

GRADVIS

VISUALIZATION AND SECOND ORDER ANALYSIS
OF OPTIMIZATION SURFACES

REF

- ▶ Chatzimichailidis, Avraam, et al. "GradVis: Visualization and Second Order Analysis of Optimization Surfaces during the Training of Deep Neural Networks." 2019 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC). IEEE, 2019.
- ▶ R-operator, <https://cswhjiang.github.io/2015/10/13/Roperator/>

ABSTRACT

ABSTRACT

- ▶ DNN의 학습은
- ▶ 고차원의 볼록하지 않은(non-convex) 최적화 문제의 해결
- ▶ 확률적 경사 하강법

ABSTRACT

- ▶ 최적화 표면(optimization surface, loss landscape)의
 - ▶ 수렴성과 일반화 보장 분석이 필요
 - ▶ 여러 연구에서 시각화 및 분석 기법들 제시
- ▶ 기존 시각화 및 분석 기법들은 고비용
- ▶ 큰 네트워크에 적용하기 어려움

ABSTRACT

- ▶ GradVis 툴박스
- ▶ 효율적이고 확장가능한 시각화 및 분석 라이브러리
- ▶ 효율적인 수학적 기법에 기반
- ▶ novel한 병렬화 스킴

ABSTRACT

- ▶ 최적화 표면 및 궤적의 2D와 3D 사영 가능
- ▶ 큰 네트워크에 대한 고해상도 이차(2차) 경사 정보를 얻을 수 있음
- ▶ Second Order

INTRODUCTION

INTRODUCTION

- ▶ 대부분의 경우 확률적 경사 하강법으로 지역적 미니마에 도달 가능
- ▶ Stochastic Gradient Descent (SGD)
- ▶ 여러 (이론적) 논의가 이어지는 중

INTRODUCTION

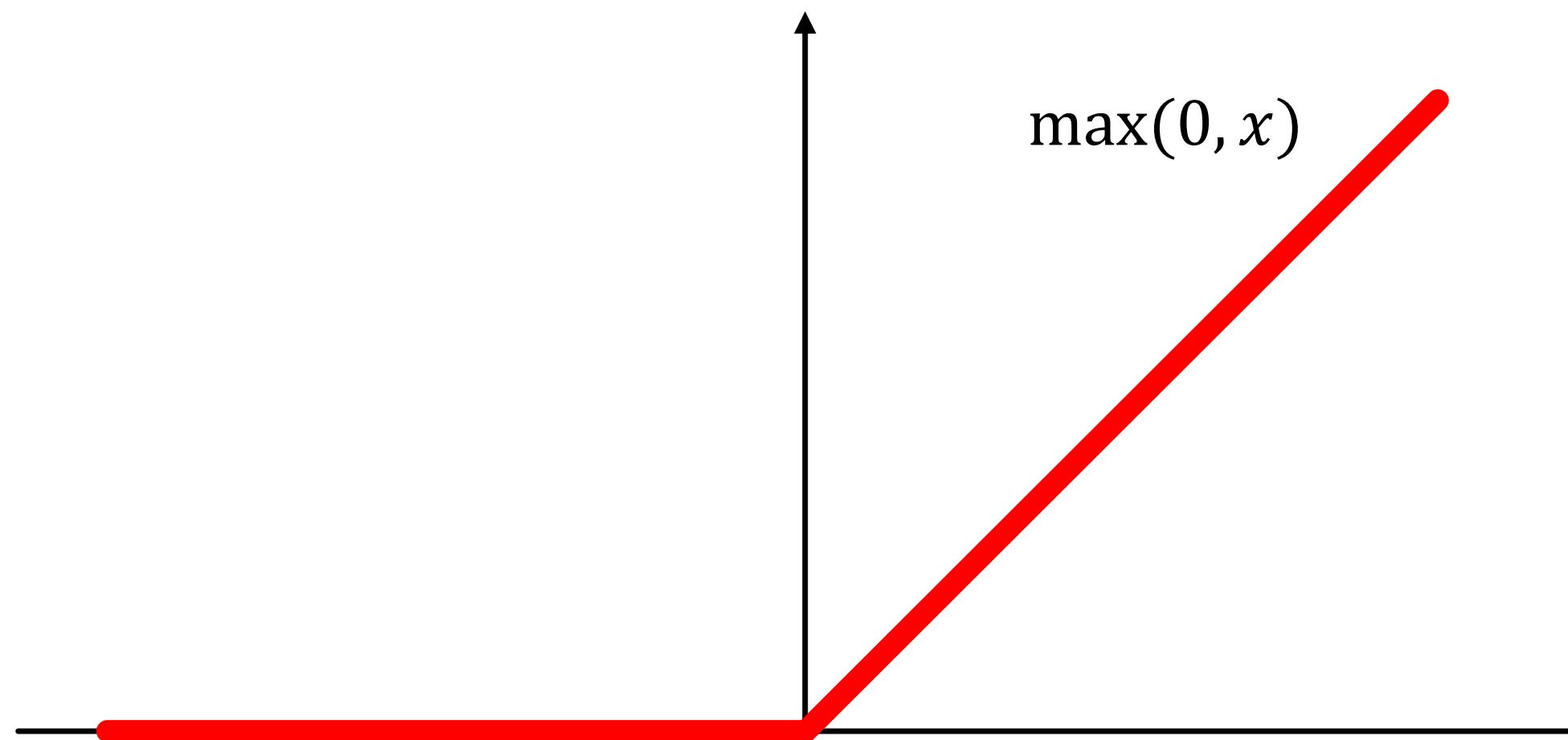
- ▶ 가정: 손실 표면의 지역적으로 넓은 미니마가 보다 나은 일반화 정도를 가짐
- ▶ 사실일까?
- ▶ 어떤 논문에서는 그렇다, 어떤 논문에서는 그렇지 않다를 보임

INTRODUCTION

- ▶ 이러한 모순은 각 논문마다 평평한 정도를 해석하는 것이 다르기 때문
- ▶ 엄밀하게는 방법에 따라 달라지기 때문
- ▶ 손실 함수의 값에 변화 없이 신경망의 가중치를 재-파라미터화 할 수 있음
- ▶ 스케일 불변성(scale invariance)

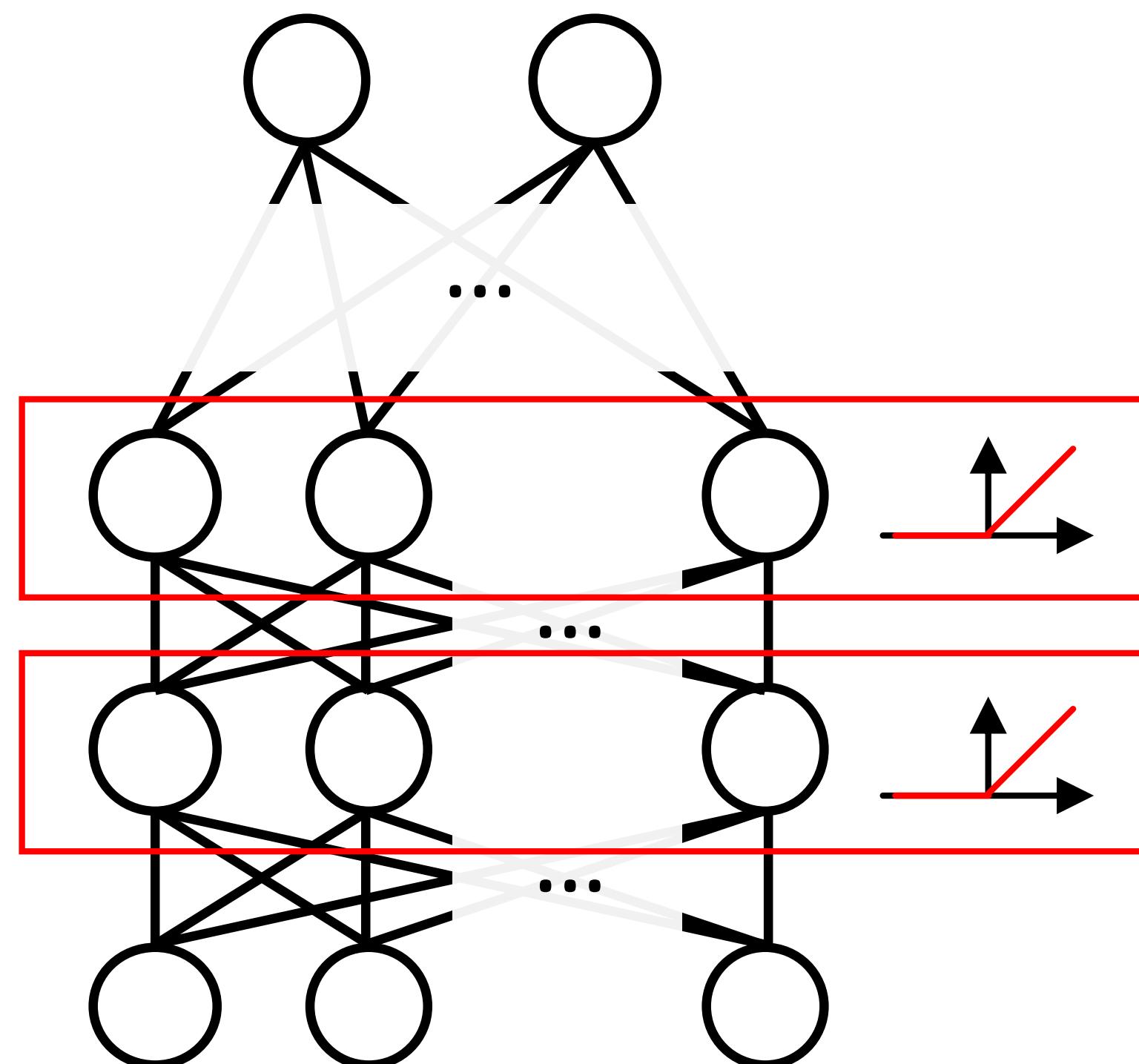
SCALE INVARIANCE

- ▶ ReLU 비선형 활성화함수의 사례:



SCALE INVARIANCE

- ▶ ReLU 비선형 활성화함수의 사례:
 - ▶ 네트워크는 변하지 않음



$$w_{j1}, w_{j2}, \dots \div 10$$
$$w_{i1}, w_{i2}, \dots \times 10$$

SCALE INVARIANCE

- ▶ 스케일 불변성으로 인해 plot 간 유의미한 비교가 힘들어짐
- ▶ 큰 가중치를 가진 신경망은 부드럽고 천천히 변화하는 손실 함수를 가짐
 - ▶ 한 단위(unit)의 혼란(perturbing)은 네트워크 성능에 매우 작은 영향을 끼침
 - ▶ 가중치가 훨씬 더 큰 스케일의 세상에 존재하기 때문
 - ▶ 반면 작은 가중치를 가진 신경망은 같은 변화에도 난리가 남

INTRODUCTION

- ▶ 이러한 문제를 다루기 위해서는 이론적인 영감이 필요함
- ▶ 본 강의에서는 디테일 생략
- ▶ 원 논문 및 references 참고

RELATED WORK

RELATED WORK: LOSS SURFACE VISUALIZATION

- ▶ Li, Hao, et al. "Visualizing the loss landscape of neural nets." Advances in Neural Information Processing Systems. 2018.
- ▶ 주 contribution은 Filter-wise normalization
 - ▶ 가중치의 불변성을 잘 취급
 - ▶ 또 하나는 PCA

RELATED WORK: LOSS SURFACE VISUALIZATION

- ▶ 그러나 PCA의 복잡도는 샘플의 수와 차원의 세 제곱에 비례
- ▶ 많은 양의 메모리를 필요로 함
- ▶ 큰 네트워크에서는 활용하기 어려움

RELATED WORK: LOSS SURFACE VISUALIZATION

- ▶ 또한, 고차원의 랜덤 워크에 대한 PCA는 항상 리사주(Lissajous) 궤적을 그림
- ▶ 물리학에서, 서로 수직 방향으로 진동하는 단진동을 합성하였을 때,
- ▶ 그 궤도가 그리는 도형
- ▶ 따라서 무의미하다는 비판이 있음

RELATED WORK: SECOND ORDER PROPERTIES OF THE LOSS SURFACE

- ▶ 헤세(헤시안) 행렬

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

RELATED WORK: SECOND ORDER PROPERTIES OF THE LOSS SURFACE

- ▶ 헤세 행렬
- ▶ 함수의 이차 미분을 다룸
- ▶ 곡률과 관련 있음

RELATED WORK: SECOND ORDER PROPERTIES OF THE LOSS SURFACE

- ▶ 헤세 행렬의 고유값 (eigenvalue of the Hessian)
- ▶ 임계점이 존재하는 경우
 - ▶ 그 임계점에서 대응되는 고유벡터에 대해
 - ▶ 헤세 행렬의 고유값들이 모두 양수면 그 임계점은 극소점
 - ▶ 헤세 행렬의 고유값들이 모두 음수면 그 임계점은 극대점
 - ▶ 양수도 있고 음수도 있으면 안장점
 - ▶ 크기는 경사의 정도

RELATED WORK: SECOND ORDER PROPERTIES OF THE LOSS SURFACE

- ▶ 신경망의 헤세 행렬의 고유값 계산은 $O(N^3)$ 의 복잡도
- ▶ 보통 신경망이 10^6 에서 10^8 개의 파라미터를 가지므로
- ▶ 실질적으로 계산하기는 불가능
- ▶ 최적화 단계 동안 10^5 개의 점에서 계산하므로 더 어려움

RELATED WORK: SECOND ORDER PROPERTIES OF THE LOSS SURFACE

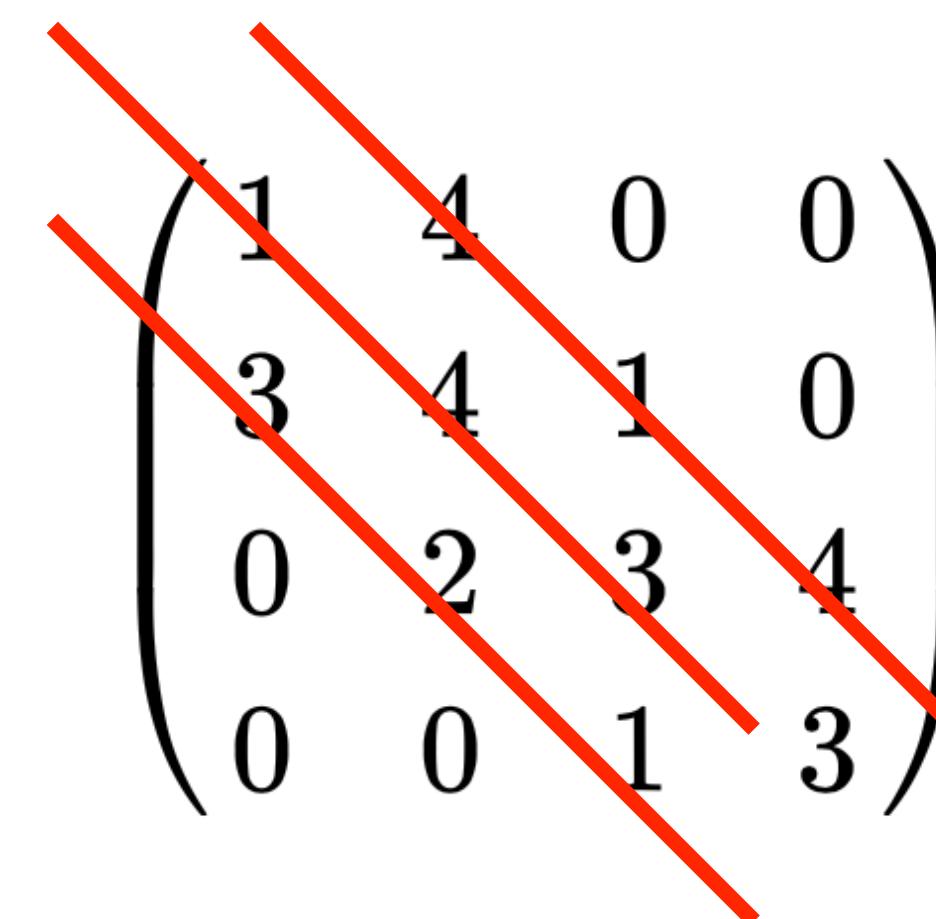
- ▶ 모든 고유값을 다 계산하기보다는, 트릭을 사용해야 함
- ▶ R-operator
- ▶ 란초스(Lanczos) 알고리즘

RELATED WORK: SECOND ORDER PROPERTIES OF THE LOSS SURFACE

- ▶ R-operator
 - ▶ 해세 행렬 · 벡터 곱을 효율적으로 계산
 - ▶ $O(N^2)$ 대신에 $O(N)$ 만으로
 - ▶ 해세 행렬을 저장할 필요도 없음
 - ▶ 어떻게 가능한가? 직접 계산하지 않고 역전파를 활용하기 때문

RELATED WORK: SECOND ORDER PROPERTIES OF THE LOSS SURFACE

- ▶ 란초스(Lanczos) 알고리즘
 - ▶ 고유값 계산에 헤세-벡터 곱만을 필요로 함
 - ▶ 모든 고유값 스펙트럼을 m 회의 란초스 반복 과정만으로 근사할 수 있음
 - ▶ 결과를 $m \times m$ tridiagonal 행렬로 대각화

$$\begin{pmatrix} 1 & 4 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ 0 & 2 & 3 & 4 \\ 0 & 0 & 1 & 3 \end{pmatrix}$$


RELATED WORK: SECOND ORDER PROPERTIES OF THE LOSS SURFACE

- ▶ 서로 다른 시작 벡터로 k 회 반복하면
 - ▶ 이들 결과로부터 가우시안(Gaussian)을 통해
 - ▶ 전체 고유값 스펙트럼을 높은 확률로 근사
- ▶ 결과적으로 $O(Nmk)$

CONTRIBUTION

CONTRIBUTION

- ▶ GradVis
 - ▶ <https://github.com/cc-hpc-itwm/GradVis>
 - ▶ 신경망 모델의 궤적 시각화 및 헤세 행렬의 고유값을 구할 수 있는 툴
 - ▶ novel한 병렬화

CONTRIBUTION

- ▶ GradVis
 - ▶ PCA 기반 시각화의 한계 극복
 - ▶ 고유값 밀도 스펙트럼을 신경망 학습의 매 반복마다 계산해
 - ▶ 고해상도 웨이브 비디오 및
 - ▶ 이차 미분 정보를 뽑아냄

METHODS

STOCHASTIC LANCZOS QUADRATURE ALGORITHM

- ▶ 확률적 랜초스 구적 알고리즘 (Stochastic Lanczos quadrature)
- ▶ 매우 큰 행렬의 고유값 밀도를 근사하는 기법

STOCHASTIC LANCZOS QUADRATURE ALGORITHM

- ▶ 고유값 밀도 스펙트럼 $\phi(t)$

$$\phi(t) = \frac{1}{N} \sum_{i=1}^N \delta(t - \lambda_i)$$

- ▶ N: 네트워크의 파라미터 개수
- ▶ λ_i : 헤시 행렬의 i-번째 고유값

STOCHASTIC LANCZOS QUADRATURE ALGORITHM

- ▶ 고유값 밀도 스펙트럼을 가우시안 함수의 합으로 근사 가능

$$\phi(t) = \frac{1}{N} \sum_{i=1}^N \delta(t - \lambda_i) \quad \longrightarrow \quad \phi_\sigma(t) = \frac{1}{N} \sum_{i=1}^N f(\lambda_i, t, \sigma^2)$$

- ▶ 가우시안 함수

$$f(\lambda_i, t, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t - \lambda_i)^2}{2\sigma^2}\right)$$

STOCHASTIC LANCZOS QUADRATURE ALGORITHM

- ▶ 랜초스 알고리즘과 전(full) 재대각화를 사용
- ▶ 헤세 행렬의 고유값과 고유벡터를 계산하기 위해서
- ▶ 고유벡터 간 직교성을 보장하기 위해서

STOCHASTIC LANCZOS QUADRATURE ALGORITHM

- ▶ 란초스 알고리즘이 Tridiagonal 행렬을 반환하므로
- ▶ 그 행렬을 대각화:

▶

$$T = U \Lambda U^T$$

고유값
 l_1 l_2
고유벡터
 $w_1 w_2$ $w_1 w_2$

- ▶ 결과로 나온 고유값과 고유벡터로 진짜 고유값 밀도 스펙트럼을 근사

STOCHASTIC LANCZOS QUADRATURE ALGORITHM

- ▶ 결과로 나온 고유값과 고유벡터로 진짜 고유값 밀도 스펙트럼을 근사

$$\hat{\phi}^{(v_i)}(t) = \sum_{i=1}^m \omega_i f(l_i, t, \sigma^2) \quad \longrightarrow \quad \hat{\phi}_\sigma(t) = \frac{1}{k} \sum_{i=1}^k \hat{\phi}^{(v_i)}(t)$$

- ▶ k 개의 가우시안 벡터

PARALLELIZATION

PARALLELIZATION

- ▶ Data-Parallel Visualization
 - ▶ 평가할 grid를 분할해 다른 worker에게 분배함으로써
 - ▶ 자명하게 병렬화 가능

PARALLELIZATION

- ▶ Data-Parallel Lanczos
 - ▶ 란초스 알고리즘은 고유값 계산을 위해
 - ▶ 전체 헤세 행렬 대신
 - ▶ 헤세-벡터 곱만을 요구
 - ▶ 병렬화 가능
 - ▶ 각자 R-operator로 계산

PARALLELIZATION

- ▶ Novel Iteration-Parallel Lanczos
 - ▶ 란초스 기법을 병렬화 하는, 저자들이 제안하는 또 다른 기법
 - ▶ 확률적 란초스 구적 알고리즘 응용

PARALLELIZATION

- ▶ Novel Iteration-Parallel Lanczos
 - ▶ 확률적 랜초스 구적 알고리즘을
 - ▶ 서로 다른 초기값(가우시안 벡터)으로 병렬 수행 후
 - ▶ 병합

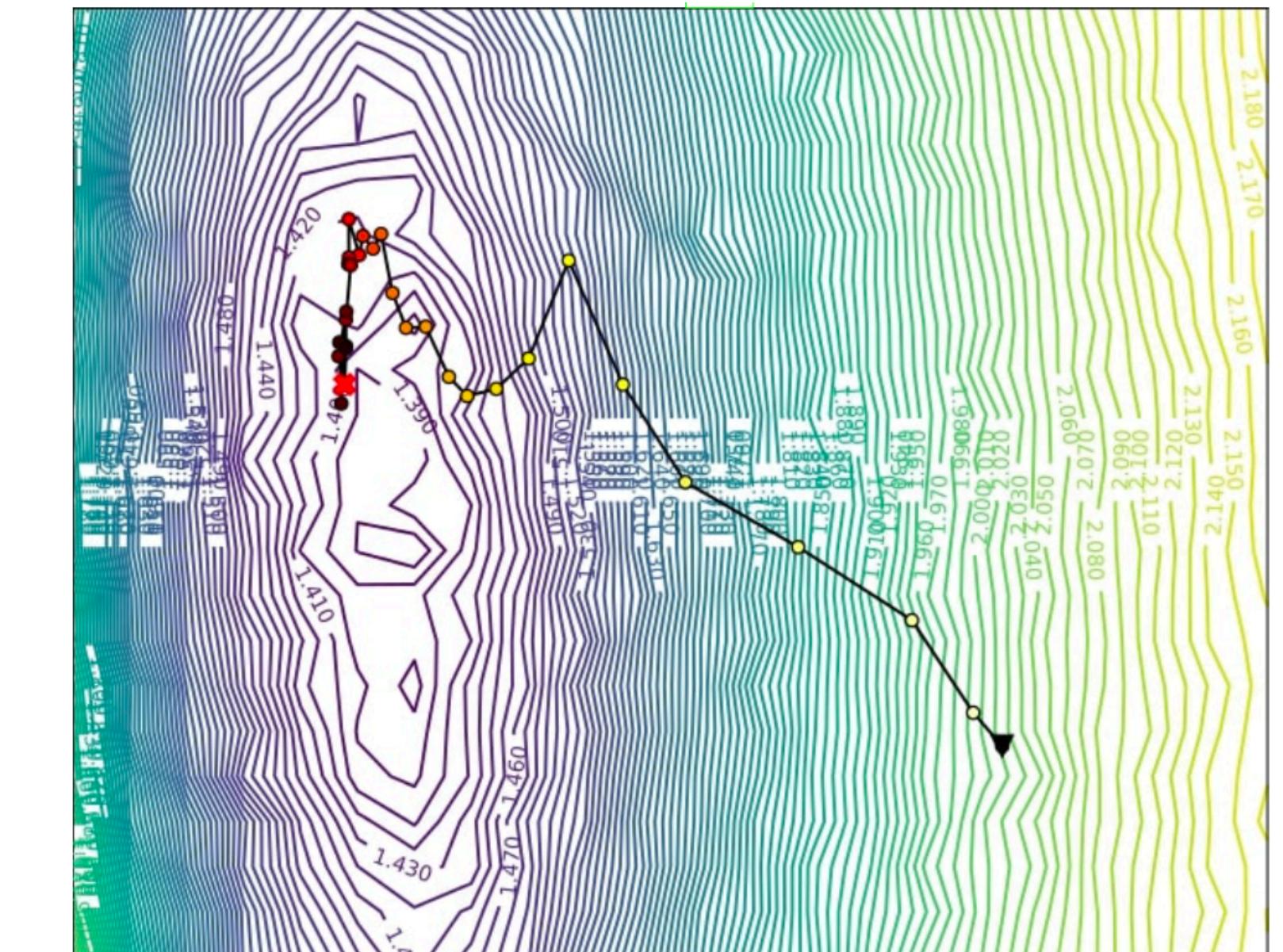
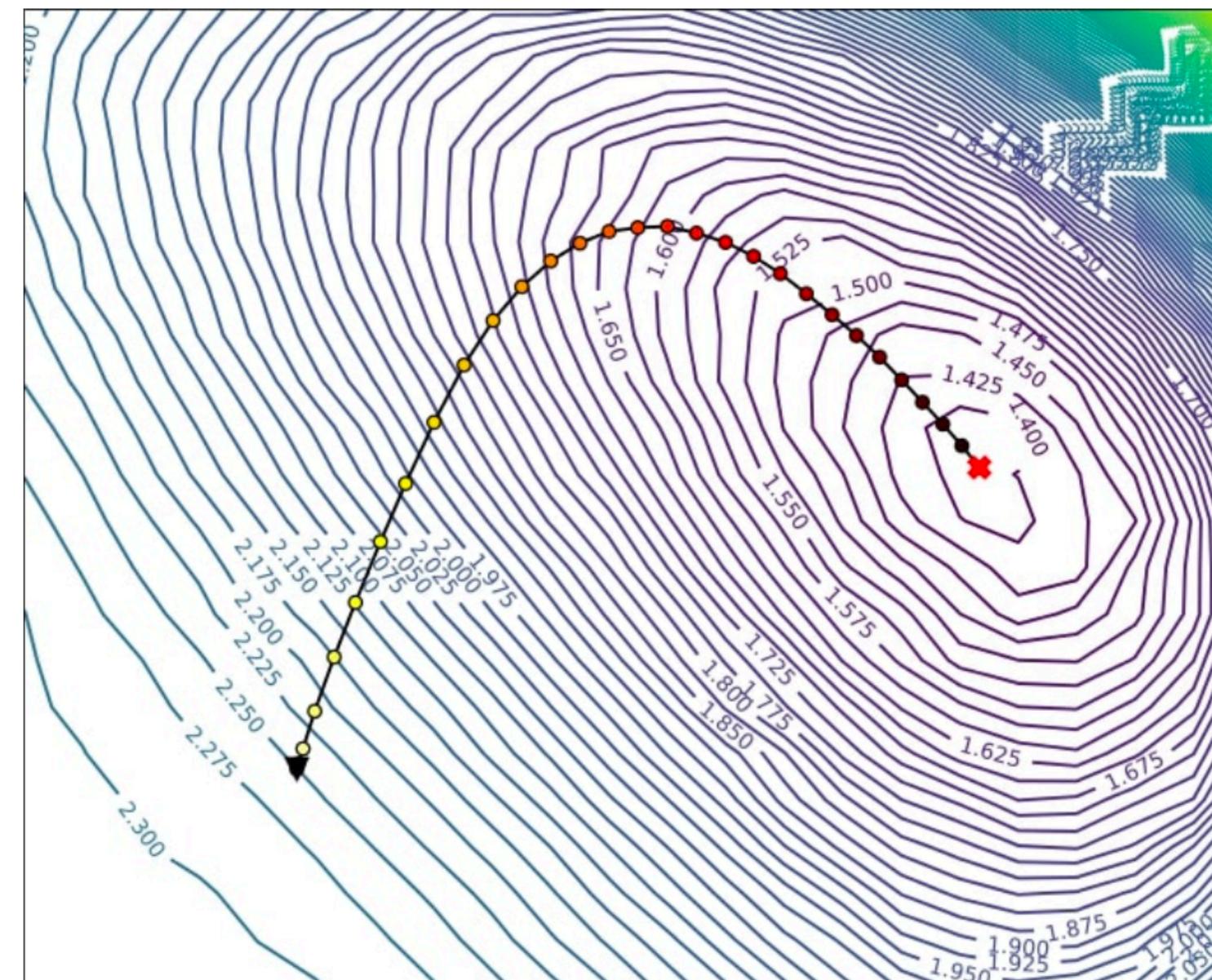
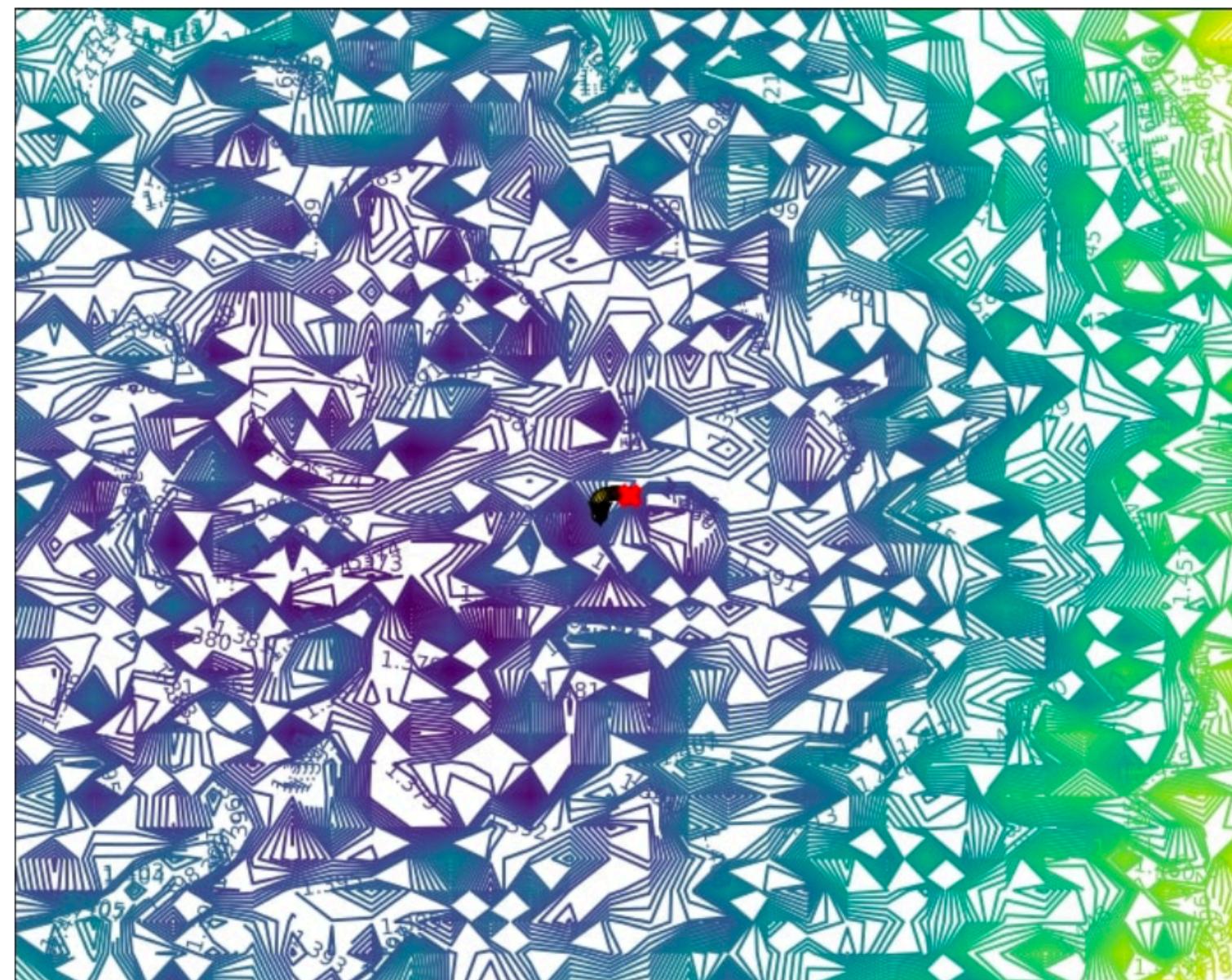
FINDING DIRECTIONS

FINDING DIRECTIONS FOR VISUALIZATION

- ▶ PCA 방법에 문제가 있으므로
- ▶ 서로 다른 고유벡터를 통한 손실 함수 및 궤적의 시각화 방법 제안

FINDING DIRECTIONS FOR VISUALIZATION

- ▶ 고유벡터(가운데) 방법을 랜덤 방향 벡터(좌)와 PCA로 구한 방향 벡터(우)와 비교



- ▶ LeNet, CIFAR10

FINDING DIRECTIONS FOR VISUALIZATION

- ▶ 고유벡터에 따른 방향 벡터
- ▶ 궤적이 경사에 따라 어떻게 움직이는지
- ▶ 저점으로 어떤 속도로 이동하는지
- ▶ 확인 가능

FINDING DIRECTIONS FOR VISUALIZATION

- ▶ 어느 고유벡터에 대응하는 고유값은 음수가 되거나 0이 되는 등
- ▶ 이 방향에 대해서는 훈련에 “실패”하는 것 처럼 보임
- ▶ 훈련의 어느 시점에 어느 방향에 대해 무엇을 하는지 모니터링 가능

EIGENVALUE DENSITY

LOSS LANDSCAPE WITH TRAJECTORY AND EIGENVALUE DENSITY

- ▶ 손실 함수와 궤적, 고유값 밀도를 함께 살펴보면
- ▶ 더 많은 인사이트를 얻을 수 있음
 - ▶ 본 강의에서는 디테일 생략
 - ▶ 원 논문 및 references 참고

LOSS LANDSCAPE WITH TRAJECTORY AND EIGENVALUE DENSITY

- ▶ 시각화 자료
 - ▶ 헤세 행렬에서 가장 큰 고유값을 가지는 고유벡터 두 개로 시각화
- ▶ 고유값 밀도
 - ▶ 확률적 랜초스 구적 알고리즘 $k=10$ 회 반복
 - ▶ 랜초스 알고리즘 $m=80$ 회 반복

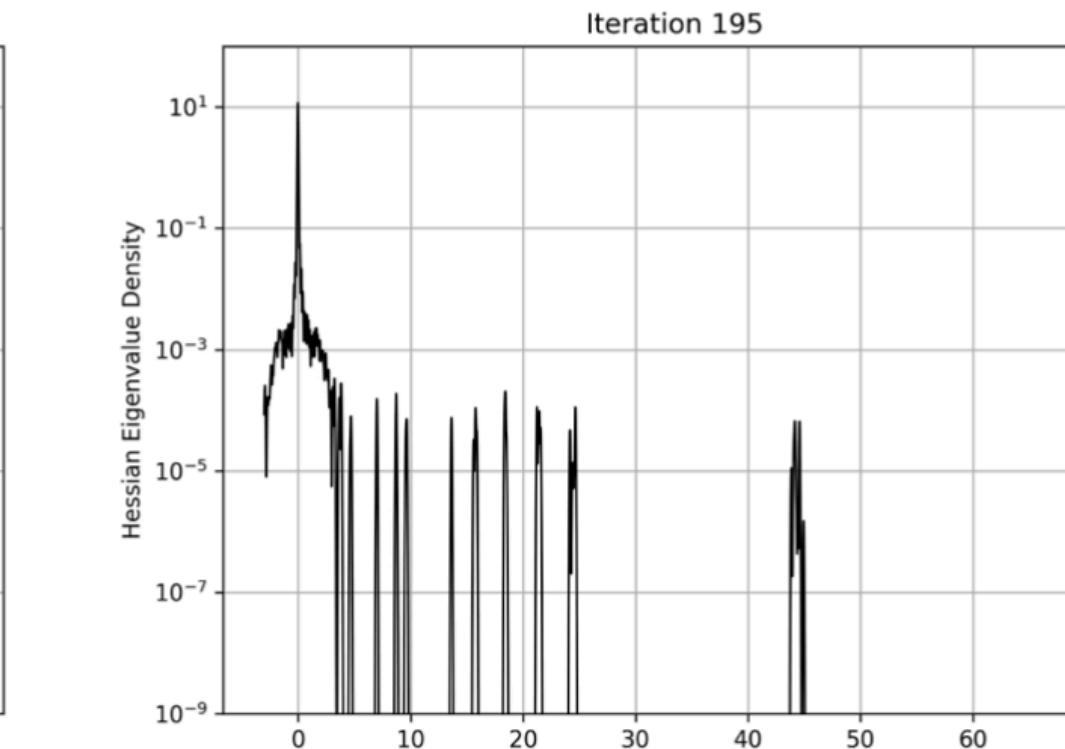
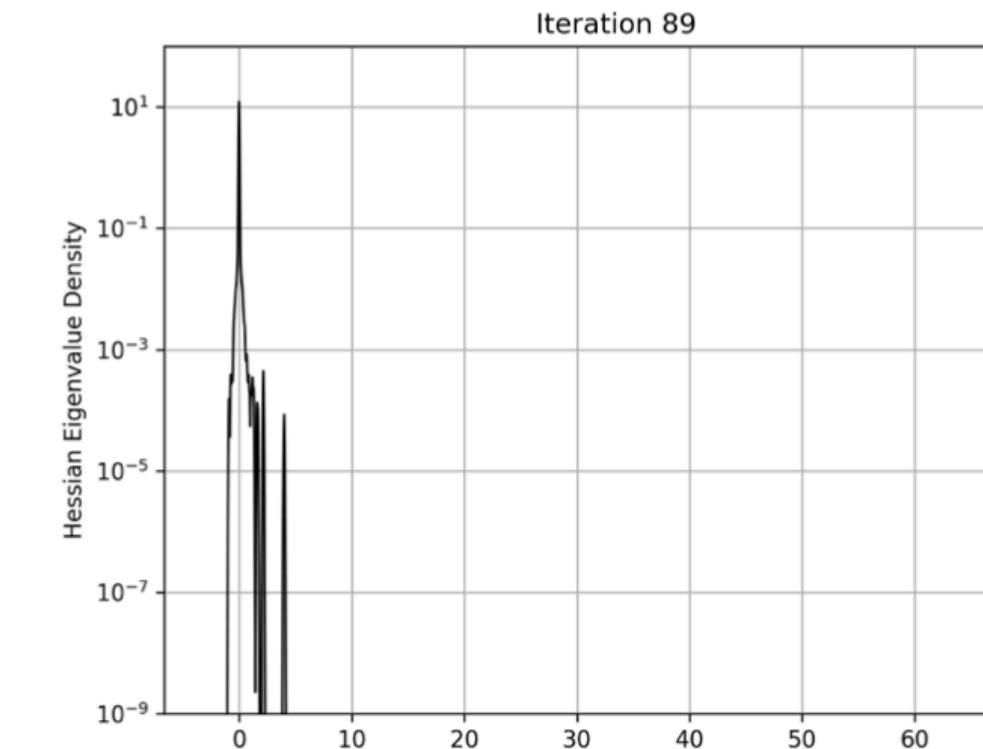
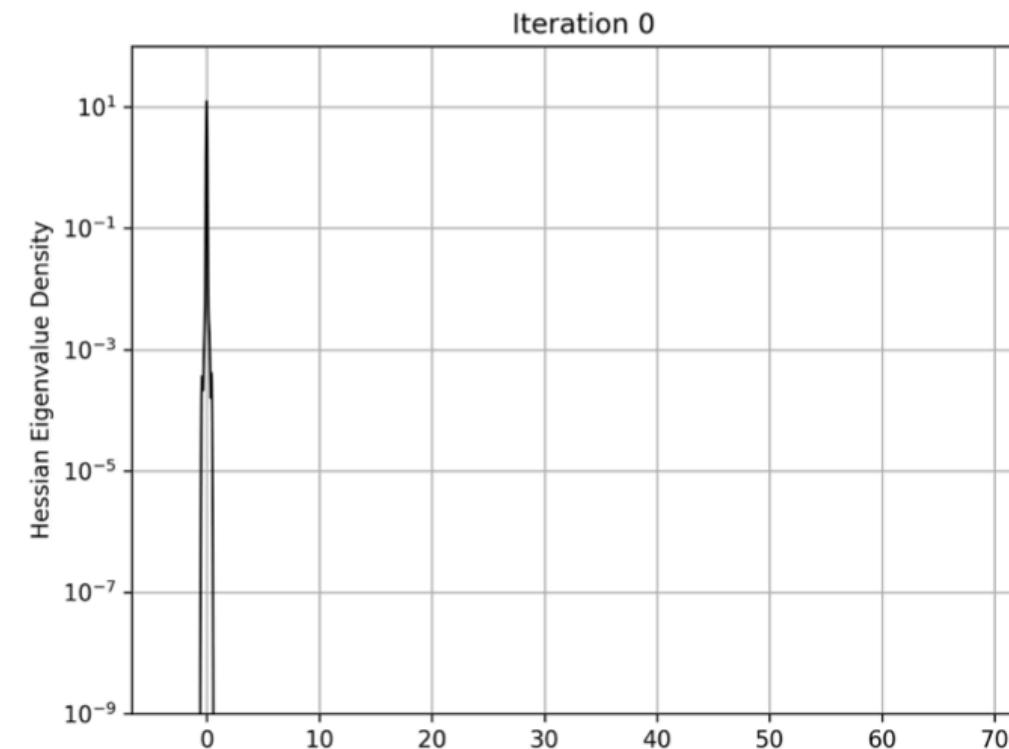
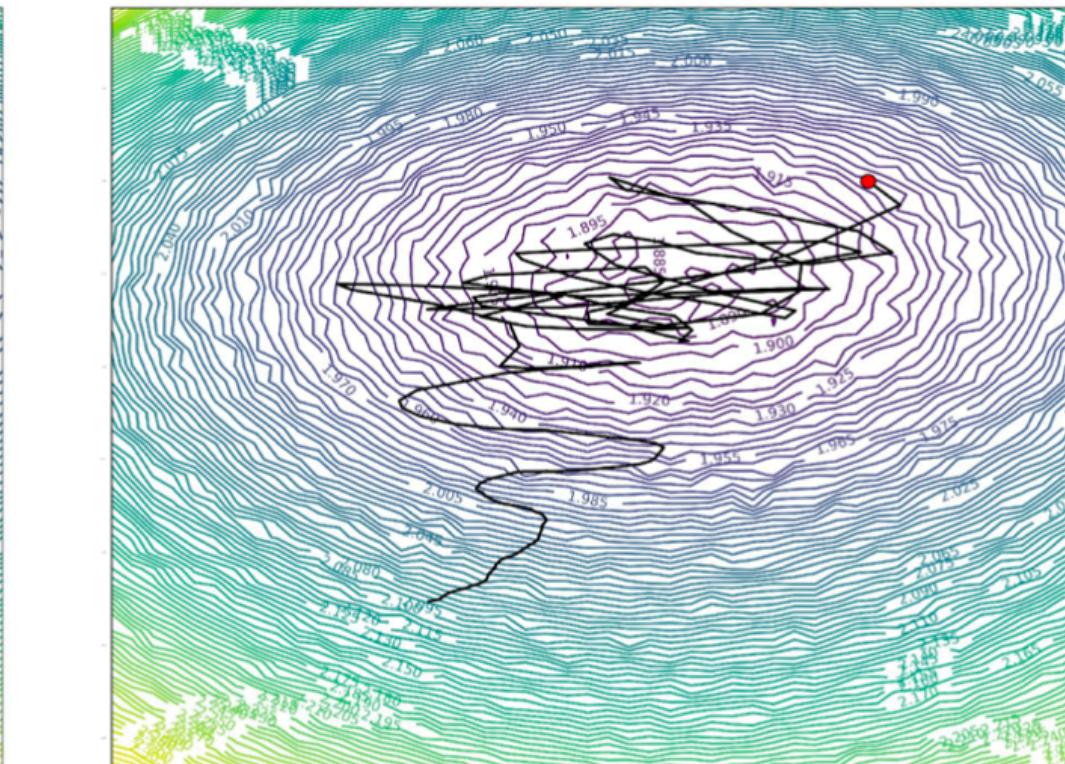
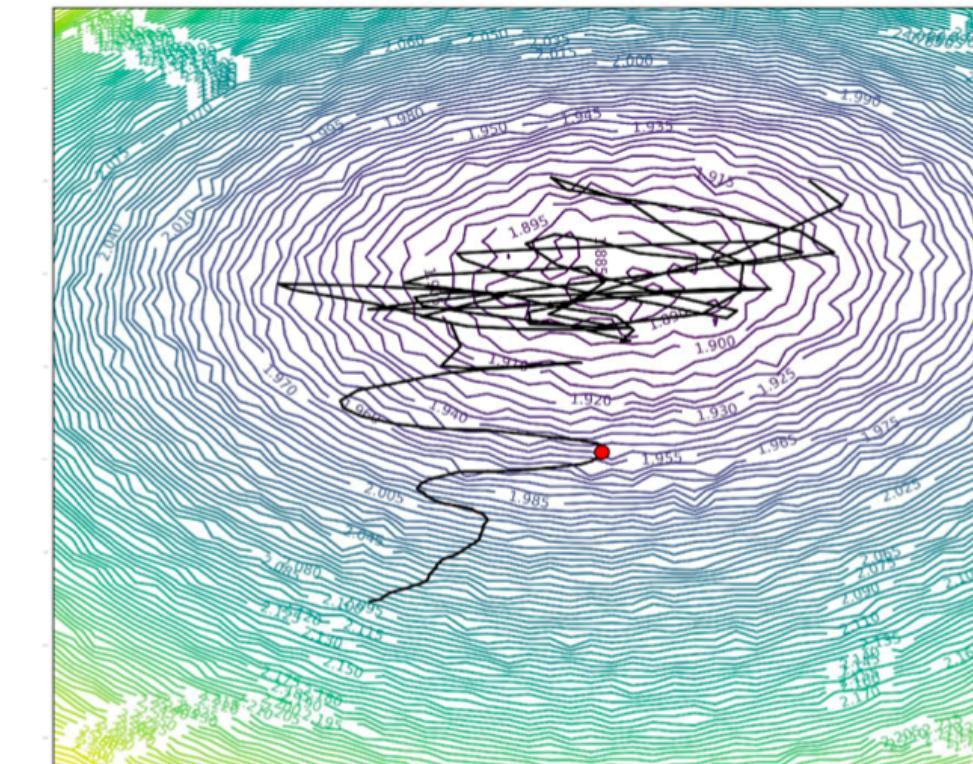
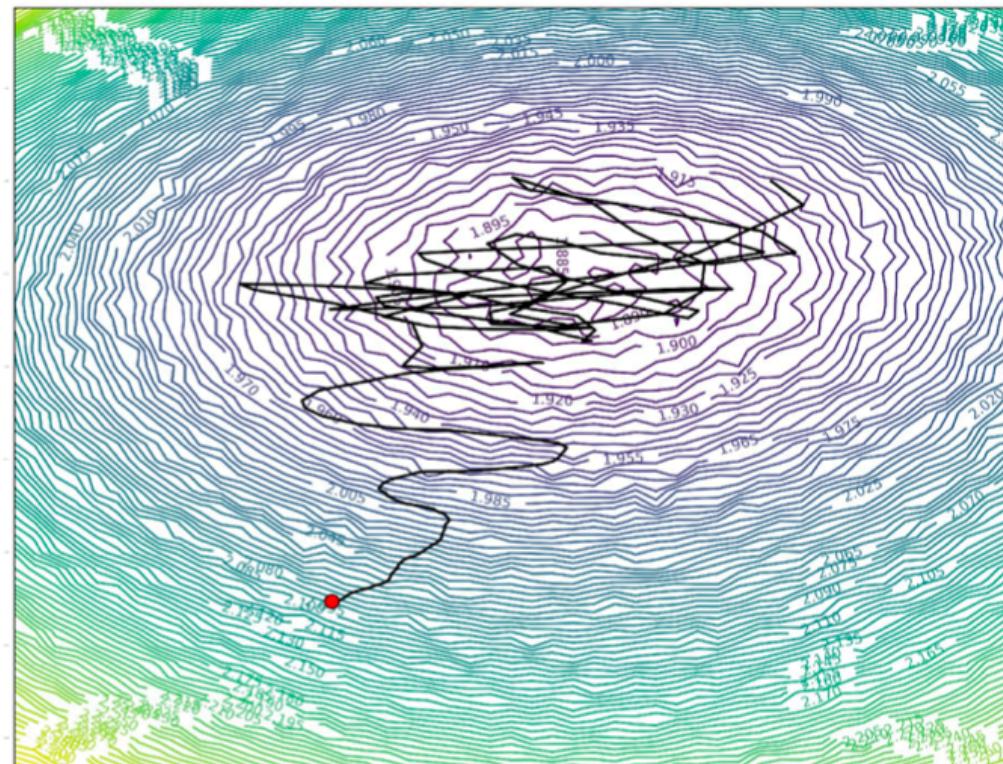
LOSS LANDSCAPE WITH TRAJECTORY AND EIGENVALUE DENSITY

- ▶ LeNet, CIFAR10
- ▶ 배치 사이즈 256
- ▶ 학습률 0.001
- ▶ SGD 사용
- ▶ 모멘텀 0.9
- ▶ 1 에폭 196회 반복

LOSS LANDSCAPE WITH TRAJECTORY AND EIGENVALUE DENSITY

▶ (좌) 0회 / (가운데) 89회 / (우) 195회 반복에서의 자료

▶ <https://youtu.be/OAKSjp-SHlo>



CONCLUSION

CONCLUSION

- ▶ GradVis 툴박스
- ▶ 신경망 학습 전반에 걸친
 - ▶ 손실 공간의 고해상도 시각화
 - ▶ 경사 궤적
 - ▶ 헤세 함수의 고유값 분포
 - ▶ 추적이 가능

CONCLUSION

- ▶ 랜초스 알고리즘의 확장성 높은 병렬화
- ▶ 더 다양한, 더 큰 네트워크를 더 많은 데이터셋에 대해 평가함이 다음 목표
 - ▶ ResNet
 - ▶ ImageNet 등

GRADVIS

VISUALIZATION AND SECOND ORDER ANALYSIS
OF OPTIMIZATION SURFACES