How regularization affects the critical points in linear networks

Authored by:

Amirhossein Taghvaei
Jin W. Kim
Prashant Mehta

Abstract

This paper is concerned with the problem of representing and learning a linear transformation using a linear neural network. In recent years, there is a growing interest in the study of such networks, in part due to the successes of deep learning. The main question of this body of research (and also of our paper) is related to the existence and optimality properties of the critical points of the mean-squared loss function. An additional primary concern of our paper pertains to the robustness of these critical points in the face of (a small amount of) regularization. An optimal control model is introduced for this purpose and a learning algorithm (backprop with weight decay) derived for the same using the Hamilton's formulation of optimal control. The formulation is used to provide a complete characterization of the critical points in terms of the solutions of a nonlinear matrix-valued equation, referred to as the characteristic equation. Analytical and numerical tools from bifurcation theory are used to compute the critical points via the solutions of the characteristic equation.

1 Paper Body

This paper is concerned with the problem of representing and learning a linear transformation with a linear neural network. Although a classical problem (Baldi and Hornik [1989, 1995]), there has been a renewed interest in such networks (Saxe et al. [2013], Kawaguchi [2016], Hardt and Ma [2016], Gunasekar et al. [2017]) because of the successes of deep learning. The motivation for studying linear networks is to gain insight into the optimization problem for the more general nonlinear networks. A ?

Financial support from the NSF CMMI grant 1462773 is gratefully acknowledged.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

focus of the recent research on these (and also nonlinear) networks has been on the analysis of the critical points of the non-convex loss function (Dauphin et al. [2014], Choromanska et al. [2015a,b], Soudry and Carmon [2016], Bhojanapalli et al. [2016]). This is also the focus of our paper. Problem: The input-output model is assumed to be of the following linear form: Z = RX0 + ?

(1)

where X0 ? Rd?1 is the input, Z ? Rd?1 is the output, and ? ? Rd?1 is the noise. The input X0 is modeled as a random variable whose distribution is denoted as p0 . Its second moment is denoted as ?0 := E[X0 X0¿] and assumed to be finite. The noise ? is assumed to be independent of X0 , with zero mean and finite variance. The linear transformation R ? Md (R) is assumed to satisfy a property (P1) introduced in Sec. 3 (Md (R) denotes the set of d ? d matrices). The problem is to learn the weights of a linear neural network from i.i.d. input-output samples {(X0k , Z k)}K k=1 . Solution architecture: is a continuous-time linear feedforward neural network model: dXt = At Xt dt

(2)

where At ? Md (R) are the network weights indexed by continuous-time (surrogate for layer) t ? [0, T], and X0 is the initial condition at time t = 0 (same as the input data). The parameter T denotes the network depth. The optimization problem is to choose the weights At over the time-horizon [0, T] to minimize the mean-squared loss function: E[—XT ? Z—2]

(3)

This problem is referred to as the [? = 0] problem. Backprop is a stochastic gradient descent algorithm for learning the weights At . In general, one obtains (asymptotic) convergence of the learning algorithm to a (local) minimum of the optimization problem Lee et al. [2016], Ge et al. [2015]. This has spurred investigation of the critical points of the loss function (3) and the optimality properties (local vs. global minima, saddle points) of these points. For linear multilayer (discrete) neural networks (MNN), strong conclusions have been obtained under rather mild conditions: every local minimum is a global minimum and every critical point that is not a local minimum is a saddle point Kawaguchi [2016], Baldi and Hornik [1989]. For the discrete counterpart of the [? = 0] problem (referred to as the linear residual network in Hardt and Ma [2016]), an even stronger conclusion is possible: all critical points of the [? = 0] problem are global minimum. In experiments, some of these properties are also empirically observed in deep nonlinear networks; cf., Choromanska et al. [2015b], Dauphin et al. [2014], Saxe et al. [2013]. In this paper, we consider the following regularized form of the optimization problem: Z T 1 1 Minimize: J[A] = E[? tr (A¿ —XT ? Z—2] t At) dt + A 2 0 2 dXt Subject to: = At Xt , X0 ? p0 dt

(4)

where ? ? R+ := {x ? R : x ? 0} is a regularization parameter. In literature, this form of regularization is referred to as weight decay [Goodfellow et al., 2016, Sec. 7.1.1]. Eq. (4) is an example of an optimal control problem and is referred to as such. The limit ? ? 0 is referred to as [? = 0+] problem.

The symbol tr(?) and superscript ¿ are used to denote matrix trace and matrix transpose, respectively. The regularized problem is important because of the following reasons: 2

(i) The learning algorithms are believed to converge to the critical points of the regularized [? = 0+] problem, a phenomenon known as implicit regularization Neyshabur et al. [2014], Zhang et al. [2016], Gunasekar et al. [2017]. (ii) It is shown in the paper that the stochastic gradient descent (for the functional J) yields the following learning algorithm for the weights At : (k+1)

At

(k)

= At

(k)

+ ?k (??At

+ backprop update)

(5)

for k = 1, 2, . . ., where ?k is the learning rate parameter. Thus, the parameter ? models dissipation (or weight decay) in backprop. In an implementation of backprop, one would expect to obtain critical points of the [? = 0+] problem. The outline of the remainder of this paper is as follows: The Hamilton?s formulation is introduced for the optimal control problem (4) in Sec. 2; cf., LeCun et al. [1988], Farotimi et al. [1991] for related constructions. The Hamilton?s equations are used to obtain a formula for the gradient of J, and subsequently derive the stochastic gradient descent learning algorithm of the form (5). The equations for the critical points of J are obtained by applying the Maximum Principle of optimal control (Prop. 1). Remarkably, the Hamilton?s equations for the critical points can be solved in closed-form to obtain a characterization of the critical points in terms of the solutions of a nonlinear matrix-valued equation, referred to as the characteristic equation (Prop. 2). For a certain special case, where the matrix R is normal, analytical results are obtained based on the use of the implicit function theorem (Thm. 2). Numerical continuation is employed to compute the solutions for this and the more general non-normal cases (Examples 1 and 2).

2

Hamilton?s formulation and the learning algorithm

Definition 1. The control Hamiltonian is the function ? tr(B ¿ B) (6) 2 where x ? Rd is the state, y ? Rd is the co-state, and B ? Md (R) is the weight matrix. The ?H ?H ¿ partial derivatives are denoted as ?H ?x (x, y, B) := B y, ?y (x, y, B) := Bx, and ?B (x, y, B) := ¿ yx ? ?B. H(x, y, B) = y ¿ Bx ?

Pontryagin?s Maximum Principle (MP) is used to obtain the Hamilton?s equations for the solution of the optimal control problem (4). The MP represents a necessary condition satisfied by any minimizer. Conversely, a solution of the Hamilton?s equation is a critical point of the functional J. The proof of the following proposition appears in the supplementary material. Proposition 1. Consider the terminal cost optimal control At is the minimizer and Xt is the corresponding trajectory. Y : [0, T] ? Rd such that dXt ?H =+ (Xt , Yt , At

) = +At Xt , dt ?y dYt ?H =? (Xt , Yt , At) = ?A¿ t Yt , dt ?x and At maximizes the expected value of the Hamiltonian At = arg max

E[H(Xt , Yt , B)]

B ? Md (R)

problem (4) with ? ? 0. Suppose Then there exists a random process

X0 ? p0

(7)

YT = Z ? XT

(8)

1 E[Yt Xt¿] ?

(9)

(?¿0)

=

Conversely, if there exists At and the pair (Xt , Yt) such that equations (7)-(8)-(9) are satisfied, then At is a critical point of the optimization problem (4). 3

Remark 1. The Maximum Principle can also be used to derive analogous (difference) equations in discrete-time as well as nonlinear settings. It is equivalent to the method of Lagrange multipliers that is used to derive the backprop algorithm in MNN, e.g., LeCun et al. [1988]. The continuous-time limit is considered here because the computations are simpler and the results are more insightful. Similar considerations have also motivated the study of continuous-time limit of other types of optimization algorithms, e.g., Su et al. [2014], Wibisono et al. [2016]. The Hamiltonian is also used to express the first order variation in the functional J. For this purpose, define the Hilbert space of matrix-valued functions L2 ([0, T]; Md (R)) := {A : [0, T] ? Md (R) — RT RT ¿ 2 tr(A¿ t At) dt ¡ ?} with the inner product hA, V iL2 := 0 tr(At Vt) dt. For any A ? L , 0 the gradient of the functional J evaluated at A is denoted as ?J[A] ? L2 . It is defined using the directional derivative formula: J(A + V) ? J(A) h?J[A], V iL2 := lim ?0

2 where V ? L prescribes the direction (variation) along which the derivative is being computed. The explicit formula for ?J is given by

?H ?J[A] := ?E (Xt , Yt , At) = ?At ? E Yt Xt¿ (10) ?B where Xt and Yt are the obtained by solving the Hamilton?s equations (7)-(8) with the prescribed (not necessarily optimal) weight matrix A ? L2 . The significance of the formula is that the steepest descent in the objective function J is obtained by moving in the direction of the steepest (for each fixed t ? [0, T]) ascent in the Hamiltonian H. Consequently, a stochastic gradient descent algorithm to learn the weights is as follows: (k+1)

At

(k)

= At

(k)

? ?k (?At (k) Xt

(k)

? Yt

(k) ¿
Xt
),
(11)
(k) Yt

where ?k is the step-size at iteration k and and are obtained by solving the Hamilton?s equations (7)-(8): d (k) (k) (k) (k) (Forward propagation) X = +At Xt , with init. cond. X0 (12) dt t d (k) (k)¿ (k) (k) (k) Y = ?At Yt , YT = Z (k) ? XT (Backward propagation) (13) dt t — {z } error

(k)
(k)

based on the sample input-output (X , Z). Note the forward-backward structure of the algorithm: (k) (k) In the forward pass, the network output XT is obtained given the input X0 ; In the backward (k) pass, the error between the network output XT and true output Z (k) is computed and propagated backwards. The regularization parameter is also interpreted as the dissipation or the weight decay parameter. By setting ? = 0, the standard backprop algorithm is obtained. A convergence result for the learning algorithm for the [? = 0] case appears as part of the supplementary material. In the remainder of this paper, the focus is on the analysis of the critical points.

3

Critical points

For continuous-time networks, the critical points of the [? = 0] problem are all global minimizers (An analogous result for residual MNN appears in [Hardt and Ma, 2016, Thm. 2.3]). Theorem 1. Consider the [? = 0] optimization problem (4) with non-singular ?0 . For this problem (provided a minimizer exists) every critical point is a global minimizer. That is, ?J[A] = 0

??
J(A) = J? := min J[A] A

4

Moreover, for any given (not necessarily optimal) A ? L2 , RT ? ¿ k?J[A]k2L2 ? T e?2 0 tr(At At) dt ?min (?0)(J(A) ? J?)

(14)

where ?min (?0) is the smallest eigenvalue of ?0 . Proof. (Sketch) For the linear system (2), the fundamental solution matrix is denoted as ?t;t0 . The solutions of the Hamilton?s equations (7)-(8) are given by Yt = ?¿ T ;t (Z ? XT)

Xt = ?t;0 X0 ,

Using the formula (10) upon taking an expectation ¿ ?J[A] = ??¿ T ;t (R ? ?T ;0)?0 ?t;0

which (because ? is invertible) proves that: ?J[A] = 0

??
??
?T ;0 = R
J(A) = J? := min J[A] A

The derivation of the bound (14) is equally straightforward and appears as part of the supplementary material. Although the result is attractive, the conclusion is somewhat misleading because (as we will demonstrate with examples) even a small amount of regularization can lead to local (but not global) minimum as well as saddle point solutions. Assumption: The following assumption is made throughout the remainder of this paper: (i) Property P1: The matrix R has no eigenvalues on R? := {x ? R : x ? 0}. The matrix R is non-derogatory. That is, no eigenvalue of R appears in more than one Jordan block. For the scalar (d = 1) case, this property means R is strictly positive. For the scalar case, the RT A dt t fundamental solution is given by the closed form formula ?T,0 = e 0 . Thus, the positivity of R is seen to be necessary to obtain a meaningful solution. For the vector case, this property represents a sufficient condition such that log(R) can be defined as a real-valued matrix. That is, under property (P1), there exists a (not necessarily unique2) matrix log(R) ? Md (R) whose matrix exponential elog(R) = R; cf., Culver [1966], Higham [2014]. The logarithm is trivially a minimum for the [? = 0] problem. Indeed, At ? T1 log(R) gives log(R)

Xt = e T t X0 and thus XT = elog(R) X0 = RX0 . This shows At can be made arbitrarily small by choosing a large enough depth T of the network. An analogous result for the linear residual MNN appears in [Hardt and Ma, 2016, Thm. 2.1]. The question then is whether the constant solution At ? T1 log(R) is also obtained as a critical point for the [? = 0+] problem? The following proposition provides a complete characterization of the critical points (for the general ? ? R+ problem) in terms of the solutions of a matrix-valued characteristic equation: Proposition 2. The general solution of the Hamilton?s equations (7)-(9) is given by ¿

Xt = e2t? etC X0 Yt = e

2t?

At = e

2t?

e

(T ?t)C

Ce

?2t?

2

(15) ?2T ?

e

(Z ? XT)

(16) (17)

Under Property (P1), log(R) is uniquely defined if and only if all the eigenvalues of R are positive. When not unique there are countably many matrix logarithms, all denoted as log(R). The principal logarithm of R is the unique such matrix whose eigenvalues lie in the strip {z ? C : ?? ¡ Im(z) ¡ ?}.

5

where C ? Md (R) is an arbitrary solution of the characteristic equation ?C = F ¿ (R ? F)?0

(18)

¿

where F := e2T ? eT C and the matrix ? := 12 (C ? C¿) is the skew-symmetric component of C. The associated cost is given by 1 1 ?T tr C¿ C + tr (F ? R)¿ (F ? R)?0 + E[—?—2] J[A] = 2 2 2 And the following holds: (?0 =I)

At ? C ?? C is normal =? R is normal Proof. (Sketch) Differentiating both sides of (9) with respect to t and using the Hamilton?s equations (7)-(8), one obtains dAt ¿ = ?A¿ t At + At At dt whose general solution is given by (17). The remainder of the analysis is straightforward and appears as part of the supplementary material. Remark 2. Prop. 2 shows that the answer to the question posed above concerning the constant solution At ? T1 log(R) is false in general for the [? = 0+] problem: For ? ¿ 0 and ?0 = I, a constant solution is a critical point only if R is a normal matrix. For the generic case of non-normal R, any critical point is necessarily non-constant for any positive choice of the parameter ?. Some of these non-constant critical points are described as part of the Example 2. Remark 3. The linear structure of the input-output model (1) is not necessary to derive the results in Prop. 2. For correlated input-output random variables (X, Z), the general form of the characteristic equation is as follows: ?C = F ¿ (E[ZX0¿] ? F ?0) ¿

where (as before) ?0 = E[X0 X0¿], and F := e2T ? eT C where ? := 12 (C ? C¿). Prop. 2 is useful because it helps reduce the infinite-dimensional problem to a finite-dimensional characteristic equation (18). The solutions C of the characteristic equation fully parametrize the solutions of the Hamilton?s equations (7)-(9) which in turn represent the critical points of the optimal control problem (4). The matrix-valued nonlinear characteristic equation (18) is still formidable. To gain analytical and numerical insight into the matrix case, the following strategy is employed: (i) A solution C is obtained by setting ? = 0 in the characteristic equation. The corresponding equation is ¿ ¿ eT (C?C) eT C = R This solution is denoted as C(0). (ii) Implicit function theorem is used to establish (local) existence of a solution branch C(?) in a neighborhood of the ? = 0 solution. (iii) Numerical continuation is used to compute the solution C(?) as a function of the parameter ?. The following theorem provides a characterization of normal solutions C for the case where R is assumed to be a normal matrix and ? = I. Its proof appears as part of the supplementary material. Theorem 2. Consider the characteristic equation (18) where R is assumed to be a normal matrix that satisfies the Property (P1) and ?0 = I. 6

Figure 1: (a) Critical points in Example 1 (the (2, 1) entry of the solution matrix C(?; n) is depicted for n = 0, ?1, ?2); (b) The cost J[A] for these solutions. (i) For ? = 0 the normal solutions of (18) are given by

1 T

log(R).

(ii) For each such solution, there exists a neighborhood N ? R+ of ? = 0 such that the solution of the characteristic equation (18) is well-defined as a continuous map from ? ? N ? C(?) ? Md (R) with C(0) = T1 log(R). This solution is given by the asymptotic formula C(?) =

? 1 log(R) ? 2 (RR¿)?1 log(R) + O(?2) T T

Remark 4. For the scalar case log(?) is a single-valued function. Therefore, At ? C = T1 log(R) is the unique critical point (minimizer) for the [? = 0+] problem. While the [? = 0+] problem admits a unique minimizer, the [? = 0] problem does not. In fact, any At of the form At = T1 log(R) + A?t RT where 0 A?t dt = 0 is also a minimizer of the [? = 0] problem. So, while there are infinitely many minimizers of the [? = 0] problem, only one of these survives with even a small amount of regularization. A global characterization of critical points as a function of parameters (?, R, ?0 , T) ? R+ ? R+ ? R+ ? R+ is possible and appears as part of the supplementary material. " # 0 ?1 Example 1 (Normal matrix case). Consider the characteristic equation (18) with R = 1 0 (rotation in the plane by ?/2), ?0 = I and T = 1. For ? = 0, the normal solutions of the characteristic equation are given by the multi-valued matrix logarithm function: " # 0 ?1 log(R) = (?/2 + 2n?) =: C(0; n), n = 0, ?1, ?2, . . . 1 0 It is easy to verify that eC(0;n) = R. C(0; 0) is referred to as the principal logarithm. The software package PyDSTool Clewley et al. [2007] is used to numerically continue the solution C(?; n) as a function of the parameter ?. Fig. 1(a) depicts the solutions branches in terms of the (2, 1) entry of the matrix C(?; n) for n = 0, ?1, ?2. The following observations are made concerning these solutions: ? n) for which there exist two solutions, a local (i) For each fixed n 6= 0, there exist a range (0, ? ? n , there is a qualitative change minimum and a saddle point. At the limit (turning) point ? = ? in the solution from a minimum to a saddle point. ? ?1 , only a single ? n decreases monotonically as —n— increases. For ? ¿ ? (ii) As a function of n, ? solution, the principal branch C(?; 0) was found using numerical continuation. 7

Figure 2: (a) Numerical continuation of the solution in Example 2; (b) The cost J[A] for the critical point (minimum) and the constant T1 log(R) solution. (iii) Along the branch with a fixed n 6= 0, as ? ? 0, the saddle point solution"escapes to infinity. # ?? ?1 That is as ? ? 0, the saddle point solution C(?; n) ? (?/2 + (2n ? 1)?) . The 1 ?? associated cost J[A] ? 1 (The cost of global minimizer J ? = 0). (iv) Among the numerically obtained solution branches, the principal branch C(?; 0) has the lowest cost. Fig. 1 (b) depicts the cost for the solutions depicted in Fig. 1 (a). The numerical calculations indicate that while the [? = 0] problem has infinitely many critical points (all global minimizers), only a finitely many critical points persist for any finite positive value of ?. Moreover, there exists both local (but not global) minimum as well as saddle points for this case. Among the solutions computed, the principal branch (continued from the principal logarithm C(0; 0)) has the minimum cost. Example 2 (Non-normal " # matrix case). Numerical continuation is used to obtain solutions for non0 ?1 normal R = , where ? is a continuation parameter and T = 1. Fig. 2(a) depicts a solution 1 ? branch as a function of parameter ?. The solution is initialized with the normal solution C(0; 0) described in Example 1. By varying" ?, the# solution is continued " to ?# = ?/2 (indicated as in 0 0 0 ?1 part (a)). This way, the solution C = ? is found for R = . It is easy to verify that C 0 1 ?2 2 is a solution of the characteristic equation (18) " for ? = 0 and T = 1. For this # solution, the critical ?? sin(?t) ? cos(?t) ? ? point of the optimal control problem At = is non-constant. The ? cos(?t) + ? ? sin(?t)

" # ? ?? tan ? ?? sec ? principal logarithm log(R) = , where ? = sin?1 . The regularization 4 ? sec ? ? tan ? cost for the non-constant solution At is strictly smaller than the constant T1 log(R) solution: Z 1 Z 1 Z 1 ?2 ¿ ¡ 3.76 = tr(log(R) log(R)¿) dt tr(At A¿) dt = tr(CC) dt = t 4 0 0 0 Next, the parameter ? = ?2 is fixed, and the solution continued in the parameter ?. Fig. 2(b) depicts the cost J[A] for the resulting solution branch of critical points (minimum). The cost with the constant T1 log(R) is also depicted. It is noted that the latter is not a critical point of the optimal control problem for any positive value of ?. 8

4

Conclusions and directions for future work

In this paper, we studied the optimization problem of learning the weights of a linear neural network with mean-squared loss function. In order to do so, we introduced a novel formulation: (i) The linear network is modeled as a continuous time (surrogate for layer) optimal control problem; (ii) A weight decay type regularization is considered where the interest is in the limit as the regularization parameter ? ? 0 (the limit is referred to as the [? = 0+] problem). The Maximum Principle of optimal control theory is used to derive the Hamilton?s equations for the critical points. A remarkable result of our paper is that the critical point solutions of the infinite-dimensional problem are completely characterized via the solutions of a finite-dimensional characteristic equation (Eq. (18)). That such a reduction is possible is unexpected because the weight update equation is nonlinear (even in the settings of linear networks). Based on the analysis of the characteristic equation, several conclusions are obtained3 : (i) It has been noted in literature that, for linear networks, all critical points are global minimum. While this is also true here for the [? = 0] and the [? = 0+] problems, even a small amount of regularization alters the picture, e.g., saddle points emerge (Example 1). (ii) The critical points of the regularized [? = 0+] problem is qualitatively very different compared to the non-regularized [? = 0] problem (Remark 4). Several quantitative results on the critical points of the regularized problem are described in Theorem 2 and Examples 1 and 2. (iii) The study of the characteristic equation revealed an unexpected qualitative difference in the critical points between the two cases where R := E[ZX0¿] is a normal or non-normal matrix. In the latter (generic) case, the network weights are necessarily non-constant (Prop. 2). We believe that the ideas and tools introduced in this paper will be useful for the researchers working on the analysis of deep learning. In particular, the paper is expected to highlight and spur work on implicit regularization. Some directions for future work are briefly noted next: (i) Non-normal solutions of the characteristic equation: Analysis of the non-normal solutions of the characteristic equation remains an open problem. The non-normal solutions are important because of the following empirical observation (summarized as part of the supplementary material): In numerical experiments with learning, the weights can get stuck at non-normal critical points before eventually converging to a ?good? minimum. (ii) Generalization error: With a finite number of samples (X0i , Z i)N i=1 , the characteristic equation (N)

?C = F ¿ (R ? F)?0

+ F ¿ Q(N) PN PN ¿ ¿ (N) where ?0 := N1 i=1 X0i X0i and Q(N) := N1 i=1 X0i ? i . Sensitivity analysis of the (N) solution of the characteristic equation, with respect to variations in ?0 and Q(N) , can shed light on the generalization error for different critical points. (iii) Second order analysis: The paper does not contain second order analysis of the critical points ? to determine whether they are local minimum or saddle points. Based on certain preliminary results for the scalar case, it is conjectured that the second order analysis is possible in terms of the first order variation for the characteristic equation. 3

Qualitative aspects of some of the conclusions may be obvious to experts in Deep Learning. The objective here is to obtain quantitative characterization in the (relatively tractable) setting of linear networks.

9

2 References

P. F. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. Neural networks, 2(1):53?58, 1989. P. F. Baldi and K. Hornik. Learning in linear neural networks: A survey. IEEE Transactions on neural networks, 6(4):837?858, 1995. S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In Advances in Neural Information Processing Systems, pages 3873?3881, 2016. A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surfaces of multilayer networks. In AISTATS, 2015a. A. Choromanska, Y. LeCun, and G. B. Arous. Open problem: The landscape of the loss surfaces of multilayer networks. In COLT, pages 1756?1760, 2015b. R. Clewley, W. E. Sherwood, M. D. LaMar, and J. Guckenheimer. Pydstool, a software environment for dynamical systems modeling, 2007. URL http://pydstool.sourceforge.net. W. J. Culver. On the existence and uniqueness of the real logarithm of a matrix. Proceedings of the American Mathematical Society, 17(5):1146?1151, 1966. Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Advances in neural information processing systems, pages 2933?2941, 2014. O. Farotimi, A. Dembo, and T. Kailath. A general weight matrix formulation using optimal control. IEEE Transactions on neural networks, 2(3):378?394, 1991. R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping From Saddle Points ? Online Stochastic Gradient for Tensor Decomposition. arXiv:1503.02101, March 2015. I. Goodfellow, Y. Bengio, and A. Courville. Deep learning. MIT press, 2016. S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. arXiv preprint arXiv:1705.09280, 2017. M. Hardt and T. Ma. Identity matters in deep learning. arXiv:1611.04231, November 2016. N. J. Higham. Functions of matrices. CRC Press, 2014. K. Kawaguchi. Deep learning without poor local minima. In Advances In Neural Information Processing Systems, pages 586?594, 2016. Y. LeCun, D. Touresky, G. Hinton, and T. Sejnowski. A theoretical framework for back-propagation. In

The Connectionist Models Summer School, volume 1, pages 21?28, 1988. J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient Descent Converges to Minimizers. arXiv:1602.04915, February 2016. B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. arXiv preprint arXiv:1412.6614, 2014. 10

A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv:1312.6120, December 2013. D. Soudry and Y. Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. arXiv:1605.08361, May 2016. W. Su, S. Boyd, and E. Candes. A differential equation for modeling nesterov?s accelerated gradient method: Theory and insights. In Advances in Neural Information Processing Systems, pages 2510?2518, 2014. A. Wibisono, A. Wilson, and M. Jordan. A variational perspective on accelerated methods in optimization. Proceedings of the National Academy of Sciences, page 201614734, 2016. C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530, 2016.

11