Dropout as a Low-Rank Regularizer for Matrix Factorization

Jacopo Cavazza[1,⋆], Benjamin D. Haeffele[2,⋆] Connor Lane[2], Pietro Morerio[1], Vittorio Murino[1], René Vidal[2]

[1] Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Genova, 16163, Italy

[2] Center for Imaging Science, Johns Hopkins University, Baltimore, MD 21218, USA

⋆ *The two first authors contributed equally to the work.*

Abstract

Dropout is a simple yet effective regularization technique that has been applied to various machine learning tasks, including linear classification, matrix factorization (MF) and deep learning. However, despite its solid empirical performance, the theoretical properties of dropout as a regularizer remain quite elusive. In this paper, we present a theoretical analysis of dropout for MF, where Bernoulli random variables are used to drop columns of the factors. We demonstrate the equivalence between dropout and a fully deterministic model for MF in which the factors are regularized by the sum of the product of squared Euclidean norms of the columns. Additionally, we investigate the case of a variable sized factorization and we prove that dropout is equivalent to a convex approximation problem with (squared) nuclear norm regularization. As a consequence, we conclude that dropout induces a low-rank regularizer that results in a data dependent singular-value thresholding.

1 INTRODUCTION

In many problems in machine learning and artificial intelligence, relevant patterns and information often lie in a low-dimensional manifold. In order to capture this structure, linear subspaces have become very popular, arguably due to their efficiency and versatility.

The problem of learning a linear subspace from data points $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n$ is usually formulated as follows. Let $\mathbf{X}$ be the $m \times n$ matrix containing the data points as its rows. In the *matrix approximation* formulation of subspace learning, the goal is to find an $m \times n$ matrix $\mathbf{A}$ that is close to $\mathbf{X}$ and satisfies certain properties (e.g., being low rank). This problem can be formulated as

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{A}\|_F^2 + \gamma\, \Xi(\mathbf{A}), \tag{1}$$

where the Frobenius norm $\|\cdot\|_F$ measures the approximation error between $\mathbf{X}$ and $\mathbf{A}$, the regularization function $\Xi(\mathbf{A})$ encourages the desired properties on $\mathbf{A}$, and $\gamma > 0$ is a trade-off parameter. An important property of the formulation in (1) is that the optimization problem is convex when Ξ is convex, which greatly facilitates finding a global minimum. At the same time, the formulation in (1) may not be adequate for large scale problems, as it requires solving for $m \cdot n$ variables.

In the *matrix factorization* (MF) formulation of subspace learning the goal is to find two matrices (or factors) $\mathbf{U} \in \mathbb{R}^{m \times d}$ and $\mathbf{V} \in \mathbb{R}^{n \times d}$, where d is the dimension of the subspace, such that $\mathbf{X} \approx \mathbf{U}\mathbf{V}^\top$. This problem can be formulated as

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|_F^2 + \lambda \Omega(\mathbf{U}, \mathbf{V}), \tag{2}$$

where $\|\cdot\|_F$ measures the approximation error between $\mathbf{X}$ and $\mathbf{U}\mathbf{V}^\top$, the regularization function $\Omega(\mathbf{U}, \mathbf{V})$ encourages some desired properties on the factors (e.g., orthonormality, sparsity, etc.), and $\lambda > 0$ is a trade-off parameter. The formulation in (2) has two main advantages. First, the optimization is carried out directly on the factors achieving a structured decomposition of $\mathbf{X}$ that depends on the choice of the regularizer Ω. In contrast, the optimal solution $\mathbf{A}^{\text{opt}}$ of (1) may not have such structure. Second, the number of variables to be optimized in (2) scales linearly with respect to $m + n$, ensuring the applicability of MF even in the big data regime. Unfortunately, an important shortcoming of (2) is that, while the problem of optimizing for $\mathbf{V}$ when $\mathbf{U}$ is fixed is convex when Ω is convex in $\mathbf{V}$ for a fixed $\mathbf{U}$, and vice versa, the problem in (2) is not convex when optimizing on $\mathbf{U}$ and $\mathbf{V}$ jointly. As a result, many

Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

challenges arise both in verifying whether a given local solution is globally optimal and whether algorithms are guaranteed to find globally optimal solutions.

These challenges have motivated a rich literature on the connections between matrix approximation (1) and matrix factorization (2) [21, 10, 4, 3, 17, 15, 18, 16]. For example, [16] derives condition under which a local minimum for (2) gives a global minimum for (1) and (2). In this paper, we further strengthen these connections by providing a theoretical analysis for MF with a particular type of regularizer called *dropout* [22, 32].

Dropout is a popular algorithm for training neural networks that, at each training iteration, sets to zero the outputs of a fraction of the neural units and updates only the weights of the remaining units. Specifically, during dropout training each neural unit is associated with a Bernoulli random variable that specifies whether the unit is retained or suppressed. The expected value θ of the random variables is called the "retain probability". At each iteration of dropout training, a new training example/minibatch and a new set of Bernoulli random variables are drawn, the outputs of the suppressed units are set to zero, and the network weights corresponding to retained units are updated using a back-propagation step, while the remaining weights remain unchanged. During inference, no unit suppression is performed and all weights are rescaled by θ. The latter stage can be interpreted as a model average up to certain approximations [32, 5, 6].

Motivated by the significant efforts made to understand dropout as an implicit regularizer [36, 5, 6, 14], in this paper, following [39, 19], we study the theoretical properties of dropout for MF. Specifically, we study the following problem: given a data matrix $\mathbf{X}$, we search for a factorization $\mathbf{X} \approx \mathbf{U}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{m\times d}$ and $\mathbf{V}^{n\times d}$, that solves the following optimization problem

$$\min_{\mathbf{U},\mathbf{V}} \mathbb{E}_{\mathbf{r}} \left\|\mathbf{X} - \tfrac{1}{\theta}\mathbf{U}\mathrm{diag}(\mathbf{r})\mathbf{V}^\top\right\|_F^2. \tag{3}$$

where $\mathbf{r} \in \mathbb{R}^d$ is a random vector whose entries are i.i.d. Bernoulli(θ) and $\mathbb{E}_{\mathbf{r}}$ denotes the expected value with respect to $\mathbf{r}$. The dropout formulation of MF in (3) takes direct inspiration from the idea of suppressing "hidden units" in a neural network, and in the linear case of (3) we suppress "columns" of the factorization. While in practice dropout for MF has shown solid performance [39, 19], it is still unclear what sort of regularization it induces.

The contributions of this paper are the following:

1. We show that the standard dropout algorithm is a stochastic gradient descent method for solving (3).
2. We show that dropout for MF (3) is equivalent to the following deterministic regularized MF problem

$$\min_{\mathbf{U},\mathbf{V}} \left[\|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|_F^2 + \tfrac{1-\theta}{\theta}\Omega_{\text{dropout}}(\mathbf{U},\mathbf{V})\right], \tag{4}$$

   where $\Omega_{\text{dropout}}(\mathbf{U},\mathbf{V}) = \sum_{k=1}^d \|\mathbf{u}_k\|_2^2\|\mathbf{v}_k\|_2^2$.
3. We show that if in the optimization problem in (4) we also minimize with respect to the size d of $\mathbf{U}$ and $\mathbf{V}$, while keeping the retain probability θ constant, then Ω_{dropout} promotes over-sized factorizations, that is, the larger d, the smaller the objective value.
4. We show that if the dropout rate is chosen as a particular increasing function of d, then Ω_{dropout} acts as a low-rank regularization strategy. Specifically, we show that the optimization problem in (3) is related to the following matrix approximation problem

$$\min_{\mathbf{A}} \left[\|\mathbf{X} - \mathbf{A}\|_F^2 + \gamma\|\mathbf{A}\|_\star^2\right], \tag{5}$$

   where the squared nuclear norm $\|\mathbf{A}\|_\star^2$ is used to induce low-rank factorizations.
5. We show that if $\mathbf{U}^{\text{opt}}$ and $\mathbf{V}^{\text{opt}}$ are globally optimal factors of (3) using a dropout rate that depends on d with the size of d learned via the optimization, then $\mathbf{A}^{\text{opt}} = (\mathbf{U}^{\text{opt}})(\mathbf{V}^{\text{opt}})^\top$ is a global optimum of (5).

Paper outline. Section 2 briefly reviews the literature related to dropout. Sections 3, 4 and 5 present our theoretical analysis. Section 6 presents numerical simulations and Section 7 provides concluding remarks.

2 RELATED WORK

The origins of dropout can be traced back to the literature on learning representations from input data corrupted by noise [9, 8, 30]. Since its original formulation [22, 32], many algorithmic variations have been proposed [25, 7, 37, 24, 29, 1, 26]. Further, the empirical success of dropout for neural network training has motivated several works investigating its formal properties from a theoretical point of view. Wager et al. [36] analyze dropout applied to the logistic loss for generalized linear models. Hembold and Long [20] discuss mathematical properties of the dropout regularizer (such as non-monotonicity and non-convexity) and derive a sufficient condition to guarantee a unique minimizer for the dropout criterion. Baldi and Sadowski [5, 6] consider dropout applied to deep neural networks with sigmoid activations and prove that the weighted geometric mean of all of the sub-networks associated with the retained units at each iteration can be computed with a single forward pass. Wager et al. [35] investigate the impact of dropout on the

generalization error in terms of the bias-variance tradeoff. Gal and Ghahramani [14] study the connections between dropout training and inference for deep Gaussian processes. Many of these prior theoretical results required simplifying assumptions, and thus the results hold only in an approximate sense [36, 20, 5, 6, 14]. In contrast, our work characterizes the regularizer induced by dropout for MF in an analytical manner.

In the context of MF, only a few works have investigated the dropout criterion. He et al. [19] leverage the formal analogy between MF and shallow neural networks, which inspires the use of dropout as a regularizer. Zhai and Zhang [39] provide some theoretical analysis for dropout for MF, but only as an argument to unify MF and encoder-decoder architectures. None of these works study the formal properties of dropout as a regularizer and our paper aims at filling this gap.

3 DROPOUT FOR MATRIX FACTORIZATION

Given an $m \times n$ matrix $\mathbf{X}$, we consider the problem of approximating $\mathbf{X}$ as the product $\mathbf{U}\mathbf{V}^\top$, where $\mathbf{U}$ is $m \times d$ and $\mathbf{V}$ is $n \times d$, and the size d of the factors is assumed fixed throughout this section. Following (3), we formulate this MF problem as one of minimizing

$$f(\mathbf{U}, \mathbf{V}) = \mathbb{E}_{\mathbf{r}} \left\| \mathbf{X} - \tfrac{1}{\theta}\mathbf{U}\mathrm{diag}(\mathbf{r})\mathbf{V}^\top \right\|_F^2, \qquad (6)$$

where $\mathbf{r} = [r_1, \ldots, r_d]$ is a random vector whose elements are assumed to be i.i.d. as $r_i \sim \text{Bernoulli}(\theta)$. We assume $0 < \theta < 1$ to avoid trivial or degenerate cases.[1]

To see why the minimization of f can be achieved using a strategy akin to dropout for neural network training, observe that if we use a gradient descent strategy to minimize f, the gradient of the expected value is equal to the expected value of the gradient, i.e.

$$\nabla f(\mathbf{U}, \mathbf{V}) = \mathbb{E}_{\mathbf{r}} \nabla \left\| \mathbf{X} - \tfrac{1}{\theta}\mathbf{U}\mathrm{diag}(\mathbf{r})\mathbf{V}^\top \right\|_F^2. \qquad (7)$$

Therefore, the evaluation of $\nabla \left\| \mathbf{X} - \frac{1}{\theta}\mathbf{U}\mathrm{diag}(\mathbf{r})\mathbf{V}^\top \right\|_F^2$ at a particular sample of $\mathbf{r}$ provides an unbiased estimate of ∇f. This leads to the stochastic gradient descent (SGD) scheme shown in Algorithm 1, where the expected gradient at each iteration is replaced by the gradient for a fixed sample $\mathbf{r}$. Notice that columns of $\mathbf{U}$ and $\mathbf{V}$ for which $r_i = 0$ are suppressed when evaluating the matrix product. Notice also that columns of $\mathbf{U}$ and $\mathbf{V}$ for which $r_i = 1$ are updated according to the SGD rule, while columns of $\mathbf{U}$ and $\mathbf{V}$ for which $r_i = 0$ are not updated. This is analogous to dropout training for a neural network with a single hidden layer.

[1] Note that this assumption is less restrictive than currently adopted practices for dropout training where $\theta > 0.5$ (see [32, Appendix A.4] for a list of typical values).

Algorithm 1: Dropout training for MF with fixed d

for $t = 1, 2, \ldots$ **do**
- Sample $\mathbf{r}^t$ element-wise from a Bernoulli(θ).
- Compute gradient directions

$$\begin{bmatrix} \partial \mathbf{U}^t \\ \partial \mathbf{V}^t \end{bmatrix} = \begin{bmatrix} (\frac{1}{\theta}\mathbf{U}^t \mathrm{diag}(\mathbf{r}^t)\mathbf{V}^{t\top} - \mathbf{X})\mathbf{V}^t \\ (\frac{1}{\theta}\mathbf{U}^t \mathrm{diag}(\mathbf{r}^t)\mathbf{V}^{t\top} - \mathbf{X})^\top \mathbf{U}^t \end{bmatrix} \qquad (8)$$

  with respect to $\mathbf{U}$ and $\mathbf{V}$, respectively.
- Update the factors

$$\begin{bmatrix} \mathbf{U}^{t+1} \\ \mathbf{V}^{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{U}^t \\ \mathbf{V}^t \end{bmatrix} - \frac{2\epsilon_t}{\theta} \begin{bmatrix} \partial \mathbf{U}^t \\ \partial \mathbf{V}^t \end{bmatrix} \mathrm{diag}(\mathbf{r}^t), \qquad (9)$$

end

The discussion so far shows that Algorithm 1 is an SGD scheme applied to the MF problem in (3) that is akin to the dropout algorithm proposed in [22, 32]. Since during training Algorithm 1 drops columns of the factorization, it is natural to ask whether this induces some redundancy or low-rank regularization on the solution, or perhaps it induces some ensemble averaging as in the case of neural networks.

In what follows, we seek to study the theoretical properties of dropout for MF, with the goal of understanding what type of regularization it induces. The following theorem shows that the optimization problem solved by dropout (3) is equivalent to the deterministic MF problem in (2) with $\lambda = \lambda_{\text{dropout}} \doteq \frac{1-\theta}{\theta}$ and the regularizer Ω chosen as the sum of the product of the squared Euclidean norms of the columns of $\mathbf{U}$ and $\mathbf{V}$.

Theorem 1. *For any θ, $\mathbf{U}$,$\mathbf{V}$ and $\mathbf{X}$, we have*

$$\mathbb{E}_{\mathbf{r}} \left\| \mathbf{X} - \frac{1}{\theta}\mathbf{U}\mathrm{diag}(\mathbf{r})\mathbf{V}^\top \right\|_F^2 \qquad (10)$$
$$= \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|_F^2 + \frac{1-\theta}{\theta} \underbrace{\sum_{k=1}^{d} \|\mathbf{u}_k\|_2^2 \|\mathbf{v}_k\|_2^2}_{\Omega_{dropout}(\mathbf{U},\mathbf{V})}.$$

Proof. The well known equality $\mathbb{E}(a^2) = \mathbb{E}(a)^2 + \mathbb{V}(a)$ for a scalar random variable a can be extended to matrices as $\mathbb{E}(\|\mathbf{A}\|_F^2) = \|\mathbb{E}(\mathbf{A})\|_F^2 + \mathbf{1}^\top \mathbb{V}(\mathbf{A})\mathbf{1}$ provided the entries of $\mathbf{A}$ are independent. Applying this to $\mathbf{A} = \mathbf{X} - \frac{1}{\theta}\mathbf{U}\mathrm{diag}(\mathbf{r})\mathbf{V}^\top$ and noticing that $\mathbb{E}(\mathrm{diag}(\mathbf{r})) = \theta I$, we obtain $\mathbb{E}(\mathbf{A}) = \mathbf{X} - \mathbf{U}\mathbf{V}^\top$. Since $\mathbb{V}(\mathbf{A}) = \frac{1}{\theta^2}\mathbb{V}(\mathbf{U}\mathrm{diag}(\mathbf{r})\mathbf{V}^\top)$ and $\mathbf{U}\mathrm{diag}(\mathbf{r})\mathbf{V}^\top = \sum_k \mathbf{u}_k \mathbf{v}_k^\top r_k$, we have $\theta^2 \mathbf{1}^\top \mathbb{V}(A)\mathbf{1} = \sum_{ijk} \mathbb{V}(u_{ik} v_{jk} r_k) = \sum_{ijk} u_{ik}^2 v_{jk}^2 \mathbb{V}(r_k) = \theta(1-\theta)\sum_k \|\mathbf{u}_k\|_2^2 \|\mathbf{v}_k\|_2^2$ because the r_k's are independent. □

The implications of Theorem 1 are clear: For $\lambda = \lambda_{\text{dropout}}$ and $\Omega = \Omega_{\text{dropout}}$ problems (2) and (3) are equivalent, hence both problems have the same global minima. At the same time, a key challenge is that the optimization problem is non-convex, and hence there

is no guarantee that dropout will converge to a local, let alone a global minimizer. Recent work has established guarantees of global optimality for MF problems regularized by the sum of the product of the norms of the columns of the factors [27, 17, 15, 16]. However, such conditions do not apply to the regularizer Ω_{dropout} which involves the square of the norms. While the use of squared norms may appear like a very subtle modification, it has an important implication on whether it induces low-rank factorizations or not when the size of the factorization is allowed to vary, as discussed in the next section.

4 CONNECTIONS BETWEEN DROPOUT AND NUCLEAR NORM MINIMIZATION

In this section, we study properties of the dropout regularizer Ω_{dropout} for MF in the case where the size d of the factorization is allowed to vary and is controlled by the Ω_{dropout} regularizer. In particular, we show that when the retain probability is constant, the regularizer Ω_{dropout} promotes over-sized factorizations.

To understand the motivation for allowing the size d to be controlled by the regularization, let us first recall that the nuclear norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, also termed the trace norm or Schatten-Von Neumann 1-norm, is defined as the sum of its singular values $\|\mathbf{A}\|_\star = \sum_{i=1}^{\min(m,n)} \sigma_i(\mathbf{A})$. The nuclear norm is a popular regularizer for many many machine learning problems [38, 2, 12, 11, 23, 13, 27, 33], especially to induce low-rank structure. The connection between $\|\cdot\|_\star$ and Ω_{dropout} becomes clearer when considering the following variational form of the nuclear norm [31, 28], where d is one of the optimization variables:

$$\|\mathbf{A}\|_\star = \inf_{\substack{d,\mathbf{U}\in\mathbb{R}^{m\times d},\mathbf{V}\in\mathbb{R}^{n\times d}\\ \mathbf{A}=\mathbf{U}\mathbf{V}^\top}} \quad \sum_{k=1}^{d} \|\mathbf{u}_k\|_2\|\mathbf{v}_k\|_2. \tag{11}$$

This fact is used in [4, 3, 17, 16] to show that the convex optimization problem $\min_{\mathbf{A}} \|\mathbf{X}-\mathbf{A}\|_F^2 + \lambda\|\mathbf{A}\|_\star$ is equivalent to the non-convex optimization problem

$$\min_{d,\mathbf{U}\in\mathbb{R}^{m\times d},\mathbf{V}\in\mathbb{R}^{n\times d}} \quad \|\mathbf{X}-\mathbf{U}\mathbf{V}^\top\|_F^2 + \lambda\sum_{k=1}^{d}\|\mathbf{u}_k\|_2\|\mathbf{v}_k\|_2 \tag{12}$$

in the sense that if $(\mathbf{U},\mathbf{V})$ is a local minimizer of the non-convex problem such that for some k we have $\mathbf{u}_k = \mathbf{0}$ and $\mathbf{v}_k = \mathbf{0}$, then $(\mathbf{U},\mathbf{V})$ is a global minimizer of the non-convex problem and $\mathbf{A} = \mathbf{U}\mathbf{V}^\top$ is a global minimizer of the convex problem.

But what does the variational form of the nuclear norm tell us about the regularizer Ω_{dropout} induced by dropout? Notice the extreme similarity between the functional optimized in (11) and Ω_{dropout}: the only difference is that the Euclidean norms of the columns of $\mathbf{U}$ and $\mathbf{V}$ are squared in Ω_{dropout}. Naively, one could argue that such difference is extremely subtle and interpret dropout as a low-rank regularizer for MF. However, this is <u>not</u> the case, as we show next.

As an example, suppose we are given an arbitrary factorization of $\mathbf{A} = \mathbf{U}\mathbf{V}^\top$ of size d. Then, we can construct a new factorization of $\mathbf{A}$ of size $2d$ which reduces the dropout regularizer by a factor of two. Specifically,

$$\Omega_{\text{dropout}}\left(\tfrac{\sqrt{2}}{2}[\mathbf{U},\mathbf{U}], \tfrac{\sqrt{2}}{2}[\mathbf{V},\mathbf{V}]\right) = \frac{1}{2}\Omega_{\text{dropout}}(\mathbf{U},\mathbf{V}). \tag{13}$$

This shows that the regularizer Ω_{dropout} does not penalize the size of the factorization. On the contrary, it encourages factorizations with a large number of columns, as we can always reduce the value of Ω_{dropout} by increasing the number of columns. This provides the main argument to prove the following proposition.

Proposition 1. *The infimum of the dropout regularizer for a variable size factorization is equal to zero, i.e.,*

$$0 = \inf_{\substack{d,\mathbf{U}\in\mathbb{R}^{m\times d},\mathbf{V}\in\mathbb{R}^{n\times d}\\ \mathbf{A}=\mathbf{U}\mathbf{V}^\top}} \quad \sum_{k=1}^{d}\|\mathbf{u}_k\|_2^2\|\mathbf{v}_k\|_2^2. \tag{14}$$

As a consequence, in the context of MF with variable size d, using dropout with a constant retain probability θ does nothing to limit the size of the factorization. This is because the optimization problem solved by dropout (3) is equivalent to the regularized factorization problem (4), which is always reduced in value by increasing the number of columns in $(\mathbf{U},\mathbf{V})$. To address this issue, in the next section we study dropout using a retain probability $\theta(d)$ that decreases with d, and show that in this case $\frac{1-\theta}{\theta}\Omega_{\text{dropout}}$ induces low-rank regularization via the square of the nuclear norm $\|\mathbf{A}\|_\star^2$.

Remark 1. *The drawback of dropout regularization discussed in Proposition 1 does not contradict the excellent performance that dropout shows in practice [39, 19]. This is because the drawback occurs in the context of MF with variable d, while prior work has used dropout in the context of a fixed d. Elucidating the properties of $\Omega_{dropout}$ for a fixed d remains an open problem.*

Remark 2. *While in this section we discussed the properties of dropout for MF with variable d, we did not discuss how to modify dropout to also find the optimal d. This is because this paper focuses on the theoretical connections between different formulations for MF, not on algorithms. That being said, in our experiments we draw inspiration from the meta-algorithm for MF presented in [16], where the d is increased until a globally*

optimal factorization is found. This requires, however, a method for checking that the optimal d has been found, which is briefly discussed in the next section.

5 VARIABLE DROPOUT RATE FOR LOW-RANK REGULARIZATION

In this section, we establish a connection between the matrix approximation problem in (1) and dropout for MF, which, as explained in the previous section, can be formulated either as (3) or as its fully deterministic counterpart (2) with $\Omega = \Omega_{\text{dropout}}$. To make this connection, we explore the question of whether there exists a way to choose the retain probability θ in such a way that the regularization $\frac{1-\theta}{\theta}\Omega_{\text{dropout}}$ limits the size of the factorization, d, as in nuclear norm regularization, and hence avoid the pathological situation described in Proposition 1 that promotes over-sized factorizations.

Specifically, for a given p, $0 < p < 1$, we define

$$\theta(d) = \frac{p}{d - (d-1)p}, \tag{15}$$

where d refers to the number of columns in $\mathbf{U}$ and $\mathbf{V}$ in the factorization $\mathbf{A} = \mathbf{U}\mathbf{V}^\top$. Here, we have modified the retain probability θ to be a function of d, while also depending on a hyper-parameter $p \in (0, 1)$. With this choice for θ, the parameter $\lambda = \frac{1-\theta}{\theta}$ becomes $\lambda = d\frac{1-p}{p}$. In other words, the regularization weight λ increases linearly with the size of the factorization. It is easily shown that this choice corrects the pathology in (14).

Beyond this basic result, we can also show that a variable dropout rate induces low-rank regularization. Specifically, we have the following result.

Proposition 2. *For $\theta = \theta(d)$ as defined in* (15)*, then the lower convex envelope of*

$$\Lambda(\mathbf{A}) = \inf_{\substack{d, \mathbf{U}\in\mathbb{R}^{m\times d}, \mathbf{V}\in\mathbb{R}^{n\times d} \\ \mathbf{A}=\mathbf{U}\mathbf{V}^\top}} \frac{1-\theta(d)}{2\theta(d)}\Omega_{dropout}(\mathbf{U}, \mathbf{V}) \tag{16}$$

is given by $\frac{1-p}{2p}\|\mathbf{A}\|_\star^2$.

Proof. Recall that the convex envelope of a function f is the largest closed, convex function g such that $g(x) \leq f(x)$ for all x and is given by $g = (f^*)^*$, where f^* denotes the Fenchel dual of f, defined as $f^*(q) \equiv \sup_x \langle q, x\rangle - f(x)$. Note that $\Lambda(\mathbf{A})$ can be equivalently written as

$$\Lambda(\mathbf{A}) = \inf_{\substack{d\geq\rho(\mathbf{A}) \\ \mathbf{U}\in\mathbb{R}^{m\times d} \\ \mathbf{V}\in\mathbb{R}^{n\times d} \\ \mathbf{w}\in\mathbb{R}^d}} \frac{\lambda_d}{2}\|\mathbf{w}\|_2^2 \;\text{ s.t. } \sum_{k=1}^d w_k \mathbf{u}_k\mathbf{v}_k^T = \mathbf{A} \;\; (\|\mathbf{u}_k\|_2, \|\mathbf{v}_k\|_2) \leq (1,1) \;\; \forall k. \tag{17}$$

where $\lambda_d = d(1-p)/p$. This gives the Fenchel dual

$$\Lambda^*(\mathbf{Q}) = \sup_{\substack{d,\ \mathbf{w}\in\mathbb{R}^d \\ \mathbf{U}\in\mathbb{R}^{m\times d} \\ \mathbf{V}\in\mathbb{R}^{n\times d}}} \sum_{k=1}^d w_k \left\langle \mathbf{Q}, \mathbf{u}_k\mathbf{v}_k^T \right\rangle - \frac{\lambda_d}{2}\|\mathbf{w}\|_2^2 \quad \text{s.t. } (\|\mathbf{u}_k\|_2, \|\mathbf{v}_k\|_2) \leq (1,1) \;\; \forall k. \tag{18}$$

If we define the vector $\mathbf{B}_d(\mathbf{U}, \mathbf{V}) \in \mathbb{R}^d$ as

$$\mathbf{B}_d(\mathbf{U}, \mathbf{V}) = \begin{bmatrix} \left\langle \mathbf{Q}, \mathbf{u}_1\mathbf{v}_1^T \right\rangle \\ \dots \\ \left\langle \mathbf{Q}, \mathbf{u}_d\mathbf{v}_d^T \right\rangle \end{bmatrix}, \tag{19}$$

then from (18) we have that

$$\Lambda^*(\mathbf{Q}) = \sup_{\substack{d,\ \mathbf{w}\in\mathbb{R}^d \\ \mathbf{U}\in\mathbb{R}^{m\times d} \\ \mathbf{V}\in\mathbb{R}^{n\times d}}} \langle \mathbf{B}_d(\mathbf{U}, \mathbf{V}), \mathbf{w}\rangle - \frac{\lambda_d}{2}\|\mathbf{w}\|_2^2 \tag{20}$$

$$\text{s.t. } (\|\mathbf{u}_k\|_2, \|\mathbf{v}_k\|_2) \leq (1,1) \;\; \forall k$$

$$= \sup_{\substack{d \\ \mathbf{U}\in\mathbb{R}^{m\times d} \\ \mathbf{V}\in\mathbb{R}^{n\times d}}} \frac{1}{2\lambda_d}\|\mathbf{B}_d(\mathbf{U}, \mathbf{V})\|_2^2 \tag{21}$$

$$\text{s.t. } (\|\mathbf{u}_k\|_2, \|\mathbf{v}_k\|_2) \leq (1,1) \;\; \forall k.$$

where the final equality comes from noting that the supremum w.r.t. $\mathbf{w}$ is the definition of the Fenchel dual of the squared ℓ_2 norm evaluated at $\mathbf{B}_d(\mathbf{U}, \mathbf{V})$.

Now, from(21) and the definition of $\mathbf{B}_d(\mathbf{U}, \mathbf{V})$ note that for a fixed value of d, (21) is optimized w.r.t. $(\mathbf{U}, \mathbf{V})$ by choosing all the columns of $(\mathbf{U}, \mathbf{V})$ to be equal to the maximum singular vector pair, given by the solution to

$$\sup_{\mathbf{u}\in\mathbb{R}^m, \mathbf{v}\in\mathbb{R}^n} \left\langle \mathbf{Q}, \mathbf{u}\mathbf{v}^T \right\rangle \;\text{ s.t. } (\|\mathbf{u}\|_2, \|\mathbf{v}\|_2) \leq (1,1). \tag{22}$$

Note also that for this optimal choice of $(\mathbf{U}, \mathbf{V})$ we have that $\mathbf{B}_d(\mathbf{U}, \mathbf{V}) = \sigma(\mathbf{Q})\mathbf{1}_d$ where $\sigma(\mathbf{Q})$ denotes the largest singular value of $\mathbf{Q}$ and $\mathbf{1}_d$ is a vector of all ones of size d. Plugging this in (21) and recalling the definition of $\lambda_d = d(1-p)/p$ gives

$$\Lambda^*(\mathbf{Q}) = \sup_d \frac{1}{2\lambda_d}\|\sigma(\mathbf{Q})\mathbf{1}_d\|_2^2 = \sup_d \frac{\sigma^2(\mathbf{Q})d}{2\lambda_d} = \left(\frac{p}{1-p}\right)\frac{\sigma^2(\mathbf{Q})}{2}. \tag{23}$$

The result then follows by exploiting the well-known duality between the spectral norm (largest singular value) and the nuclear norm, as well as the basic properties of the Fenchel dual. □

Recall from (14) that for a fixed dropout rate, Ω_{dropout} does not act to constrain the number of factors in

$\mathbf{U}, \mathbf{V}$. However, as shown by the above result, once the dropout rate is adjusted as a function of d as given by (15) then the regularization is globally lower-bounded by the (squared) nuclear norm, ensuring that the regularization will never trivially be identically 0.

Additionally, by exploiting the link between the squared nuclear norm and dropout regularization provided by Proposition 2, we can provide a stronger theoretical result, which, establishes a direct connection between dropout for MF with variable size and squared nuclear norm regularization.

Theorem 2. *Let* $\mathbf{U}^{\text{opt}} \in \mathbb{R}^{m\times d^{\text{opt}}}$ *and* $\mathbf{V} \in \mathbb{R}^{n\times d^{\text{opt}}}$ *be the optimal factors that achieve the global minimum of the dropout MF problem given by*

$$\min_{\mathbf{U},\mathbf{V},d} \mathbb{E}_{\mathbf{r}} \left\| \mathbf{X} - \tfrac{1}{\theta(d)} \mathbf{U}\text{diag}(\mathbf{r})\mathbf{V}^\top \right\|_F^2 = \tag{24}$$

$$\min_{\mathbf{U},\mathbf{V},d} \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|_F^2 + \frac{1-\theta(d)}{\theta(d)} \sum_{k=1}^{d} \|\mathbf{u}_k\|_2^2 \|\mathbf{v}_k\|_2^2 \tag{25}$$

with $\theta = \theta(d)$ *as in* (15) *for some fixed hyper-parameter* $p \in (0,1)$. *Then* $\mathbf{A}^{\text{opt}} = (\mathbf{U}^{\text{opt}})(\mathbf{V}^{\text{opt}})^\top$ *is the global minimizer of*

$$\min_{\mathbf{A}} \left[\|\mathbf{X} - \mathbf{A}\|_F^2 + \frac{1-p}{p} \|\mathbf{A}\|_\star^2 \right]. \tag{26}$$

Theorem 2 provides not only a link between matrix factorization (2) and matrix approximation (1), but also insight into the effects of dropout regularization in promoting low-rank solutions. In particular, we note that this result implies that if the dropout rate is adapted to the number of columns in $\mathbf{U}$, $\mathbf{V}$ then solutions to the dropout regularized problems will be equivalent to regularization via the squared nuclear norm.

As a final remark, since the objective function of (26) is strictly convex, the existence and uniqueness of the global minimizer of (26) is guaranteed and, moreover, it can be expressed through the following closed form solution.

Theorem 3. *Let* $\mathbf{X} = \mathbf{L}\mathbf{\Sigma}\mathbf{R}^\top$ *be the singular value decomposition of* $\mathbf{X}$. *The optimal solution* $\mathbf{A}^{\text{opt}}$ *to* (26) *is given by*

$$\mathbf{A}^{\text{opt}} = \mathbf{L}\mathcal{S}_\mu(\mathbf{\Sigma})\mathbf{R}^\top, \tag{27}$$

where $\mathcal{S}_\mu(\sigma) = \max(\sigma - \mu, 0)$ *defines the shrinkage thresholding operator*[2] *[34] applied entrywise to the singular values* $\sigma_i(\mathbf{X})$ *of* $\mathbf{X}$ *and*

$$\mu = \frac{1-p}{p + (1-p)\bar{d}} \sum_{i=1}^{\bar{d}} \sigma_i(\mathbf{X}) \tag{28}$$

where $\bar{d}$ *denotes the largest integer such that*

$$\sigma_{\bar{d}}(\mathbf{X}) > \frac{1-p}{p + (1-p)\bar{d}} \sum_{i=1}^{\bar{d}} \sigma_i(\mathbf{X}). \tag{29}$$

The convex lower bound (26) to dropout for MF allows a closed-form solution for the dropout problem in terms of the singular value decomposition of $\mathbf{X}$, which we exploit to verify our analysis experimentally in the next section. In particular, this result implies that for dropout regularization in MF problems, one solution is to take the singular vectors of $\mathbf{X}$ and shrink singular values via the shrinkage thresholding operator $\mathcal{S}_\mu$ where μ is data dependent. Moreover, in this computation one must also find the optimal value of $\bar{d}$, which likewise corresponds to the optimal size of the $\mathbf{U}$, $\mathbf{V}$ factors.

We can interpret the latter points as follows: using dropout regularization for MF where the size of the factors is controlled via regularization acts as a dimensionality reduction technique very closely to related to PCA [34]. However, two differences arise: first, the number of principal components is not (heuristically) fixed but dropout learns it to be the value of $d^{\text{opt}} = \bar{d}$ in (29). Second, the top $\bar{d}$ singular values are not directly used for the projection, but, instead, we shrink them in a way that is adaptively induced by the data itself as in standard nuclear norm regularization. From this connection between dropout for MF and this data dependent adaptive PCA described above, we conclude that dropping out columns in the factors acts as a regularizer which promotes spectral sparsity for low-rank solutions.

6 NUMERICAL SIMULATIONS

This section validates our theoretical predictions about the connections between the stochastic, deterministic, and squared nuclear norm regularized formulations of MF through experiments on synthetic and real data.

Stochastic vs. deterministic reformulations of dropout. In this first experiment, we verify the equivalence between the stochastic optimization problem (3) and its deterministic counterpart (2), in which $\Omega = \Omega_{\text{dropout}}$. To do so, we constructed a synthetic data matrix $\mathbf{X} = \mathbf{U}_0\mathbf{V}_0^\top$, where $\mathbf{U}_0 \in \mathbb{R}^{m\times d}$, $\mathbf{V}_0 \in \mathbb{R}^{n\times d}$, $m = n = 100$ and $d = 160$. The entries of $\mathbf{U}_0$ and $\mathbf{V}_0$ were sampled from a $\mathcal{N}(0, \varsigma^2)$ Gaussian distribution with standard deviation 0.1. Both the stochastic and

[2]For a general scalar x, one usually defines $\mathcal{S}_\mu(x) = \text{sgn}(x)\max(|x| - \mu, 0)$, but, here, due to the non-negativity of the singular values $\sigma > 0$, we will exploit the simplified expression $\mathcal{S}_\mu(\sigma) = \max(\sigma - \mu, 0)$.

deterministic formulations of dropout were solved by 10,000 iterations of gradient descent with diminishing $O(\frac{1}{t})$ lengths for the step size. In the stochastic setting, we approximated the objective in (3) and its gradient by sampling a new Bernoulli vector $\mathbf{r}$ at each iteration of Algorithm 1.

Figure 1 plots the objective curves for the stochastic and deterministic dropout formulations for different choices of the retain rate $\theta = 0.1, 0.3, 0.5, 0.7, 0.9$. We observe that across all choices of θ, the deterministic objective (2) tracks the apparent expected value that is computed in (3). This validates the claim of Theorem 1 that the two formulations are equivalent.

Evaluating connections to the nuclear norm. In a second experiment, we validate the equivalence between Ω_{dropout} and squared nuclear norm regularization for variable size factorizations (Theorem 2). To do so, we constructed a synthetic dataset $\mathbf{X}$ consisting of a low-rank matrix combined with dense Gaussian noise. Specifically, we let $\mathbf{X} = \mathbf{U}_0\mathbf{V}_0^\top + \mathbf{Z}_0$ where $\mathbf{U}_0, \mathbf{V}_0 \in \mathbb{R}^{100\times 10}$ and $\mathbf{Z}_0 \in \mathbb{R}^{100\times 100}$ contain entries drawn from a normal distribution $\mathcal{N}(0, \varsigma^2)$, with $\varsigma = 0.1$ for $\mathbf{U}_0, \mathbf{V}_0$ and $\varsigma = 0.01$ for $\mathbf{Z}_0$. We then compared the regularization performance of the squared nuclear norm with that of Ω_{dropout}, both with a fixed $\theta = 0.9$ and an adaptive $\theta = \theta(d)$ ($p = 0.9$). Algorithm 1 was used to solve the Ω_{dropout} regularized problems, while for the squared nuclear norm we computed the closed form solution of Theorem 3.

Figure 2 plots the singular values for the optimal solution to each of the three problems. We observe first that without adjusting θ, dropout regularization has little effect on the rank of the solution. The smallest singular values are still relatively large and not shifted significantly compared to the singular values of the original data. This is consistent with the idea of Proposition 1 that traditional dropout is a poor regularizer when d is allowed to vary. On the other hand, by adjusting the dropout rate to the size of the factorization we find that consistent with Theorem 2, the regularization behavior of Ω_{dropout} closely matches that of the squared nuclear norm, and moreover both formulations are able to recover the rank of the noise-free data ($\text{rank}(\mathbf{U}_0\mathbf{V}_0^\top) = 10$). Furthermore, across the choices for d, the relative Frobenius distances between the solutions of these two methods are very small (between 10^{-6} and 10^{-2}). Taken together, our theoretical predictions and experimental results suggest that adapting the dropout rate to the size of the factorization is potentially critical to ensuring the effectiveness of dropout as a regularizer and in limiting the degrees of freedom of the model.

Matrix factorization meets approximation with dropout. In a final experiment, we compared the quality of Ω_{dropout} and squared nuclear norm regularization in a low-rank approximation task using images of handwritten digits. Here the data matrix $\mathbf{X}$ contains 55K 28×28 images from the MNIST training set. Preprocessing steps include min-max normalization and vectorization. In this experiment we fixed the size of the factorization to $d = 40$.

We solved the Ω_{dropout} regularized problem (4) by full gradient descent with a fixed learning rate $\epsilon = 10^{-4}$. In order to better cope with the non-convexity of the problem, we updated the factors $\mathbf{U}$ and $\mathbf{V}$ in an alternating fashion. We performed the following updating scheme: we applied 50 gradient updates to $\mathbf{U}$ with $\mathbf{V}$ fixed, followed by 50 updates to $\mathbf{V}$ with $\mathbf{U}$ fixed. The previous updating scheme was repeated for 1000 times overall. We computed the solution to the squared nuclear norm problem (26) in closed form following Theorem 3.

In Figure 3 we compare samples of the original MNIST data to their reconstructions obtained through either Ω_{dropout} or the equivalent squared nuclear norm regularization, using fixed dropout rates $\theta = 0.5$ and $\theta = 0.8$. Visually, the two sets of reconstructions are nearly identical. Numerically, mean squared difference between the factorizations $\mathbf{U}\mathbf{V}^\top$ for the two formulations are within 10^{-3}. Moreover, both formulations are able to represent the data reasonably well, achieving mean reconstruction error at most 10^{-2}, despite the simplicity of the linear low-rank approximation model.

7 CONCLUSIONS

We have presented a theoretical analysis of dropout as a regularization strategy for matrix factorization (MF). We showed that the dropout algorithm for MF is a stochastic gradient descent strategy applied to a stochastic objective in which Bernoulli random variables are used to drop columns of the factors. We also showed that the expected value of the stochastic objective is equal to a purely deterministic objective with a regularizer which is equal to the sum of the product of the squared norms of the columns of the factors. When the factorization size is allowed to vary, we showed that using dropout with a fixed dropout rate is not sufficient to limit the size of the factorization. To address this issue, we proposed a dropout strategy that adjusts the dropout rate based on the size of the factorization, and showed that this induces a regularizer that is closely related to the squared nuclear norm. Finally, we presented experimental results that confirmed our theoretical predictions.

Acknowledgements. The authors acknowledge the financial support of the grants NSF 1618485, ARO MURI W911NF-17-1-0304, and IARPA 54663-Z9108203.

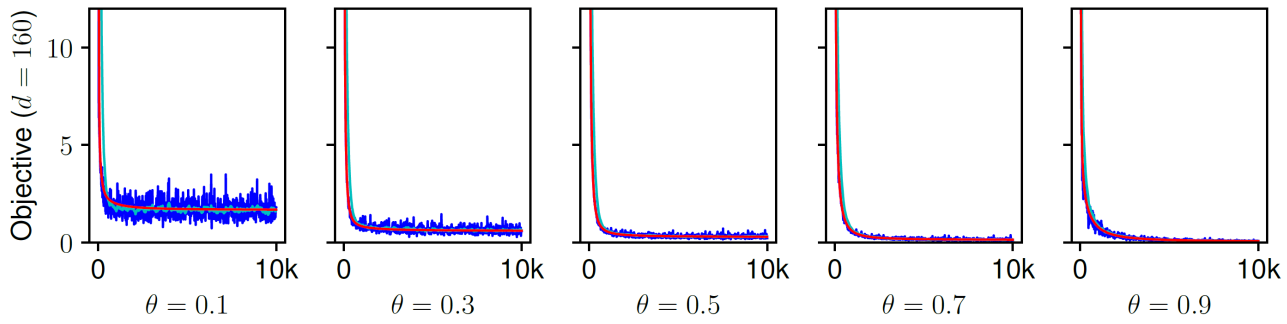

Figure 1: For $\theta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $d = 160$ we compare dropout for MF (3) (blue) and its deterministic counterpart (2) (red). The exponential moving average of the stochastic objective is in cyan. Best viewed in color.

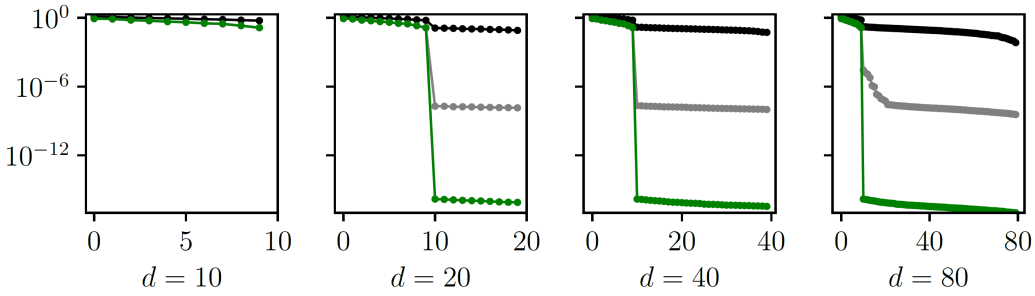

Figure 2: Singular values in log-scale corresponding to the optimal solutions of the three regularization schemes considered: fixed dropout rate of $\theta = 0.9$ (black), adaptive dropout $\theta = \theta(d)$ as (15) with $p = 0.9$ (gray), and the nuclear-norm squared closed-form optimization as in Proposition 2 (green). Best viewed in color.

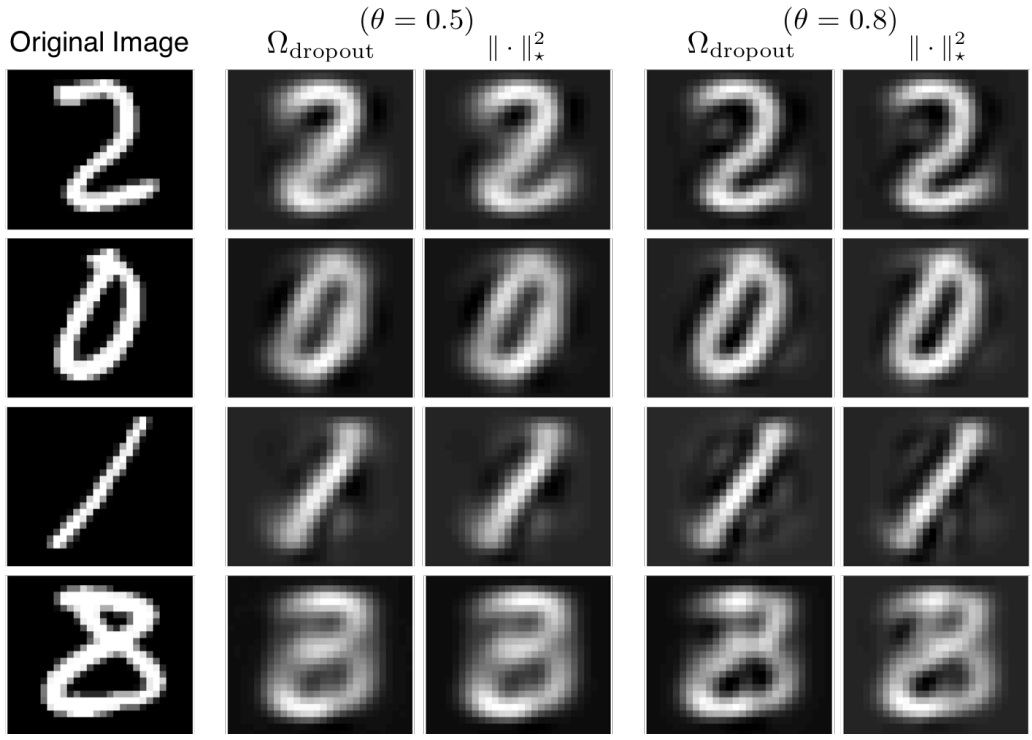

Figure 3: Low-rank approximation of MNIST digits using dropout (Ω_{dropout}) and squared nuclear norm ($\|\cdot\|_\star^2$) regularization, with $\theta = 0.5$ and $\theta = 0.8$.

References

[1] A. Achille and S. Soatto. Information Dropout: learning optimal representations through noisy computation. *ArXiv e-prints*, November 2016.

[2] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, Dec 2008.

[3] F. Bach. Convex relaxations of structured matrix factorizations. In *CoRR:1309.3117v1*, 2013.

[4] F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. In *CoRR:0812.1869v1*, 2008.

[5] P. Baldi and P. Sadowski. Understanding dropout. In *Neural Information Processing Systems*, 2013.

[6] P. Baldi and P. Sadowski. The dropout learning algorithm. In *Artificial Intelligence*, 2014.

[7] Justin Bayer, Christian Osendorfer, and Nutan Chen. On fast dropout and its applicability to recurrent networks. In *CoRR:1311.0701*, 2013.

[8] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference in Machine Learning*, 2009.

[9] Chris M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

[10] A. Del Bue, J. Xavier, L. Agapito, and M. Paladini. Bilinear modeling via augmented lagrange multipliers (BALM). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1496–1508, 2012.

[11] Ricardo S Cabral, Fernando Torre, João P Costeira, and Alexandre Bernardino. Matrix completion for multi-label image classification. In *Advances in Neural Information Processing Systems*, pages 190–198, 2011.

[12] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, Apr 2009.

[13] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[14] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML-16)*, 2016.

[15] Benjamin D. Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. In *CoRR abs/1506.07540*, 2015.

[16] Benjamin D. Haeffele and René Vidal. Structured low-rank matrix factorization: Global optimality, algorithms, and applications. In *CoRR abs/1708.07850*, 2017.

[17] Benjamin D. Haeffele, Eric Young, and René Vidal. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *International Conference in Machine Learning*, 2014.

[18] Benjamin D. Haeffele, Eric Young, and René Vidal. Global optimality in neural network training. In *IEEE Conference in Computer Vision and Pattern Recognition*, 2017.

[19] Zhicheng He, Jie Liu, Caihua Liu, Yuan Wang, Airu Yin, and Yalou Huang. *Dropout Non-negative Matrix Factorization for Independent Feature Learning*, pages 201–212. Springer, Cham, 2016.

[20] David P. Helmbold and Philip M. Long. On the inductive bias of dropout. *Journal of Machine Learning Research*, 16:3403–3454, 2015.

[21] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[22] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[23] Cho-Jui Hsieh and Peder Olsen. Nuclear norm minimization via active subspace selection. In *International Conference in Machine Learning*, pages 575–583, 2014.

[24] Ba Jimmy and Brendan Frey. Adaptive dropout for training deep neural networks. In *NIPS*, 2016.

[25] Zhe Gong Li and Tianbao Boqing Yang. Improved dropout for shallow and deep learning. In *Neural Information Processing Systems*, 2016.

[26] Pietro Morerio, Jacopo Cavazza, Riccardo Volpi, René Vidal, and Vittorio Murino. Curriculum dropout. In *International Conference on Computer Vision*, 2017.

[27] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[28] Jasson D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *ICML*, 2005.

[29] Steven J. Rennie, Vaibhava Goel, and Samuel Thomas. Annealed dropout training of deep networks. In *Proceedings onf the IEEE Workshop on SLT*, pages 159–164, 2014.

[30] Salah Rifai, Xavier Glorot, Bengio Yoshua, and Pascal Vincent. Adding noise to the input of a model trained with a regularized objective. In *CoRR:1104.3250v1*, 2011.

[31] Nathan Srebro, Jason DM Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *NIPS*, 2004.

[32] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.

[33] R. Vidal and P. Favaro. Low rank subspace clustering (LRSC). *Pattern Recognition Letters*, 43:47–61, 2014.

[34] Rene Vidal, Yi Ma, and S. S. Sastry. *Generalized Principal Component Analysis*. Springer, 1st edition, 2016.

[35] Stefan Wager, William Fithian, Sida Wang, and Percy S. Liang. Altitude training: Strong bounds for single-layer dropout. In Z. Ghahramani, M. Welling, C. Cortes, N.d. Lawrence, and K.q. Weinberger, editors, *Neural Information Processing Systems*, pages 100–108, 2014.

[36] Stefan Wager, Sida Wang, and Percy S Liang. Dropout training as adaptive regularization. In *Neural Information Processing Systems*, 2013.

[37] Haibing Wu and Xiaodong Gu. Towards dropout training for convolutional neural networks. *Neural Networks*, 71:1–10, 2015.

[38] Ming Yuan, Ali Ekici, Zhaosong Lu, and Renato Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society*, 69(3):329–346, 2007.

[39] Shangfei Zhai and Zhongfei Zhang. Dropout training of matrix factorization and autoencoders for link prediction in sparse graphs. In *CoRR:1512.04483v1*, 2015.