 8,*

$$\lim_{w_i \to \infty} \mathbf{reg}_{D,q}(\underbrace{0, \ldots, 0}_{i-1}, w_i, \underbrace{0, \ldots, 0}_{n-i}) = \mathbf{Pr}_D(x_i \neq 0)(1-p)\ln(2).$$

Note that this bound on the regularization penalty depends neither on the range nor expectation of x_i. In particular, it has a far different character than the approximation of Equation (8).

In Theorem 8 the other weights are fixed at 0 as w_i goes to infinity. An additional assumption implies that the regularization penalty remains bounded even when the other components are non-zero. Let $\mathbf{w}$ be a weight vector such that for all $\mathbf{x}$ in the support of D and dropout noise vectors $\boldsymbol{\nu}$ we have $|\sum_{j \neq i} w_j(x_j + \nu_j)| \leq M$ for some bound M (this implies that $|\sum_{j \neq i} w_j x_j| \leq M$ also). Then

$$\begin{aligned}
\mathbf{reg}_{D,q}(\mathbf{w}) &= \mathbf{E}_{\mathbf{x},\boldsymbol{\nu}}\left(\left(\frac{\exp(\frac{\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu})}{2})+\exp(-\frac{\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu})}{2})}{\exp(\frac{\mathbf{w}\cdot\mathbf{x}}{2})+\exp(-\frac{\mathbf{w}\cdot\mathbf{x}}{2})}\right)\right) \\
&\leq \mathbf{E}_{x_i,\nu_i}\left(\log\left(\frac{\exp(\frac{M-w_i(x_i+\nu_i)}{2}+\exp(\frac{M+w_i(x_i+\nu_i)}{2})}{\exp(-\frac{M-w_ix_i}{2}+\exp(-\frac{M+w_ix_i}{2})}\right)\right) \\
&\leq M + \mathbf{E}_{x_i,\nu_i}\left(\log\left(\frac{\exp(-\frac{w_ix_i+\nu_i}{2})+\exp(\frac{w_i(x_i+\nu_i)}{2})}{\exp(\frac{-w_ix_i}{2})+\exp(\frac{w_ix_i}{2})}\right)\right).
\end{aligned} \tag{11}$$

Using (11) instead of the first line in Theorem 8's proof gives the following.

Proposition 10 *Under the conditions of Theorem 8, if the weight vector* $\mathbf{w}$ *has the property that* $|\sum_{j \neq i} w_j(x_j + \nu_j)| \leq M$ *for each* $\mathbf{x}$ *in the support of* D *and all of its corresponding dropout noise vectors* $\boldsymbol{\nu}$ *then*

$$\sup_{\omega} \mathbf{reg}_{D,q}(w_1, w_2, \ldots, w_{i-1}, \omega, w_{i+1}, \ldots, w_n) \leq M + \mathbf{Pr}_D(x_i \neq 0)(1-p)\ln(2).$$

Proposition 10 shows that the regularization penalty starting from a non-zero initial weight vector remains bounded as any one of its components goes to infinity. On the other hand, unless M is small, the bound will be larger than the dropout criterion for the zero vector. This is a natural consequence as the starting weight vector $\mathbf{w}$ could already have a large regularization penalty.

The derivative of (10) in the proof of Theorem 8 implies that the dropout regularization penalty is monotonic in $|w_i|$ when the other weights are zero. Surprisingly, this is does *not* hold in general. The dropout regularization penalty due to a single example (as in Proposition 6) can be written as

$$\mathbf{E}_{\boldsymbol{\nu}} \left(\ln \left(\exp(\tfrac{\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu})}{2}) + \exp(\tfrac{-\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu})}{2}) \right) \right) - \ln \left(\exp(\tfrac{\mathbf{w}\cdot\mathbf{x}}{2}) + \exp(\tfrac{-\mathbf{w}\cdot\mathbf{x}}{2}) \right).$$

Therefore if increasing a weight makes the second logarithm increase faster than the expectation of the first, then the regularization penalty decreases even as the weight increases. This happens when the $w_i x_i$ products tend to have the same sign. The regularization penalty as a function of w_1 for the single example $\mathbf{x} = (1,1)$, $p = 1/2$, and w_2 set to various values is plotted in Figure 1.[2] This gives us the following.

Proposition 11 *Unlike p-norm regularizers, the dropout regularization penalty* $\mathbf{reg}_{D,q}(\mathbf{w})$ *is <u>not</u> always monotonic in the individual weights.*

In fact, the dropout regularization penalty can decrease as weights move up from 0.

Proposition 12 *Fix* $p = 1/2$, $w_2 > 0$, *and an arbitrary* $\mathbf{x} \in (0,\infty)^2$. *Let D be the distribution concentrated on* $\mathbf{x}$. *Then* $\mathbf{reg}_{D,q}(w_1, w_2)$ *locally <u>decreases</u> as* w_1 *<u>increases</u> from 0.*

Proposition 12 is proved in Appendix A.

We now turn to the dropout regularization's behavior when two weights vary together. If any features are always zero then their weights can go to ±∞ without affecting either the predictions or $\mathbf{reg}_{D,q}(\mathbf{w})$. Two linearly dependent features might as well be one feature. After ruling out degeneracies like these, we arrive at the following theorem, which is proved in Appendix B.

Theorem 13 *Fix an arbitrary distribution D with support in* $\mathbf{R}^2$, *weight vector* $\mathbf{w} \in \mathbf{R}^2$, *and non-dropout probability p. If there is an* $\mathbf{x}$ *with positive probability under D such that* $w_1 x_1$ *and* $w_2 x_2$ *are both non-zero and have different signs, then the regularization penalty* $\mathbf{reg}_{D,q}(\omega \mathbf{w})$ *goes to infinity as* ω *goes to* $\pm\infty$.

The theorem can be straightforwardly generalized to the case $n > 2$; except in degenerate cases, sending two weights to infinity together will lead to a regularization penalty approaching infinity.

Theorem 13 immediately leads to the following corollary.

[2] Setting $\mathbf{x} = (1,1)$ is in some sense without loss of generality as the prediction and dropout regularization values for any $\mathbf{w}$, $\mathbf{x}$ pair are identical to the values for $\tilde{\mathbf{w}}$, $\mathbf{1}$ when each $\tilde{w}_i = w_i x_i$.

Corollary 14 *For a distribution D with support in $\mathbf{R}^2$, if there is an $\mathbf{x}$ with positive probability under D such that $x_1 \neq 0$ and $x_2 \neq 0$, then there is a $\mathbf{w}$ such that for any $q \in (0, 1)$, the regularization penalty $\mathbf{reg}_{D,q}(\omega\mathbf{w})$ goes to infinity with ω.*

For any $\mathbf{w} \in \mathbf{R}^2$ with both components nonzero, there is a distribution D over $\mathbf{R}^2$ with bounded support such that the regularization penalty $\mathbf{reg}_{D,q}(\omega\mathbf{w})$ goes to infinity with ω.

Together Theorems 8 and 13 demonstrate that $\mathbf{reg}_{D,q}(\mathbf{w})$ is *not* convex (see also Figure 1). In fact, $\mathbf{reg}_{D,q}(\mathbf{w})$ cannot be approximated to within any factor by a convex function, even if a dependence on n and p is allowed. For example, Theorem 8 shows that, for all D with bounded support, both $\mathbf{reg}_{D,q}(0, \omega)$ and $\mathbf{reg}_{D,q}(\omega, 0)$ remain bounded as ω goes to infinity, whereas Theorem 13 shows that there is such a D such that $\mathbf{reg}_{D,q}(\omega/2, \omega/2)$ is unbounded as ω goes to infinity.

Theorem 13 relies on the $w_i x_i$ products having different signs. The following shows that $\mathbf{reg}_{D,q}(\mathbf{w})$ does remain bounded when multiple components of $\mathbf{w}$ go to infinity if the corresponding features are compatible in the sense that the signs of $w_i x_i$ are always in alignment.

Theorem 15 *Let $\mathbf{w}$ be a weight vector and D be a discrete distribution such that $w_i x_i \geq 0$ for each index i and all $\mathbf{x}$ in the support of D. The limit of $\mathbf{reg}_{D,q}(\omega\mathbf{w})$ as ω goes to infinity is bounded by $\ln(2)(1-p)\mathbf{P}_{\mathbf{x}\sim D}(\mathbf{w} \cdot \mathbf{x} \neq 0)$.*

The proof of Theorem 15 (which is Appendix C) easily generalizes to alternative conditions where $\omega \to -\infty$ and/or $w_i x_i \leq 0$ for each $i \leq k$ and $\mathbf{x}$ in the support of D.

Taken together Theorems 15 and 13 give an almost complete characterization of when multiple weights can go to infinity while maintaining a finite dropout regularization penalty.

3.1 Discussion

The bounds in the preceding theorems and propositions suggest several properties of the dropout regularizer. First, the $1 - p$ factors indicate that the strength of regularization grows linearly with dropout probability $q = 1 - p$. Second, the $\mathbf{P}_{\mathbf{x}\sim D}(x_i \neq 0)$ factors in several of the bounds suggest that weights for rare features are encouraged by being penalized less strongly than weights for frequent features. This preference for rare features is sometimes seen in algorithms like the Second-Order Perceptron (Cesa-Bianchi et al., 2002) and AdaGrad (Duchi et al., 2011). Wager et al. (2013) discussed the relationship between dropout and these algorithms, based on approximation (8). Empirical results indicate that dropout performs well in domains like document classification where rare features can have high discriminative value (Wang and Manning, 2013). The theorems of this section suggest that the exact dropout regularizer minimally penalizes the use of rare features. Finally, Theorem 13 suggests that dropout limits co-adaptation by strongly penalizing large weights if the $w_i x_i$ products often have different signs. On the other hand, if the $w_i x_i$ products usually have the same sign, then Proposition 12 indicates that dropout encourages increasing the smaller weights to help share the prediction responsibility. This intuition is reinforced by Figure 1, where the dropout penalty for two large weights is much less then a single large weight when the features are highly correlated.

4. A Definition of Separation

Now we turn to illustrating the inductive bias of dropout by contrasting it with L_2 regularization. For this, we will use a definition of separation between pairs of regularizers.

Each regularizer has a regularization parameter that governs how strongly it regularizes. If we want to describe qualitatively what is preferred by one regularizer over another, we need to control for the amount of regularization.

Let $\mathrm{er}_P(\mathbf{w}) = \mathbf{Pr}_{(\mathbf{x},y)\sim P}(\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y)$, and recall that $\mathbf{w}^*$ and $\mathbf{v}$ are the minimizers of the dropout and L_2-regularized criteria respectively.

Say that sources P and Q C-separate L_2 and dropout if there exist q and λ such that both $\frac{\mathrm{er}_P(\mathbf{w}^*(P,q))}{\mathrm{er}_P(\mathbf{v}(P,\lambda))} > C$ and $\frac{\mathrm{er}_Q(\mathbf{v}(Q,\lambda))}{\mathrm{er}_Q(\mathbf{w}^*(Q,q))} > C$. Say that indexed families $\mathcal{P} = \{P_\alpha\}$ and $\mathcal{Q} = \{Q_\alpha\}$ *strongly separate* L_2 and dropout if pairs of distributions in the family C-separate them for arbitrarily large C. We provide strong separations, using both $n = 2$ and larger n.

5. A Source Preferred by L_2

Consider the joint distribution P_5 defined as follows[3]:

x_1	x_2	y	$\mathbf{Pr}(\mathbf{x}, y)$
10	-1	1	1/3
1.1	-1	1	1/3
-1	1.1	1	1/3

(12)

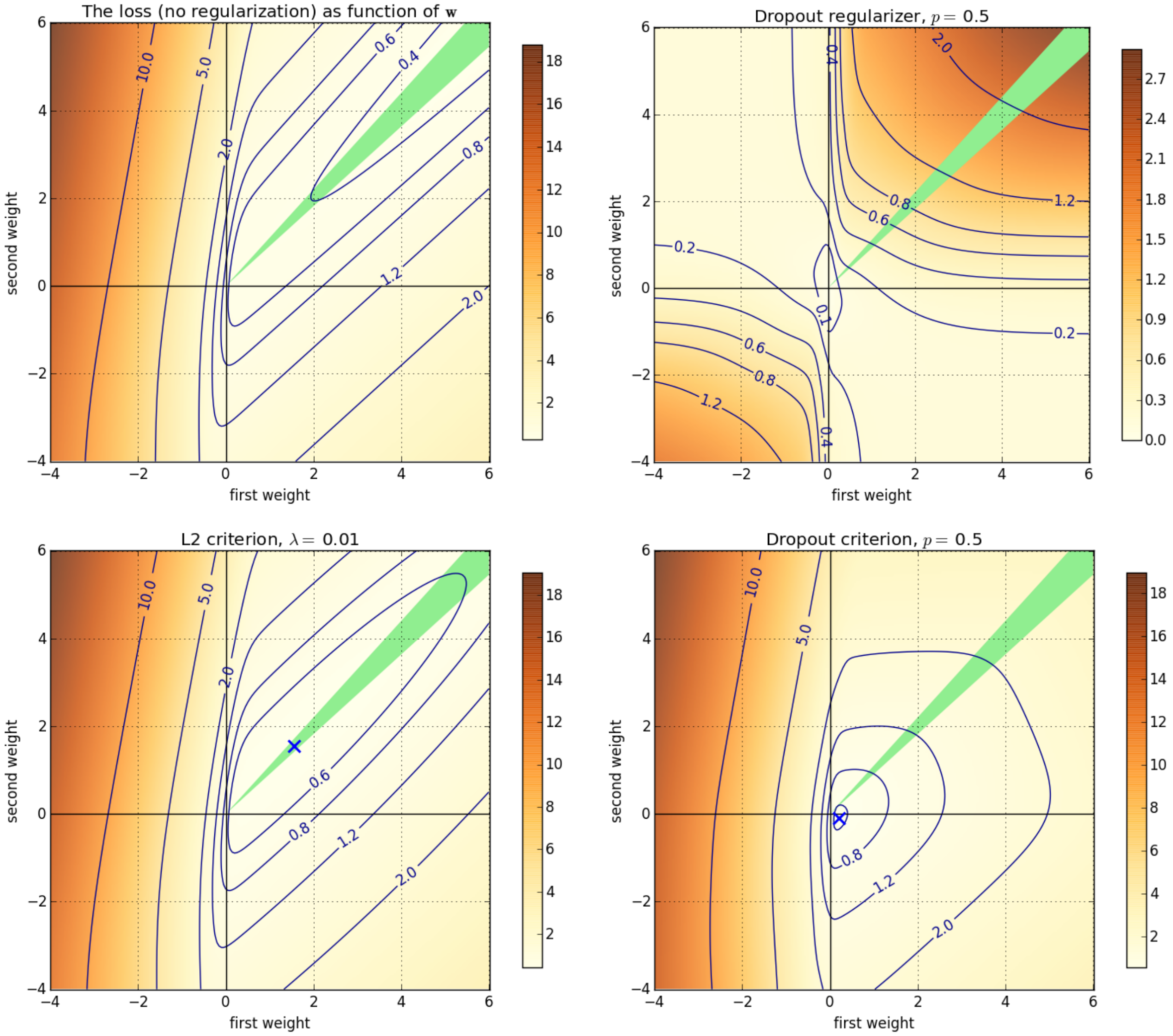
This distribution has weight vectors that classify examples perfectly (the green shaded region in Figure 2). For this distribution, optimizing an L_2-regularized criterion leads to a perfect hypothesis[4], while the weight vectors optimizing the dropout criterion make prediction errors on one-third of the distribution.

The intuition behind this behavior for the distribution described in (12) is that weight vectors that are positive multiples of $(1, 1)$ classify all of the data correctly. However, with dropout regularization the $(10, -1)$ and $(1.1, -1)$ data points encourage the second weight to be negative when the first component is dropped out. This negative push on the second weight is strong enough to prevent the minimizer of the dropout-regularized criterion from correctly classifying the $(-1, 1.1)$ data point. Figure 2 illustrates the loss, dropout regularization, and dropout and L_2 criterion for this data source.[5]

[3] Although several of our sources have all positive instances, that is not essential for the construction. The probability on each $(\mathbf{x}, y)$ example can be split evenly between the original $(\mathbf{x}, y)$ and its negatively-labeled counterpart $(-\mathbf{x}, -y)$. Note that for any $\mathbf{w}$, both $(\mathbf{x}, y)$ and its counterpart $(-\mathbf{x}, -y)$ make the same contribution to both the loss and dropout regularization. After splitting all of the examples, both labels will be equally represented in the distribution. Furthermore, with such paired examples, convexity implies that the weight on any non-dropped out bias input will be 0 when the criterion is minimized.

[4] Having the labels of this distribution be consistent with a linear threshold function eases discussion, but is not essential. Adding a fourth inconsistent point with sufficiently small probability would preserve the property that the L_2-regularized criterion leads to a minimum error linear threshold hypothesis while the error of dropout's hypothesis is significantly larger.

[5] The contours in this and the subsequent figures are not evenly spaced, but chosen to emphasize interesting aspects of the surfaces while minimizing clutter.


The loss (no regularization) as function of $\mathbf{w}$

Dropout regularizer, $p =$ 0.5

L2 criterion, $\lambda =$ 0.01

Dropout criterion, $p =$ 0.5

Figure 2: Using data favoring L_2 in (12). The expected loss is plotted in the upper-left, the dropout regularizer in the upper-right, the L_2 regularized criterion as in (5) in the lower-left and the dropout criterion as in (3) in the lower-right, all as functions of the weight vector. The Bayes-optimal weight vectors are in the green region, and "×" marks show the optimizers of the criteria.

We first show that distribution P_5 of (12) is compatible with mild enough L_2 regularization. Recall that $\mathbf{v}(P_5, \lambda)$ is weight vector found by minimizing the L_2 regularized criterion (5).

Theorem 16 *If* $0 < \lambda \leq 1/50$, *then* $\mathrm{er}_{P_5}(\mathbf{v}(P_5, \lambda)) = 0$ *for the distribution* P_5 *defined in (12).*

In contrast, the $\mathbf{w}^*(P_5, q)$ minimizing the dropout criterion (3) has error rate at least 1/3.

Theorem 17 *If* $q \geq 1/3$ *then* $\mathrm{er}_{P_5}(\mathbf{w}^*(P_5, q)) \geq 1/3$ *for the distribution* P_5 *defined in (12).*

The proofs of Theorem 16 and 17 are in Appendices D and E.

6. A Source Preferred by Dropout

In this section, consider the joint distribution P_6 defined by

x_1	x_2	y	$\mathbf{Pr}(\mathbf{x}, y)$
1	0	1	$3/7$
$-1/1000$	1	1	$3/7$
$1/10$	-1	1	$1/7$

(13)

The intuition behind this distribution is that the $(1, 0)$ data point encourages a large weight on the first feature. This means that the negative pressure on the second weight due to the $(1/10, -1)$ data point is much smaller (especially given its lower probability) than the positive pressure on the second weight due to the $(-1/1000, 1)$ example. The L_2 regularized criterion emphasizes short vectors, and prevents the first weight from growing large enough (relative to the second weight) to correctly classify the $(1/10, -1)$ data point. On the other hand, the first feature is nearly perfect; it only has the wrong sign on the second example where it is $-\epsilon = -1/1000$. This means that, in light of Theorem 8 and Proposition 10, dropout will be much more willing to use a large weight for x_1, giving it an advantage for this source over L_2. The plots in Figure 3 illustrate this intuition.

Theorem 18 *If* $1/100 \leq \lambda \leq 1$, *then* $\mathrm{er}_{P_6}(\mathbf{v}(P_6, \lambda)) \geq 1/7$ *for the distribution* P_6 *defined in (13).*

In contrast, the minimizer of the dropout criterion is able to generalize perfectly.

Theorem 19 *If* $q \leq 1/2$, *then* $\mathrm{er}_{P_6}(\mathbf{w}^*(P_6, q)) = 0$ *for the distribution* P_6 *defined in (13).*

Theorems 18 and 19 are proved in Appendices F and G.

The results in this and the previous section show that the distributions defined in (12) and (13) strongly separate dropout and L_2 regularization. Theorem 19 shows that for distribution P analyzed in this section $\mathrm{er}_P(\mathbf{w}^*(P, q)) = 0$ for all $q \leq 1/2$ while Theorem 18 shows that for the same distribution $\mathrm{er}_P(\mathbf{v}(P, \lambda) \geq 1/7$ whenever $\lambda \geq 1/100$. In contrast, when Q is the distribution defined in the previous section, Theorem 16 shows $\mathrm{er}_Q(\mathbf{v}(Q, \lambda)) = 0$ whenever $\lambda \leq 1/50$. For this same distribution Q, Theorem 17 shows that $\mathrm{er}_Q(\mathbf{w}^*(Q, q)) \geq 1/3$ whenever $q \geq 1/3$.

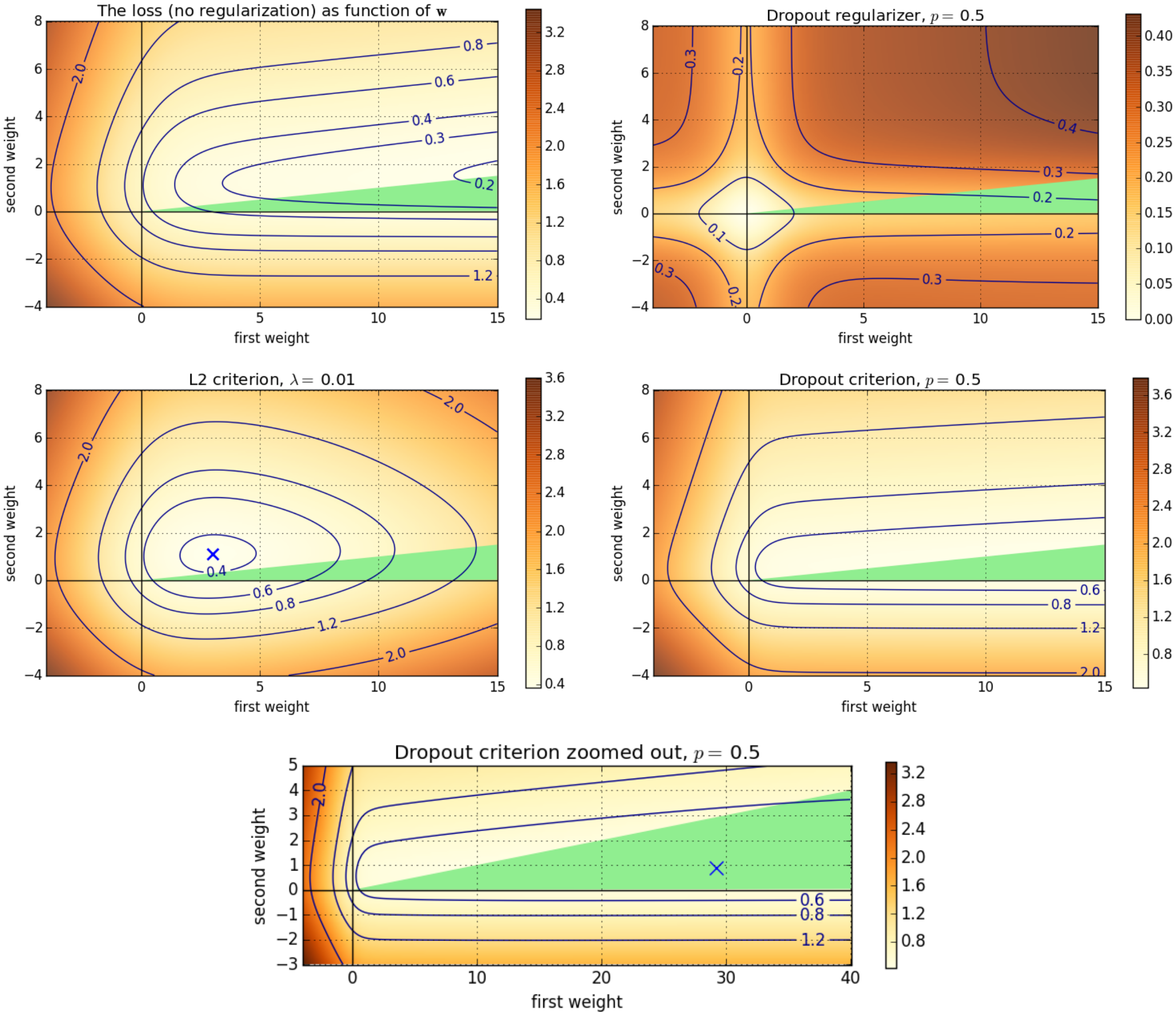


Figure 3: For the source from (13) favoring the dropout, the expected loss is plotted in the upper-left, the dropout regularizer in the upper-right, the expected loss plus L_2 regularization as in (5) in the lower-left and the dropout criterion as in (3) in the lower-right, all as functions of the weight vector. The Bayes-optimal weight vectors are in the green region, and "×" marks show the optimizers of the criteria. Note that the minimizer of the dropout criterion lies outside the middle-right plot and is shown on the bottom plot (which has a different range and scale than the others.)

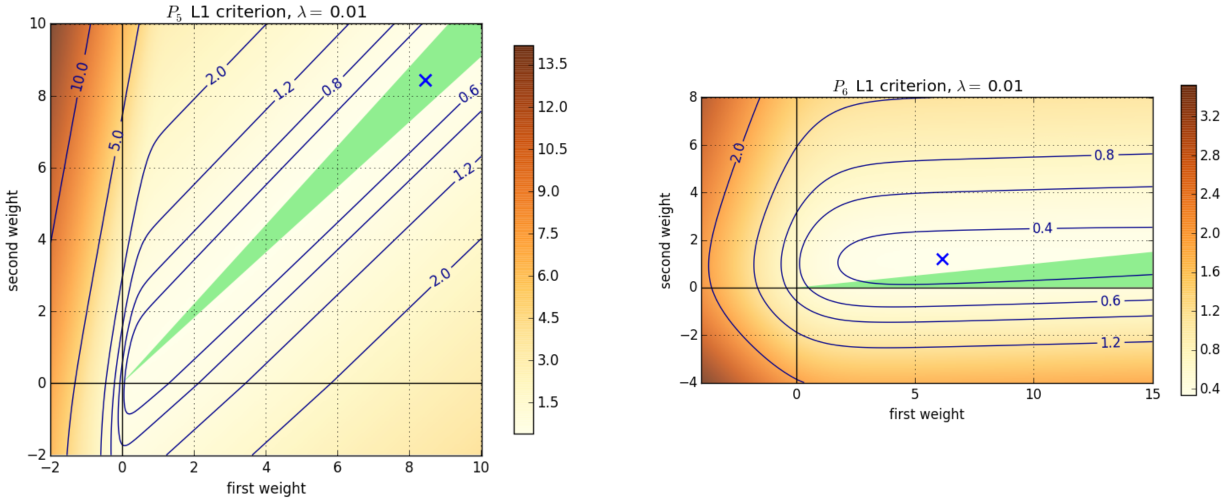


Figure 4: A plot of the L_1 criterion with $\lambda = 0.01$ for distributions P_5 defined in Section 5 (left) and P_6 defined in Section 6 (right). As before, the Bayes optimal classifiers are denoted by the region shaded in green and the minimizer of the criterion is denoted with an x.

7. L_1 Regularization

In this section, we show that the same P_5 and P_6 distributions that separate dropout from L_2 regularization also separate dropout from L_1 regularization: the algorithm the minimizes

$$\mathbf{E}_{(\mathbf{x},y)\sim P}(\ell(y\mathbf{w}\cdot\mathbf{x})) + \lambda||\mathbf{w}||_1. \tag{14}$$

As in Sections 5 and 6, we set $\lambda = 1/100$. Figure 4 plots the L_1 criterion (14) for the distributions P_5 defined in (12) and P_6 defined in (13). Like L_2 regularization, L_1 regularization produces a Bayes-optimal classifier on P_5, but not on P_6. Therefore the same argument shows that these distributions also strongly separate dropout and L_1 regularization.

8. Dropout and Co-adaptation

Hinton et al. (2012) and Srivastava et al. (2014) give evidence that dropout helps prevent the co-adaptation of units in neural networks, encouraging individual units to learn simpler functions of their inputs. In this section we provide a definition of co-adaptation and illustrate how dropout training can restrict the co-adaptation of weights.

We say that two weights w_i and w_j are co-adapted in a weight vector $\mathbf{w}$ if either alone increases the loss, but both together decrease the loss. More formally, let "$\mathbf{w} \setminus i$" denote vector $\mathbf{w}$ modified by replacing w_i with 0, and "$\mathbf{w} \setminus i, j$" denote the resulting vector when both w_i and w_j are replaced by 0. If all of:

1. The loss of $\mathbf{w}$ is less than the loss of $\mathbf{w} \setminus i, j$,

2. The loss of $\mathbf{w} \setminus i, j$ is less than the loss of $\mathbf{w} \setminus i$, and

3. The loss of $\mathbf{w} \setminus i, j$ is less than the loss of $\mathbf{w} \setminus j$,

then we say that weights w_i and w_j are *co-adapted* in $\mathbf{w}$.

For example, consider the case when features x_1 and x_2 tend to have the same sign, but x_1 is usually a little bigger than x_2 when the label is $+$, and x_2 tends to be larger when the label is $-$. Then the difference $x_1 - x_2$ usually has the same sign as the label, and making w_1 large and w_2 negative with a similar-magnitude is likely to decrease the loss. This is similar to constructing the new good feature $x_1 - x_2$ and giving it large weight. However, if neither feature x_1 nor feature x_2 is strongly correlated with the label, then using a large magnitude weight on just x_1 or just x_2 is likely to result in many badly misclassified examples, and greater loss than if w_1 and w_2 were both set to zero. (Note that similar co-adaptation situations arise when x_1 and x_2 have different signs, but their sum tends to have the same sign, or tends to have the opposite sign, as the label.)

Theorem 13 shows that the dropout regularization penalty goes to infinity as the opposite-signed weights given to x_1 and x_2 in the situation described. Furthermore, the dropout penalty for weight vector $\mathbf{w}$ includes terms for the loss of $\mathbf{w} \setminus i$ and $\mathbf{w} \setminus j$, so if these grow too large, then $\mathbf{w}$ cannot be the minimizer of the dropout criterion. This suggests that dropout training minimizes co-adaptation. The following example gives a more concrete illustration of this behavior.

Consider the joint distribution P_8 defined as follows:

x_1	x_2	y	$\mathbf{Pr}(\mathbf{x}, y)$
10	9	1	0.64
9	10	-1	0.32
1.0	-0.35	1	0.03
-0.35	1.0	-1	0.01

(15)

The loss and dropout regularization for P_8 are plotted in Figure 5. To obtain small loss, the hypothesis must give weights a similar large magnitude with w_1 positive while w_2 is negative. On the other hand, almost all of the probability is on the first two examples, and giving the weights different signs satisfies the conditions of Theorem 13 for them, and the dropout penalty quickly increases. The low probability points will also make the dropout regularization for weight vectors $\mathbf{w} = (a, a)$ go to infinity as a goes to infinity, but the small probabilities keeps the penalty small until a becomes very large (e.g. for $\mathbf{w} = (30, 30)$, the penalty is still less than 0.4). Omitting these points has a nearly indistinguishable effect on the first plots: their presence, as well as the different probabilities for the points, will be more important later, when we introduce the alternative labeling P_8'.

The "checkerboard" pattern of the regularization in Figure 5 shows that common patterns in the data can strongly shape the dropout regularizer, making it discriminate against certain directions. In Figure 6 we plot the L_1, L_2, and dropout regularized criteria for source P_8, illustrating that the dropout regularizer forces the weight vector away from the Bayes optimal region. In fact, the regularization is so strong that both weights are positive at the minimizer of the dropout criterion.

We can verify that the minimizing $\mathbf{v} \approx (2.8, -2.75)$ for the L_2 criterion exhibits co-adaptation. The loss of $\mathbf{v}$ is about 0.06, the loss of $\mathbf{v} \setminus 1 \approx 15$, the loss of $\mathbf{w} \setminus 2 \approx 8$, and the loss of $\mathbf{v} \setminus 1, 2 = \ln 2 \approx 0.69$. The co-adaptation is even more dramatic for the L_1 criterion.

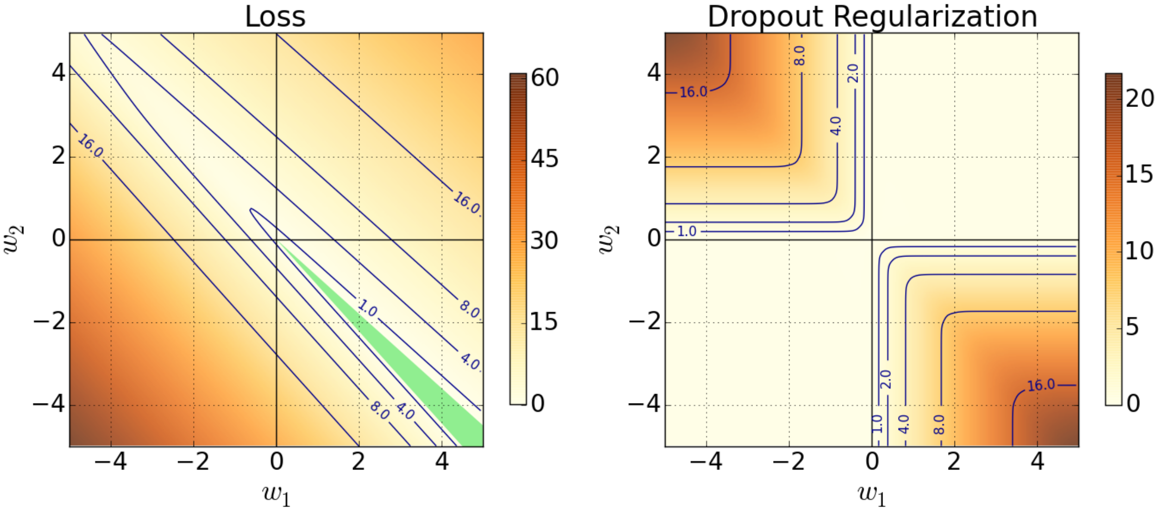


Figure 5: Plots of the loss and $p = 1/2$ dropout regularization for distribution P_8. Note that the regularization penalty increases quickly when the weights have opposite signs, but much more slowly when they have the same sign. In the loss plot, the green region indicates the Bayes optimal classifiers.

On the other hand, the weight vector $\mathbf{w}^* \approx (0.035, 0.014)$ minimizing the dropout criterion is not co-adapted. The losses of $\mathbf{w}^* \setminus 1$ and $\mathbf{w}^* \setminus 2$ are both greater than the loss of $\mathbf{w}^*$, but both are also *less* than the loss of $\mathbf{w}^* \setminus 1, 2$.

Although minimizing the dropout criterion fails to yield a Bayes optimal weight vector for P_8, the situation reverses when we consider the modified distribution P_8' with the same feature vectors and probabilities as P_8, but with all all labels set to 1. When all the labels are positive, the heavier points on the right pull the weight vector in that direction. If it is pulled far enough, then the (-0.3, 1) point will be misclassified.

Since the dropout regularization penalty depends only on the instance probabilities and not on the labels, P_8 and P_8' have the same regularization penalty function. The difference is that P_8' with its modified labels has low loss when both weights are large, a situation compatible with the dropout regularization. See Figure 7 for plots of the loss and various criteria for the modified P_8' source.

The plots in Figures 6 and 7 show that distributions P_8 and P_8' also strongly separate dropout from both L_2 and L_1 regularization. Since the two distributions have the same marginal distribution over feature vectors (and thus use the same dropout regularization penalty function), they provide vivid evidence of how dropout shapes the landscape, encouraging some directions while heavily penalizing others.

9. A High-Dimensional Source Preferred by L_2

In this section we exhibit a source where L_2 regularization leads to a perfect predictor while dropout regularization creates a predictor with a constant error rate.

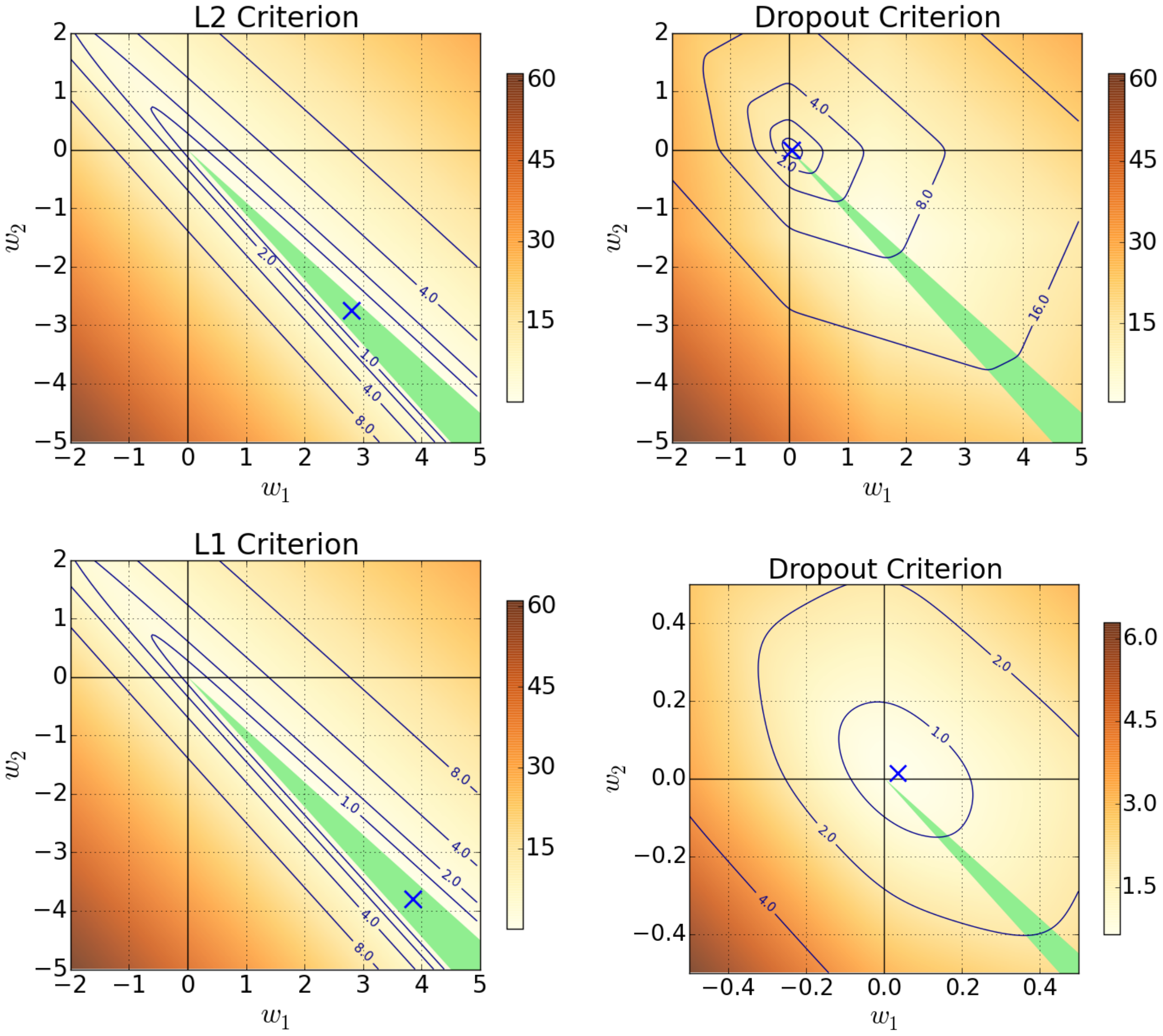


Figure 6: Plots of the criteria and their minimizers for the source P_8. The L_2 and L_1 criteria with $\lambda = 1/100$ are plotted on the right, and the $p = 1/2$ dropout criterion at the same scale and "zoomed in" are shown on the left. As before, the green region indicates the Bayes optimal classifiers.

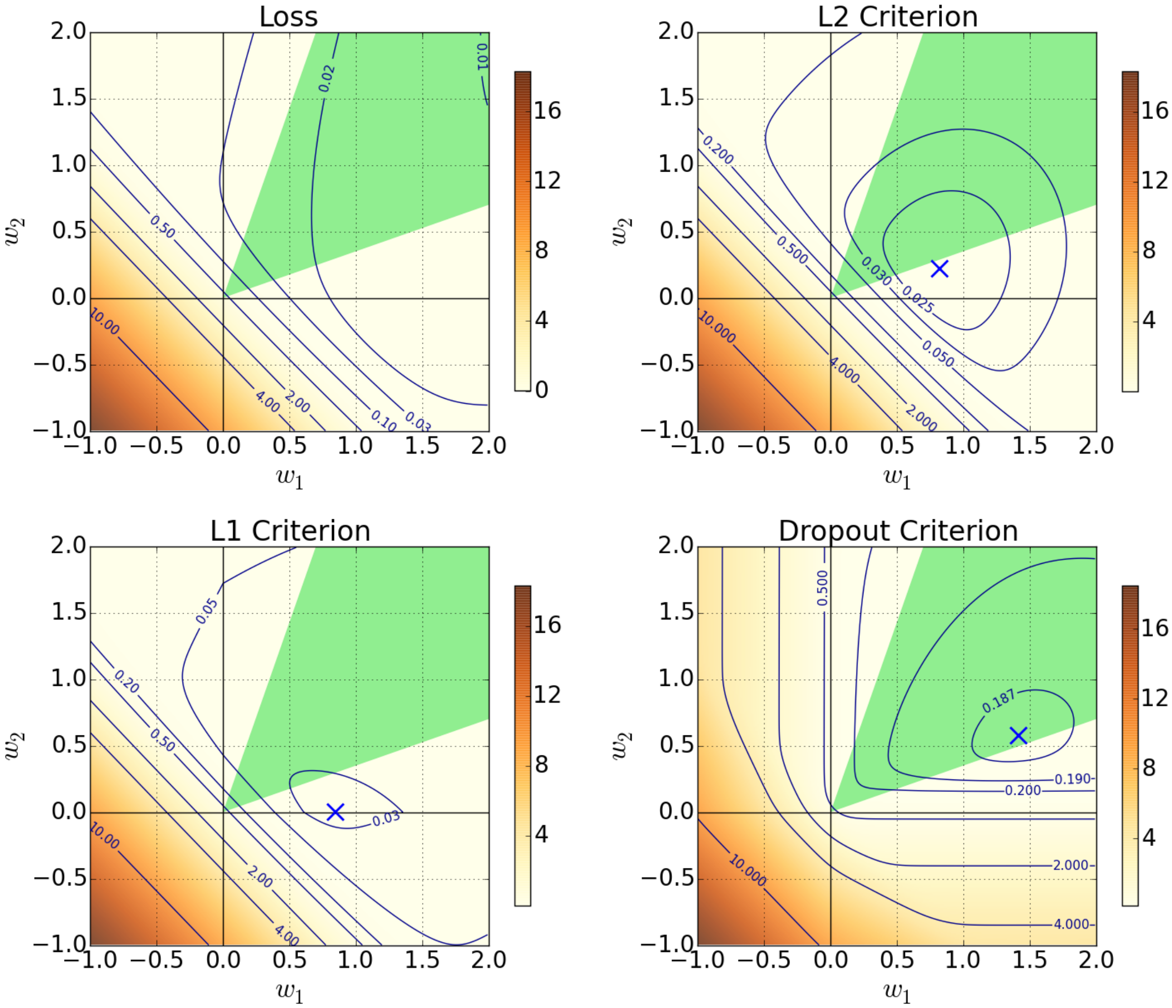


Figure 7: Plots of the loss and various criteria and minimizers for the source P'_8, the modification of P_8 where all the labels are set to 1. As before, $p = 1/2$ for dropout, $\lambda = 1/100$ for the other regularizers, and the green region indicates the Bayes optimal classifiers.

Consider the source P_9 defined as follows. The number n of features is even. All examples are labeled 1. A random example is drawn as follows: the first feature takes the value 1 with probability 9/10 and -1 otherwise, and a subset of exactly $n/2$ of the remaining $n-1$ features (chosen uniformly at random) takes the value 1, and the remaining $n/2-1$ of those first $n-1$ features take the value -1.

A majority vote over the last $n-1$ features achieves perfect prediction accuracy. This is despite the first feature (which does not participate in the vote) being more strongly correlated with the label than any of the voters in the optimal ensemble. Dropout, with its bias for single good features and discrimination against multiple disagreeing features, puts too much weight on this first feature. In contrast, L_2 regularization leads to the Bayes optimal classifier by placing less weight on the first feature than on any of the others.

Theorem 20 *If $\lambda \leq \frac{1}{30n}$ then the weight vector $v(P_9, \lambda)$ optimizing the L_2 criterion has perfect prediction accuracy: $\mathrm{er}_{P_9}(v(P_9, \lambda)) = 0$.*

When $n > 125$, dropout with $q = 1/2$ fails to find the Bayes optimal hypothesis. In particular, we have the following theorem.

Theorem 21 *If the dropout probability $q = 1/2$ and the number of features is an even $n > 125$ then the weight vector $\mathbf{w}^*(P_9, q)$ optimizing the dropout criterion has prediction error rate $\mathrm{er}_{P_9}(\mathbf{w}^*(P_9, q)) \geq 1/10$.*

We conjecture that dropout fails on P_9 for all $n \geq 4$. As evidence, we analyze the $n = 4$ case.

Theorem 22 *If dropout probability $q = 1/2$ and the number of features is $n = 4$ then the minimizer of the dropout criteria $\mathbf{w}^*(P_9, q)$ has has prediction error rate $\mathrm{er}_{P_9}(\mathbf{w}^*(P_9, q)) \geq 1/10$.*

Theorems 20, 21 and 22 are proved in Appendices H, I and J.

10. A High-Dimensional Source Preferred by Dropout

Define the source P_{10}, which depends on (small) positive real parameters η, α, and β, as follows. A random label y is generated first, with both of $+1$ and -1 equally likely. The features $x_1, ..., x_n$ are conditionally independent given y. The first feature tends to be accurate but small: $x_1 = \alpha y$ with probability $1 - \eta$, and is $-\alpha y$ with probability η. The remaining features are larger but less accurate: for $2 \leq i \leq n$, feature x_i is y with probability $1/2 + \beta$, and $-y$ otherwise.

When η is small enough relative to β, the Bayes' optimal prediction is to predict with the first feature. When α is small, this requires concentrating the weight on w_1 to outvote the other features. Dropout is capable of making this one weight large while L_2 regularization is not.

Theorem 23 *If $q = 1/2$, $n \geq 100$, $\alpha > 0$, $\beta = 1/(10\sqrt{n-1})$, and $\eta \leq \frac{1}{2+\exp(54\sqrt{n})}$, then $\mathrm{er}_{P_{10}}(\mathbf{w}^*(P_{10}, q)) = \eta$.*

Theorem 24 *If $\beta = 1/(10\sqrt{n-1})$, $\lambda = \frac{1}{30n}$, $\alpha < \beta\lambda$, and n is a large enough even number, then for any $\eta \in [0,1]$, $\mathrm{er}_{P_{10}}(\mathbf{v}(P_{10}, \lambda)) \geq 3/10$.*

Theorems 23 and 24 are proved in Appendices K and L.

Let $\tilde{n}$ be a large enough even number in the sense of Theorem 24. Let P_η be the distribution defined at the start of Section 10 with number of features $n = \tilde{n}$, $\beta = 1/(10\sqrt{n-1})$, $\alpha = 1/(300n\sqrt{n})$, and $0 < \eta < 1/(2 + \exp(54\sqrt{n}))$ is a free parameter. Theorem 23 shows that $\mathrm{er}_{P_\eta}(\mathbf{w}^*(P_\eta, q)) = \eta$ when dropout probability $q = 1/2$. For this same distribution, Theorem 24 shows $\mathrm{er}_{P_\eta}(\mathbf{v}(P_\eta, \lambda)) \geq 3/10$ when $\lambda = 1/30n$. Therefore

$$\frac{\mathrm{er}_{P_\eta}(\mathbf{w}^*(P_\eta, 1/2))}{\mathrm{er}_{P_\eta}(\mathbf{v}(P, 1/30\tilde{n}))}$$

goes to 0 as $\eta \to 0$.

The distribution defined at the start of Section 9, which we call Q here, provides contrasting behavior when $n = \tilde{n}$. Theorem 21 shows that the error $\mathrm{er}_Q(\mathbf{w}^*(Q, 1/2)) \geq 1/10$ while Theorem 20 shows that $\mathrm{er}_Q(v(Q, 1/30\tilde{n}) = 0$. Therefore the P_η and Q distributions strongly separate dropout and L_2 regularization for parameters $q = 1/2$ and $\lambda = 1/30n$.

11. Conclusions

We have built on the interpretation of dropout as a regularizer in Wager et al. (2013) to prove several interesting properties of the dropout regularizer. This interpretation decomposes the dropout criterion minimized by training into a loss term plus a regularization penalty that depends on the feature vectors in the training set (but not the labels). We started with a characterization of when the dropout criterion has a unique minimum, and then turn to properties of the dropout regularization penalty. We verified that the dropout regularization penalty has some desirable properties of a regularizer: it is 0 at the zero vector, and the contribution of each feature vector in the training set is non-negative.

On the other hand, the dropout regularization penalty does not behave like standard regularizers. In particular, we have shown:

1. Although the dropout "loss plus regularization penalty" criterion is convex in the weights $\mathbf{w}$, the regularization penalty imposed by dropout training is *not* convex.

2. Starting from an arbitrary weight vector, any single weight can go to infinity while the dropout regularization penalty remains bounded.

3. In some cases, multiple weights can simultaneously go to infinity while the regularization penalty remains bounded.

4. The regularization penalty can *decrease* as weights increase from 0 when the features are correlated.

These are in stark contrast to standard norm-based regularizers that always diverge as any weight goes to infinity, and are non-decreasing in each individual weight.

In most cases the dropout regularization penalty *does* diverge as multiple weights go to infinity. We characterize when sending two weights to infinity causes the dropout regularization penalty to diverge, and when it will remain finite. In particular, dropout is willing to put a large weights on multiple features if the $w_i x_i$ products tend to have the same sign.

The form of our analytical bounds suggest that the strength of the regularizer grows linearly with the dropout probability q, and provide additional support for the claim (Wager et al., 2013) that dropout favors rare features.

We found it important to check our intuition by working through small examples. To make this more rigorous we needed a definition of when a source favored dropout regularization over a more standard regularizer like L_2. Such a definition needs to deal with the strength of regularization, a difficulty complicated by the fact that dropout regularization is parameterized by the dropout probability $q \in [0, 1]$ while L_2 regularization is parameterized by $\lambda \in [0, \infty]$. Our solution is to consider pairs of sources P and Q. We then say the pair *separates* the dropout and L_2 if dropout with a particular parameter q performs better then L_2 with a particular parameter λ on source P, while L_2 (with the same λ) performs better than dropout (with the same q) on source Q. Our definition uses generalization error as the most natural interpretation of "performs better".

Sections 5 through 10 are devoted to proving that dropout and L_2 are strongly separated by certain pairs of distributions. Section 7 shows that dropout and L_1 regularization are also strongly separated, and Section 8 describes a separation illustrating dropout's bias against co-adaptation of weights. Proving strong separation is non-trivial even after one finds the right distributions. This is due to several factors: the minimizers of the criteria do not have closed forms, we wish to prove separation for ranges of the regularization values, and the binomial distributions induced by dropout are not amenable to exact analysis. Despite these difficulties, the separation results reinforce the intuition that dropout is more willing to use a large weight in order to better fit the training data than L_2 regularization. However, if two features often have both the same and different signs (as in Theorem 13) then dropout is less willing to put even moderate weight on both features.

As a side benefit of these analyses, the plots in Figure 2 and Figure 3 provide a dramatic illustration of the dropout regularizer's non-convexity and its preference for making only a single weight large, and the checkerboard pattern of the dropout regularizer in Figure 5 illustrates its bias against co-adaptation of weights. This is consistent with the insight provided by Theorems 13 and 15.

Some feature transformations appear to have substantially different effects on dropout and L_2. For example, suppose we replace a boolean feature x_i with a batch of features $x_{i,1}, ..., x_{i,k}$, and,

- when $x_i = 0$, we set $x_{i,1} = ... = x_{i,k} = 0$ and
- when $x_i = 1$, we set $x_{i,j'} = 1$ for j' chosen uniformly at random from $\{1, ..., k\}$, and $x_{i,j} = 0$ for $j \neq j'$.

We can think of $x_{i,1}, ..., x_{i,k}$ as a "partition" of x_i. This kind of transformation can arise in document classification when words have alternate spellings, or a single feature representing a set of synonyms is split into features for the individual words (assuming that each document uses only one of the synonyms).

The inductive bias of dropout is apparently not affected by such feature partitioning. For any weight vector $\mathbf{w}$ on the original features, the modified weight vector which copies w_i for each feature in the partition of x_i makes the same predictions and has the same dropout regularization penalty. On the other hand, the L_2 regularization penalty increases. If an algorithm creates k copies of the weight w_i to have the same behavior on the modified data, this increases the penalty arising from this feature by a factor of k, providing an incentive for the algorithm to use other features instead.

Dropout's relative affinity with partitioned features could be another basis of separation with L_2. It suggests that dropout might be able to more effectively exploit rare primitive features, while L_2 regularization benefits from having more frequent higher-level features. This is a potential subject for future research.

Now suppose that, instead of partitioning x_i, we set $x_{i,1}, ..., x_{i,k}$ to be k copies of x_i. In this case, an L_2-regularized algorithm could split weight w_i into k parts, putting weight w_i/k on each copy of x_i. This will classify the transformed data the same way as the original data while *reducing* the L_2 regularization cost of using the feature by a factor of k (since $\sum_j (w_i/k)^2 = (1/k)w_i^2$). Although such feature cloning can also reduce the dropout regularization penalty (see Figure 1 and Proposition 12), we conjecture that the reduction is at most an additive constant.

If this conjecture were true, then L_2-regularized algorithms make heavier use of duplicated features than dropout-regularized algorithms. This in turn suggest that dropout confers resistance to paying undue attention to groups of mostly redundant features. This possibility is another potential subject for future research.

The aim of our analysis has been to aid general understanding of what kinds of problems are well-suited to dropout. A more authoritative idea of whether dropout confers an advantage in a particular case can be gained experimentally.

Linear classifiers are often learned with a bias term, creating a classifier of the form $\text{sign}(\mathbf{w} \cdot \mathbf{x} - b)$. Here the bias b is also learned, but not regularized. We have focused on the case $b = 0$ to keep the analysis simple, and our constructions can be easily modified so that the optimal bias is 0 (see footnote 3). The effect of a non-zero bias term on the general properties in Section 3 can be more subtle, and is a potential subject for future research.

Our analysis is for the logistic regression case corresponding to a single output node. It would be very interesting to have similar analysis for multi-layer neural networks. However, dealing with non-convex loss of such networks will be a major challenge. Another open problem suggested by this work is how the definition of separation can be used to gain insight about other regularizers and settings.

Acknowledgments

We thank the reviewers for their thoughtful comments and corrections.

Appendix A. Proof of Proposition 12

> **Proposition 12.** Fix $p = 1/2$, $w_2 > 0$, and an arbitrary $\mathbf{x} \in (0, \infty)^2$. Let D be the distribution concentrated on $\mathbf{x}$. Then $\mathbf{reg}_{D,q}(w_1, w_2)$ locally decreases as w_1 increases from 0.

First, we show that assuming $\mathbf{x} = (2, 2)$ is without loss of generality. When D concentrates all of its probability on a single $\mathbf{x}$, let us denote $\mathbf{reg}_{D,1/2}$ by $\mathbf{reg}_{\mathbf{x},1/2}$. Since anyplace w_1 appears in the expression for $\mathbf{reg}_{\mathbf{x},1/2}$, it is multiplied by x_1, if we multiply w_1 by some constant c and divide x_1 by c, we do not change w_1x_1, and therefore do not change $\mathbf{reg}_{\mathbf{x},1/2}$. The same holds for w_2. Thus

$$\mathbf{reg}_{\mathbf{x},1/2}(\mathbf{w}) = \mathbf{reg}_{(2,2),1/2}(w_1x_1/2, w_2x_2/2).$$

If we change variables and let $\tilde{w}_1 = w_1x_1/2$ and $\tilde{w}_2 = w_2x_2/2$, then since x_1 and x_2 are both positive, $\tilde{w}_2$ is positive iff w_2 is, and $\mathbf{reg}_{\mathbf{x},1/2}(\mathbf{w})$ is increasing with w_1 iff $\mathbf{reg}_{(2,2),1/2}(\tilde{\mathbf{w}})$ is increasing with $\tilde{w}_1$.

We continue assuming $\mathbf{x} = (2, 2)$. It suffices to show $\partial\mathbf{reg}_{D,q}(w_1, w_2)/\partial w_1|_{w_1=0} < 0$. This derivative is

$$\frac{3e^{w_2} + e^{-3w_2} - 3e^{-w_2} - e^{3w_2}}{2(e^{w_2} + e^{-w_2})(e^{2w_2} + e^{-2w_2})}. \tag{16}$$

The sign depends only on the numerator, which is 0 when $w_2 = 0$. The derivative of the numerator with respect to w_2 is $3e^{w_2} - 3e^{-3w_2} + 3e^{-w_2} - 3e^{3w_2}$, which is negative for $w_2 > 0$, since $e^z + e^{-z}$ is an increasing function in z. Thus the numerator in (16) is decreasing in w_2. Therefore (16) is negative when $w_2 > 0$, and the regularization penalty is (locally) decreasing as w_1 increases from 0.

(Note: Proposition 12 may be generalized with slight modifications to apply whenever $\mathbf{x}$ has two nonzero components. What is needed is that x_1w_1 and x_2w_2 have the same sign. For example, if x_1 is negative but x_2w_2 is positive, then moving w_1 from 0 in the negative direction decreases $\mathbf{reg}_{D,q}(\mathbf{w})$.)

Appendix B. Proof of Theorem 13

> **Theorem 13.** Fix an arbitrary distribution D with support in $\mathbf{R}^2$, weight vector $\mathbf{w} \in \mathbf{R}^2$, and non-dropout probability p. If there is an $\mathbf{x}$ with positive probability under D such that w_1x_1 and w_2x_2 are both non-zero and have different signs, then the regularization penalty $\mathbf{reg}_{D,q}(\omega\mathbf{w})$ goes to infinity as ω goes to $\pm\infty$.

Fix an $\mathbf{x}$ satisfying the conditions of the theorem.

$$\begin{aligned}
\mathbf{reg}_{D,q}(\omega\mathbf{w}) &\geq D(\mathbf{x})\mathbf{E}_{\boldsymbol{\nu}}\left(\ln\left(\frac{\exp(-\frac{\omega\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu})}{2}) + \exp(\frac{\omega\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu})}{2})}{\exp(\frac{-\omega\mathbf{w}\cdot\mathbf{x}}{2}) + \exp(\frac{\omega\mathbf{w}\cdot\mathbf{x}}{2})}\right)\right) \\
&> D(\mathbf{x})\mathbf{E}_{\boldsymbol{\nu}}\left(\ln\left(\frac{\exp(\frac{|\omega\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu})|}{2})}{2\exp(\frac{|\omega\mathbf{w}\cdot\mathbf{x}|}{2})}\right)\right) \\
&= D(\mathbf{x})\mathbf{E}_{\boldsymbol{\nu}}\left(-\ln(2) + \frac{|\omega\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu})|}{2} - \frac{|\omega\mathbf{w}\cdot\mathbf{x}|}{2}\right).
\end{aligned} \tag{17}$$

We now examine the expectation over $\boldsymbol{\nu}$ of the term that depends on $\boldsymbol{\nu}$. We assume that $|w_1x_1| \geq |w_2x_2|$ so $|\mathbf{w} \cdot \mathbf{x}| = |w_1x_1| - |w_2x_2|$; the other case is symmetrical.

$$\begin{aligned}\mathbf{E}_{\boldsymbol{\nu}}(|\omega\mathbf{w} \cdot (\mathbf{x} + \boldsymbol{\nu})|) &= |\omega|\big(p^2|\mathbf{w} \cdot \mathbf{x}/p| + p(1-p)|w_1x_1/p| + p(1-p)|w_2x_2/p|\big)\\ &= |\omega|\big(p|\mathbf{w} \cdot \mathbf{x}| + (1-p)(|w_1x_1| - |w_2x_2| + |w_2x_2|) + (1-p)|w_2x_2|\big)\\ &= |\omega|(|\mathbf{w} \cdot \mathbf{x}| + 2(1-p)|w_2x_2|).\end{aligned}$$

Plugging this into (17) gives:

$$\mathbf{reg}_{D,q}(\omega\mathbf{w}) > D(\mathbf{x})\left(-\ln 2 + (1-p)|\omega||w_2x_2|\right)$$

which goes to infinity as ω goes to $\pm\infty$.

Appendix C. Proof of Theorem 15

> **Theorem 15.** Let $\mathbf{w}$ be a weight vector and D be a discrete distribution such that $w_ix_i \geq 0$ for each index i and all $\mathbf{x}$ in the support of D. The limit of $\mathbf{reg}_{D,q}(\omega\mathbf{w})$ as ω goes to infinity is bounded by $\ln(2)(1-p)\mathbf{P}_{\mathbf{x}\sim D}(\mathbf{w} \cdot \mathbf{x} \neq 0)$.

First note that If $\mathbf{w}$ and D are such that $\mathbf{w} \cdot \mathbf{x} = 0$ for all $\mathbf{x}$ in the support of D, then $\mathbf{reg}_{D,q}(\mathbf{w}) = \mathbf{reg}_{D,q}(\omega\mathbf{w}) = 0$. We now analyze the general case.

$$\begin{aligned}\mathbf{reg}_{D,q}(\omega\mathbf{w}) &= \mathbf{E}_{\mathbf{x},\boldsymbol{\nu}}\left(\ln\left(\frac{\exp(\frac{\omega\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu})}{2}) + \exp(\frac{-\omega\mathbf{w}\cdot(\mathbf{x}+\boldsymbol{\nu})}{2})}{\exp(\frac{\omega\mathbf{w}\cdot\mathbf{x}}{2}) + \exp(\frac{-\omega\mathbf{w}\cdot\mathbf{x}}{2})}\right)\right)\\ &= \mathbf{E}_{\mathbf{x},\boldsymbol{\nu}}\left(\ln\left(\frac{\exp(\frac{\omega\mathbf{w}\cdot(\mathbf{x}+\nu)}{2})(1 + \exp(-\omega\mathbf{w} \cdot (\mathbf{x} + \boldsymbol{\nu})))}{\exp(\frac{\omega\mathbf{w}\cdot\mathbf{x}}{2})(1 + \exp(-\omega\mathbf{w} \cdot \mathbf{x}))}\right)\right)\\ &= \mathbf{E}_{\mathbf{x},\boldsymbol{\nu}}\Big((\omega\mathbf{w} \cdot (\mathbf{x} + \boldsymbol{\nu})/2) + \ln\left(1 + \exp\left(-\omega\mathbf{w} \cdot (\mathbf{x} + \boldsymbol{\nu})\right)\right)\\ &\qquad - (\omega\mathbf{w} \cdot \mathbf{x}/2) - \ln\left(1 + \exp\left(-\omega\mathbf{w}\mathbf{x}\right)\right)\Big). \end{aligned} \tag{18}$$

Of the four terms inside the expectation in Equation (18), the first and third cancel since the expectation of $\boldsymbol{\nu}$ is $\mathbf{0}$. Therefore:

$$\mathbf{reg}_{D,q}(\omega\mathbf{w}) = \mathbf{E}_{\mathbf{x}}\big(\mathbf{E}_{\boldsymbol{\nu}}\big(\ln(1 + \exp(-\omega\mathbf{w} \cdot (\mathbf{x} + \boldsymbol{\nu}))) - \ln(1 + \exp(-\omega\mathbf{w}\mathbf{x}))\big)\big). \tag{19}$$

Define $\text{nez}(\mathbf{w}, \mathbf{x})$ to be the number of indices i where $w_ix_i \neq 0$. We now consider cases based on $\text{nez}(\mathbf{w}, \mathbf{x})$.

Whenever $\text{nez}(\mathbf{w}, \mathbf{x}) = 0$ then both $\mathbf{w} \cdot \mathbf{x} = 0$ and $\mathbf{w} \cdot (\mathbf{x} + \boldsymbol{\nu}) = 0$. Therefore the contribution of these $\mathbf{x}$ to the expectation in (19) is $\ln(2) - \ln(2) = 0$.

If $\text{nez}(\mathbf{w}, \mathbf{x}) > 0$ then $\mathbf{w} \cdot \mathbf{x} > 0$ (since each $w_ix_i \geq 0$), and the second term of (19) goes to zero as ω goes to infinity. The first term of (19) also goes to zero, *unless* all of the $\text{nez}(\mathbf{w}, \mathbf{x})$ components where $w_ix_i > 0$ are dropped out. If they are all dropped out, then the first term becomes $\ln(2)$. The probability that all $\text{nez}(\mathbf{w}, \mathbf{x})$ non-zero components are simultaneously dropped out is $(1-p)^{\mathrm{nez}(\mathbf{w},\mathbf{x})}$. With this reasoning we get from (19) that:

$$\begin{aligned}
&\lim_{\omega \to \infty} \mathbf{reg}_{D,q}(\omega \mathbf{w}) \\
&= \sum_{k=1}^{n} \mathbf{P}_{\mathbf{x} \sim D}(\mathrm{nez}(\mathbf{w},\mathbf{x}) = k) \left(\ln(2)(1-p)^k \right) \qquad (20) \\
&\leq \sum_{k=1}^{n} \mathbf{P}_{\mathbf{x} \sim D}(\mathrm{nez}(\mathbf{w},\mathbf{x}) = k) \left(\ln(2)(1-p) \right) \\
&= \ln(2)(1-p)\mathbf{P}(\mathbf{w} \cdot \mathbf{x} \neq 0)
\end{aligned}$$

as desired.

(Note that Equation 20 gives a precise, but more complex expression for the limit.)

Appendix D. Proof of Theorem 16

Theorem 16. If $0 < \lambda \leq 1/50$, then $\mathrm{er}_{P_5}(\mathbf{v}(P_5, \lambda)) = 0$ for the distribution P_5 defined in (12).

To keep the notation clean let us abbreviate P_5 as just P throughout this proof.

By scaling the L_2 criterion we can obtain cancellation in the expectation. Let $\mathbf{v}$ be weight vector found by minimizing the following L_2 regularized criterion J:

$$J(\mathbf{w}) = 3 \left(\mathbf{E}_{(\mathbf{x},y) \sim P}(\ell(y(\mathbf{w} \cdot \mathbf{x}))) + (\lambda/2)||\mathbf{w}||^2 \right). \qquad (21)$$

Note the factor of 3 is to simplify the expressions and doesn't affect the minimizing $\mathbf{v}$.

We will prove Theorem 16 with a series of lemmas.

But first, let's take some partial derivatives:

$$\frac{\partial J}{\partial w_1} = \frac{-10}{1+\exp(10w_1 - w_2)} + \frac{-1.1}{1+\exp(1.1w_1 - w_2)} + \frac{1}{1+\exp(-w_1 + 1.1w_2)} + 3\lambda w_1 \qquad (22)$$

$$\frac{\partial J}{\partial w_2} = \frac{1}{1+\exp(10w_1 - w_2)} + \frac{1}{1+\exp(1.1w_1 - w_2)} + \frac{-1.1}{1+\exp(-w_1 + 1.1w_2)} + 3\lambda w_2. \qquad (23)$$

We will repeatedly use the following basic, well-known, lemma.

Lemma 25 *For any convex, differentiable function ψ defined on $\mathbf{R}^n$ with a unique minimum $\mathbf{w}^*$, for any $\mathbf{w} \in \mathbf{R}^n$, if $g(\mathbf{w})$ is the gradient of ψ at $\mathbf{w}$ then $\mathbf{w}^*$ is contained in the closed halfspace whose separating hyperplane goes through $\mathbf{w}$, and whose normal vector is $-g(\mathbf{w})$; i.e., $\mathbf{w}^* \cdot g(\mathbf{w}) \leq \mathbf{w} \cdot g(\mathbf{w})$. Furthermore, if $g(\mathbf{w}) \neq \mathbf{0}$ then $\mathbf{w}^* \cdot g(\mathbf{w}) < \mathbf{w} \cdot g(\mathbf{w})$.*

Now we're ready to start our analysis of P.

Lemma 26 *If $0 \leq \lambda$, the optimizing v_1 is positive.*

Proof: By Lemma 25, it suffices to show that there is a point $(0, a_2)$ where both $\frac{\partial J}{\partial w_1}\big|_{(0,a_2)} < 0$ and $\frac{\partial J}{\partial w_2}\big|_{(0,a_2)} = 0$.

From Equation (22):

$$\frac{\partial J}{\partial w_1}\bigg|_{(0,a_2)} = \frac{-11.1}{1+\exp(-a_2)} + \frac{1}{1+\exp(1.1a_2)}$$

and each term is decreasing as a_2 increases. Since it is negative when $a_2 = -2$, we have $\frac{\partial J}{\partial w_1}\Big|_{(0,a_2)} < 0$ for all $a_2 > -2$. So, to prove the lemma, if suffices to show that there is a $a_2 \in (-2, \infty)$ such that the other derivative $\frac{\partial J}{\partial w_2}\Big|_{(0,a_2)} = 0$.

From equation (23):

$$\frac{\partial J}{\partial w_2}\bigg|_{(0,a_2)} = \frac{2}{1+\exp(-a_2)} + \frac{-1.1}{1+\exp(1.1a_2)} + 3\lambda a_2$$

and each term is continuously increasing in a_2. When $a_2 = -2$, $\frac{\partial J}{\partial w_2}\Big|_{(0,a_2)}$ is negative. On the other hand, $\frac{\partial J}{\partial w_2}\big|_{(0,0)}$ is positive. Therefore for some $a_2 \in (-2, 0)$ we have $\frac{\partial J}{\partial w_2}\big|_{(0,a_2)} = 0$ as desired. ∎

Lemma 27 *There is a real $a > 0$ such that*

$$\frac{\partial J(\mathbf{w})}{\partial w_1}\bigg|_{(a,a)} + \frac{\partial J(\mathbf{w})}{\partial w_2}\bigg|_{(a,a)} = 0.$$

Proof: Applying (22) and (23), we get

$$b \stackrel{\text{def}}{=} \frac{\partial J(\mathbf{w})}{\partial w_1}\bigg|_{(a,a)} + \frac{\partial J(\mathbf{w})}{\partial w_2}\bigg|_{(a,a)} = \frac{-9}{1+\exp(9a)} + \frac{-0.2}{1+\exp(a/10)} + 6\lambda a.$$

Since b is negative when $a = 0$ and is a continuous function of a, and $\lim_{a\to\infty} b > \infty$, the lemma holds. ∎

Lemma 28 *$v_1 \geq v_2$.*

Proof: Let a be the value from Lemma 27, and let $\mathbf{g} = (g_1, g_2)$ be the gradient of J at (a, a). Lemma 25 implies that $\mathbf{v}$ lies in the halfspace through (a, a) in the direction of $-\mathbf{g}$. Lemma 27 implies that

$$g_1 = \frac{\partial J(\mathbf{w})}{\partial w_1}\bigg|_{(a,a)} = -\frac{\partial J(\mathbf{w})}{\partial w_2}\bigg|_{(a,a)} = -g_2.$$

Examination of the derivatives (22) and (23) at (a, a) shows that the first term of (22) is negative and the first term of (23) is positive while the last three terms match (although in a different order). Therefore $g_1 < 0$ and $g_2 = -g_1$ is positive. Applying Lemma 25 completes the proof. ∎

Lemma 28 implies that $\mathbf{v}$ correctly classifies $(10, -1)$ and $(11/10, -1)$. It remains to show that $\mathbf{v}$ correctly classifies $(-1, 11/10)$, that is, that v_1 is not *too* much bigger than v_2.

Lemma 29 *If $v_2 \geq 0.6$ and $\lambda > 0$ then $v_1 < 11v_2/10$.*

Proof: Combining $\frac{\partial J}{\partial w_1}\Big|_{\mathbf{v}} = 0$ with (22), we get

$$3\lambda v_1 = \frac{10}{1+\exp(10v_1 - v_2)} + \frac{1.1}{1+\exp(1.1v_1 - v_2)} + \frac{-1}{1+\exp(-v_1 + 1.1v_2)}$$

and, similarly,

$$3\lambda v_2 = \frac{-1}{1+\exp(10v_1 - v_2)} + \frac{-1}{1+\exp(1.1v_1 - v_2)} + \frac{1.1}{1+\exp(-v_1 + 1.1v_2)}.$$

Thus

$$3\lambda(10v_1 - 11v_2) = \frac{111}{1+\exp(10v_1 - v_2)} + \frac{22}{1+\exp(1.1v_1 - v_2)} - \frac{22.1}{1+\exp(-v_1 + 1.1v_2)}. \quad (24)$$

Assume for contraction that $v_1 \geq 11v_2/10$. Then $10v_1 - v_2 \geq 10v_2$, $1.1v_1 - v_2 \geq 0.21v_2$, and $-v_1 + 1.1v_2 \leq 0$, so

$$3\lambda(10v_1 - 11v_2) \leq \frac{111}{1+\exp(10v_2)} + \frac{22}{1+\exp(0.21v_2)} - 11.05.$$

However, $10v_1 - 11v_2 \geq 0$ and (since $v_2 \geq 0.6$) the RHS is negative, giving the desired contradiction. ∎

Lemma 30 *If $0 < \lambda \leq 1/50$ then $v_2 \geq 0.6$.*

Proof: It suffices to show that there is a point $(x, 0.6)$ where the partial w.r.t. w_1 is 0 and the partial w.r.t w_2 is negative.

$$\frac{\partial J}{\partial w_1}\Big|_{(x,0.6)} = \frac{-10}{1+\exp(10x - 0.6)} + \frac{-1.1}{1+\exp(1.1x - 0.6)} + \frac{1}{1+\exp(-x + 0.66)} + 3\lambda x$$

and is increasing in x and λ (assuming $x > 0$) and becomes positive as x goes to infinity. It is negative when evaluated at $x = 0.6$ and $\lambda = 1/50$, so for all $\lambda \leq 1/50$ there is an $x > 0.6$ such that $\partial J/\partial w_+\big|_{(x,1)} = 0$.

$$\frac{\partial J}{\partial w_2}\Big|_{(x,0.6)} = \frac{1}{1+\exp(10x - 0.6)} + \frac{1}{1+\exp(1.1x - 0.6)} + \frac{-1.1}{1+\exp(-x + 0.66)} + 1.8\lambda$$

and is decreasing in x and increasing in λ. It is negative when $x = 0.6$ and $\lambda = 1/50$, so it will remain negative for all $x > 0.6$ and $0 \leq \lambda \leq 1/50$, as desired. ∎

So, we have shown that, if $\lambda \leq 1/50$, then all examples are classified correctly by $\mathbf{v}$, which proves Theorem 16.

Appendix E. Proof of Theorem 17

Theorem 17. If $q \geq 1/3$ then $\mathrm{er}_{P_5}(\mathbf{w}^*(P_5, q)) \geq 1/3$ for the distribution P_5 defined in (12).

Throughout this proof we also abbreviate P_5 as just P.

For this subsection, let us define the scaled dropout criterion

$$J(\mathbf{w}) = 3\, \mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y(\mathbf{w} \cdot (\mathbf{r} \odot \mathbf{x})))) \tag{25}$$

where the components of $\mathbf{r}$ are independent samples from a Bernoulli distribution with parameter $p = 1 - q > 0$. Again, the factor of 3 is to simplify the expectation and doesn't change the minimizing $\mathbf{w}$. Let $\mathbf{w}^{\circledast}$ be the minimizer of this $J(\mathbf{w})$, so that Equation (4) implies that the optimizer $\mathbf{w}^*$ of the dropout criterion is $p\mathbf{w}^{\circledast}$. Note that $\mathbf{w}^*$ classifies an example correctly if and only if $\mathbf{w}^{\circledast}$ does.

Next, note that we may assume without loss of generality that both components of $\mathbf{w}^{\circledast}$ are positive, since, if either is negative, one of $(-1, 1.1)$ or $(1.1, -1)$ is misclassified and we are done.

We will prove Theorem 17 by proving that, when $q \geq 1/3$, $\mathbf{w}^{\circledast}$ misclassifies $(-1, 1.1)$, or, equivalently, that $w_1^{\circledast} > (11/10)w_2^{\circledast}$.

First, let us evaluate some partial derivatives. (Note that, if x_i is dropped out, the value of w_i does not matter.)

$$\begin{aligned}\frac{\partial J}{\partial w_1} &= (1-q)^2 \left(\frac{-10}{1+\exp(10w_1 - w_2)} + \frac{-1.1}{1+\exp(1.1w_1 - w_2)} + \frac{1}{1+\exp(-w_1 + 1.1w_2)} \right) \qquad (26)\\ &\quad + (1-q)q \left(\frac{-10}{1+\exp(10w_1)} + \frac{-1.1}{1+\exp(1.1w_1)} + \frac{1}{1+\exp(-w_1)} \right)\\ \frac{\partial J}{\partial w_2} &= (1-q)^2 \left(\frac{1}{1+\exp(10w_1 - w_2)} + \frac{1}{1+\exp(1.1w_1 - w_2)} + \frac{-1.1}{1+\exp(-w_1 + 1.1w_2)} \right) \qquad (27)\\ &\quad + q(1-q) \left(\frac{1}{1+\exp(-w_2)} + \frac{1}{1+\exp(-w_2)} + \frac{-1.1}{1+\exp(1.1w_2)} \right).\end{aligned}$$

The following is the key lemma. As before, it is useful since, for any $\mathbf{w}$, if $g(\mathbf{w})$ is nonzero, then $\mathbf{w}^{\circledast}$ lies in the open halfspace through $\mathbf{w}$ whose normal vector is the negative gradient.

Lemma 31 *For all $a > 0$ and $q \geq 1/3$,*

$$\left.\frac{\partial J}{\partial w_2}\right|_{(a,10a/11)} > 0. \tag{28}$$

Proof: We have

$$\begin{aligned}\left.\frac{\partial J}{\partial w_2}\right|_{(a,10a/11)} =& (1-q)^2 \left(\frac{1}{1+\exp(100a/11)} + \frac{1}{1+\exp(21a/110)} + \frac{-1.1}{2} \right)\\ &+ q(1-q) \left(\frac{2}{1+\exp(-10a/11)} + \frac{-1.1}{1+\exp(a)} \right).\end{aligned}$$

Note that this derivative is positive if and only if

$$\begin{aligned}
&f(q,a)\\
&\quad= \left(\frac{1}{1-q}\right) \left.\frac{\partial J}{\partial w_2}\right|_{(a,10a/11)}\\
&=q\left(\frac{11}{20}+\frac{2}{1+\exp(-10a/11)}+\frac{-1}{1+\exp(21a/110)}+\frac{-1}{1+\exp(100a/11)}+\frac{-11/10}{1+\exp(a)}\right)\\
&\qquad+\frac{1}{1+\exp(21a/110)}+\frac{1}{1+\exp(100a/11)}+\frac{-11}{20}
\end{aligned}$$

is positive, as $0<q<1$. Note that the terms multiplying q are increasing in a and sum to 0 when $a=0$. On the other hand, the terms not multiplied by q are decreasing in a and turn negative when a is just over 1/4. Thus both parts are positive when $a \leq 1/4$. Note that $f(q,a)$ can be underestimated by underestimating a on the q-terms and overestimating a on the other terms.

For $1/4 \leq a \leq 2$,

$$\begin{aligned}
&f(q,a)\\
&\geq q\left(\frac{11}{20}+\frac{2}{1+\exp(-10/44)}+\frac{-1}{1+\exp(21/440)}+\frac{-1}{1+\exp(100/44)}+\frac{-11/10}{1+\exp(1/4)}\right)\\
&+\frac{1}{1+\exp(42/110)}+\frac{1}{1+\exp(200/11)}+\frac{-11}{20}\\
&\geq 0.5q-0.15
\end{aligned}$$

and is positive whenever $q \geq 1/3$.

For $a \geq 2$,

$$\begin{aligned}
&f(q,a)\\
&\geq q\left(\frac{11}{20}+\frac{2}{1+\exp(-20/11)}+\frac{-1}{1+\exp(42/110)}+\frac{-1}{1+\exp(200/11)}+\frac{-11/10}{1+\exp(2)}\right)+\frac{-11}{20}\\
&\geq 1.7q-11/20
\end{aligned}$$

and is also positive whenever $q \geq 1/3$. ∎

Proof of Theorem 17: Let $\mathbf{g}=(g_1,g_2)$ be the gradient of J at $(w_1^{\circledast}, 10w_1^{\circledast}/11)$. Lemma 31 shows $\mathbf{g}$ is not $\mathbf{0}$, so by convexity

$$\mathbf{w}^{\circledast}\cdot\mathbf{g} < (w_1^{\circledast}, 10w_1^{\circledast}/11)\cdot\mathbf{g}$$

which implies

$$w_2^{\circledast}\, g_2 < (10w_1^{\circledast}/11)\, g_2.$$

Since $g_2>0$ (Lemma 31), this implies

$$w_2^{\circledast} < (10w_1^{\circledast}/11)$$

and the $(-1, 11/10)$ example is misclassified by $\mathbf{w}^{\circledast}$, and therefore by $\mathbf{w}^{*}$, completing the proof. ∎

Appendix F. Proof of Theorem 18

Theorem 18. If $1/100 \le \lambda \le 1$, then $\mathrm{er}_{P_6}(\mathbf{v}(P_6, \lambda)) \ge 1/7$ for the distribution P_6 defined in (13).

To keep the notation clean, in this section let us abbreviate P_6 simply as P.

As the reader might expect, we will prove Theorem 18 by proving that $\mathbf{v}$ fails to correctly classify $(1/10, -1)$, that is, by proving that $v_1 < 10v_2$.

We may assume that $v_1 > 0$, since, otherwise, $(1, 0)$ is misclassified.

To obtain cancellation in the expectation, we work with the scaled L_2 criterion

$$J(\mathbf{w}) = 7\mathbf{E}_{(\mathbf{x},y)\sim P}(\ell(y(\mathbf{w} \cdot \mathbf{x}))) + (7\lambda/2)||\mathbf{w}||^2. \tag{29}$$

and let $\mathbf{v}(P, \lambda)$ be the vector minimizing this J, which we often abbreviate as simply $\mathbf{v}$, leaving it implicitly a function of λ. Note that this scaling of the criteria does not change the minimizing $\mathbf{v}$.

Taking derivatives,

$$\frac{\partial J}{\partial w_1} = \frac{-3}{1+\exp(w_1)} + \frac{3\epsilon}{1+\exp(-\epsilon w_1 + w_2)} + \frac{-0.1}{1+\exp(w_1/10 - w_2)} + 7\lambda w_1 \tag{30}$$

$$\frac{\partial J}{\partial w_2} = \frac{-3}{1+\exp(-\epsilon w_1 + w_2)} + \frac{1}{1+\exp(w_1/10 - w_2)} + 7\lambda w_2. \tag{31}$$

Lemma 32 *If either: $\lambda \ge 1/100$ and $a \ge 1/3$, or $\lambda \ge 1/4$ and $a \ge 1/15$ then*

$$\left.\frac{\partial J(\mathbf{w})}{\partial w_1}\right|_{(10a,a)} > 0.$$

Proof: We have

$$\begin{aligned} &\left.\frac{\partial J(\mathbf{w})}{\partial w_1}\right|_{(10a,a)} \\ &= \frac{-3}{(1+\exp(10a))} + \frac{3\epsilon}{1+\exp((1-10\epsilon)a)} + \frac{-1}{20} + 70\lambda a \\ &> \frac{-3}{(1+\exp(10a))} + \frac{-1}{20} + 70\lambda a. \end{aligned}$$

Each term of the RHS is non-decreasing in a and λ, and the RHS is positive when either $\lambda = 1/100$ and $a = 1/3$ or $\lambda = 1/4$ and $a = 1/15$. ∎

To apply this, we want to show that v_2 is large enough, which we do next.

Lemma 33 *If $\lambda \le 1/4$ then $v_2 \ge 1/3$ and if $\lambda \le 1$ then $v_2 \ge 1/15$.*

Proof: Assume to the contrary that $\lambda \le 1/4$ but $v_2 < 1/3$. From (31), and using that $v_1 > 0$, we have

$$\left.\frac{\partial J}{\partial w_2}\right|_{\mathbf{v}} < \frac{-3}{1+\exp(v_2)} + \frac{1}{1+\exp(-v_2)} + 7\lambda v_2, \tag{32}$$

x_1r_1	x_2r_2	y	seven times probability			$\mathbf{w}^{\circledast} \cdot (\mathbf{r} \odot \mathbf{x})$ over-estimate
0	0	1	$3q$	$+3q^2$	$+q^2$	0
1	0	1	$3(1-q)$			∞
0	1	1		$3q(1-q)$		w_2
$-1/1000$	0	1		$3q(1-q)$		0
$-1/1000$	1	1		$3(1-q)^2$		w_2
0	-1	1			$q(1-q)$	∞
$1/10$	0	1			$q(1-q)$	∞
$1/10$	-1	1			$(1-q)^2$	∞

Table 2: Seven times the dropout distribution. The three probability sub-columns correspond to the original examples (1,0), (-1/1000, 1), (1/10, -1), and the final column is the over-estimate used in Lemma 36.

a bound that is increasing in v_2 and λ. Since $\left.\frac{\partial J}{\partial w_2}\right|_{\mathbf{v}} = 0$, the bound must be positive. However, when $v_2 \leq 1/3$ and $\lambda \leq 1/4$, it is negative, giving the desired contradiction.

Since the bound (32) is also negative at $v_2 = 1/15$ and $\lambda = 1$, a similar contradiction proves the other half of the lemma. ∎

Proof: (of Theorem 18): Lemmas 32 and 33 imply that $(10v_2, v_2)$ is not the minimizing $\mathbf{v}$ (when $\lambda \geq 1/100$), so by convexity,

$$J(10v_2, v_1) + \big((v_1, v_2) - (10v_2, v_2)\big) \cdot \nabla J(10v_2, v_2) < J(v_1, v_2) \tag{33}$$

$$(v1 - 10v_2) \left.\frac{\partial J}{\partial w_2}\right|_{(10v_2, v_2)} < 0. \tag{34}$$

If $1/100 \leq \lambda \leq 1/4$ then Lemma 33 shows that $v_2 \geq 1/3$ and if $1/4 \leq \lambda \leq 1$ then it shows that $v_2 \geq 1/15$. In either case, Lemma 32 shows that that $\left.\frac{\partial J}{\partial w_2}\right|_{(10v_2, v_2)} > 0$. Therefore,

$$v_1 < 10v_2$$

and $(0.1, -1)$ is misclassified by $\mathbf{v}$, completing the proof. ∎

Appendix G. Proof of Theorem 19

Theorem 19. If $q \leq 1/2$, then $\mathrm{er}_{P_6}(\mathbf{w}^*(P_6, q)) = 0$ for the distribution P_6 defined in (13).

In this proof, let us abbreviate P_6 with just P, and use ϵ to denote $1/1000$.

For this section, let us define the scaled dropout criterion

$$J(\mathbf{w}) = 7\mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y(\mathbf{w} \cdot (\mathbf{r} \odot \mathbf{x})))), \tag{35}$$

where, as earlier, the components of $\mathbf{r}$ are independent samples from a Bernoulli distribution with parameter $p = 1-q = 1/2 > 0$. (Note that, similarly to before, scaling up the objective function by 7 does not change the minimizer of J.) See Table 2 for a tabular representation of the distribution after dropout. Let $\mathbf{w}^{\circledast}$ be the minimizer of J, so that $\mathbf{w}^* = p\mathbf{w}^{\circledast}$ (see Equation (4)).

First, let us evaluate some partial derivatives (note that $1-q = (1-q)^2 + q(1-q)$).

$$\begin{aligned}\frac{\partial J}{\partial w_1} &= (1-q)^2 \left(\frac{-3}{1+\exp(w_1)} + \frac{3\epsilon}{1+\exp(-\epsilon w_1 + w_2)} + \frac{-0.1}{1+\exp(0.1 w_1 - w_2)} \right) \qquad (36)\\ &\quad + (1-q)q \left(\frac{-3}{1+\exp(w_1)} + \frac{3\epsilon}{1+\exp(-\epsilon w_1)} + \frac{-0.1}{1+\exp(0.1 w_1)} \right)\\ \frac{\partial J}{\partial w_2} &= (1-q)^2 \left(\frac{-3}{1+\exp(-\epsilon w_1 + w_2)} + \frac{1}{1+\exp(0.1 w_1 - w_2)} \right) \qquad (37)\\ &\quad + q(1-q) \left(\frac{-3}{1+\exp(w_2)} + \frac{1}{1+\exp(-w_2)} \right).\end{aligned}$$

Let's get started by showing that $\mathbf{w}^{\circledast}$ correctly classifies $(1,0)$.

Lemma 34 $w_1^{\circledast} > 0$.

Proof: As before, it suffices to show that there is a point $(0, a_2)$ where both $\frac{\partial J}{\partial w_1}\big|_{(0,a_2)} < 0$ and $\frac{\partial J}{\partial w_2}\big|_{(0,a_2)} = 0$.

From Equation (36):

$$\frac{\partial J}{\partial w_1}\Big|_{(0,a_2)} = (1-q)^2 \left(\frac{-3}{2} + \frac{3\epsilon}{1+\exp(a_2)} + \frac{-0.1}{1+\exp(-a_2)} \right) + \frac{(1-q)q}{2}(-3.1 + 3\epsilon)$$

which is decreasing in a_2, and negative even as a_2 approaches $-\infty$ (recalling $\epsilon = 1/1000$), so $\frac{\partial J}{\partial w_1}\Big|_{(0,a_2)}$ is always negative.

Equation (37) implies

$$\begin{aligned}&\frac{\partial J}{\partial w_2}\Big|_{(0,a_2)}\\ &= (1-q)^2 \left(\frac{-3}{1+\exp(a_2)} + \frac{1}{1+\exp(-a_2)} \right) + q(1-q) \left(\frac{-3}{1+\exp(a_2)} + \frac{1}{1+\exp(-a_2)} \right).\end{aligned}$$

This is negative when $a_2 = 0$, approaches $1-q$ as a_2 goes to infinity, and is continuous, so there is a a_2 such that $\frac{\partial J}{\partial w_2}\Big|_{(0,a_2)} = 0$. Since $\frac{\partial J}{\partial w_1}\Big|_{(0,a_2)} < 0$, this proves the lemma. ∎

Next, we'll start to work on showing that $\mathbf{w}^{\circledast}$ correctly classifies $(-\epsilon, 1)$.

Lemma 35 *For all $a > 1/10$,*

$$\frac{\partial J}{\partial w_1}\Bigg|_{(a/\epsilon,a)} > 0.$$

Proof: From (36), we have

$$\begin{aligned}\frac{\partial J}{\partial w_1}\Big|_{(a/\epsilon,a)} =&(1-q)^2 \left(\frac{-3}{1+\exp(a/\epsilon)} + \frac{3\epsilon}{1+\exp(0)} + \frac{-0.1}{1+\exp(0.1(a/\epsilon)-a)} \right)\\ &+ q(1-q) \left(\frac{-3}{1+\exp(a/\epsilon)} + \frac{3\epsilon}{1+\exp(-a)} + \frac{-0.1}{1+\exp(a/10\epsilon)} \right)\end{aligned}$$

which is positive if $a > 1/10$ as the positive terms (even with the ϵ factors) dominate the negative ones. ■

Lemma 36

$$w_2^{\circledast} > 1/4.$$

Proof: Assuming $w_1 \geq 0$, the estimates in Table 2 along with the facts that $\ell(z)$ is positive and decreasing show :

$$J(\mathbf{w}) \geq 3(1-q)\ln(1+\exp(-w_2)) + 6q\ln(2) + q^2\ln(2) \tag{38}$$

which is decreasing in w_2. If $w_2^{\circledast} \leq 1/4$, then bound (38) and the fact that $w_1^{\circledast} > 0$ (Lemma 34) imply that

$$J(\mathbf{w}^{\circledast}) \geq 0.69q^2 + 2.4q + 1.7.$$

On the other hand,

$$J(100, 2) \leq -1.5q^2 + 6q + 0.42,$$

and the upper bound on $J(100, 2)$ is less than the lower bound on $J(\mathbf{w}^{\circledast})$ when $0 \leq q \leq 1/2$, giving the desired contradiction. ■

Now, we're ready to show that $\mathbf{w}^{\circledast}$ correctly classifies $(-\epsilon, 1)$.

Lemma 37 $\epsilon w_1^{\circledast} < w_2^{\circledast}$.

Proof: Let $\mathbf{g}$ be the gradient of J evaluated at $(w_2^{\circledast}/\epsilon, w_2^{\circledast})$. Combining Lemmas 35 and 36, $\mathbf{g} \neq (0, 0)$, so

$$\mathbf{w}^{\circledast} \cdot \mathbf{g} < (w_2^{\circledast}/\epsilon, w_2^{\circledast}) \cdot \mathbf{g}.$$

This implies

$$w_1^{\circledast} \left.\frac{\partial J}{\partial w_1}\right|_{(w_2^{\circledast}/\epsilon, w_2^{\circledast})} < \frac{w_2^{\circledast}}{\epsilon} \left.\frac{\partial J}{\partial w_1}\right|_{(w_2^{\circledast}/\epsilon, w_2^{\circledast})}.$$

Since Lemmas 35 and 36 imply that $g(w_2^{\circledast}/\epsilon, w_2^{\circledast})_1 > 0$, this completes the proof. ■

Finally, we are ready to work on showing that $(1/10, -1)$ is correctly classified by $\mathbf{w}^{\circledast}$, i.e. that $w_1^{\circledast} > 10 w_2^{\circledast}$.

Lemma 38 *For all* $a \in \mathbf{R}$*,*

$$\left.\frac{\partial J}{\partial w_1}\right|_{(10a, a)} < 0.$$

Proof: Choose $a \in R$. From (36), we have

$$\begin{aligned}
\left.\frac{\partial J}{\partial w_1}\right|_{(10a, a)} &= q(1-q)\left(\frac{-3}{1+\exp(10a)} + \frac{3\epsilon}{1+\exp(-10\epsilon a)} + \frac{-1}{10(1+\exp(a))}\right) \\
&\quad + (1-q)^2\left(\frac{-3}{1+\exp(10a)} + \frac{3\epsilon}{1+\exp(a - 10\epsilon a)} + \frac{-1}{20}\right) \\
&\leq (1-q)^2\left(6\epsilon + \frac{-1}{20}\right) < 0
\end{aligned}$$

using $q \leq 1/2$ and $\epsilon = 1/1000$. ■

Lemma 39 $w_1^{\circledast} > 10 w_2^{\circledast}$.

Proof: Let $\mathbf{g}$ be the gradient of J evaluated at $\mathbf{u} = (10w_2^{\circledast}, w_2^{\circledast})$. Lemma 38 implies that $\mathbf{g} \neq (0,0)$, i.e. that $w_1^{\circledast} \neq 10w_2^{\circledast}$. Therefore,

$$\mathbf{w}^{\circledast} \cdot \mathbf{g} < \mathbf{u} \cdot \mathbf{g}$$

which, since $u_2 = w_2^{\circledast}$, implies

$$w_1^{\circledast} \frac{\partial J}{\partial w_1}\Big|_{\mathbf{u}} < 10 w_2^{\circledast} \frac{\partial J}{\partial w_1}\Big|_{\mathbf{u}}.$$

Since Lemma 38 implies that $\partial J / \partial w_1 \Big|_{\mathbf{u}} < 0$, this in turn implies

$$w_1^{\circledast} > 10 w_2^{\circledast},$$

completing the proof. ∎

Now we have all the pieces to prove that dropout succeeds on P.

Proof (of Theorem 19): Lemma 34 implies that $(1,0)$ is classified correctly by $\mathbf{w}^{\circledast}$, and therefore by $\mathbf{w}^* = p\mathbf{w}^{\circledast}$. Lemma 37 implies that $(-\epsilon, 1)$ is classified correctly. Lemma 39 implies that $(1/10, -1)$ is classified correctly, completing the proof. ∎

Appendix H. Proof of Theorem 20

Theorem 20. If $\lambda \leq \frac{1}{30n}$ then the weight vector $v(P_9, \lambda)$ optimizing the L_2 criterion has perfect prediction accuracy: $\mathrm{er}_{P_9}(v(P_9,\lambda)) = 0$.

In this proof, let us abbreviate P_9 as just P.

By symmetry and convexity, the optimizing $\mathbf{v}$ is of the form $(v_1, v_2, v_2, \ldots, v_2)$ with the last $n-1$ components being equal. Thus for this distribution minimizing the L_2 criterion is equivalent to minimizing the simpler criterion $K(w_1, w_2)$ defined by:

$$K(w_1,w_2) = \frac{9}{10} \ln\left(1 + \exp(-w_1 - w_2)\right) + \frac{1}{10} \ln\left(1 + \exp(w_1 - w_2)\right) + \frac{\lambda}{2}\left(w_1^2 + (n-1) w_2^2\right).$$

Let (v_1, v_2) be the minimizing vector of $K()$, retaining an implicit dependence on n and λ. We will be making frequent use of the partial derivatives of K:

$$\frac{\partial K}{\partial w_1} = \frac{-9}{10(1+\exp(w_1+w_2))} + \frac{1}{10(1+\exp(-w_1+w_2))} + \lambda w_1 \tag{39}$$

$$\frac{\partial K}{\partial w_2} = \frac{-9}{10(1+\exp(w_1+w_2))} + \frac{-1}{10(1+\exp(-w_1+w_2))} + (n-1)\lambda w_2. \tag{40}$$

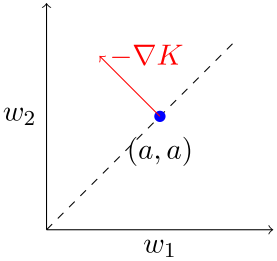
It suffices to show that $0 \leq v_1 < v_2$ so that the first feature does not perturb the majority vote of the others.

To see $0 \leq v_1$, notice that $\partial K / \partial w_1 \big|_{(0,w_2)}$ is negative for all w_2, including when $w_2 = v_2$.

To prove $v_1 < v_2$ we show the existence of a point (a,a) such that

$$\left.\frac{\partial K}{\partial w_1}\right|_{(a,a)} = -\left.\frac{\partial K}{\partial w_2}\right|_{(a,a)} > 0, \tag{41}$$

so that Lemma 25 implies that the optimizing (v_1, v_2) lies above the $w_1 = w_2$ diagonal.



We have

$$\left.\frac{\partial K}{\partial w_1}\right|_{(a,a)} = \frac{-9}{10(1+\exp(2a))} + \frac{1}{20} + \lambda a$$

which is increasing in a, negative when $a = 0$ and goes to infinity with a. It turns positive at some $a < 1.5$ (exactly where depends on λ).

On the other hand,

$$\left.\frac{\partial K}{\partial w_2}\right|_{(a,a)} = \frac{-9}{10(1+\exp(2a))} + \frac{-1}{20} + \lambda(n-1)a$$

and is also increasing in a and goes to infinity. However, $\partial K/\partial w_2\big|_{(a,a)}$ is negative at $a = 1.5$ whenever $1.5\lambda(n-1) \leq 1/20$, which is implied by the premise of the theorem.

Both partial derivatives are negative when $a = 0$, continuously go to infinity with a, and $\partial K/\partial w_1\big|_{(a,a)}$ crosses zero first. From the point where $\partial K/\partial w_1\big|_{(a,a)}$ crosses zero until $\partial K/\partial w_2\big|_{(a,a)}$ does, the magnitude of $\partial K/\partial w_1\big|_{(a,a)}$ is increasing, starting at 0, and the magnitude of $\partial K/\partial w_2\big|_{(a,a)}$ is decreasing until it reaches 0. When they meet, Equation (41) holds, completing the proof.

Appendix I. Proof of Theorem 21

Theorem 21. If the dropout probability $q = 1/2$ and the number of features is an even $n > 125$ then the weight vector $\mathbf{w}^*(P_9, q)$ optimizing the dropout criterion has prediction error rate $\mathrm{er}_{P_9}(\mathbf{w}^*(P_9, q)) \geq 1/10$.

In this proof, we again abbreviate, using P for P_9.

The complicated form of the criterion optimized by dropout makes analyzing it difficult. Here we make use of Jensen's inequality. However, a straightforward application of it is fruitless, and a key step is to apply Jensen's inequality on just half the distribution resulting from dropout.

Similarly to before, let

$$J(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y(\mathbf{w} \cdot (\mathbf{r} \odot \mathbf{x})))), \tag{42}$$

and let $\mathbf{w}^{\circledast}$ minimize J, so that $\mathbf{w}^* = p\mathbf{w}^{\circledast}$.

Again using symmetry and convexity, the last $n-1$ components of the optimizing $\mathbf{w}^{\circledast}$ are equal, so $\mathbf{w}^{\circledast}$ is of the form $(w_1^{\circledast}, w_2^{\circledast}, w_2^{\circledast}, \ldots, w_2^{\circledast})$.

Lemma 40 *The minimizing $w_1^{\circledast}$ of (42) is positive.*

Proof: Let $\widetilde{P,\mathbf{r}}$ be the marginal distribution of the last $n-1$ components after dropout and $\tilde{\mathbf{x}}$ denote these last $n-1$ components of the dropped-out feature vector. Then, recalling y is always 1 in our distribution (and p is the probability that the first feature is *not* dropped out),

$$\frac{\partial J(w)}{\partial w_1} = \mathbf{E}_{(r_2,\ldots,r_n)}\left(\frac{9p}{10}\mathbf{E}_{\tilde{\mathbf{x}}\sim\widetilde{P,\mathbf{r}}}(\ell'(\mathbf{w}\cdot(1,\tilde{\mathbf{x}}))) - \frac{p}{10}\mathbf{E}_{\tilde{\mathbf{x}}\sim\widetilde{P,\mathbf{r}}}(\ell'(\mathbf{w}\cdot(-1,\tilde{\mathbf{x}})))\right)$$

which is negative whenever $w_1 = 0$, since $\ell'()$ is negative and the two inner expectations become identical when $w_1 = 0$. Therefore the optimizing $w_1^{\circledast}$ is positive. ∎

To show that dropout fails, we want to show that $w_1^{\circledast} > w_2^{\circledast}$, i.e. that $w_1^{\circledast} \leq w_2^{\circledast}$ leads to a contradiction, so we begin to explore the consequences of $w_1^{\circledast} \leq w_2^{\circledast}$.

Lemma 41 *If $q = 1/2$ and $w_1^{\circledast} \leq w_2^{\circledast}$ then $w_2^{\circledast} > 4/9$.*

Proof: Assume to the contrary that $w_1^{\circledast} \leq w_2^{\circledast} \leq 4/9$.

Using Jensen's inequality,

$$J(\mathbf{w}^{\circledast}) \geq \ell(\mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(y(\mathbf{w}^{\circledast}\cdot\mathbf{x})))$$

and the inner expectation is $8w_1^{\circledast}/20 + w_2^{\circledast}/2 \leq 9w_2^{\circledast}/10$ as $w_1^{\circledast} \leq w_2^{\circledast}$. Therefore, since $w_2^{\circledast} \leq 4/9$,

$$J(\mathbf{w}^{\circledast}) \geq \ell(0.4) > 0.51.$$

However,

$$J(2.1, 0, 0, \ldots, 0) = \frac{\ln(2)}{2} + \frac{9\ln(1+e^{-2.1})}{20} + \frac{\ln(1+e^{2.1})}{20} < 0.51$$

contradicting the optimality of $\mathbf{w}^{\circledast}$. ∎

Lemma 42 *If $q = 1/2$ and $w_1^{\circledast} \leq w_2^{\circledast}$ then $J(\mathbf{w}^{\circledast}) \geq \mathbf{E}_{k\sim B(n,1/2)}\ell(w_2^{\circledast}(k-(n/2)+1))$ where $B(n, 1/2)$ is the binomial distribution.*

Proof: Consider the modified distribution P_1 over $(\mathbf{x}, y)$ examples where y is always 1, x_2, ..., x_n are uniformly distributed over the the vectors with $n/2$ ones and $(n/2)-1$ negative ones (as in P), but x_1 is always one. Since $0 < w_1^\circledast \leq w_2^\circledast$ and the label $y = 1$ under P and P_1,

$$\begin{aligned} J(\mathbf{w}^\circledast) &= \mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(\mathbf{w}^\circledast \cdot \mathbf{x})) \\ &> \mathbf{E}_{(\mathbf{x},y)\sim P_1,\mathbf{r}}(\ell(\mathbf{w}^\circledast \cdot \mathbf{x})) \\ &= \mathbf{E}_{(\mathbf{x},y)\sim P_1,\mathbf{r}}\left(\ell\big(w_2^\circledast(\mathbf{1}\cdot(\mathbf{x}\odot\mathbf{r}))\big)\right) \\ &= \mathbf{E}_{(\mathbf{x},y)\sim P_1,\mathbf{r}}\left(\ell\big(w_2^\circledast(\mathbf{x}\cdot\mathbf{r})\big)\right). \end{aligned}$$

Every $\mathbf{x}$ in the support of P_1 has exactly $(n/2)+1$ components that are 1, and the remaining $(n/2)-1$ components are -1. Call a component a *success* if it is either -1 and dropped out or 1 and not dropped out. Now, $\mathbf{x}\cdot\mathbf{r}$ is exactly $1-(n/2)$ plus the number of successes. Furthermore, the number of successes is distributed according to the binomial distribution $B(n, 1/2)$. Therefore

$$\mathbf{E}_{(\mathbf{x},y)\sim P_1,\mathbf{r}}(w_2^\circledast(\mathbf{x}\cdot\mathbf{r})) = \mathbf{E}_{k\sim B(n,1/2)}(\ell(w_2^\circledast(k-(n/2)+1)))$$

giving the desired bound. ∎

Lemma 43 *For even* $n \geq 6$, $\mathbf{E}_{k\sim B(n,1/2)}(\ell(w_2^\circledast(k-(n/2)+1))) \geq \frac{1}{3}\ell\left(w_2^\circledast - \frac{w_2^\circledast\sqrt{2n}}{4}\right)$.

Proof: Let $\alpha = \sum_{i=0}^{n/2-1}\binom{n}{i}$, so α is slightly less than 2^{n-1}.

$$\begin{aligned} \mathbf{E}_{k\sim B(n,1/2)}(\ell(w_2^\circledast(k-(n/2)+1))) &= \frac{1}{2^n}\sum_k \binom{n}{k}\ell(w_2^\circledast(k+1-(n/2))) \\ &> \frac{\alpha}{2^n}\sum_{k=0}^{n/2-1}\frac{1}{\alpha}\binom{n}{k}\ell(w_2^\circledast(1+k-(n/2))) \\ &> \frac{\alpha}{2^n}\ell\left(\sum_{k=0}^{n/2-1}\frac{1}{\alpha}\binom{n}{k}w_2^\circledast(1+k-(n/2))\right) \end{aligned}$$

where the last step uses Jensen's inequality. Continuing,

$$\mathbf{E}_{k\sim B(n,1/2)}(\ell(w_2^\circledast(k-(n/2)+1))) > \frac{\alpha}{2^n}\ell\left(w_2^\circledast + \frac{w_2^\circledast}{\alpha}\sum_{k=0}^{n/2-1}\binom{n}{k}(k-(n/2))\right).$$

Equation (5.18) of Concrete Mathematics (Graham et al., 1989) and the bound $\binom{n}{n/2} \geq \frac{2^n}{\sqrt{2n}}$ give

$$\sum_{k=0}^{n/2-1}\binom{n}{k}(k-(n/2)) = \frac{-n}{4}\binom{n}{n/2} \leq \frac{-\sqrt{2n}\,2^{n-1}}{4}.$$

Therefore, recalling that $\alpha < 2^{n-1}$ and noting $\alpha/2^n > 1/3$ when $n \geq 6$,

$$\mathbf{E}_{k\sim B(n,1/2)}(\ell(w_2^{\circledast}(k-(n/2)+1))) > \frac{\alpha}{2^n}\ell\left(w_2^{\circledast} - \frac{w_2^{\circledast}}{\alpha}\frac{2^{n-1}\sqrt{2n}}{4}\right)$$
$$> \frac{1}{3}\ell\left(w_2^{\circledast} - \frac{w_2^{\circledast}\sqrt{2n}}{4}\right).$$

∎

We now have the necessary tools to prove Theorem 21.

Proof: (of Theorem 21) If $w_1^{\circledast} > w_2^{\circledast}$ then the first feature will dominate the majority vote of the others and the optimizing $\mathbf{w}^{\circledast}$ has prediction error rate 1/10 . We now assume to the contrary that $w_1^{\circledast} \leq w_2^{\circledast}$. When $n > 125$ and $w_2^{\circledast} \geq 4/9$ (from Lemma 41) we have

$$w_2^{\circledast} - \frac{w_2^{\circledast}\sqrt{2n}}{4} \leq -1.31$$

and $\ell(w_2^{\circledast} - \frac{w_2^{\circledast}\sqrt{2n}}{4}) > 1.54$.

Lemmas 42 and 43 now imply that $J(\mathbf{w}^{\circledast}) > 0.51$, but (as in Lemma 41) $J(2.1, 0, \ldots 0) < 0.51$, contradicting the optimality of $\mathbf{w}^{\circledast}$. ∎

Many of the approximations used to prove Theorem 21 are quite loose, resulting in large values of n being needed to obtain the contradiction. For this class of distributions and $q = 1/2$ we conjecture that optimizing the dropout criterion fails to produce the Bayes optimal hypothesis for every even $n \geq 4$.

Appendix J. Proof of Theorem 22

Theorem 22. If dropout probability $q = 1/2$ and the number of features is $n = 4$ then the minimizer of the dropout criteria $\mathbf{w}^*(P_9, q)$ has has prediction error rate $\mathrm{er}_{P_9}(\mathbf{w}^*(P_9,q)) \geq 1/10$.

In this proof, let us also refer to P_9 as just P and let $\mathbf{w}^{\circledast}$ be the minimizer of (42).

As before, the optimizing $\mathbf{w}^{\circledast}$ has the form $(w_1^{\circledast}, w_2^{\circledast}, w_2^{\circledast}, w_2^{\circledast})$ by symmetry and convexity. Recalling that the label y is always 1 under distribution P, we can use the equivalent criterion

$$K(w_1,w_2) = \mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y(\mathbf{w}\cdot\mathbf{x}))) = \mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}\left(\ell\left(w_1 x_1 r_1 + w_2\sum_{i=2}^{4} x_i r_i\right)\right).$$

This expectation can be written with 12 terms, one for each pairing of the three possible $x_1 r_1$ values with the four possible $\sum_{i=2}^4 x_i r_i \in \{-1, 0, 1, 2\}$ values (see Table 3).

Taking them in order, we have

$$\begin{aligned} K(w_1,w_2) =& \frac{9}{160}\ell(w_1+2w_2) + \frac{27}{160}\ell(w_1+w_2) + \frac{27}{160}\ell(w_1) + \frac{9}{160}\ell(w_1-w_2) \\ &+ \frac{10}{160}\ell(2w_2) + \frac{30}{160}\ell(w_2) + \frac{30}{160}\ell(0) + \frac{10}{160}\ell(w_2) \\ &+ \frac{1}{160}\ell(-w_1+2w_2) + \frac{3}{160}\ell(-w_1+w_2) + \frac{3}{160}\ell(-w_1) + \frac{1}{160}\ell(-w_1-w_2). \end{aligned}$$

x_1r_1	probability	$\sum_{i=2}^4 x_ir_i$	probability
1	9/20	2	1/8
0	1/2	1	3/8
-1	1/20	0	3/8
		-1	1/8

Table 3: Probabilities of x_1r_1 and $\sum_{i=2}^4 x_ir_i$ values assuming dropout probability $q = 1/2$.

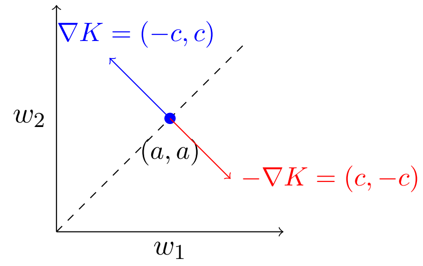


Figure 8: If ∇K at some (a, a) is $(-c, c)$ for some $c > 0$ then $w_1^{\circledast} > w_2^{\circledast}$.

So, when $p = q = 1/2$, the derivatives are:

$$\frac{\partial K}{\partial w_1}$$
$$= \frac{1}{160}\left(\frac{-9}{1+\exp(w_1+2w_2)} + \frac{-27}{1+\exp(w_1+w_2)} + \frac{-27}{1+\exp(w_1)} + \frac{-9}{1+\exp(w_1-w_2)}\right.$$
$$\left. + \frac{1}{1+\exp(-w_1+2w_2)} + \frac{3}{1+\exp(-w_1+w_2)} + \frac{3}{1+\exp(-w_1)} + \frac{1}{1+\exp(-w_1-w_2)}\right),$$
$$\frac{\partial K}{\partial w_2}$$
$$= \frac{1}{160}\left(\frac{-18}{1+\exp(w_1+2w_2)} + \frac{-27}{1+\exp(w_1+w_2)} + \frac{9}{1+\exp(w_1-w_2)}\right.$$
$$+ \frac{-20}{1+\exp(2w_2)} + \frac{-30}{1+\exp(w_2)} + \frac{10}{1+\exp(-w_2)}$$
$$\left. + \frac{-2}{1+\exp(-w_1+2w_2)} + \frac{-3}{1+\exp(-w_1+w_2)} + \frac{1}{1+\exp(-w_1-w_2)}\right).$$

If $w_1^{\circledast} > w_2^{\circledast}$, then dropout will have prediction error rate 1/10 as $w_1^{\circledast}$ will dominate the vote of the other three components. We show that $w_1^{\circledast} > w_2^{\circledast}$ by proving that there is a point (a, a) in weight space such that the gradient at (a, a) is of the form $(-c, c)$ for some $c > 0$ (see Figure 8).

The derivatives when evaluated at (a, a) are:

$$\left.\frac{\partial K}{\partial w_1}\right|_{(a,a)}$$
$$= \frac{1}{160}\left(\frac{-9}{1+\exp(3a)} + \frac{-27}{1+\exp(2a)} + \frac{-26}{1+\exp(a)} - 3 + \frac{3}{1+\exp(-a)} + \frac{1}{1+\exp(-2a)}\right)$$

and

$$\left.\frac{\partial K}{\partial w_2}\right|_{(a,a)}$$
$$= \frac{1}{160}\left(\frac{-18}{1+\exp(3a)} + \frac{-47}{1+\exp(2a)} + \frac{-32}{1+\exp(a)} + 3 + \frac{10}{1+\exp(-a)} + \frac{1}{1+\exp(-2a)}\right).$$

Note that both of these derivatives are increasing in a, positive for large a, and negative when $a = 0$. At $a = 2\ln(2)$, derivative $\partial K/\partial w_1\big|_{(a,a)}$ is still negative, while $\partial K/\partial w_1\big|_{(a,a)}$ has turned positive, so $\partial K/\partial w_1\big|_{(a,a)}$ crosses 0 first. The continuity of the partial derivatives now implies the existence of an (a, a) where ∇K has the form $(-c, c)$, completing the proof. ∎

Appendix K. Proof of Theorem 23

Theorem 23. If $q = 1/2$, $n \geq 100$, $\alpha > 0$, $\beta = 1/(10\sqrt{n-1})$, and $\eta \leq \frac{1}{2+\exp(54\sqrt{n})}$, then $\mathrm{er}_{P_{10}}(\mathbf{w}^*(P_{10}, q)) = \eta$.

For this subsection, let $P = P_{10}$ and define the scaled dropout criterion

$$J(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim P,\mathbf{r}}(\ell(y\mathbf{w}\cdot(\mathbf{r}\odot\mathbf{x}))),$$

where, as earlier, the components of $\mathbf{r}$ are independent samples from a Bernoulli distribution with parameter $p = 1 - q = 1/2 > 0$. Let $\mathbf{w}^{\circledast}$ be the minimizer of J, so that $\mathbf{w}^* = p\mathbf{w}^{\circledast}$.

Note that, by symmetry, the contribution to J from the cases where y is -1 and 1 respectively are the same, so the value of J is not affected if we clamp y at 1. Let us use this form to express J, and let D be the marginal distribution of feature vector $\mathbf{x}$ conditioned on the label $y = 1$.

Let $B = \{2, ..., n\}$. By symmetry, $w_i^{\circledast}$ is identical for all $i \in B$ so $\mathbf{w}^{\circledast}$ is the minimum of J over weight vectors satisfying this constraint. Let $K(w_1, w_2) = J(w_1, w_2, ..., w_2)$; note that $w_1^{\circledast}, w_2^{\circledast}$ minimizes K defined by

$$K(w_1, w_2) = \mathbf{E}_{\mathbf{x}\sim D,\mathbf{r}}(\ell(w_1 r_1 x_1 + w_2 \sum_{i\in B} r_i x_i)).$$

To prove Theorem 23, it suffices to show that

$$w_1^{\circledast} > (n-1)w_2^{\circledast}/\alpha > 0, \tag{43}$$

since when (43) holds, $\mathbf{w}^{\circledast}$ always outputs x_1.

We have

$$\frac{\partial K}{\partial w_1} = \frac{1}{2}\mathbf{E}_{\mathbf{x}\sim D,\mathbf{r}}\left(\frac{-x_1}{1+\exp(w_1 x_1 + w_2 \sum_{i\in B} r_i x_i)}\right) \tag{44}$$

$$\frac{\partial K}{\partial w_2} = \mathbf{E}_{\mathbf{x}\sim D,\mathbf{r}}\left(\frac{-\sum_{i\in B} r_i x_i}{1+\exp(w_1 r_1 x_1 + w_2 \sum_{i\in B} r_i x_i)}\right). \tag{45}$$

(Note that, in (44), we have marginalized out r_1.)

Lemma 44 $w_2^{\circledast} > 0$.

As before, it suffices to show that there is a point $(a_1, 0)$ where both $\frac{\partial K}{\partial w_2}\big|_{(a_1,0)} < 0$ and $\frac{\partial K}{\partial w_1}\big|_{(a_1,0)} = 0$. From equation (45),

$$\frac{\partial K}{\partial w_2}\Big|_{(a_1,0)} = \mathbf{E}_{\mathbf{x}\sim D,\mathbf{r}}\left(\frac{-\sum_{i\in B} r_i x_i}{1+\exp(a_1 r_1 x_1)}\right) < 0$$

for all real a_1.

Now, evaluating (44), dividing into cases based on x_1, we get

$$\frac{\partial K}{\partial w_1}\Big|_{(a_1,0)} = (\eta/2)\left(\frac{\alpha}{1+\exp(-\alpha a_1)}\right) + ((1-\eta)/2)\left(\frac{-\alpha}{1+\exp(\alpha a_1)}\right).$$

This approaches $-\alpha((1-\eta)/2)$ as a_1 approaches $-\infty$, and it approaches $\alpha\eta/2$ as a_1 approaches ∞. Since it is a continuous function of a_1, there must be a value of a_1 such that $\frac{\partial K}{\partial w_1}\big|_{(a_1,0)} = 0$. Putting this together with $\frac{\partial K}{\partial w_2}\big|_{(a_1,0)} < 0$ completes the proof. ∎

To show the sufficient inequalities (43), it will be useful to prove an upper bound on $w_2^{\circledast}$. (This upper bound will make it easier to show, informally, that $w_1^{\circledast}$ is needed.) In order to bound the size of $w_2^{\circledast}$, we will prove a lower bound on K in terms of w_2. For this, we want to show that, if w_2 is too large, then the algorithm will pay too much when it makes large-margin errors. For *this*, we need a lower bound on the probability of a large-margin error. For this, we can adapt an analysis that provided a lower bound on the probability of an error from (Helmbold and Long, 2012).

To simplify the proof, we will first provide a lower bound on the dropout risk in terms of the risk without dropout. We will actually prove something somewhat more general, for possible future reference.

Lemma 45 *Let* $\mathbf{r}$ *and* $\mathbf{x}$ *be independent,* $\mathbf{R}^N$*-valued random variables; let* ϕ *be convex function of a scalar real variable. Then*

$$\mathbf{E}_{\mathbf{r},\mathbf{x}}\left(\phi\left(\sum_i x_i r_i\right)\right) \geq \mathbf{E}_{\mathbf{x}}\left(\phi\left(\sum_i x_i \mathbf{E}_{\mathbf{r}}(r_i)\right)\right).$$

Proof: Since $\mathbf{x}$ and $\mathbf{r}$ are independent,

$$\begin{aligned}
&\mathbf{E}_{\mathbf{r},\mathbf{x}}(\phi(\sum_i x_i r_i)) \\
&= \mathbf{E}_{\mathbf{x}}(\mathbf{E}_{\mathbf{r}}(\phi(\sum_i x_i r_i))) \\
&\geq \mathbf{E}_{\mathbf{x}}(\phi(\mathbf{E}_{\mathbf{r}}(\sum_i x_i r_i))) \text{ (by Jensen's Inequality)} \\
&= \mathbf{E}_{\mathbf{x}}(\phi(\sum_i x_i \mathbf{E}_{\mathbf{r}}(r_i))),
\end{aligned}$$

completing the proof. ∎

Now, it is enough to lower bound the probability of a large-margin error with respect to the original distribution. Recall $B = \{2, \ldots, n\}$.

Lemma 46 $\mathbf{Pr}\left(\frac{1}{n-1}\sum_{i\in B} x_i < -2\beta\right) \geq \frac{3}{10}$.

Proof: If Z is a standard normal random variable and R is a binomial (ℓ, p) random variable with $p \leq 1/2$, then for $\ell(1-p) \leq j \leq \ell p$, Slud's inequality (Slud, 1977) gives

$$\mathbf{Pr}(R \geq j) \geq \mathbf{Pr}\left(Z \geq \frac{j - \ell p}{\sqrt{\ell p(1-p)}}\right), \tag{46}$$

as worked out in Lemma 23 of (Helmbold and Long, 2012).

Now, we have

$$\begin{aligned}
\mathbf{Pr}\left(\frac{1}{n-1}\sum_{i\in B} x_i < -2\beta\right) &= \mathbf{Pr}\left(\sum_{i\in B} x_i/2 < -(n-1)\beta\right) \\
&= \mathbf{Pr}\left(\sum_{i\in B}(x_i+1)/2 < (n-1)/2 - (n-1)\beta\right) \\
&= \mathbf{Pr}\left(\sum_{i\in B} z_i < (n-1)(1/2-\beta)\right)
\end{aligned}$$

where the z_i's are independent $\{0,1\}$-valued variables with $\mathbf{Pr}(z_i = 1) = 1/2+\beta$. Let $\bar{z}_i$ be $1 - z_i$, so $\sum_{i\in B} \bar{z}_i$ is a Binomial $(n-1, 1/2-\beta)$ random variable. Furthermore,

$$\begin{aligned}
\mathbf{Pr}\left(\sum_{i\in B} z_i < (n-1)(1/2-\beta)\right) &= \mathbf{Pr}\left(\sum_{i\in B} \bar{z}_i > (n-1) - (n-1)(1/2-\beta)\right) \\
&= \mathbf{Pr}\left(\sum_{i\in B} \bar{z}_i > (n-1)(1/2+\beta)\right).
\end{aligned}$$

Using (46) with $j = (n-1)(1/2+\beta)$, $\ell = (n-1)$, and $p = 1/2 - \beta$ gives:

$$\mathbf{Pr}\left(\sum_{i \in B} \bar{z}_i > (n-1)(1/2+\beta)\right) \geq \mathbf{Pr}\left(Z \geq \frac{(n-1)(1/2+\beta) - (n-1)(1/2-\beta)}{\sqrt{(n-1)(1/4-\beta^2)}}\right)$$
$$= \mathbf{Pr}\left(Z \geq \frac{2(n-1)\beta}{\sqrt{(n-1)(1/4-\beta^2)}}\right).$$

Since $\beta = 1/(10\sqrt{n})$ and $n \geq 100$, this implies

$$\mathbf{Pr}\left(\frac{1}{n-1}\sum_{i \in B} x_i < -2\beta\right) \geq \mathbf{Pr}\left(Z \geq 1/2\right).$$

Since the density of Z is always at most $1/\sqrt{2\pi}$, we have

$$\mathbf{Pr}\left(\frac{1}{n-1}\sum_{i \in B} x_i < -2\beta\right) \geq \mathbf{Pr}(Z \geq 0) - \mathbf{Pr}(Z \in (0, 1/2)) > \frac{1}{2} - \frac{1}{2\sqrt{2\pi}} > 3/10,$$

completing the proof. ∎

Now we are ready for the lower bound on the dropout risk in terms of w_2.

Lemma 47 *For all w_1,*

$$K(w_1, w_2) > \frac{w_2\sqrt{n-1}}{67}.$$

Proof: Considering only the case in which x_1 is dropped out (i.e. $r_1 = 0$), we have

$$K(w_1, w_2) \geq \frac{1}{2}\mathbf{E}\left(\ell\left(w_2 \sum_i r_i x_i\right)\right).$$

Applying Lemma 45, we get

$$K(w_1, w_2) \geq \frac{1}{2}\mathbf{E}\left(\ell\left((w_2/2)\sum_{i \in B} x_i\right)\right).$$

Since ℓ is non-increasing and non-negative, we have

$$K(w_1, w_2) \geq \frac{1}{2}\ell(-w_2\beta(n-1))\mathbf{Pr}\left(\frac{1}{n-1}\sum_{i \in B} x_i < -2\beta\right),$$

and applying Lemma 46 gives

$$K(w_1, w_2) \geq \frac{3\ell(-w_2\beta(n-1))}{20}.$$

Since $\ell(z) > -z$, we have

$$K(w_1, w_2) \geq \frac{3w_2\beta(n-1)}{20}$$

and, using $\beta = \frac{1}{10\sqrt{n-1}}$, we get

$$K(w_1, w_2) \geq \frac{3w_2\sqrt{n-1}}{200},$$

completing the proof. ∎

Lemma 48 $w_2^{\circledast} < \frac{27}{\sqrt{n-1}}$.

Proof: Note that

$$K(w, 0) = \ell(0)/2 + (1/2)(\eta\ell(-\alpha w) + (1-\eta)\ell(\alpha w)),$$

is increasing in η so that

$$K(w_1^{\circledast}, w_2^{\circledast}) \leq K(5/\alpha, 0) < \ell(0)/2 + 1/35 \tag{47}$$

since $\eta < 1/100$.

On the other hand, Lemma 47 gives

$$K(w_1^{\circledast}, w_2^{\circledast}) > \frac{w_2\sqrt{n-1}}{67}.$$

Solving for $w_2^{\circledast}$ completes the proof. ∎

Lemma 49 *For all* $0 < u < \frac{27}{\sqrt{n-1}}$, *we have*

$$\frac{\partial K}{\partial w_1}\Big|_{((n-1)u/\alpha, u)} < 0.$$

Proof: From (44), we have

$$\begin{aligned}
&2\frac{\partial K}{\partial w_1}\Big|_{(nu/\alpha, u)} \\
&= \mathbf{E}_{\mathbf{x}\sim D,\mathbf{r}}\left(\frac{-x_1}{1+\exp((n-1)ux_1/\alpha + u\sum_{i\in B} r_i x_i)}\right) \\
&= \eta\mathbf{E}_{\mathbf{x}\sim D,\mathbf{r}}\left(\frac{\alpha}{1+\exp(-(n-1)u + u\sum_{i\in B} r_i x_i)}\right) \\
&\qquad + (1-\eta)\mathbf{E}_{\mathbf{x}\sim D,\mathbf{r}}\left(\frac{-\alpha}{1+\exp((n-1)u + u\sum_{i\in B} r_i x_i)}\right) \\
&< \eta\alpha + (1-\eta)\mathbf{E}_{\mathbf{x}\sim D,\mathbf{r}}\left(\frac{-\alpha}{1+\exp((n-1)u + u\sum_{i\in B} r_i x_i)}\right) \\
&< \alpha\left(\eta + \frac{-(1-\eta)}{1+\exp(2(n-1)u)}\right) \quad (\text{since } \textstyle\sum_{i\in B} r_i x_i \leq n-1) \\
&< \alpha\left(\eta + \frac{-(1-\eta)}{1+\exp(54\sqrt{n-1}}\right) \quad (\text{since } u < 27/\sqrt{n-1}) \\
&< 0
\end{aligned}$$

since $\eta \leq 1/(2+\exp(54\sqrt{n}))$, completing the proof. ∎

Recall that, to prove Theorem 23, since we already showed $w_2^{\circledast} > 0$, all we needed was to show that $\alpha w_1^{\circledast} > (n-1)w_2^{\circledast}$. We do this next.

Lemma 50 $\alpha w_1^{\circledast} > (n-1) w_2^{\circledast}$.

Proof: Let $\mathbf{g}$ be the gradient of J evaluated at $\mathbf{u} = ((n-1)w_2^{\circledast}/\alpha, w_2^{\circledast})$. Lemmas 48 and 49 implies that $\mathbf{g} \neq (0,0)$. By convexity

$$\mathbf{w}^{\circledast} \cdot \mathbf{g} < \mathbf{u} \cdot \mathbf{g}$$

which, since $u_2 = w_2^{\circledast}$, implies

$$w_1^{\circledast} g_1 < (n-1) w_2^{\circledast} g_1 / \alpha.$$

Since, by Lemmas 48 and 49, $g_1 < 0$,

$$w_1^{\circledast} > (n-1) w_2^{\circledast} / \alpha$$

completing the proof. ∎

Appendix L. Proof of Theorem 24

> **Theorem 24.** If $\beta = 1/(10\sqrt{n-1})$, $\lambda = \frac{1}{30n}$, $\alpha < \beta\lambda$, and n is a large enough even number, then for any $\eta \in [0,1]$, $\mathrm{er}_{P_{10}}(\mathbf{v}(P_{10}, \lambda)) \geq 3/10$.

In this proof, let us also abbreviate P_{10} with P and use J to denote the L_2 regularized criterion in Equation (5) specialized for the distribution of this P.

As before, the contribution to the L_2 criterion from the cases where y is -1 and 1 respectively are the same, so the value of the criterion is not affected if we clamp y at 1. Furthermore, we leave the dependency on λ implicit and (since the source is fixed) use the more succinct $\mathbf{v}$ for $\mathbf{v}(P, \lambda)$.

Also, if, as before, we let $B = \{2, ..., n\}$, then by symmetry, v_i is identical for all $i \in B$ so $\mathbf{v}$ is the minimum of J over weight vectors satisfying this constraint. Let $K(w_1, w_2) = J(w_1, w_2, ..., w_2)$ so that (v_1, v_2) minimizes K. Recall that D is the marginal distribution of $\mathbf{x}$ under P conditioned on $y = 1$.

$$K(w_1, w_2) = \mathbf{E}_{\mathbf{x} \sim D}\left(\ell\left(w_1 x_1 + w_2 \sum_{i \in B} x_i\right)\right) + \frac{\lambda}{2}(w_1^2 + (n-1) w_2^2).$$

Lemma 46, together with the fact that $|x_1| = \alpha$, implies that,

$$\alpha v_1 < 2\beta(n-1) v_2 \tag{48}$$

suffices to prove Theorem 24, so we set this as our subtask.

We have

$$\frac{\partial K}{\partial w_1} = \mathbf{E}_{\mathbf{x} \sim D}\left(\frac{-x_1}{1 + \exp(w_1 x_1 + w_2 \sum_{i \in B} x_i)}\right) + \lambda w_1 \tag{49}$$

$$\frac{\partial K}{\partial w_2} = \mathbf{E}_{\mathbf{x} \sim D}\left(\frac{-\sum_{i \in B} x_i}{1 + \exp(w_1 x_1 + w_2 \sum_{i \in B} x_i)}\right) + \lambda (n-1) w_2. \tag{50}$$

First, we need a rough bound on v_1.

Lemma 51 $|v_1| \leq \frac{\alpha}{\lambda} < \beta$.

Proof: The second inequality follows from the constraint on α. From (49), we get

$$|v_1| \leq \frac{1}{\lambda} \mathbf{E}_{\mathbf{x} \sim D} \left(\left| \frac{x_1}{1 + \exp(v_1 x_1 + v_2 \sum_{i \in B} x_i)} \right| \right)$$

and the facts $|x_1| \leq \alpha$ and $0 < \frac{1}{1+\exp(v_1 x_1 + v_2 \sum_{i \in B} x_i)} \leq 1$ then imply $|v_1| \leq \alpha / \lambda$. ∎

Lemma 52 *For large enough n,*

$$\mathbf{Pr} \left(\sum_{i \in B} x_i \in [\beta(n-1), 3\beta(n-1)] \right) \geq \frac{1}{13}.$$

Proof: Let $\Phi(z) = \mathbf{Pr}(Z \leq z)$ for a standard normal random variable Z and let $S = \sum_{i \in B} x_i$. Note that $\mathbf{E}(x_i) = 2\beta$, $\mathbf{var}(x_i) = 1 - 4\beta^2$, and the third moment $\mathbf{E}(|x_i - \mathbf{E}(x_i)|^3) = 1 - 16\beta^4$. The Berry-Esseen inequality (DasGupta, 2008, Theorem 11.1) relates binomial distributions to the normal distribution using these moments, and directly implies that

$$\sup_z \left| \mathbf{Pr} \left(\frac{S}{n-1} - 2\beta \leq \sqrt{\frac{1-4\beta^2}{n-1}} \times z \right) - \Phi(z) \right| \leq \frac{C(1-16\beta^4)}{(1-4\beta^2)^{3/2}\sqrt{n-1}} < \frac{1}{\sqrt{n-1}}$$

where the last inequality follows from the facts that the Berry-Esseen global constant $C \leq 0.8$ and $\beta < 1/10$.

Using the change of variable $s = \sqrt{(1-4\beta^2)(n-1)}\, z + 2\beta(n-1)$ this can be restated:

$$\sup_s \left| \mathbf{Pr}(S \leq s) - \Phi \left(\frac{s - 2\beta(n-1)}{\sqrt{(1-4\beta^2)(n-1)}} \right) \right| \leq \frac{1}{\sqrt{n-1}},$$

so

$$\begin{aligned}
&\mathbf{Pr}(S \in [\beta(n-1), 3\beta(n-1)]) \\
&\geq \mathbf{Pr}_{z \in N(0,1)} \left(z \in \left[-\beta\sqrt{\frac{n-1}{1-4\beta^2}}, \beta\sqrt{\frac{n-1}{1-4\beta^2}} \right] \right) - \frac{2}{\sqrt{n-1}} \\
&\geq \mathbf{Pr}_{z \in N(0,1)} \left(z \in \left[\frac{-1}{10}, \frac{1}{10} \right] \right) - \frac{2}{\sqrt{n-1}} \\
&\geq \frac{1}{13},
\end{aligned}$$

for large enough n. ∎

Recent work shows that the Berry-Esseen constant C is less then $1/2$, this allows us to replace the $2\sqrt{n-1}$ with $1/\sqrt{n-1}$, but it still requires n on the order of 150,000 to get the 1/13 bound. Reducing the bound to 1/50 would make n as small as 300 sufficient.

Next, we need a rough bound on v_2.

Lemma 53 $v_2 \geq \frac{1}{n-1}$.

Proof: From (50), we have

$$v_2 = \frac{1}{\lambda(n-1)} \mathbf{E}_{\mathbf{x} \sim D} \left(\frac{\sum_{i \in B} x_i}{1 + \exp(v_1 x_1 + v_2 \sum_{i \in B} x_i)} \right).$$

If we denote $\sum_{i \in B} x_i$ by S, then

$$v_2 = \frac{1}{\lambda(n-1)} \mathbf{E}_{\mathbf{x} \sim D} \left(\frac{S}{1 + \exp(v_1 x_1 + v_2 S)} \right).$$

Since, for all odd[6] $s > 0$

$$\frac{\mathbf{Pr}(S = s)}{\mathbf{Pr}(S = -s)} = \left(\frac{1 + 2\beta}{1 - 2\beta} \right)^s$$

so $\mathbf{Pr}(S = -s) = \mathbf{Pr}(S = s) \left(\frac{1-2\beta}{1+2\beta} \right)^s$. Analyzing the contributions of s and $-s$ together we have

$$\begin{aligned} & v_2 \lambda(n-1) \\ & = \sum_{s=1}^{n-1} \mathbf{Pr}(S = s) \Big((1-\eta) \frac{s}{1 + \exp(v_1 \alpha + v_2 s)} + \eta \frac{s}{1 + \exp(-v_1 \alpha + v_2 s)} \\ & + \left((1-\eta) \frac{-s}{1 + \exp(v_1 \alpha - v_2 s)} + \eta \frac{-s}{1 + \exp(-v_1 \alpha - v_2 s)} \right) \left(\frac{1 - 2\beta}{1 + 2\beta} \right)^s \Big). \end{aligned}$$

Recalling that $|v_1| \leq \alpha/\lambda$ (Lemma 51), and using the minimizing value in this range for each term gives

$$\begin{aligned} & v_2 \lambda(n-1) \\ & \geq \sum_{s=1}^{n-1} \mathbf{Pr}(S = s) \left(\frac{s}{1 + \exp(\alpha^2/\lambda + v_2 s)} + \left(\frac{-s}{1 + \exp(-\alpha^2/\lambda - v_2 s)} \right) \left(\frac{1 - 2\beta}{1 + 2\beta} \right)^s \right) \\ & = \sum_{s=1}^{n-1} \mathbf{Pr}(S = s) s \left(\frac{1 - \exp(\alpha^2/\lambda + v_2 s) \left(\frac{1-2\beta}{1+2\beta} \right)^s}{1 + \exp(\alpha^2/\lambda + v_2 s)} \right) \\ & \geq \sum_{s=1}^{n-1} \mathbf{Pr}(S = s) s \left(\frac{1 - \exp(\alpha^2/\lambda + v_2 s - 4\beta s)}{1 + \exp(\alpha^2/\lambda + v_2 s)} \right). \end{aligned}$$

[6] S is the sum of an odd number of ±1's, and thus cannot be even.

Assume for contradiction that $v_2 < 1/(n-1)$. Then,

$$\begin{aligned}
&v_2 \lambda(n-1) \\
&\geq \sum_{s=1}^{n-1} \mathbf{Pr}(S=s) s \left(\frac{1-\exp(\alpha^2/\lambda + s/(n-1) - 4\beta s)}{1+\exp(\alpha^2/\lambda + s/(n-1))} \right) \\
&\geq \sum_{s=1}^{n-1} \mathbf{Pr}(S=s) s \left(\frac{1-\exp(s/(n-1) - 3\beta s)}{1+\exp(\beta^2 \lambda + s/(n-1))} \right) \quad \text{(since } \alpha \leq \beta\lambda\text{)} \\
&\geq \sum_{s=1}^{n-1} \mathbf{Pr}(S=s) s \left(\frac{1-\exp(-2\beta s)}{1+\exp(\beta^2 \lambda + s/(n-1))} \right) \quad \text{(for large enough } n\text{)} \\
&\geq \sum_{s \in [\beta(n-1), 3\beta(n-1)]} \mathbf{Pr}(S=s) s \left(\frac{1-\exp(-2\beta s)}{1+\exp(\beta^2 \lambda + s/(n-1))} \right),
\end{aligned}$$

since each term is positive. Taking the worst-case among $[\beta(n-1), 3\beta(n-1)]$ for each instance of s, and applying Lemma 52, we get

$$\begin{aligned}
v_2 &\geq \frac{1}{\lambda(n-1)} \times \frac{1}{13} \times \beta(n-1) \left(\frac{1-\exp(-2\beta^2(n-1))}{1+\exp(\beta^2\lambda + 3\beta)} \right) \\
&= \frac{30\sqrt{n-1}}{130} \left(\frac{1-\exp(-1/50)}{1+\exp(3/(10\sqrt{n-1}) + 1/(3000n(n-1)))} \right). \qquad (51)
\end{aligned}$$

Thus $v_2 = \Omega(\sqrt{n-1})$, which, for large enough n, contradicts our assumption that $v_2 < 1/(n-1)$, completing the proof. ∎

Not that even with the many approximations made, Inequality (51) gives the desired contradiction at $n = 60$. Even when the weaker bound of $1/50$ discussed following Lemma 52 is used, $n = 145$ still suffices to give the desired contradiction.

Now we're ready to put everything together.

Proof (of Theorem 24): Recall that, by Lemma 46, if $v_1 < 2\beta(n-1)v_2$, then

$$\mathrm{er}_P(\mathbf{v}(P,\lambda)) \geq 3/10.$$

Lemma 51 gives $v_1 < \beta$. Lemma 53 implies $(n-1)v_2 \geq 1$. Therefore $v_1 < \beta(n-1)v_2$, completing the proof. ∎

Using the $1/50$ version of Lemma 52 leads to a proof of the theorem for all even $n \geq 300$.

References

J. Abernethy, C. Lee, A. Sinha, and A. Tewari. Online Linear Optimization Via Smoothing. *COLT*, pages 807–823, 2014.

P. Bachman, O. Alsharif, and D. Precup. Learning with Pseudo-ensembles. *NIPS*, 2014.

P. Baldi and P. Sadowski. The Dropout Learning Algorithm. *Artificial intelligence*, 210: 78–122, 2014.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

L. Breiman. Some Infinity Theory for Predictor Ensembles. *Annals of Statistics*, 32(1): 1–11, 2004.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. A Second-order Perceptron Algorithm. *COLT*, 2002.

G. E. Dahl. Deep Learning How I Did It: Merck 1st Place Interview, 2012. http://blog.kaggle.com.

G. E. Dahl, T. N. Sainath, and G. E. Hinton. Improving Deep Neural Networks for LVCSR Using Rectified Linear Units and Dropout. *ICASSP*, 2013.

A. DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer, 2008.

L. Deng, J. Li, J. Huang, K. Yao, D. Yu, F. Seide, M. L. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero. Recent Advances in Deep Learning for Speech Research at Microsoft. *ICASSP*, 2013.

J. Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *JMLR*, 12:2121–2159, 2011.

R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, 1989.

D. P. Helmbold and P. M. Long. On the Necessity of Irrelevant Variables. *JMLR*, 13: 2145–2170, 2012.

G. E. Hinton. Dropout: a Simple and Effective Way to Improve Neural Networks, 2012. videolectures.net.

G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving Neural Networks by Preventing Co-adaptation of Feature Detectors, 2012. Arxiv, arXiv:1207.0580v1.

H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer, 1997.

P. M. Long and R. A. Servedio. Random Classification Noise Defeats All Convex Potential Boosters. *Machine Learning*, 78(3):287–304, 2010.

E. Slud. Distribution Inequalities for the Binomial Law. *Annals of Probability*, 5:404–412, 1977.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks From Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

T. Van Erven, W. Kotowski, and M. K. Warmuth. Follow the Leader with Dropout Perturbations. *Annual ACM Workshop on Computational Learning Theory*, pages 949–974, 2014.

S. Wager, S. Wang, and P. Liang. Dropout Training as Adaptive Regularization. *NIPS*, 2013.

S. Wager, W. Fithian, S. Wang, and P. S. Liang. Altitude Training: Strong Bounds for Single-layer Dropout. *NIPS*, 2014.

L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus. Regularization of Neural Networks using DropConnect. In *ICML*, pages 1058–1066, 2013.

S. Wang and C. Manning. Fast Dropout Training. In *ICML*, pages 118–126, 2013.

T. Zhang. Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization. *Annals of Statistics*, 32(1):56–85, 2004.