CS282 Final Project

Yiran Liu
ShanghaiTech University
2021533003
liuyr1@shanghaitech.edu.cn

Zhehao Shen
ShanghaiTech University
2021533110
shenzhh@shanghaitech.edu.cn

Shouchen Zhou
ShanghaiTech University
2021533042
zhoushch@shanghaitech.edu.cn

Abstract

We explored a technique by applying an approximate metric for achieving the flat minima for low-rank matrix recovery, which has better performance on generalization. We have clarified the context and basic principles of the article. Some personal insights are inspired based on this paper and been explained. Some possible new methods with these insights were proposed, and correspondence experimental verification was conducted.

1 Introduction

Recent years have witnessed a proliferation of machine learning and artificial intelligence techniques [1], [2], [3], among which deep neural networks (DNNs) have been particularly prominent. The effectiveness of these methods often hinges on fitting highly overparameterized models that contain more parameters than the training data points, enabling them to achieve near-zero training errors on specific tasks.

Overparameterized models often achieve zero training error; however, while some generalize effectively, others do not. Traditional machine learning methodologies, including those taught in courses, have introduced various techniques to prevent overfitting and enhance model generalization. Techniques such as regularization, batch normalization, and dropout can somewhat overfit mitigate. Other seminal works on the interpretability of model generalization [4], [5], [6] have primarily focused on quantifying stability, capacity, and margin bounds.

Existing literature [7], [8] highlights two intriguing properties: "small norm" and "flat landscape" that correlate with generalization. While the former has been extensively discussed in many traditional machine-learning approaches, our project's paper focuses on the latter. We aim to maintain the predictive capability of models while adhering to the "flat landscape" property through the method of low-rank matrix factorization. This approach reduces the number of model parameters and enhances generalization due to the mathematical properties of the "flat landscape" afforded by low-rank matrices.

2 Related Work

[9] introduces a tailored version of Adam that incorporates a regularization term to harmonize with weight decay and introduces a first moment on weight decay, enhancing its regularization effect. [10] introduces Projected Weight Regularization (PWR) to enhance DNN generalization by

balancing the eigenvalues of weight matrices, demonstrating improved performance on CIFAR10 with VGG16. [11] critiques traditional weight decay in batch-normalized deep neural networks (BN-DNNs) and proposes a Weight Rescaling (WRS) scheme to better regulate weight norms and enhance generalization across various computer vision tasks. These methods primarily focus on the "small norm" property, which is crucial for generalization. By employing various forms of weight decay and regularization techniques, they effectively ensure that the model maintains its fitting capability while keeping the weights' norms small. This indeed enhances the model's generalization capacity to a certain extent and yields good results, yet there is a lack of discussion on the "flat landscape" property.

3 Problem Modeling

Setting the stage, consider a ground truth matrix $M_\natural \in \mathbb{R}^{d_1 \times d_2}$ with rank $r_\natural$. The goal is to recover $M_\natural$ from the observed measurements $b = \mathcal{A}(M_\natural)$ under a linear measurement map $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$. A common approach to this task is through the non-convex optimization problem:

$$\min_{L,R} \quad f(L,R) := \left\| \mathcal{A}\left(LR^\top\right) - b \right\|_2^2 \quad \text{with } L \in \mathbb{R}^{d_1 \times k} \text{ and } R \in \mathbb{R}^{d_2 \times k}. \tag{1}$$

The set of minimizers of f, which we denote by $\mathcal{S}$, consists of all solutions to the equation $\mathcal{A}\left(LR^\top\right) = b$. To model overparameterization, we focus on the rank-overparameterized setting $k \geq r_\natural$; indeed k can be arbitrarily large. The three notions discussed so far can be formally defined for pairs $(L,R) \in \mathcal{S}$ as follows.

- (L,R) is **norm-minimal** if it minimizes over $\mathcal{S}$ the square Frobenius norm $\|L\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2$.
- (L,R) is **balanced** if it satisfies $L^\top L = R^\top R$.
- (L,R) is **flat** if it minimizes over $\mathcal{S}$ the "scaled trace" of the Hessian, $\operatorname{str}\left(D^2 f(L,R)\right)$.

Thus being norm-minimal means that (L,R) is the closest pair from $\mathcal{S}$ to the origin in Frobenius norm. Being balanced amounts to requiring L and R to have the same singular values and right singular vectors. Flat solutions are defined in terms of the "scaled trace" of the bilinear form $D^2 f(L,R)$ defined as

$$\begin{aligned} \operatorname{str}\left(D^2 f(L,R)\right) := & \frac{1}{d_1} \sum_{i \leq d_1, j \in [k]} D^2 f(L,R)\left[\left(e_i e_j^\top, 0_{d_2 \times k}\right)\right] \\ & + \frac{1}{d_2} \sum_{i > d_1, j \in [k]} D^2 f(L,R)\left[\left(0_{d_1 \times k}, e_i e_j^\top\right)\right]. \end{aligned}$$

where e_i and e_j are the unit coordinate vectors in $\mathbb{R}^{d_1+d_2}$ and $\mathbb{R}^k$, respectively. In the square setting $d_1 = d_2 = d$, the scaled trace reduces to the usual trace divided by d.

This paper subsequently proves that under a linear measurement map, the properties of norm-minimal, balanced, and flat in LR^T are approximately equivalent. This indicates that the more tractable norm-minimal and balanced properties can be effectively utilized to ensure the model's flat properties.

$$\mathcal{A}(X) = \left(\langle A_1, X\rangle, \langle A_2, X\rangle \ldots, \langle A_m, X\rangle\right),$$

where $A_i \in \mathbb{R}^{d_1 \times d_2}$ are some matrices. As always, $\mathcal{S}$ denotes the set of solutions to the equation $\mathcal{A}\left(LR^\top\right) = b$. We will make use of the following two "rescaling" matrices:

$$D_1 = \left(\frac{1}{md_2} \sum_{i=1}^m A_i A_i^\top\right)^{\frac{1}{2}} \quad \text{and} \quad D_2 = \left(\frac{1}{md_1} \sum_{i=1}^m A_i^\top A_i\right)^{\frac{1}{2}}.$$

We can then give the problem as:

$$\min_{L,R} \frac{1}{2}\left(\|D_1 L\|_{\mathrm{F}}^2 + \|D_2 R\|_{\mathrm{F}}^2\right) \quad \text{subject to} \quad \mathcal{A}\left(LR^\top\right) = b. \tag{2}$$

This problem is equivalent to minimizing the nuclear norm over rank-constrained matrices:

$$\min_{X\in\mathbb{R}^{d_1\times d_2}:\mathrm{rank}(X)\le k} \|X\|_* \quad \text{subject to} \quad \mathcal{A}\left(D_1^{-1}XD_2^{-1}\right)=b. \tag{3}$$

Therefore, a natural convex relaxation for finding the flattest solution drops the rank constraint:

$$\min_{X\in\mathbb{R}^{d}\times d_2} \|X\|_* \quad \text{subject to} \quad \mathcal{A}\left(D_1^{-1}XD_2^{-1}\right)=b. \tag{4}$$

With the denote above, we can prove the following theorems:

- $$\mathrm{str}\left(D^2 f(L,R)\right) = 2m\left(\|D_1L\|_{\mathrm{F}}^2+\|D_2R\|_{\mathrm{F}}^2\right) \tag{5}$$

  Proof: An elementary computation yields for any (L,R) the expression

  $$D^2f(L,R)[U,V] = 4\left\langle \mathcal{A}\left(LR^\top\right)-b, \mathcal{A}\left(LV^\top+UR^\top\right)\right\rangle + 2\left\|\mathcal{A}\left(LV^\top+UR^\top\right)\right\|_2^2.$$

  $$\Rightarrow \mathrm{str}(\mathcal{H}(L,R)) = \frac{2}{d_1}\sum_{i=1}^{d_1}\sum_{j=1}^{k}\left\|\mathcal{A}\left(e_ie_j^\top R^\top\right)\right\|_2^2+\frac{2}{d_2}\sum_{i=1}^{d_2}\sum_{j=1}^{k}\left\|\mathcal{A}\left(Le_je_i^\top\right)\right\|_2^2.$$

  Let us analyze the second term on the right. Letting $A_{l,i}$ denote the i 'th column of A_l, we compute

  $$\begin{aligned}\sum_{i=1}^{d_2}\sum_{j=1}^{k}\left\|\mathcal{A}\left(Le_je_i^\top\right)\right\|_2^2 &= \sum_{i=1}^{d_2}\sum_{j=1}^{k}\sum_{l=1}^{m}\left\langle A_l, Le_je_i^\top\right\rangle^2 = \sum_{i=1}^{d_2}\sum_{j=1}^{k}\sum_{l=1}^{m}\left\langle A_{l,i}, L_j\right\rangle^2\\ &= \sum_{i=1}^{d_2}\sum_{j=1}^{k}\sum_{l=1}^{m}\left\langle A_{l,i}A_{l,i}^\top, L_jL_j^\top\right\rangle\\ &= \sum_{l=1}^{m}\left\langle\sum_{i=1}^{d_2}A_{l,i}A_{l,i}^\top, \sum_{j=1}^{k}L_jL_j^\top\right\rangle\\ &= \sum_{l=1}^{m}\left\langle A_lA_l^\top, LL^\top\right\rangle = md_2\left\|D_1L\right\|_F^2.\end{aligned}$$

- The optimal values of (1) and (2) are equal.
- If L, R solves (1), then X = $D_1LR^TD_2$ is a minimizer of (2)
- If X = $D_1M_\natural D_2$ is the unique minimizer of the problem (3), then any flat solution (L, R) satisfies $LR^T = M_\natural$

  Proof: Since

  $$\|X\|_* = \min_{X=LR^\top}\|L\|_{\mathrm{F}}\|R\|_{\mathrm{F}} = \min_{X=LR^\top}\frac{1}{2}\left(\|L\|_{\mathrm{F}}^2+\|R\|_{\mathrm{F}}^2\right). \tag{6}$$

  We make a variable substitution $L' = D_1L$ and $R' = D_2R$ and using the equation (6), and then we can prove the three claims above.

- Suppose that there exist constants $\alpha_1,\alpha_2>0$ satisfying $\alpha_1 I \le D_i \le \alpha_2 I$ for each $i\in\{1,2\}$. Define the constant $\kappa := \frac{\alpha_2}{\alpha_1}$. Then any flat solution (L_f,R_f) of (1) satisfies the following properties:

  1. The pair (L_f,R_f) is approximately norm-minimal:

  $$\|L_f\|_F^2+\|R_f\|_F^2 \le \kappa^2\cdot\left(\min_{\mathcal{A}(LR^\top)=b}\|L\|_F^2+\|R\|_F^2\right). \tag{7}$$

  2. The pair (L_f,R_f) is approximately balanced:

  $$\left\|L_f^\top L_f - R_f^\top R_f\right\|_* \le 2\left(\kappa^2-1\right)\|M_\natural\|_*. \tag{8}$$

The proof of the theorem relies on the following simple linear algebraic lemma.
Lemma. Consider two symmetric matrices $Q_1 \in R^{d_1 \times d_1}$ and $Q_2 \in R^{d_2 \times d_2}$. Suppose that there exist constants α_1, $\alpha_2 > 0$ satisfying $\alpha_1 \mathrm{I} \preceq Q_i \preceq \alpha_2 \mathrm{I}$ for each i$\in \{1, 2\}$. Define the constant $\kappa = \frac{\alpha_2}{\alpha_1}$.Then given any matrix X $\in R^{d_1 \times d_2}$, any minimizer (L, R) of the problem

$$\min_{\overline{L},\overline{R}: Q_q \overline{L}\overline{R}^T} \frac{1}{2}(||Q_1\overline{L}||_F^2 + ||Q_2\overline{R}||_F^2), \tag{9}$$

satisfies the inequality:

$$||L^T L - R^T R||_* \leq (1 - \kappa^{-2})(||L||_F^2 + ||R||_F^2) \tag{10}$$

Proof. the pair ($Q_1 L$, $Q_2 R$) is balanced, meaning $L^L Q_1^2 L = R^T Q_2^2 R$ Hence, we may decompose $L^T L - R^T R$ following way:

$$L^T L - R^T R = (L^T L - \frac{L^T Q_1^2 L}{\alpha_2^2}) + (\frac{R^T Q_2^2 R}{\alpha_2^2} - R^T R) \tag{11}$$

We bound the first term on the right as follows,

$$||L^T(I - \frac{1}{\alpha_2^2}Q_1^2)L||_* \leq ||L^T(I - \frac{1}{\alpha_2^2}Q_1^2)||_F ||L||_F \tag{12}$$

$$\leq ||(I - \frac{1}{\alpha_2^2}Q_1^2)||_{op} ||L||_F^2 \tag{13}$$

$$\leq (1 - \kappa^{-2})||L||_F^2 \tag{14}$$

A similar argument yields the inequality, and then the claimed estimate (9) follows immediately We are now ready to prove the Theorem
Proof of Theorem We first prove inequality (7). To this end, for any $(L, R) \in$ S , we successfully estimate:

$$\alpha_1^2(||L_f||_F^2 + ||R_f||_f^2) \leq ||D_1 L_f||_F^2 + ||D_2 R_f||_F^2 \leq ||D_1 L||_F^2 + ||D_2 R||_F^2 \tag{15}$$

$$\leq \alpha_2^2(||L||_F^2 + ||R||_F^2) \tag{16}$$

We next verify (10). To this end, define the matrix X=$D_1 L_f R_f^T D_2$,Then clearly (L_f , R_f) is a minimizer of the problem

$$\min_{\overline{L},\overline{R}: D_1\overline{L}\overline{R}^T D_2 = X} \frac{1}{2}(||D_1\overline{L}||_F^2 + ||D_2\overline{R}||_F^2) \tag{17}$$

Lemma therefore guarantees the estimate

$$||L_f^T L_f - R_f^T R_f||_* \leq (1 - \kappa^{-2})(||L||_F^2 + ||R||_F^2) \tag{18}$$

The already-established estimate ensures

$$(||L||_F^2 + ||R||_F^2) \leq \kappa^2(||L||^2 + ||R||^2) \tag{19}$$

The proof is complete.

4 Method

4.1 Flat minima under RIP conditions: matrix and bilinear sensing

Restricted isometry property(RIP) shows that: a linear $\max \mathcal{A}$:$\mathbb{R}_1^d \times d_2 \to \mathbb{R}^m$ satisfies an ℓ_p/ℓ_2 RIP with parameters (r, δ_1, δ_2)if the estimate

$$\delta_1 \|X\|_{\mathrm{F}} \leq \frac{\|\mathcal{A}(X)\|_p}{m^{1/p}} \leq \delta_2 \|X\|_{\mathrm{F}},$$

holds for all matrices $X \in \mathbb{R}_1^d \times d_2$ with rank at most r.

With these properties, it could be proved that the exact recovery in bilinear sensing, which is:

Suppose that $\mathcal{A}$ is a Gaussian bilinear ensemble. Then for any $\delta \in (0,1)$ there exist numerical constants $c, C, c_1, c_2, c_3, c_4 > 0$ depending only on δ such that in the regime $m \geq cr_\natural(d_1+d_2)$ and $\log(m) \leq c_4 d_{\min}$, with probability at least $1 - c_3\exp(-Cd_{\min})$ any flat solution L_f, R_f a function's flat solutions satisfies $L_f R_f^\top = M_\natural$ and is automatically nearly norm-minimal and nearly balanced:

$$\|L_f\|_F^2 + \|R_f\|_F^2 \leq \left(\frac{1+\delta}{1-\delta}\right)^2 \cdot \left(\min_{\mathcal{A}(LR^\top)=b} \|L\|_F^2 + \|R\|_F^2\right)$$

$$\left\|L_f^\top L_f - R_f^\top R_f\right\|_* \leq 2\left(\left(\frac{1+\delta}{1-\delta}\right)^2 - 1\right)\|M_\natural\|_*.$$

4.2 Matrix completion and approximate recovery

For the matrix completion problem, it could be proven that for each $i \in [d_1]$ and $j \in [d_2]$, let ξ_{ij} be independent Bernoulli random variables with success probability p. The linear map $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ is then defined by the relation

$$[\mathcal{A}(Z)]_{ij} = Z_{ij}\xi_{ij} \quad \text{for any } (i,j) \in [d_1] \times [d_2].$$

The difficulty of recovering the matrix $M_\natural$ is typically measured by an incoherence parameter, which we now define. Given a singular value decomposition $M_\natural = U_\natural \Sigma_\natural V_\natural^\top$ with $\Sigma_\natural \in \mathbb{R}^{r_\natural \times r_\natural}$, the incoherence parameter is the smallest $\mu > 0$ satisfying

$$\|U_\natural\|_{2,\infty} \leq \sqrt{\frac{\mu r_\natural}{d_1}}, \quad \text{and} \quad \|V_\natural\|_{2,\infty} \leq \sqrt{\frac{\mu r_\natural}{d_2}}.$$

Where $\|A\|_2, \infty$ denotes the maximal ℓ_2-norm of the rows of the matrix A. The strategies outlined in the previous section do not directly apply to analyzing flat minima of the matrix completion problem because the linear map $\mathcal{A}(D_1^{-1} \cdot D_2^{-1})$ does not satisfy RIP type conditions. So a weaker recovery result is settled.

5 Our opinion

We want to get the balanced or norm-minimal minima which has been proven to have good generalization ability. Some previous work [12] had already discussed the balance of the loss and the norm-minimal loss. However, we have not yet found any research related to the balanced loss. We analyzed the possible reasons as follows: First, the loss is a matrix instead of a constant number and is hard to characterize. Suppose we use the norm of the matrix. The result is very complex and hard to solve. Second $L^T L = R^t R$ is not a common manifold, so it is also hard to solve(We will discuss it later)So we list the following problem:
At the condition of

$$L^T L = R^T R$$

We want to minimize of

$$||LR^T - M||_F^2$$

So it's a type of manifold optimization, so we use Manopt(a library in Matlab) to solve this problem. However, as $L^T L = R^T R$ isn't a common condition we can't use it strictly. So we add some conditions to make it easier to solve. We find matrices B, C, and D such that: $L = BC$, $R = BD$ and $B^T B = I$, $D^T D = I$,which can ensure $L^T L = R^T R$ and easy to minimize in Manopt. However, this problem isn't equivalent to the original one. So the result isn't very good. But we believe this could be a possible future research direction

6 Conclusion

This paper proposes a highly effective low-rank matrix decomposition strategy that can significantly reduce the number of model parameters while ensuring the model's generalization ability. The authors introduce the concept of flat, combining it with Norm-minima and balance, and provide detailed mathematical proofs. We believe there are many areas in this paper that warrant further research.

References

[1] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Mia Chen, Dehao Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. Gpipe: Efficient training of giant neural networks using pipeline parallelism. 2019.

[2] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

[3] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[4] B PL. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

[5] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[6] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

[7] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

[8] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[9] Xixi Jia, Xiangchu Feng, Hongwei Yong, and Deyu Meng. Weight decay with tailored adam on scale-invariant weights for better generalization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5):6936–6947, 2024.

[10] Guoqiang Zhang, Kenta Niwa, and W. Bastiaan Kleijn. Projected weight regularization to improve neural network generalization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4242–4246, 2020.

[11] Ziquan Liu, Yufei Cui, Jia Wan, Yu Mao, and Antoni B. Chan. Weight rescaling: Effective and robust regularization for deep neural networks with batch normalization, 2022.

[12] Ching pei Lee, Ling Liang, Tianyun Tang, and Kim-Chuan Toh. Accelerating nuclear-norm regularized low-rank matrix optimization through burer-monteiro decomposition, 2023.