

ReLU 기반 모델에서의 대각 Hessian을 활용한 전역 수렴 보장형 2차 최적화 방법

윤인성

Abstract

본 논문에서는 ReLU 계열 신경망에서 Hessian 대각값의 음수 절댓값이 δ 로 상한 되며 이 δ 가 양의 상수 M 에 비해 매우 작다고 가정($-\delta \leq H_{ii} \leq M$, $\delta \ll M$)하고, 이를 활용한 경량 2차 최적화 알고리즘을 제안한다. 제안 기법은 먼저 대각 성분만을 사용하여 근사 Hessian을 구성하고, 여기에 땡평 항 λI 를 추가하여 strict PD(Positive Definite) 조건을 확보한다. 이를 통해 전형적인 2차 최적화법의 계산량을 크게 줄이면 서도, 이론적으로 전역 선형 수렴성을 보장함을 수학적으로 증명하였다. 나아가, 다양한 규모의 다층 퍼셉트론(Multilayer Perceptron) 구조에 제안 알고리즘을 적용하여, 기존의 1차 최적화 기법들과 비교·분석한 결과 유의미한 성능 향상을 확인하였다.

Contents

1 서론	2
2 관련 연구	3
3 문제 정의 및 수학적 배경	4
3.1 최적화 문제 공식화	4
3.2 모델 구조 및 기호 정의	5
3.3 핵심 가정	5
3.4 연구 질문 및 기여	6
4 제안 방법	6
4.1 알고리즘 및 모델 구조	7
4.2 2차 미분 기반 학습률 계산	8
4.3 동적 러닝레이트 및 모멘텀 업데이트	9
4.4 전체 알고리즘 흐름	11
5 이론적 보장 및 설계 직관	13
5.1 이론적 보장	13
5.2 설계 아이디어 및 직관	14
6 실험 설정	15
6.1 모델 아키텍처	15
6.2 실험 단계 구분	15
6.3 실험 환경	16
6.4 비교군 옵티마이저 및 하이퍼파라미터	17
6.5 데이터셋 및 평가 지표	17

7 실험 결과	18
7.1 순수 MLP 실험 (정규화 OFF)	18
7.2 정규화 기법 실험 (Custom: Dropout 0.004 + Label Smoothing 0.025 Base-line: BatchNorm ON + Dropout + LS)	19
7.2.1 데이터셋별 열위 사례	20
7.3 실측 학습 시간 비교 및 GPU 활용도 분석	21
7.4 하이퍼파라미터 민감도 기반 최종 값 선택	21
8 결론 및 향후 과제	21
8.1 결론 요약	21
8.2 한계점	22
8.3 향후 과제	22
A Newton 방법의 전역 수렴 증명	24
B 다변수 라그랑주 나머지항 정리	26
C 일변수 라그랑주 나머지항 정리	28
D 데이터셋 출처	33
E 실험 결과 (250 epochs) 부록	34
E.1 Hidden7 No Regulation	34
E.2 Hidden7 Regulation	37
E.3 Hidden10 No Regulation	40
E.4 Hidden10 Regulation	43
E.5 하이퍼파라미터 민감도 분석 히트맵	46

1 서론

딥러닝 최적화 분야에서 2차 정보(Hessian)를 활용한 접근은 빠른 수렴성과 높은 안정성을 보장하는 이론적 장점을 갖는다. 그러나 일반적인 2차 최적화 방법은 계산량이 크고, 실용적인 구현이 어려워 대부분의 실전 모델에서는 1차 정보 기반의 Adam, Momentum 등의 알고리즘이 널리 사용되고 있다. 본 논문에서는 2차 최적화 기법들이 이론상 빠름에도 불구하고 대규모 병렬처리 환경에서의 확장성에 한계가 있음을 확인하였으며, 이에 대비하여 본 연구에서 제안하는 경량 2차 정보 기반 커스텀 옵티마이저의 병렬 효율성을 집중적으로 평가하기 위해 기존의 대표적 1차 최적화 기법(Adam, AdamW, AdaBelief 등)과의 직접 비교만을 수행한다.

본 논문의 기여는 다음과 같다.

- **Diagonal-only 2nd-order update:** ReLU/Leaky-ReLU에서 관측된 δ -bounded Hessian 대각 성분만으로 strict PD를 확보하는 경량 2차 최적화 기법을 제안한다.
- **Global convergence proof:** 변환 함수 기반 탬핑과 모멘텀 보정을 결합하여 전역 선형 수렴성을 이론적으로 보장한다.
- **Empirical validation:** 다양한 MLP 구조에서 제안 기법이 Adam 계열 대비 수렴 속도 및 성능 면에서 일관된 개선을 보임을 실험적으로 입증한다.

연구 범위 및 제외 사항. 본 연구에서는 해시안 행렬의 대각 성분만을 활용한 경량 2차 최적화 기법에 중점을 두었습니다. 따라서 Shampoo, K-FAC, AdaHessian 등과 같은 복잡한 2차 최적화 기법들과의 비교 실험은 본 논문의 실험 범위에서 제외하였습니다. 이러한 고급 기법들과의 종합적인 벤치마킹은 후속 연구에서 수행할 예정입니다.

논문의 구성은 다음과 같다.

- 제2장에서는 관련 연구들을 간단히 정리하고, 본 연구의 차별점을 설명한다.
- 제3장에서는 문제 정의 및 수학적 가정을 제시한다.
- 제4장에서는 제안하는 최적화 알고리즘의 수식 및 구조를 상세히 기술한다.
- 제5장에서는 제안한 최적화 알고리즘의 전역 선형 수렴성을 수학적으로 증명하고, 변환 함수 기반 평탄화 및 local minima 탈출 원리 등 수렴 유도 아이디어를 함께 제시한다.
- 제6장에서는 실험 설정(모델 · 하이퍼파라미터 · 환경 · 데이터 · 지표)을 기술한다.
- 제7장에서는 실험 결과 및 분석(성능 비교 · GPU 활용도 · 민감도 분석)을 제시한다.
- 제8장에서는 결론을 요약하고, 실무 기여도 · 한계점 및 향후 연구 방향을 논의한다.

Hyperparameter Note. 본 논문에서 주요하게 사용하는 두 개의 하이퍼파라미터는 diff (fixed finite-difference step)와 square (차수 $2k + 1$ 변환 시의 k)이다.

$$\text{diff} = 0.25, \quad \text{square} = 7 \quad (\text{i.e. } 2k + 1 = 7 \Rightarrow k = 3).$$

이 값들은 이후 실험 섹션에서 설명할 바와 같이, 실험적 탐색을 통해 일반화 성능과 수렴 속도 간에 최적의 Trade-off를 보이는 것으로 확인였다.

2 관련 연구

본 연구는 2차 최적화의 효과를 갖추면서도 계산 비용을 최소화하고, 안장점(saddle point) 및 비볼록 구간에서도 안정적인 수렴을 유도하는 알고리즘을 제안한다. 비교 대상은 대표적인 1차 최적화 알고리즘인 Adam, AdamW, 그리고 최근 제안된 AdaBelief로 한정된다.

Adam [5]은 1차 모멘텀과 적응적 학습률(adaptive learning rate)을 결합하여 실용성과 수렴 속도 모두에서 널리 사용되는 옵티마이저이다. AdamW [6]는 weight decay를 정규화 항과 분리하여 일반화 성능을 향상시킨 변형이다. AdaBelief [12]는 Adam의 2차 모멘텀 추정 방식에서 분산 대신 제곱편차를 사용하는 방법으로, 빠른 수렴성과 일반화 성능 개선을 동시에 추구한다.

최근 경량 2차 정보 활용 동향 1차 옵티마이저만으로는 포착하기 어려운 curvature 정보를 저비용으로 활용하기 위해, 최근에는 대각 Hessian 또는 저순위 근사 기반 2차 기법들이 활발히 제안되고 있다. 예컨대 HiZOO [11], SOAA [9] 등은 메모리 · 연산량을 크게 낮추면서도 2차 효과를 일정 부분 유지한다.

경량 2차 정보 활용 기법 비교 은닉 유닛 수를 n 이라 할 때, 한 층의 weight 행렬은 $n \times n$ 이므로 전체 파라미터 수 N 은

$$N = n^2.$$

이를 기준으로 각 기법의 복잡도를 정리하면 다음과 같다.

- **Shampoo** [4], **K-FAC** [7]: 층별로 $n \times n$ 크기의 두 전처리 행렬을 다루므로 메모리 $O(n^2)$, 시간 $O(n^3)$ (즉 $O(N^{3/2})$)이 필요하다.
- **AdaHessian** [10]: Hutchinson 샘플링으로 대각 근사를 수행하여 메모리 $O(n^2)$, 시간 $O(k n^2) \approx O(n^2)$ 을 달성하지만, 프로브 수 k 가 작을수록 분산(variance) 이슈가 발생한다.
- **본 연구의 대각 Hessian 기법**: inverse 연산이나 샘플링 없이 오직 대각 성분만을 tf.matmul/tf.tensordot 배치 처리로 계산하여 메모리 $O(N) = O(n^2)$, 시간 $O(n^3)$ (은닉 유닛 수 n 기준) 를 유지한다. XLA/JIT 최적화 하에 Adam 대비 학습 시간 단축을 실험적으로 검증했다.

본 논문의 차별점은 다음과 같다.

- 기존의 full Hessian 또는 block 근사 방식 대신, **Hessian의 대각 성분만을 이용하여 계산 비용을 획기적으로 줄이면서도 전역 수렴성을 확보하였다.**
- 손실 함수에 대해

$$g(x) = (f(x) - bl)^{2k+1}$$

형태의 변환을 적용하여 함수의 국소 굴곡(curvature)을 평탄화함으로써, 뉴턴 업데이트에서 **2차 정보(헤시안)를 모멘텀 형태로 결합한 보정만으로도** local trap을 벗어나도록 설계하였다. 실제 실험에서는 $k = 3$ 을 택해 7제곱 손실을 사용하였으며, 이를 통해 평탄화된 구간에서의 수렴 안정성과 전역 최소값 도달 성능이 향상되었음을 확인했다.

- 이러한 대각 Hessian 위에 **1차 모멘텀만을 적용하여**, 안장점(saddle point)이나 고차 비선형 영역에서도 빠르게 탈출하는 구조를 만들었다. 이는 기존 1차 기법이 가지는 saddle point 정체 문제를 해결하는 방향성과 맞닿는다 [1].

3 문제 정의 및 수학적 배경

3.1 최적화 문제 공식화

본 논문에서 다루는 다층 페셉트론(Multilayer Perceptron, MLP)의 학습은 다음의 비볼록 최적화 문제로 정식화된다:

$$\min_P \mathcal{L}(P) = \min_{\{W^{(l)}, b^{(l)}\}_{l=1}^r} \frac{1}{N} \sum_{i=1}^N \ell(f(X_i; P), Y_i),$$

여기서 $P = \{P^{(l)}\}_{l=1}^r$, $P^{(l)} = [W^{(l)} \ b^{(l)}]$ 는 l 번째 층의 가중치와 편향을 통합한 파라미터이며, $f(X; P)$ 는 Leaky ReLU 활성함수 f_r 와 softmax 출력을 갖는 MLP 모델, $\ell(\hat{y}, y) = -\sum_k y_k \log \hat{y}_k$ 는 categorical cross-entropy 손실이다.

Baseline Shift (ε 고정) 전략

제안 알고리즘에서는 매 에폭마다

$$bl \leftarrow f(x_t) - \varepsilon \implies \varepsilon = f(x_t) - bl \quad (\varepsilon \ll 1)$$

로 업데이트하여, 변환 함수 $(f(x) - bl)^{2k+1}$ 의 평탄화(flattening) 효과를 일관되게 유지한다.

3.2 모델 구조 및 기호 정의

본 논문 전반에서 사용하는 기호는 다음과 같다:

- $X^{(1)} \in \mathbb{R}^{n_1 \times \lambda}$: 입력층 출력 (미니배치 크기 $\lambda = B$)
- $Z^{(l)} = W^{(l)}X^{(l)} + b^{(l)}$, $X^{(l+1)} = f_r(Z^{(l)})$ (f_r : Leaky ReLU, $l = 1, \dots, r-1$)
- $\hat{Y} = f(Z^{(r)}) \in \mathbb{R}^{n_r \times B}$: softmax 출력
- $\tilde{X}^{(l)} = [X^{(l)}; 1] \in \mathbb{R}^{(n_{l-1}+1) \times B}$: 편향 포함 입력
- B : 배치 크기 (batch size)
- M : 에폭 수 (epochs)
- S : 스텝 수 (#steps/epoch)
- r : 가중치 레벨 수 (number of weight matrices)
- n_l : l 번째 레이어의 뉴런 수, $n := \max_l n_l$

3.3 핵심 가정

제안한 2차 최적화 기법의 전역 수렴 보장을 위해 다음과 같은 가정을 둔다.

Assumption 1 (δ -Bounded Negativity of Hessian Diagonal). 임의의 파라미터 P 에 대해 손실 함수의 Hessian 대각 성분은

$$-\delta \leq H_{ii}(P) \leq M, \quad 0 < \delta \ll M.$$

즉, 음(負) 대각값의 절댓값은 상수 δ 로 엄격히 상한(上限)된다.

본 가정은 7-layer Leaky-ReLU MLP(은닉층 7개, 각 64 노드, 배치 128) 구조를 대상으로 MNIST에서 수행한 파일럿 테스트를 통해 실험적으로 뒷받침된다. 세 가지 고정 랜덤 시드(42, 123, 2025)로 150 epoch 학습을 반복한 뒤, 전 epoch에 걸쳐 관측된 Hessian 대각 최소값 δ 와 최대값 M , 그리고 비율 δ/M 을 측정한 결과는 Table 1와 같다.

테이블에서 보는 바와 같이, 세 run 중 최대 δ/M 는 약 0.13에 불과하며 대부분의 경우 0.1 이하로 관측되었다. 이는 $\delta \ll M$ 가정이 경험적으로도 충족됨을 강하게 뒷받침한다. “Sagun et al. (2016)에서 “대부분의 Hessian 고유값이 0 근처의 bulk에 집중되고, 소수의 outlier eigenvalues만 bulk 밖으로 튀어나온다”고 보고하였으며 [8], Ghorbani et al. (2019) 역시 “대규모 딥 네트워크에서도 bulk 집중 현상이 유지됨”을 확인하였다 [3]. 따라서 본 논문의 $\delta \ll M$ 가정은 경험적·문헌적 근거 모두에 부합한다.

Table 1: 파일럿 테스트 결과: 7-layer Leaky-ReLU MLP(64×7 노드) 전 epoch global δ , M , δ/M , 및 검증 손실

Seed	δ (min H_{ii})	M (max H_{ii})	δ/M	Val Loss
42	-3.2263	92.354	0.0349	0.1039
123	-5.0300	39.097	0.1287	0.0998
2025	-18.143	282.430	0.0642	0.0943
Overall max				0.1287

Remark (Rayleigh 봇에 의한 대각성분 경계). 임의의 대칭행렬 H 와 표준기저벡터 e_i 에 대해, Rayleigh 봇 성질로

$$\lambda_{\min}(H) \leq \frac{e_i^\top H e_i}{e_i^\top e_i} = H_{ii} \leq \lambda_{\max}(H)$$

가 성립한다. 논문 가정 하에 $\lambda_{\min}(H) \geq -\delta$ 및 $\lambda_{\max}(H) \leq M$ 이므로

$$-\delta \leq H_{ii} \leq M$$

가 보장된다.

Assumption 2 (Local gradient 및 Hessian bound). 각 층 가중치 $\|W^{(l)}\| \leq R$ 및 입력 $\|x\| \leq B$ 를 가정하면, 손실 함수 $\mathcal{L}(P)$ 에 대하여

$$\|\nabla \mathcal{L}(x)\| \leq G_{\text{loc}}, \quad \nabla^2 \mathcal{L}(x) \preceq M_{\text{loc}} I.$$

여기서

G_{loc} 및 M_{loc} 은 [2]의 §IV–§V에서 유도된 값입니다.

3.4 연구 질문 및 기여

본 논문은 위 최적화 문제에 대해, 대각 Hessian만을 이용한 경량화된 2차 최적화 알고리즘을 제안하며 다음 질문에 답한다.

- **수렴 개선:** 해시안 대각 성분만을 활용한 동적 학습률이 기존 1차 옵티마이저(Adam, AdamW, AdaBelief) 대비 수렴 속도를 얼마나 개선하는가?
- **깊이 확장성:** 은닉층 개수 r 가 커질 때도 계산 비용 및 수렴 안정성이 어떻게 변화하는가?
- **통계적 검정:** 제안 방법이 기존 옵티마이저 대비 성능에서 통계적으로 유의미한 차이를 보이는가?

이어서 제4장에서는 제안하는 2차 미분 기반 학습률 계산과 알고리즘 구조를 상세히 기술한다.

4 제안 방법

Assumption 1에 따라, 모든 반복에서 손실 함수 Hessian의 대각 성분 H_{ii} 는

$$-\delta \leq H_{ii} \leq M, \quad 0 < \delta \ll M.$$

을 만족한다. 즉, 음(負) 대각값의 절댓값은 상수 δ 로 엄격히 상한(上限)된다.

변환 함수의 Hessian 양정(positivization) 변환 함수

$$g(x) = (f(x) - bl)^{2k+1}$$

를 도입하면,

$$\nabla^2 g(x) = 2k(2k+1)(f(x) - bl)^{2k-1} \nabla f(x) \nabla f(x)^\top + (2k+1)(f(x) - bl)^{2k} \nabla^2 f(x).$$

따라서 대각 근사 $\nabla^2 g(x)_{ii}$ 는

$$\nabla^2 g(x)_{ii} \geq - (2k+1)(f(x) - bl)^{2k} \delta,$$

이고, 추가 뎀핑 상수 $\lambda \gg (2k+1)(f(x) - bl)^{2k} \delta$ 를 더해주면,

$$\tilde{H}_{ii} = |\nabla^2 g(x)_{ii}| + \lambda \geq \lambda - (2k+1)(f(x) - bl)^{2k} \delta > 0.$$

따라서 $\text{diag}(\tilde{H})$ 는 strict PD가 보장된다.

절댓값 변환에 따른 스펙트럼 오차 절댓값 처리 후 생기는 대각 원소 오차는

$$|\nabla^2 g(x)_{ii}| - \nabla^2 g(x)_{ii} \leq 2(2k+1)(f(x) - bl)^{2k} \delta.$$

여기서 $\varepsilon = f(x) - bl$ 라 놓으면, 절댓값 변환에 따른 스펙트럼 노름 오차는

$$2(2k+1)\varepsilon^{2k} \delta = \mathcal{O}(\varepsilon^{2k} \delta).$$

따라서 $\varepsilon \ll 1$ 을 가정할 때, 절댓값 변환이 도입하는 스펙트럼 노름 오차는 $\mathcal{O}(\varepsilon^{2k} \delta)$ 로 충분히 작음을 보장할 수 있으며, $\mathcal{O}(\varepsilon^{2k} \delta) \ll \delta$ 이므로, Assumption 1의 δ -bounded negativity 가정(음수 바운드 $-\delta$)을 결코 위반하지 않는다.

4.1 알고리즘 및 모델 구조

- 입력층: $X^{(1)} \in \mathbb{R}^{n_1 \times \lambda}$

- l 번째 레이어:

$$Z^{(l)} = W^{(l)} X^{(l)} + b^{(l)}, \quad X^{(l+1)} = f_r(Z^{(l)})$$

여기서 f_r 은 Leaky ReLU.

- 출력층:

$$\hat{Y} = f(Z^{(r)})$$

여기서 f 는 softmax.

- 파라미터 통합:

$$P^{(l)} = [W^{(l)} \ b^{(l)}], \quad \tilde{X}^{(l)} = \begin{bmatrix} X^{(l)} \\ 1 \end{bmatrix}.$$

- 전체 배치 크기: B

4.2 2차 미분 기반 학습률 계산

1. 헤시안 대각 \mathcal{L} 에 대한 파라미터 $P_{ab}^{(l)}$ 에 대한 2차 편미분을

$$H_{ab}^{(l)} = \frac{1}{B} \sum_k \frac{\partial^2 \mathcal{L}}{\partial (P_{ab}^{(l)})^2}$$

로 정의한다.

2. 2차 미분 항 전개

$$\frac{\partial^2 \mathcal{L}}{\partial (P_{ab}^{(r)})^2} = -\frac{1}{B} \sum_k \frac{\partial \hat{y}_{ak}}{\partial Z_{ak}^{(r)}} (\tilde{X}_{bk}^{(r)})^2, \quad (1)$$

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial (P_{ab}^{(r-1)})^2} &= \frac{1}{B} \sum_{i,k} (W^{(r)T})_{ai} \hat{y}_{ik} W_{ia}^{(r)} [f'_r(Z_{ak}^{(r-1)})]^2 (\tilde{X}_{bk}^{(r-1)})^2 \\ &\quad - \frac{1}{B} \sum_{i,j,k} (W^{(r)T})_{ai} \hat{y}_{ik} (W^{(r)T})_{aj} \hat{y}_{jk} [f'_r(Z_{ak}^{(r-1)})]^2 (\tilde{X}_{bk}^{(r-1)})^2, \end{aligned} \quad (2)$$

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial (P_{ab}^{(r-2)})^2} &= \frac{1}{B} \sum_{i,s,j,k} (W^{(r-1)T})_{as} (W^{(r-1)T})_{aj} [(W^{(r)T})_{ji} (W^{(r)T})_{si} \hat{y}_{ik}] \\ &\quad \times f'_r(Z_{sk}^{(r-1)}) f'_r(Z_{jk}^{(r-1)}) [f'_r(Z_{ak}^{(r-2)})]^2 (\tilde{X}_{bk}^{(r-2)})^2 \\ &\quad - \frac{1}{B} \sum_{i,s,k} \left[(W^{(r-1)T})_{as} [(W^{(r)T})_{si} \hat{y}_{ik} f'_r(Z_{sk}^{(r-1)})] f'_r(Z_{ak}^{(r-2)}) \right]^2 (\tilde{X}_{bk}^{(r-2)})^2 \end{aligned} \quad (3)$$

3. 일반화: $r-m$ 계층에서

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial (P_{ab}^{(r-m)})^2} &= \frac{1}{B} \sum_{s,j,k} (W^{(r-m+1)T})_{as} (W^{(r-m+1)T})_{aj} D_{sjk}^{(r-m)} [f'_r(Z_{ak}^{(r-m)})]^2 (\tilde{X}_{bk}^{(r-m)})^2 \\ &\quad - \frac{1}{B} \sum_k [J_{ak}^{(r-m)}]^2 (\tilde{X}_{bk}^{(r-m)})^2, \end{aligned} \quad (4)$$

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial (P_{ab}^{(r-m-1)})^2} &= \frac{1}{B} \sum_{\alpha,\beta,s,j,k} (W^{(r-m)T})_{a\alpha} (W^{(r-m)T})_{a\beta} (W^{(r-m+1)T})_{\alpha s} (W^{(r-m+1)T})_{\beta j} D_{sjk}^{(r-m)} \\ &\quad \times f'_r(Z_{\alpha k}^{(r-m)}) f'_r(Z_{\beta k}^{(r-m)}) [f'_r(Z_{ak}^{(r-m-1)})]^2 (\tilde{X}_{bk}^{(r-m-1)})^2 \\ &\quad - \frac{1}{B} \sum_{s,k} \left[(W^{(r-m)T})_{as} J_{sk}^{(r-m)} f'_r(Z_{ak}^{(r-m-1)}) \right]^2 (\tilde{X}_{bk}^{(r-m-1)})^2. \end{aligned} \quad (5)$$

4. 중간 텐서 정의

$$D_{sjk}^{(r-2)} = \sum_h (W^{(r)T})_{sh} (W^{(r)T})_{jh} \hat{y}_{hk}, \quad J_{ak}^{(r-1)} = \sum_i (W^{(r)T})_{ai} \hat{y}_{ik} f'_r(Z_{ak}^{(r-1)}).$$

$$D_{\alpha\beta k}^{(r-m-1)} = \sum_{s,j} (W^{(r-m+1)T})_{\alpha s} (W^{(r-m+1)T})_{\beta j} D_{sjk}^{(r-m)} f'_r(Z_{\alpha k}^{(r-m)}) f'_r(Z_{\beta k}^{(r-m)}),$$

$$J_{ak}^{(r-m-1)} = \sum_s (W^{(r-m)T})_{as} J_{sk}^{(r-m)} f'_r(Z_{ak}^{(r-m-1)}).$$

5. `Hsolve` 정의 입력으로 받은 텐서 $T \in \mathbb{R}^{n_l \times B}$, 확장된 입력 $\tilde{X} \in \mathbb{R}^{(n_{l-1}+1) \times B}$, 1차 그라디언트 행렬 $dP \in \mathbb{R}^{n_l \times (n_{l-1}+1)}$, 배치 크기 B , 계수 $\alpha_{\text{out}}, \alpha_H$ 에 대해,

$$H = \frac{1}{B} T (\tilde{X}^{\circ 2})^T, \quad \text{outer_dP} = dP^{\circ 2},$$

$$\tilde{H} = |\alpha_{\text{out}} \text{outer_dP} + \alpha_H H| + \lambda, \quad \lambda = 10^{-2} \text{ (안정화 상수)}$$

$$L = \frac{\alpha_H dP}{\tilde{H}}.$$

Remark (댐핑 상수 λ 설정 근거). 본 연구에서는 사전 탐색을 위해 $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ 범위에서 실험을 수행하였으며, $\lambda = 10^{-2}$ 에서 가장 일관된 일반화 성능을 보여 이후 모든 실험에서 해당 값을 고정하였다.

6. 가중치별 학습률

$$L_{ab} = \frac{dP_{ab} \alpha_H}{\tilde{H}_{ab}}.$$

4.3 동적 러닝레이트 및 모멘텀 업데이트

기본 식 손실 기반 감쇠와 전역 스케일 η_0 를 도입하여

$$\text{damp}(x) = \text{shifted_normal}(x), \quad \eta_0 = 2 \times 10^{-3}, \quad \text{lr} = \frac{\eta_0 \text{damp}(\mathcal{L})}{\text{diff}^{(\text{square}-1)}}.$$

감쇠 함수 정의

$$\bar{x} = \min(\max(x, 0), 2), \quad \text{damp}(x) = (M' - \text{base}) \exp\left(-\frac{(\bar{x} - \mu)^2}{2\sigma^2}\right) + \text{base},$$

where $\mu = 0.7$, $M' = 0.125$, $\sigma = 0.575$, and $\text{base} = 0.01$. 각 매개변수는 실험적으로 설정된 값으로, μ 는 peak 위치, M' 은 최대 출력값, σ 는 분포의 폭, base 는 $x \rightarrow \infty$ 에서의 수렴값을 의미한다.

학습률 분리 각 계층에서 계산된 학습률 텐서 $L \in \mathbb{R}^{n_l \times (n_{l-1}+1)}$ 를

$$d2W = L[:, 1 : n_{l-1}], \quad d2b = L[:, n_{l-1} + 1 : n_{l-1} + 2]$$

로 분리하여, 모멘텀 변수에 대응시킵니다.

모멘텀 업데이트

$$m_W \leftarrow \beta_1 m_W + (1 - \beta_1) d2W, \quad \hat{m}_W = \frac{m_W}{1 - \beta_1^t},$$

$$m_b \leftarrow \beta_1 m_b + (1 - \beta_1) d2b, \quad \hat{m}_b = \frac{m_b}{1 - \beta_1^t},$$

$$W \leftarrow W - \text{lr} \hat{m}_W, \quad b \leftarrow b - \text{lr} \hat{m}_b.$$

Table 2: 주요 하이퍼파라미터

기호	설명	기본값 또는 범위
M	에폭 수 (Epochs)	250
B	배치 크기 (Batch size)	128
S	스텝 수 (#steps per epoch)	$[N/B]$
diff	고정 차분 ($f(x) - bl = 0.25$)	0.25
square	차수 계수 (order)	7
α_{out}	차분 제곱 스케일링 계수 (diff 기반)	$\text{square}(\text{square} - 1) \text{diff}^{\text{square}-2}$
α_H	해시안 계수	$\text{square} \text{diff}^{\text{square}-1}$
β_1	1차 모멘텀 계수	0.25
η_0	전역 학습률 스케일	2×10^{-3}
λ	안정화 상수	10^{-2}

Table 3: 주요 하이퍼파라미터 및 설정 근거

기호	값	설정 근거
diff	0.25	0.1–0.5 그리드 탐색 중 손실 진동 최소화 및 수렴 속도 최적화
square	7	차수 5–9 테스트에서 플랫닝(flattening) 효과와 빠른 수렴 간 트레이드오프가 가장 우수한 차수
η_0	2×10^{-3}	초기 학습률을 10^{-4} – 10^{-2} 범위 실험 후 가장 빠른 초기 수렴 보임

이후 본문에서는 이 설정을 바탕으로 제안한 최적화 알고리즘의 수렴 및 성능을 분석합니다.

4.4 전체 알고리즘 흐름

Algorithm 1 2차 미분 기반 옵티마이저 (TensorFlow 구현 구조 반영)

Require: 데이터셋 ‘dataset’, 파라미터 $\{W^{(l)}, b^{(l)}\}$, 배치 크기 B , 에폭 수 M , 스텝 수 S

- 1: $\triangleright \ast$ 이 의사코드에서는 이해를 돋기 위해 einsum 표기를 사용했으나, 실제 구현에서는 `tf.tensordot`를 사용합니다
- 2: **for** epoch $t = 1, \dots, M$ **do**
- 3: **for** each $(X_{\text{batch}}, Y_{\text{batch}})$ in $\text{dataset.take}(S)$ **do**
- 4:

\triangleright — 순전파 및 오류 계산 —
- 5: compute $\{Z^{(l)}, X^{(l+1)}\}$ and $\hat{Y} = f(Z^{(r)})$
- 6: $J \leftarrow \hat{Y} - Y_{\text{batch}}$
- 7: Let $W^T = W^{(r)T}$. Then
- 8: $D \leftarrow \text{einsum('is,js,sk', } W^T, W^T, \hat{Y})$
- 9: $Je \leftarrow (W^T \hat{Y}) \odot f'_r(Z^{(r-1)})$
- 10: compute $\alpha_{\text{out}}, \alpha_H$ from ‘square’, ‘diff’
- 11: **for** $l = r, r-1, \dots, 1$ **do**
- 12:

\triangleright — 1차 그라디언트 계산 —
- 13: $dW \leftarrow \frac{1}{B} J X^{(l)T}, db \leftarrow \frac{1}{B} \sum J$
- 14: $dP \leftarrow [dW db]$
- 15: **if** $l > 1$ **then**
- 16:

\triangleright — 역전파: 활성화 미분 적용 —
- 17: $J \leftarrow W^{(l)T} J \odot f'_r(Z^{(l-1)})$
- 18: **end if**
- 19:

\triangleright — 해시안 행 T 분기 계산 —
- 20: **if** $l = r$ **then**
- 21: $T \leftarrow \hat{Y} - \hat{Y}^2$
- 22: **else if** $l = r-1$ **then**
- 23: $T \leftarrow ((W^{(l+1)T})^{\circ 2} \hat{Y} - (W^{(l+1)T} \hat{Y})^{\circ 2}) \odot (f'_r(Z^{(l)}))^{\circ 2}$
- 24: **else**
- 25: Let $W^T = W^{(l+1)T}, \delta = f'_r(Z^{(l)})$.
- 26: Let $E = \text{einsum('as,aj,skj->ak', } W^T, W^T, D)$.
- 27: Then $T \leftarrow (E - (W^T Je)^{\circ 2}) \odot \delta^{\circ 2}$.
- 28: $D \leftarrow \text{einsum('ij,kl,jlm,im,km->ikm', } W^T, W^T, D, \delta, \delta)$.
- 29: $Je \leftarrow (W^T Je) \odot \delta$.
- 30: **end if**
- 31:

\triangleright — 해시안 기반 학습률 계산 —
- 32: $L \leftarrow \text{_Hsolve}(T, [X^{(l)}; 1], dP, B, \alpha_{\text{out}}, \alpha_H)$
- 33: split L into $d2W$ (첫 n_{l-1} 열) and $d2b$ (마지막 열)
- 34:

\triangleright — 손실 기반 감쇠 & lr 계산 —
- 35: $damp \leftarrow \text{shifted_normal}(\mathcal{L})$
- 36: $lr \leftarrow \eta_0 \frac{damp}{\text{diff}(\text{square}-1)}$
- 37:

\triangleright — 모멘텀 업데이트 —
- 38: $m_W \leftarrow \beta_1 m_W + (1 - \beta_1) d2W, \hat{m}_W \leftarrow \frac{m_W}{1 - \beta_1^t}$
- 39: $m_b \leftarrow \beta_1 m_b + (1 - \beta_1) d2b, \hat{m}_b \leftarrow \frac{m_b}{1 - \beta_1^t}$
- 40:

\triangleright — 파라미터 갱신 —
- 41: $W^{(l)} \leftarrow W^{(l)} - lr \hat{m}_W, b^{(l)} \leftarrow b^{(l)} - lr \hat{m}_b$
- 42: **end for**
- 43: **end for**
- 44: **end for**

시간 복잡도 분석

(본 분석은 Section 4.4의 의사코드가 아닌, 실제 구현된 TensorFlow 함수 `_hessian_block_step`를 기반으로 합니다.)

함수 `_hessian_block_step` 한 배치 처리 시 주요 연산 비용을 계층 l 별로 살펴보면 다음과 같다. 여기서 r 은 파라미터 블록(가중치 매트릭스) 수 (#weight matrices), n_l 은 l 번째 은닉층의 뉴런 수, B 는 배치 크기, M 은 에폭 수로 정의한다.

표기. $n := \max_l n_l$ (은닉층 최대 너비)

1. `_tensorOfH` 연산 ($l \leq r - 2$ 일반 계층)

$$\begin{aligned} \text{tensordot} &: O(B n_l n_{l-1}^2), \\ \text{reduce_sum} &: O(B n_l n_{l-1}), \\ \text{matmul} &: O(B n_l n_{l-1}), \\ D \text{ 업데이트} (\text{batched matmul } \times 2 + \text{브로드캐스트 곱}) &: O(B n_l^2 n_{l-1}), \end{aligned}$$

따라서 계층 l 당 지배적 항은

$$O(B n_l n_{l-1}^2 + B n_l^2 n_{l-1}) \approx O(B n^3).$$

2. `_Hsolve` 연산

$$H = \frac{1}{B} T(\tilde{X}^{\circ 2})^T : O(n_l n_{l-1} B), \quad \text{요소별 연산} : O(n_l n_{l-1}), \quad L = \frac{\alpha_H dP}{\tilde{H}} : O(n_l n_{l-1}).$$

계층 l 당 총

$$O(n_l n_{l-1} B) \approx O(n^2 B).$$

3. 1차 그라디언트 & 모멘텀 업데이트

$$\text{역전파: } O(n_l n_{l-1} B), \quad \text{파라미터 갱신: } O(n_l n_{l-1}) \approx O(n^2 B).$$

전체 배치당 복잡도 모든 계층을 합산하면

$$\sum_{l=1}^r O(B n^3) = O(r B n^3).$$

에폭 수 M 를 곱하여 학습 전체 복잡도는

$$O(M B r n^3).$$

참고: Adam 옵티마이저 파라미터 수 $P \approx \sum_l n_l n_{l-1}$ 라고 할 때,

$$O(M B P) \approx O(M B r n^2),$$

즉 n^2 차수이므로, 제안 옵티마이저의 n^3 의존도보다 연산량이 훨씬 낮다.

5 이론적 보장 및 설계 직관

5.1 이론적 보장

본 절에서는 제안한 최적화 알고리즘이 변환 함수 기반 Newton 방법의 수학적 성질을 이용하여 전역적으로 수렴함을 요약한다. 자세한 수학적 유도는 부록 A에 수록되어 있다.

Assumption 1에 따라 ReLU 계열 활성화 함수를 사용하는 신경망에 대하여, 모든 파라미터 x 에서 Hessian의 대각 성분 $H_{ii} = [\nabla^2 f(x)]_{ii}$ 는

$$-\delta \leq H_{ii} \leq M, \quad 0 < \delta \ll M.$$

즉, 음(負) 대각값의 절댓값은 상수 δ 로 엄격히 상한(上限)된다.

Assumption 2에 따라, 각 층 가중치 $\|W^{(l)}\| \leq R$ 및 입력 $\|x\| \leq B$ 를 가정하면

$$\|\nabla f(x)\| \leq G_{\text{loc}} = G, \quad \nabla^2 f(x) \preceq M_{\text{loc}} I = M I$$

가 성립한다.

Theorem 1 (전역 수렴 보장). 손실 함수 $f(x)$ 에 대해 변환 함수

$$g(x) = (f(x) - bl)^{2k+1}$$

을 정의하고, 이에 대한 damped Newton 업데이트를 다음과 같이 구성한다:

$$x_{t+1} = x_t - \alpha [\nabla^2 g(x_t) + \lambda I]^{-1} \nabla g(x_t).$$

양의 준정부호 유지 및 댐핑 확보. 변환 함수

$$g(x) = (f(x) - bl)^{2k+1}$$

에 대하여,

$$\nabla^2 g(x) = 2k(2k+1)(f(x) - bl)^{2k-1} \nabla f(x) \nabla f(x)^\top + (2k+1)(f(x) - bl)^{2k} \nabla^2 f(x)$$

이다.

Note. 본 절에서는 식 전개의 간결함을 위해 $g(x) = (f(x) - bl)^{2k+1}$ 이라 표기하였으나, 실제 알고리즘 상에서는 매 반복 t 마다 $bl_t = f(x_t) - \varepsilon$ 로 갱신되므로, 엄밀히는 $g_t(x) = (f(x) - bl_t)^{2k+1}$ 로 해석됩니다.

Strict PD 확보. Assumption 1에 따라 $\nabla^2 f(x) \succeq -\delta I$ 이고, $\nabla f(x) \nabla f(x)^\top \succeq 0$ 으로

$$\begin{aligned} \nabla^2 g(x) &= 2k(2k+1)(f(x) - bl)^{2k-1} \nabla f(x) \nabla f(x)^\top \\ &\quad + (2k+1)(f(x) - bl)^{2k} \nabla^2 f(x) \\ &\succeq - (2k+1)(f(x) - bl)^{2k} \delta I. \end{aligned}$$

따라서 $\lambda \gg (2k+1)(f(x) - bl)^{2k} \delta$ 인 $\lambda > 0$ 를 더하면

$$\nabla^2 g(x) + \lambda I \succeq [\lambda - (2k+1)(f(x) - bl)^{2k} \delta] I =: \mu I \succ 0,$$

즉, $\nabla^2 g(x) + \lambda I$ 는 항상 양의 정부호를 만족한다.

상한 L_g 정의 및 댐핑 포함 Hessian 상한. 손실과 기준선 차이가 $f(x) - bl = \varepsilon$ 이고, $\varepsilon > 0$ 일 때

$$\nabla^2 g(x) = 2k(2k+1) \varepsilon^{2k-1} \nabla f(x) \nabla f(x)^\top + (2k+1) \varepsilon^{2k} \nabla^2 f(x),$$

이므로 가정으로부터

$$\nabla^2 g(x) \preceq 2k(2k+1)\varepsilon^{2k-1}G^2I + (2k+1)\varepsilon^{2k}MI.$$

따라서 뎁핑 항 λI 를 포함한 전체 Hessian은

$$\nabla^2 g(x) + \lambda I \preceq \left[2k(2k+1)\varepsilon^{2k-1}G^2 + (2k+1)\varepsilon^{2k}M + \lambda \right] I =: L_g I.$$

또한

$$\nabla^2 g(x) + \lambda I \preceq L_g I, \quad 0 \prec \mu I \preceq \nabla^2 g(x) + \lambda I.$$

따라서, $\tilde{g}(x) = g(x) + \frac{\lambda}{2}\|x\|^2$ 는 $\nabla^2 \tilde{g}(x) \succeq \mu I$ 이므로 μ -strongly convex이고, 해 x^* 가 유일하게 존재하고 $\nabla \tilde{g}(x^*) = 0$ 이라 가정하면, 부록 A에서 증명한 바와 같이 적절한 $\alpha \in (0, \frac{2\lambda^2}{L_g^2})$ 및 $\lambda > 0$ 하에서

$$\tilde{g}(x_t) - \tilde{g}(x^*) = O(\rho^t) \implies \|x_t - x^*\| \rightarrow 0 \quad (t \rightarrow \infty),$$

를 만족하는 상수 $\rho < 1$ 가 존재하며, 이는 전역 선형 수렴을 의미한다.

그때의 α 범위:

$$0 < \alpha < \frac{2\lambda^2}{L_g^2}.$$

Definition 5.1 (ε -shift 유지 가정). 본 증명에서 $\varepsilon = f(x) - bl$ 가 일정한 값으로 유지되는 것이 핵심이다. 이를 위해 $\varepsilon > 0$ 를 상수로 미리 고정하고, 각 반복 단계에서

$$bl \leftarrow f(x) - \varepsilon$$

로 shift 상수 bl 를 정의한다고 가정한다. 이러한 설정을 통해 $\varepsilon = f(x) - bl$ 는 반복 과정 전반에 걸쳐 일정하게 유지되며, 이는 수렴 정리에 사용된 $\varepsilon > 0$ 가정을 안정적으로 충족시킨다.

5.2 설계 아이디어 및 직관

2차 정보의 모멘텀 형태 결합 본 알고리즘은 Hessian 대각 성분 D_t 에 대해 지수 가중 이동 평균을 적용하여

$$m_t = \beta m_{t-1} + (1-\beta)D_t, \quad \beta \in [0, 1]$$

와 같이 과거 curvature 정보를 일정 부분 유지합니다. 이때 얻어진 m_t 를 뉴턴 업데이트의 보정 항으로 사용함으로써, 평탄해진 국소 최솟값 구간에서도 “관성(inertia)”을 확보하여 단순 로컬 기울기만으로는 탈출하기 어려운 flat trap을 효과적으로 벗어날 수 있습니다.

평탄화와 local minima 탈출 아이디어 변환 함수

$$g(x) = (f(x) - bl)^{2k+1}$$

를 도입하고, 기준선 bl 를 $f(x) - bl = \varepsilon$ ($\varepsilon \ll 1$)로 유지하면

$$\nabla^2 g(x) = 2k(2k+1)\varepsilon^{2k-1}\nabla f(x)\nabla f(x)^\top + (2k+1)\varepsilon^{2k}\nabla^2 f(x).$$

여기서 첫 항은 항상 PSD이고 크기가 $O(\varepsilon^{2k-1}\|\nabla f(x)\|^2)$ (스펙트럼 노름 기준)이어서, 원 함수의 negative curvature 구간에서도 충분히 작은 $\lambda > 0$ 만 더하면 $\nabla^2 g(x) + \lambda I \succ 0$ 이 보장됩니다.

위 수식을 바탕으로 살펴보면, 기준선 bl 를 반복마다 조정하여 항상

$$f(x) - bl = \varepsilon \quad (\varepsilon \text{ 충분히 작음})$$

이 되도록 유지하면, 변환 함수 $g(x)$ 는 로컬 최솟값 부근에서는 이미 기울기(gradient)가 0에 수렴하고, 원 함수의 Hessian 값도 일반 지점보다 작아 $O(\varepsilon^{2k-1})$ 스케일로 수축되는 것만으로도 Hessian이 거의 0에 근접하여 곡률(curvature)이 효과적으로 평탄화(flatten)됩니다. 이 상태에서 뉴턴 업데이트는 **2차 정보(헤시안)를 모멘텀 형태로 결합한 보정만으로도** 로컬 트랩을 효과적으로 탈출하여 전역 최솟값 쪽으로 이동할 수 있습니다.

전역 최솟값에 도달한 이후에도 곡률은 평탄하지만, 더 낮은 굴곡이나 다른 최솟값이 존재하지 않으므로 알고리즘은 전역 최솟값 부근으로 되돌아와 최종적으로 안정적으로 수렴합니다.

6 실험 설정

6.1 모델 아키텍처

본 실험에서 사용한 MLP의 구조는 다음과 같다.

- 은닉층 수 (#layers): 7, 10
- 은닉 노드 수: 모든 은닉층에서 64개 고정
- 활성화 함수: Leaky ReLU
- 출력층: softmax

하이퍼파라미터 본 실험에서는 §4.3에서 정의한 하이퍼파라미터를 그대로 사용하였으며, 특히 배치사이즈 $B = 128$ 와 에폭 수 $M = 250$ 을 적용하였다.

6.2 실험 단계 구분

1. 순수 MLP 실험 (정규화 off):

- 목적: 순수 MLP 구조(은닉층 7, 10)에서 정규화 없이 옵티마이저 성능 평가
- 설정: 은닉층 7, 10 / 배치 128 / 은닉 노드 64 / Leaky ReLU / softmax
- 반복: 랜덤 시드 7회
- 측정 지표: Training Loss, Validation Loss, Validation Accuracy, Macro F₁-Score, Training Time

2. 정규화 기법 비교 실험:

- 목적: 커스텀 옵티마이저 vs. baseline(Adam, AdamW, Adabelief)에서 정규화 기법이 성능에 미치는 영향
- 설정:
 - 커스텀 옵티마이저: Dropout 0.004, Label Smoothing 0.025
 - baseline: BatchNorm ON, Dropout 0.004, Label Smoothing 0.025
- 공통 설정: 은닉층 7, 10 / 배치 128 / 은닉 노드 64 / Leaky ReLU / softmax

- 반복: 랜덤 시드 7회
- 측정 지표: Training Loss, Validation Loss, Validation Accuracy, Macro F₁-Score, Training Time

3. 하이퍼파라미터 민감도 분석

- 목적 Dropout–Label Smoothing 조합에 따른 검증 손실 민감도 파악
- 고정 설정
 - BatchNorm: baseline(Adam, AdamW, AdaBelief)에만 ON
 - 은닉층 수: 7
 - epochs: 100
 - 데이터셋: WineQuality
- 변수 설정
 - Dropout: {0.002, 0.004, 0.008}
 - Label Smoothing α : {0.0125, 0.025, 0.05}
- 반복 각 조합당 랜덤 시드 3회 → 평균 검증 손실 기록
- 측정 지표 Validation Loss (평균)

6.3 실험 환경

하드웨어

- GPU: NVIDIA GeForce RTX 4060 (8 GB)
- CPU: AMD Ryzen 7 5700X 8-Core @ 3.40 GHz
- RAM: 32 GB DDR4-3200

소프트웨어

- 운영체제: Ubuntu 22.04 LTS
- Python: 3.9
- CUDA / cuDNN: CUDA 12.1 (nvcc V12.1.105), cuDNN 9.9.0
- TensorFlow: 2.15.0 (XLA JIT 컴파일러 활성화)
- 주요 라이브러리:
 - tensorflow-addons 0.22.0
 - numpy 1.26.4
 - pandas 2.2.3
 - matplotlib 3.9.4
 - scipy 1.13.1
 - scikit-learn 1.6.1

6.4 비교군 옵티마이저 및 하이퍼파라미터

본 논문에서는 제안한 2차 미분 기반 옵티마이저를 Adam, AdamW, AdaBelief 세 가지 1차 옵티마이저와 비교합니다. 비교군 Adam, AdamW, AdaBelief은 모두 `learning_rate = 0.001`, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-7}$ 을 사용하였습니다. 또한 AdamW와 AdaBelief에는 `weight_decay = 1 \times 10^{-4}`를 적용하였습니다.

6.5 데이터셋 및 평가 지표

데이터셋 본 실험에서는 다음의 8개 실제 데이터셋과 4개 합성 Gaussian 데이터셋을 사용하였다.

- **MNIST**: 60,000 train, 10,000 test, 입력 차원 784, 픽셀 값 [0,1] 정규화
- **CIFAR-10**: 50,000 train, 10,000 test, 입력 차원 $32 \times 32 \times 3$, 픽셀 값 [0,1] 정규화
- **CIFAR-100**: 50,000 train, 10,000 test, 입력 차원 $32 \times 32 \times 3$, 픽셀 값 [0,1] 정규화
- **20 Newsgroups**: 18,846 문서, train/test 80%/20%, TF-IDF 벡터화(2,000차원)
- **Imbalance**: 30,000 샘플, 4클래스(가중치 [0.7, 0.15, 0.1, 0.05]), train/test 80%/20%
- **WineQuality-Red**: 1,599 샘플, 11피쳐 표준화, 6단계 원-핫 레이블, train/test 80%/20%
- **Fashion-MNIST**: 60,000 train, 10,000 test, 입력 차원 784, 픽셀 값 [0,1] 정규화
- **HAR (UCI)**: 센서값 9채널, train/test 80%/20%, train/test .npy 파일 로드
- **합성 Gaussian**: $\text{class_sep} \in \{0.5, 1.0, 2.0\}$, $\text{clusters} \in \{1, 3, 5\}$, $\text{flip_y} \in \{0, 0.05\}$, 30,000 샘플, 12피쳐, 8클래스, train/test 80%/20%.

본 실험에서 사용한 데이터셋들은 이미지 · 텍스트 · 수치 · 시계열 · 합성 데이터 등 다양한 도메인과 분포 특성을 포괄하여, 제안 기법의 일반화 성능과 안정성을 종합적으로 평가하기 위함이다. 출처(URL · 저자 · 연도 · 라이선스)는 부록 D에 정리하였다.

평가 지표

- **Training Loss & Validation Loss**: 각 에폭 종료 시점의 categorical cross-entropy
- **Accuracy**:
$$\frac{\text{정확히 분류된 샘플 수}}{\text{전체 샘플 수}}$$
- **Macro F₁-Score**: 클래스별 F₁을 평균화

$$F1_{macro} = \frac{1}{C} \sum_{c=1}^C \frac{2 \text{Precision}_c \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

- **Time to Convergence**: 총 학습 소요 시간 (초 단위, `time.perf_counter()` 호출 전후로 직접 측정하여 `buf['time_c']`에 저장)
- **통계적 검정**:
 - Paired t-test (유의수준 $\alpha = 0.05$)
 - 효과크기 측정: Cohen's d
 - * 큰 효과: $0.8 \leq |d| < 1.2$
 - * 아주 큰 효과: $|d| \geq 1.2$

7 실험 결과

통계량 정의

- $\Delta(\%)$: $(\text{Custom} - \text{Baseline}) / \text{Baseline} \times 100$ ($\Delta > 0$ ⇒ Custom 우위, $\Delta < 0$ ⇒ Custom 열위).
- **Mean** : 12개 데이터셋(8 실제 + 4 합성)의 Δ 값을 산술평균.
- **Median**: 같은 Δ 값의 중앙값(극단값 완화 목적).

Mean · Median 모두 표에 병기하고, 효과크기(d , 정규화 실험 한정) · p -값은 부록 표에 별도 제시한다.

7.1 순수 MLP 실험 (정규화 off)

Table 4: (DropoutLabel Smoothing OFF)
Custom 대비 Adam 계열 $\Delta\%$ — Mean / Median

	Hidden	Baseline	Train	Val	Acc	F_1	Time
7		Adam Mean	-37.8%	+0.0%	+1.2%	+1.4%	-44.6%
		Adam Median	-49.6%	+7.5%	-0.1%	+0.2%	-46.5%
		AdamW Mean	-39.6%	+0.5%	+1.0%	+1.1%	-46.6%
		AdamW Median	-48.4%	+5.2%	-0.2%	-0.3%	-48.1%
		AdaBelief Mean	-39.8%	+4.0%	+0.9%	+1.1%	-47.6%
		AdaBelief Median	-49.1%	+10.6%	+0.1%	-0.0%	-49.2%
10		Adam Mean	-52.3%	+25.3%	+0.6%	+0.7%	-39.7%
		Adam Median	-89.1%	+17.8%	+0.1%	+0.2%	-41.0%
		AdamW Mean	-44.8%	+26.6%	+0.9%	+0.7%	-42.2%
		AdamW Median	-75.5%	+23.1%	+0.6%	+0.4%	-45.9%
		AdaBelief Mean	-56.0%	+30.2%	+0.2%	+0.2%	-44.3%
		AdaBelief Median	-94.9%	+25.9%	-0.1%	-0.1%	-47.2%

과적합 특징

- **훈련 손실** — Median 기준으로 최대 -95% 까지 낮추었으나
- **검증 손실** — 10층 MLP · AdaBelief에서 최대 $\text{Median } \Delta_{\text{Val}} = +25.9\%$ ($\text{Mean } \Delta_{\text{Val}} = +30.2\%$)로 크게 악화. 대체적으로 모든 커스텀 옵티마이저의 검증 손실이 baseline 대비 높게 나타나(열위) 과적합 경향을 보임.
- **정확도 · F_1** — Median 편차 $\leq \pm 0.3\%$.
- **학습 시간** — 평균 · 중앙값 모두 40–50 % 단축.

즉, 과적합 양상이 뚜렷하여 정규화 기법이 필수적임을 시사한다.

7.2 정규화 기법 실험 (Custom: Dropout 0.004 + Label Smoothing 0.025 Baseline: BatchNorm ON + Dropout + LS)

Table 5: 정규화 적용 후 Custom 대비 baseline $\Delta\% = \text{Mean} / \text{Median}$

Hidden	Baseline	Train	Val	Acc	F_1	Time
7	Adam Mean	-58.4%	-19.7%	+0.1%	-0.2%	-57.8%
	Adam Median	-74.8%	-20.6%	-0.2%	-0.2%	-60.2%
	AdamW Mean	-58.4%	-19.3%	+0.3%	-0.2%	-58.9%
	AdamW Median	-74.6%	-20.1%	-0.4%	-0.5%	-60.9%
	AdaBelief Mean	-58.5%	-20.1%	-0.0%	-0.0%	-60.1%
	AdaBelief Median	-75.4%	-20.0%	-0.1%	+0.0%	-61.9%
10	Adam Mean	-54.2%	-19.6%	-1.9%	-1.7%	-64.0%
	Adam Median	-71.9%	-21.0%	-0.8%	-0.4%	-66.3%
	AdamW Mean	-54.2%	-19.8%	-1.7%	-1.7%	-65.8%
	AdamW Median	-71.8%	-21.5%	-0.0%	-0.0%	-68.3%
	AdaBelief Mean	-54.3%	-19.4%	-1.8%	-1.6%	-67.0%
	AdaBelief Median	-71.9%	-21.2%	-0.5%	-0.3%	-69.4%

일반화·속도 개선 (정규화 실험) 드롭아웃과 스무스 라벨링을 적용한 후, Custom은 은 닉층별로 다음과 같은 $Median / Mean$ 향상을 달성하였다.

1) 검증 손실 감소

감소 폭 Median: 20.0%–21.2%; Mean: 19.3%–20.1%

⇒ baseline 대비 전반적 일반화 우위 확보.

2) 학습 시간 단축

전반적 감축 Median 평균 -64.5%, Mean 평균 -62.3% (약 60%)

⇒ BatchNorm 이 포함된 baseline보다 최대 66% 빠름.

3) 정확도 · F_1 안정

Median 편차 $\leq \pm 0.5\%$, Mean 편차 $\leq \pm 1.9\%$ 로 성능 저하 없이 일반화·속도만 개선.

7.2.1 테이터셋별 열위 사례

Table 6: Custom 열위 사례: 상대 차이 ($\Delta\%$), 통계적 유의성 (\checkmark), 효과크기 수준 (\star : $|d| \geq 1.2$, $\star\star$: $0.8 \leq |d| < 1.2$)

Dataset	Base	ΔVal	ΔAcc	ΔF_1	$p_{t\text{test}}$	d
Hidden 7						
Gauss_sep1.0_clust3_flip0.05	Adam	+10.0	-3.0	-3.0	4.99×10^{-5}	\checkmark
	AdamW	+9.1	-3.2	-3.2	8.47×10^{-5}	\checkmark
	Adabelief	+9.3	-3.0	-3.0	1.36×10^{-4}	\checkmark
20NG	Adam	+8.7	-8.5	-8.2	2.21×10^{-3}	\checkmark
	AdamW	+7.7	-8.1	-7.7	7.31×10^{-4}	\checkmark
	Adabelief	+8.3	-7.8	-7.4	5.63×10^{-4}	\checkmark
Gauss_sep1.0_clust3	Adam	+4.5	-3.0	-3.0	3.47×10^{-2}	\checkmark
	AdamW	+4.6	-3.0	-3.0	5.49×10^{-2}	+0.84
	Adabelief	+5.6	-3.1	-3.1	1.42×10^{-2}	\checkmark
WineQuality	Adam	+1.9	-1.8	-2.8	2.75×10^{-1}	+0.28
Hidden 10						
20NG	Adam	+12.7	-14.6	-14.1	3.52×10^{-6}	\checkmark
	AdamW	+12.0	-14.5	-14.1	2.91×10^{-5}	\checkmark
	Adabelief	+12.9	-13.8	-13.2	3.81×10^{-5}	\checkmark
Gauss_sep1.0_clust3_flip0.05	Adam	+3.9	-2.4	-2.4	5.35×10^{-3}	\checkmark
	AdamW	+3.7	-2.4	-2.4	8.64×10^{-3}	\checkmark
	Adabelief	+3.5	-2.7	-2.7	5.06×10^{-3}	\checkmark

기호 설명: Δ 는 Custom 대비 Baseline의 상대 차이[%]로, 양수(+)일 경우 Custom이 열위 (Val-Loss 증가, 정확도 감소)를 의미한다. \checkmark 는 $p < 0.05$ 일 때 통계적으로 유의함을 나타낸다. \star ($|d| \geq 1.2$)는 매우 큰 효과(very large effect), $\star\star$ ($0.8 \leq |d| < 1.2$)는 큰 효과(large effect)로 해석한다.

정밀 분석

- 1) **열위 클러스터**: Custom 열위($\text{Val} \uparrow$)는 20NG와 low-separation Gaussian($\text{flip}=0.05$ 포함)에 집중·반복. CIFAR-10/100, MNIST 등 이미지 계열에서는 열위 사례가 없었다.
- 2) **유의성 · 효과크기**: 20NG와 Gauss_sep1.0_clust3_flip0.05 대부분 비교에서 $p < 10^{-2}$ & $|d| \geq 1.2 \rightarrow \star$ 표시(very large). WineQuality는 $p = 0.28$, $|d| = 0.28 \rightarrow$ 통계적으로 유의하지 않고 효과도 작음.
- 3) **정확도 · F_1 동반 하락**: 열위 데이터셋에서 Acc/ F_1 도 2~14%p 감소 \rightarrow 손실 악화가 실제 성능 저하로 직결됨을 확인.
- 4) **전반적 우위 유지**: 열위 2종을 제외한 10/12 데이터셋에서 Custom은 Val-Loss를 평균 20% 이상 감소시키며 일반화 측면 우위(표 5).

세부 수치 · 에폭별 손실 곡선은 부록 E(E.1-E.4)에서 확인할 수 있다.

7.3 실측 학습 시간 비교 및 GPU 활용도 분석

제안 옵티마이저의 이론적 복잡도는 $O(B r n^3)$ 이지만, 실제 구현에서는 작은 n^2 연산을 반복 호출하지 않고 하나의 대규모 텐서 컨트랙션으로 묶어 GPU에서 병렬 실행하여 커널 런칭을 최소화했습니다. 이를 통해 Adam($O(B r n^2)$) 대비 실측 학습 시간에서 일관된 속도 우위를 확보하였습니다.

또한 GPU 활용도를 비교한 결과, baseline(Adam 계열)은 평균 약 36% 수준이었던 반면, 커스텀 옵티마이저는 평균 약 55% 이상을 기록하여 연산 집약도를 크게 향상시켰습니다.

7.4 하이퍼파라미터 민감도 기반 최종 값 선택

네 옵티마이저(Adam, AdamW, AdaBelief, Custom)에 대해 부록 E.5에 제시된 Dropout–Label Smoothing 히트맵을 분석한 결과, $\{0.002, 0.004, 0.008\} \times \{0.0125, 0.025, 0.05\}$ 조합 중 $(0.004, 0.025)$ 이 네 옵티마이저 모두에서 과도한 편차 없이 가장 **평균적인 검증 손실 성능**을 보였습니다. 따라서 보수적으로 Dropout=0.004, Label Smoothing $\alpha = 0.025$ 를 최종 하이퍼파라미터로 채택하였습니다.

8 결론 및 향후 과제

8.1 결론 요약

본 연구에서 제안한 경량 2차 정보 기반 커스텀 옵티마이저는 다음과 같은 주요 성과를 보였다.

- **순수 MLP 실험 (정규화 off):** 훈련 손실은 최대 -95%까지 대폭 감소시켰으나, 검증 손실은 다수 데이터셋에서 오히려 증가하여 과적합 경향을 보였다(표 4).
- **정규화 기법 적용 (Dropout + Label Smoothing):** 검증 손실을 평균 -19.3%--20.1%로 크게 감소시키고, 학습 시간을 약 62% 단축하여(평균 · 중앙값 약 60%) 일반화 및 속도 측면에서 모두 우수한 성능을 달성했다(표 5).
- **배치정규화 미적용 사유:** 제안 옵티마이저는 배치정규화를 적용할 경우 손실 값이 폭발하는 등의 불안정성을 보여, 통상 정규화에 필수적이라 여겨지는 배치정규화를 커스텀 측면에는 적용할 수 없었다. 이에, 비교 대상인 baseline에는 배치정규화를 포함하여도 커스텀 성능이 크게 뒤쳐지지 않음을 보이도록 설계하였다.
- **조기 종료(Early Stopping) 효과:** 부록(E.2, E.4)의 에폭별 검증 손실 곡선을 살펴 보면, 커스텀이 열위에 있던 데이터셋에서도 최적의 시점에 학습을 중단하면 baseline 보다 더 우수한 검증 손실을 달성할 수 있음을 확인할 수 있다.
- **FP16 TensorCore 파일럿 테스트:** 동일한 하드웨어 · 소프트웨어 환경에서 은닉층 7개(각 256 노드), 입력 차원 784, 배치 크기 64 구성의 7-layer Leaky-ReLU MLP를 대상으로 float16 연산과 TensorCore를 활용한 파일럿 테스트를 수행한 결과, 커스텀 옵티마이저 학습 시간이 약 201초로 Adam(≈ 286 초) 대비 약 30% 단축됨을 확인하였다.
- **중대형 규모 확장성:** 충분한 컴퓨팅 리소스를 확보할 경우, 7-layer Leaky-ReLU MLP 파일럿 테스트와 유사하게 본 연구에서 제안한 커스텀 옵티마이저는 중대형 모델에서도 baseline 대비 일관된 학습 시간 단축 효과를 기대할 수 있음을 예비적으로 확인하였다.

8.2 한계점

- 일부 데이터셋에서 조기 종료 전제 하에만 우위가 확보됨.
- FP16 파일럿 테스트 수준으로, 체계적 검증을 위한 대규모 실험이 필요.

8.3 향후 과제

- **경량 2차 기법 실험적 비교:** Shampoo, K-FAC, AdaHessian 등 경량 2차 기법과의 실험적 비교는 후속 연구로 예정한다.
- **저분리 · 노이즈 환경 추가 실험:** 20 Newsgroups 텍스트 및 클래스 경계가 불명확한 low-separation Gaussian 데이터셋에서 제안 옵티마이저의 일반화 성능을 심층 평가 한다.
- **CNN/Transformer 확장 적용:** ResNet, Vision Transformer 등 이미지 · 언어 분야의 대규모 모델에 제안 기법을 적용하여 성능 및 학습 속도 · 메모리 효율을 검증한다.
- **FP16 · TensorCore 커널 최적화:** float16 연산과 TensorCore를 적극 활용하는 맞춤 형 커널을 구현하여 학습 속도 및 에너지 효율을 극대화한다.
- **중대형 규모 스케일업:** 다양한 실세계 데이터셋 및 대규모 MLP/CNN/Transformer에서 학습 시간 · 메모리 효율성을 종합 평가 · 개선한다.

References

- [1] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2933–2941, 2014.
- [2] Behrooz Fazlyab, Carlos Okuno, Yaman Yoshida, Barnabás Póczos, and Milad Razaviyayn. Provable bounds on the hessian of neural networks. *arXiv preprint arXiv:2406.04476*, 2024.
- [3] Behnam Ghorbani, Shiva Prasad Krishnan, and Yingyu Xiao. An investigation into neural net hessian spectra in modern architectures. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2234–2243, 2019.
- [4] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2956–2964, 2018.
- [5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. OpenReview: <https://openreview.net/forum?id=Bkg6RiCqY7>.

- [7] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 2408–2417, 2015.
- [8] Levent Sagun, Léon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.
- [9] James Vo. Efficient second-order neural network optimization via adaptive trust region methods. *arXiv preprint arXiv:2410.02293*, 2024.
- [10] Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Adahessian: An adaptive second order optimizer for machine learning. In *International Conference on Learning Representations (ICLR)*, 2021.
- [11] Yanjun Zhao, Sizhe Dang, Haishan Ye, Guang Dai, Yi Qian, and Ivor Tsang. Second-order fine-tuning without pain for llms: A hessian informed zeroth-order optimizer. In *International Conference on Learning Representations (ICLR)*, 2025. Poster.
- [12] Juntang Zhuang, Tianyun Tang, Yifan Ding, Sekhar Tatikonda, Nicha C. Dvornek, Xenophon Papademetris, and James S. Duncan. Adabelief optimizer: Adapting step-sizes by the belief in observed gradients. *arXiv preprint arXiv:2010.07468*, 2020.

A Newton 방법의 전역 수렴 증명

1. 문제 설정 및 가정

함수 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 가 다음 조건을 만족한다고 가정한다.

1. **강한 볼록성 (Strong Convexity):** 어떤 상수 $m > 0$ 에 대하여

$$\nabla^2 f(x) \succeq mI, \quad \forall x \in \mathbb{R}^n.$$

즉, f 는 m -강한 볼록함을 가지며, 해 x^* 가 유일하게 존재하고

$$\nabla f(x^*) = 0$$

이다.

2. **두번 연속 미분 가정:** $f \in C^2(\mathbb{R}^n)$.

3. **해시안 상한:** 어떤 상수 $L > 0$ 에 대하여

$$\nabla^2 f(x) \preceq L I, \quad \forall x \in \mathbb{R}^n.$$

일반적인 뉴턴 업데이트

$$x_{t+1} = x_t - \alpha [\nabla^2 f(x_t)]^{-1} \nabla f(x_t)$$

대신, 해시안의 대각 성분만을 사용한 업데이트를 고려한다.

2. 대각 업데이트 알고리즘

2.1 대각 근사 행렬 D_t 의 정의

각 반복 t 에서 해시안 $\nabla^2 f(x_t)$ 의 대각 원소만을 모아 대각행렬 $D_t \in \mathbb{R}^{n \times n}$ 을 정의한다:

$$D_t := \text{diag}\left(\frac{\partial^2 f}{\partial x_1^2}(x_t), \frac{\partial^2 f}{\partial x_2^2}(x_t), \dots, \frac{\partial^2 f}{\partial x_n^2}(x_t)\right).$$

강한 볼록성 $\nabla^2 f(x) \succeq mI$ 와 해시안 상한 $\nabla^2 f(x) \preceq L I$ 로부터 모든 t 에 대해

$$m e_i^T e_i = m \leq e_i^T \nabla^2 f(x_t) e_i = \frac{\partial^2 f}{\partial x_i^2}(x_t) \leq L e_i^T e_i = L, \quad i = 1, \dots, n, \quad e_i \text{는 표준 기저 벡터.}$$

따라서

$$m I \preceq D_t \preceq L I, \quad \lambda_{\min}(D_t) = \min_{1 \leq i \leq n} D_{t,ii} \geq m, \quad \lambda_{\max}(D_t) = \max_{1 \leq i \leq n} D_{t,ii} \leq L.$$

2.2 대각 근사 업데이트 식

스칼라 스텝 크기 $\alpha > 0$ 를 고정하고, 다음과 같이 업데이트한다:

$$x_{t+1} = x_t - \alpha D_t^{-1} \nabla f(x_t).$$

오차를

$$e_t := \|x_t - x^*\|$$

로 정의할 때, 이 업데이트가 적절한 α 하에서 전역 선형 수렴을 갖는지 증명한다.

3. 증명

$f \in C^2$ 이고 $\nabla^2 f(x) \preceq L I$ 이므로 (따라서 f 는 L-smooth하다.) 임의의 $x \in \mathbb{R}^n$, $\Delta \in \mathbb{R}^n$ 에 대하여 다변수 라그랑주 나머지 항에 의해 (자세한 증명은 부록 B를 참조하라.)

$$f(x + \Delta) = f(x) + \nabla f(x)^T \Delta + \frac{1}{2} \Delta^T \nabla^2 f(c) \Delta, \quad c \in \{x + \theta \Delta : 0 < \theta < 1\}.$$

따라서

$$\Delta^T \nabla^2 f(c) \Delta \leq L \|\Delta\|^2 \implies f(x + \Delta) \leq f(x) + \nabla f(x)^T \Delta + \frac{L}{2} \|\Delta\|^2.$$

이에, $x = x_t$, $\Delta = -\alpha D_t^{-1} \nabla f(x_t)$ 를 대입하여

$$\begin{aligned} f(x_{t+1}) &= f(x_t - \alpha D_t^{-1} \nabla f(x_t)) \leq f(x_t) + \nabla f(x_t)^T (-\alpha D_t^{-1} \nabla f(x_t)) + \frac{L}{2} \|-\alpha D_t^{-1} \nabla f(x_t)\|^2 \\ &= f(x_t) - \alpha \nabla f(x_t)^T D_t^{-1} \nabla f(x_t) + \frac{L\alpha^2}{2} \|D_t^{-1} \nabla f(x_t)\|^2. \end{aligned}$$

또한 $mI \preceq D_t \preceq L I$ 므로

$$\|D_t^{-1}\| \leq \frac{1}{m}, \quad (\text{여기서 } \|\cdot\| \text{는 스펙트럼 노름임을 의미한다.})$$

$$\nabla f(x_t)^T D_t^{-1} \nabla f(x_t) \geq \frac{1}{\lambda_{\max}(D_t)} \|\nabla f(x_t)\|^2 \geq \frac{1}{L} \|\nabla f(x_t)\|^2.$$

특히,

$$\|D_t^{-1} \nabla f(x_t)\|^2 \leq \|D_t^{-1}\|^2 \|\nabla f(x_t)\|^2 \leq \frac{1}{m^2} \|\nabla f(x_t)\|^2.$$

이를 이용하여

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \alpha \nabla f(x_t)^T D_t^{-1} \nabla f(x_t) + \frac{L\alpha^2}{2} \|D_t^{-1} \nabla f(x_t)\|^2 \\ &\leq f(x_t) - \alpha \cdot \frac{1}{L} \|\nabla f(x_t)\|^2 + \frac{L\alpha^2}{2} \cdot \frac{1}{m^2} \|\nabla f(x_t)\|^2 \\ &= f(x_t) - \left(\frac{\alpha}{L} - \frac{L\alpha^2}{2m^2}\right) \|\nabla f(x_t)\|^2. \end{aligned}$$

역시 다변수 라그랑주 나머지 항에 의해

$$f(x_t) = f(x^*) + \nabla f(x^*) (x_t - x^*) + \frac{1}{2} (x_t - x^*)^T \nabla^2 f(c') (x_t - x^*)$$

가 성립한다. 여기서 $c' \in \{x^* + \theta'(x_t - x^*) : 0 < \theta' < 1\}$ 이다. $\nabla f(x^*) = 0$ 으로

$$f(x_t) - f(x^*) = \frac{1}{2} (x_t - x^*)^T \nabla^2 f(c') (x_t - x^*) \text{이다.}$$

따라서

$$\frac{m}{2} \|x_t - x^*\|^2 \leq f(x_t) - f(x^*) \leq \frac{L}{2} \|x_t - x^*\|^2 \text{이 성립한다.}$$

또한,

$$\|\nabla f(x_t)\|^2 \geq \frac{m}{2} (f(x_t) - f(x^*)),$$

가 성립하므로,

$$\begin{aligned} f(x_{t+1}) - f(x^*) &\leq f(x_t) - f(x^*) - \left(\frac{\alpha}{L} - \frac{L\alpha^2}{2m^2}\right) \frac{m}{2} (f(x_t) - f(x^*)) \\ &= \left[1 - \left(\frac{\alpha}{L} - \frac{L\alpha^2}{2m^2}\right) \frac{m}{2}\right] (f(x_t) - f(x^*)). \end{aligned}$$

$r(\alpha) := \left(\frac{\alpha}{L} - \frac{L\alpha^2}{2m^2} \right) \frac{m}{2}$ 라 정의하면

$$f(x_{t+1}) - f(x^*) \leq [1 - r(\alpha)] (f(x_t) - f(x^*)).$$

$r(\alpha) > 0$] 려면

$$0 < \alpha < \frac{2m^2}{L^2}$$

을 만족해야 한다. 특히 $\alpha = \frac{m^2}{L^2}$ 일 때

$$r(\alpha) = \left(\frac{m^2/L^2}{L} - \frac{L(m^2/L^2)^2}{2m^2} \right) \frac{m}{2} = \left(\frac{m^2}{L^3} - \frac{m^4}{2L^5} \right) \frac{m}{2} = \frac{m^3}{2L^3} - \frac{m^5}{4L^5} = \frac{m^3}{4L^3}.$$

으로 $r(\alpha)$ 는 최댓값을 갖으며 이때

$$0 < 1 - \frac{m^3}{4L^3} < 1$$

이 성립한다. 따라서

$$\rho := 1 - r(\alpha), \quad 0 < \rho < 1, \quad f(x_{t+1}) - f(x^*) \leq \rho (f(x_t) - f(x^*)),$$

즉 함수값의 차이는 선형 수렴한다.

또한

$$\frac{m}{2} \|x_t - x^*\|^2 \leq f(x_t) - f(x^*) = O(\rho^t) \implies \|x_t - x^*\| \rightarrow 0 \quad (t \rightarrow \infty).$$

이로써 x_t 는 전역 선형 수렴한다.

그때의 α 는

$$0 < \alpha < \frac{2m^2}{L^2}$$

을 만족한다.

부록: $\|\nabla f(x_t)\|^2 \geq \frac{m}{2} [f(x_t) - f(x^*)]$ 증명

$g(t) = f(x^* + t(x_t - x^*))$ 라고 하자.

$g''(t) = (x_t - x^*)^T \nabla^2 f(x^* + t(x_t - x^*)) (x_t - x^*) \geq m \|x_t - x^*\|^2 \geq 0$, (따라서 g' 는 증가함수이다).

$$g(1) - g(0) = f(x_t) - f(x^*) = g'(c) = \nabla f(x^* + c(x_t - x^*)) \cdot (x_t - x^*)$$

$$\leq g'(1) = \nabla f(x_t) \cdot (x_t - x^*) \leq \|\nabla f(x_t)\| \|x_t - x^*\|.$$

$$\therefore \frac{m}{2} \|x_t - x^*\|^2 \leq f(x_t) - f(x^*), \frac{m}{2} [f(x_t) - f(x^*)] \leq \frac{m}{2} \|\nabla f(x_t)\| \|x_t - x^*\|,$$

$$\frac{m}{2} \|x_t - x^*\| \|\nabla f(x_t)\| \leq \|\nabla f(x_t)\|^2, \quad \therefore \frac{m}{2} [f(x_t) - f(x^*)] \leq \|\nabla f(x_t)\|^2.$$

B 다변수 라그랑주 나머지항 정리

다변수 Taylor 전개 정리 (Lagrange 나머지형)

정리. 다변수 함수 $f: D \rightarrow \mathbb{R}$ 가 $D \subset \mathbb{R}^{d+1}$ 에서 C^{n+1} 이고, 임의의 $a, x \in D$ 대하여

$$\{a + t(x - a) \mid t \in [0, 1]\} \subset D$$

가 성립한다고 하자. 그러면 어떤 $\theta \in (0, 1)$ 가 존재하여 다음 식이 성립한다:

$$f(x) = \sum_{|\alpha| \leq n} \frac{D^\alpha f(a)}{\alpha!} (x-a)^\alpha + \sum_{|\alpha|=n+1} \frac{D^\alpha f(a + \theta(x-a))}{\alpha!} (x-a)^\alpha, \quad (6)$$

여기서 $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$ 는 multi-index이다.

$$|\alpha| = \alpha_1 + \dots + \alpha_d, \quad \alpha! = \alpha_1! \cdots \alpha_d!, \quad (x-a)^\alpha = (x_1 - a_1)^{\alpha_1} \cdots (x_d - a_d)^{\alpha_d},$$

$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$ 로 정의한다.

증명.

(i) 1변수 함수 도함수 계산.

우선 다음과 같이 1변수 함수 g 를 정의한다:

$$g: [0, 1] \longrightarrow \mathbb{R}, \quad g(t) = f(a + t(x-a)).$$

$f \in C^{n+1}(D)$ 이고 선분 $\{a + t(x-a) \mid t \in [0, 1]\} \subset D$ 므로 $g \in C^{n+1}[0, 1]$ 이다.

체인 룰과 multi-index 표기법을 이용하면, 임의의 $k = 0, \dots, n+1$ 에 대하여

$$g^{(k)}(t) = \sum_{|\alpha|=k} \frac{k!}{\alpha!} D^\alpha f(a + t(x-a)) (x-a)^\alpha.$$

특히 $t=0$ 일 때

$$\frac{g^{(k)}(0)}{k!} = \sum_{|\alpha|=k} \frac{D^\alpha f(a)}{\alpha!} (x-a)^\alpha.$$

(ii) 1변수 Taylor 정리 (Lagrange 나머지형).

$g \in C^{n+1}[0, 1]$ 이므로, 1변수 Taylor 정리(잔차 Lagrange 형식)를 적용하면 (자세한 증명은 부록 C를 참조하라.) $\theta \in (0, 1)$ 가 존재하여

$$g(1) = \sum_{k=0}^n \frac{g^{(k)}(0)}{k!} + \frac{g^{(n+1)}(\theta)}{(n+1)!} (1-0)^{n+1}.$$

(iii) 다변수 식으로 정리.

위 식에서 $g(1) = f(x)$ 이며, $\frac{g^{(k)}(0)}{k!}$ 과 $g^{(n+1)}(\theta)$ 를 (i)에서 구한 표현으로 대체하면

$$f(x) = \sum_{k=0}^n \sum_{|\alpha|=k} \frac{D^\alpha f(a)}{\alpha!} (x-a)^\alpha + \sum_{|\alpha|=n+1} \frac{D^\alpha f(a + \theta(x-a))}{\alpha!} (x-a)^\alpha.$$

이는 바로 (6)와 동일하다. \square

C 일변수 라그랑주 나머지항 정리

일변수 테일러 정리 (라그랑주 나머지항 형)

Taylor 정리 (Lagrange 나머지항). 함수 $f: [a, b] \rightarrow \mathbb{R}$ 가 $n+1$ 번 연속 미분 가능하다고 하자. 그러면 임의의 $x \in [a, b]$ 에 대하여

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_{n+1}(x),$$

여기서 나머지항 $R_{n+1}(x)$ 은 다음과 같이 표현된다:

$$R_{n+1}(x) = \frac{f^{(n+1)}(c)}{(n+1)!} (x-a)^{n+1}, \quad 어떤 c \in (a, x) (또는 c \in (x, a)) 가 존재한다.$$

특히 $c = a + \theta(x-a)$, $0 < \theta < 1$ 의 형태로 나타낼 수 있다.

Proof. 증명은 수학적 귀납법(induction)을 사용한다.

※ $x = a$ 경계점에 대한 언급.

만약 $x = a$ 인 경우, 우변에 보이는 모든 $(x-a)$ 인자가 0이 되어 나머지항 $R_{n+1}(a)$ 은 자명하게 0이 된다. 즉, 경계점 $x = a$ 에서는 이미

$$f(a) = f(a) + 0 + \cdots + 0 + R_{n+1}(a), \quad R_{n+1}(a) = 0$$

형태로 성립하므로, 본격적인 귀납법 증명은 $x \neq a$ 인 경우에 집중해도 무리가 없다.

1. 기본 단계 (Base Case: $n = 0$).

함수 f 가 한 번 연속 미분 가능하다고 가정하면, $n = 0$ 일 때

$$f(x) = f(a) + R_1(x), \quad R_1(x) = f(x) - f(a).$$

평균값 정리(Mean Value Theorem)를 f 에 적용하면

$$\frac{f(x) - f(a)}{x - a} = f'(\xi), \quad \xi \in (a, x),$$

따라서

$$R_1(x) = f(x) - f(a) = f'(\xi)(x-a) = \frac{f'(1)(\xi)}{1!} (x-a)^1, \quad \xi \in (a, x).$$

즉 $n = 0$ 인 경우에도 정리가 성립함을 보였다.

2. 귀납 가정 (Induction Hypothesis: $n = k-1$).

어떤 양의 정수 $k \geq 1$ 에 대하여, f 가 k 번 연속 미분 가능하고

$$f(x) = \sum_{i=0}^{k-1} \frac{f^{(i)}(a)}{i!} (x-a)^i + R_k(x),$$

여기서 잔여항 $R_k(x)$ 은

$$R_k(x) = \frac{f^{(k)}(c)}{k!} (x-a)^k, \quad c \in (a, x)$$

꼴로 표현된다고 가정한다. 이를 귀납 가정이라 한다.

3. 귀납 단계 (Induction Step: $n = k$).

함수 f 가 $(k+1)$ 번 연속 미분 가능하다고 가정하고, 먼저 귀납 가정에서 얻은 식을 기억하자:

$$f(x) = \sum_{i=0}^{k-1} \frac{f^{(i)}(a)}{i!} (x-a)^i + R_k(x), \quad R_k(x) = \frac{f^{(k)}(c)}{k!} (x-a)^k, \quad c \in (a, x).$$

이제 이 $R_k(x)$ 를 “테일러 다항식의 $i = k$ 항”과 “ $(k+1)$ 차 잔여항”으로 분리하여

$$f(x) = \sum_{i=0}^k \frac{f^{(i)}(a)}{i!} (x-a)^i + R_{k+1}(x), \quad R_{k+1}(x) = \frac{f^{(k+1)}(c')}{(k+1)!} (x-a)^{k+1}, \quad c' \in (a, x)$$

형태를 보이는 것이 목표이다. 이를 위해 두 단계를 거친다.

(Step 3-1) 잔여항 $R_k(x)$ 에 평균값 정리를 적용.

$$R_k(x) = \frac{f^{(k)}(c)}{k!} (x-a)^k, \quad c \in (a, x).$$

함수 $f^{(k)}$ 는 연속 미분 가능하므로, $a < c < x$ 구간에서 평균값 정리를 적용하면

$$\frac{f^{(k)}(c) - f^{(k)}(a)}{c-a} = f^{(k+1)}(\xi), \quad \xi \in (a, c) \subset (a, x).$$

* $x < a$ 인 경우에 대한 언급.

본 증명 전체에서 “ $\xi \in (a, x)$ ”라 표현된 부분은, 만약 $x < a$ 일 때는 “ $\xi \in (x, a)$ ”로 읽으면 된다. 즉, 평균값 정리를 적용할 때 구간의 방향이 바뀌는 것뿐이며, 논리 전개는 전혀 달라지지 않는다. 필요하다면 “ $\xi \in (\min\{a, x\}, \max\{a, x\})$ ”로 통일하여 써도 좋다.

따라서

$$f^{(k)}(c) = f^{(k)}(a) + f^{(k+1)}(\xi) (c-a).$$

또한 $c \in (a, x)$ 이므로 $c = a + \theta(x-a)$ ($0 < \theta < 1$)로 쓸 수 있다. 이때 $c-a = \theta(x-a)$ 이므로,

$$f^{(k)}(c) = f^{(k)}(a) + f^{(k+1)}(\xi) \theta(x-a), \quad \xi \in (a, c) \subset (a, x).$$

이를 $R_k(x)$ 에 대입하면

$$\begin{aligned} R_k(x) &= \frac{f^{(k)}(c)}{k!} (x-a)^k = \frac{1}{k!} \left[f^{(k)}(a) + f^{(k+1)}(\xi) \theta(x-a) \right] (x-a)^k \\ &= \underbrace{\frac{f^{(k)}(a)}{k!} (x-a)^k}_{\text{테일러 다항식의 } i=k \text{ 항}} + \underbrace{\frac{f^{(k+1)}(\xi)}{k!} \theta(x-a)^{k+1}}_{\text{남은 } (k+1)\text{차 잔여항}}. \end{aligned}$$

첫 번째 항 $\frac{f^{(k)}(a)}{k!} (x-a)^k$ 은 이미 $\sum_{i=0}^k \frac{f^{(i)}(a)}{i!} (x-a)^i$ 의 $i = k$ 항이므로 합칠 수 있다. 따라서 진정한 $(k+1)$ 차 잔여항은

$$R_{k+1}(x) = \frac{f^{(k+1)}(\xi)}{k!} \theta(x-a)^{k+1}, \quad \xi \in (a, x), \quad \theta \in (0, 1).$$

이것이 “평균값 정리만으로 유도되는 형태”이고, $\xi = a + \theta(x-a)$ 이다.

* c 와 ξ 의 역할에 대한 부연 설명.

- Step 3-1에서

$$R_k(x) = \frac{f^{(k)}(c)}{k!} (x-a)^k, \quad c \in (a, x)$$

일 때, c 는 귀납 가정으로 이미 (a, x) 구간 안에 존재하는 한 점이다. 이어서 두 번째 평균값 정리를 적용하는 과정에서

$$c = a + \theta(x-a), \quad 0 < \theta < 1, \quad \xi \in (a, c) \subset (a, x)$$

라는 두 변수가 등장한다.

- 이후 “라그랑주 나머지 형태”를 맞추려면 θ 와 ξ 를 “하나의 새로운 점 c' ”로 합치는 과정을 거친다. Step 3-2를 통해 $\theta = \frac{1}{k+1}$ 임을 보였으므로, 최종적으로

$$c' = a + \frac{1}{k+1}(x-a) \in (a, x)$$

를 취하거나, 또는 평균값 정리를 한 번 더 적용하여 (실제 존재하는) 어떤 $c' \in (a, x)$ 를 확정하면,

$$R_{k+1}(x) = \frac{f^{(k+1)}(c')}{(k+1)!} (x-a)^{k+1}$$

꼴로 정확하게 표현할 수 있다.

(Step 3-2) $(k+1)$ 차 도함수 일치로부터 $\theta = \frac{1}{k+1}$ 결정.

※ 고차 도함수 존재 범위에 대한 부연.

Leibniz 공식을 이용하여

$$\frac{d^{k+1}}{dx^{k+1}} [f^{(k+1)}(a + \theta(x-a)) (x-a)^{k+1}]$$

를 전개할 때, 겉보기에는 $G^{(j)}(x)$ 항에 $f^{(k+1+j)}$ (즉 최대 $2k+2$ 차 도함수)까지 필요할 것처럼 보인다. 그러나 실제로는 **평가 지점이 $x=a$ 하나뿐**이므로,

$$\Psi(x) = (x-a)^{k+1} \implies \Psi^{(m)}(a) = 0 \quad (0 \leq m \leq k),$$

따라서 $j \geq 1$ 항들은 모두 $(x-a)$ 인자를 하나 이상 가지고 있어 $x=a$ 에서 0으로 사라진다. 결국 $G^{(j)}(a)$ 의 존재는 $f^{(k+1)}$ 가 **연속**이라는 가정만으로 충분하므로, 함수가 $(k+1)$ 번 연속 미분 가능하다는 조건만으로 단계가 성립한다. 위에서 이미 $R_{k+1}(x) = \frac{f^{(k+1)}(\xi)}{k!} \theta (x-a)^{k+1}$ 꼴임을 알았지만, 잔여항을 “ θ 포함 형태”로 가정하고 $(k+1)$ 차 미분 과정을 통해 $\theta = \frac{1}{k+1}$ 임을 명시적으로 보이는 방법도 있다. 즉, 처음부터

$$R_{k+1}(x) = \theta \frac{f^{(k+1)}(a + \theta(x-a))}{k!} (x-a)^{k+1}, \quad 0 < \theta < 1$$

라고 가정하고, 이 식이 참이라면 양변을 $(k+1)$ 차 미분했을 때:

$$\frac{d^{k+1}}{dx^{k+1}} [f(x)] = \frac{d^{k+1}}{dx^{k+1}} \left[\sum_{i=0}^k \frac{f^{(i)}(a)}{i!} (x-a)^i + \theta \frac{f^{(k+1)}(a + \theta(x-a))}{k!} (x-a)^{k+1} \right]$$

식이 성립해야 한다. 먼저 왼변은

$$\frac{d^{k+1}}{dx^{k+1}} [f(x)] = f^{(k+1)}(x).$$

오른쪽에서 $\sum_{i=0}^k \frac{f^{(i)}(a)}{i!} (x-a)^i$ 는 다항식이므로 $(k+1)$ 회 미분 시 모두 0이 되고, 남는 것은

$$\theta \frac{d^{k+1}}{dx^{k+1}} \left[\frac{1}{k!} f^{(k+1)}(a + \theta(x-a)) (x-a)^{k+1} \right].$$

○ 부분을 계산하기 위해

$$G(x) := f^{(k+1)}(a + \theta(x-a)), \quad \Psi(x) := (x-a)^{k+1}$$

로 두자. 그러면

$$\frac{d^{k+1}}{dx^{k+1}} [G(x) \Psi(x)] = \sum_{j=0}^{k+1} \binom{k+1}{j} G^{(j)}(x) \Psi^{(k+1-j)}(x).$$

단,

$$\Psi(x) = (x-a)^{k+1} \implies \Psi^{(m)}(x) = \begin{cases} \frac{(k+1)!}{(k+1-m)!} (x-a)^{k+1-m}, & 0 \leq m \leq k, \\ (k+1)!, & m = k+1. \end{cases}$$

○므로 “ $\Psi^{(k+1-j)}(a) \neq 0$ ”가 되려면 $k+1-j = k+1$, 즉 $j=0$ 이어야 한다. 따라서 합을

$$\underbrace{\binom{k+1}{0} G^{(0)}(x) \Psi^{(k+1)}(x)}_{j=0 \text{ 항}} + \underbrace{\sum_{j=1}^{k+1} \binom{k+1}{j} G^{(j)}(x) \Psi^{(k+1-j)}(x)}_{=:g(x)}$$

로 나눌 수 있다.

- $j=0$ 때: $\binom{k+1}{0} = 1$, $G^{(0)}(x) = G(x)$, $\Psi^{(k+1)}(x) = (k+1)!$ ○므로
 $j=0$ 때 $= (k+1)! f^{(k+1)}(a + \theta(x-a)).$

- $j \geq 1$ 때:

$$g(x) := \sum_{j=1}^{k+1} \binom{k+1}{j} G^{(j)}(x) \Psi^{(k+1-j)}(x).$$

여기서 $1 \leq j \leq k$ 인 경우

$$\Psi^{(k+1-j)}(x) = \frac{(k+1)!}{j!} (x-a)^j$$

로서 $(x-a)$ 인자가 최소 하나 이상 남아 있으므로 “ $x=a$ 에서 $\Psi^{(k+1-j)}(a) = 0$ ”.
 $j=k+1$ 인 경우 $\Psi^{(k+1-(k+1))}(x) = \Psi^{(0)}(x) = (x-a)^{k+1}$ 역시 $x=a$ 에서 0이 된다.
 결론적으로 $g(a) = 0$.

따라서

$$\frac{d^{k+1}}{dx^{k+1}} [G(x) \Psi(x)] = (k+1)! f^{(k+1)}(a + \theta(x-a)) + g(x).$$

○를 원래 식에 대입하면,

$$f^{(k+1)}(x) = \theta \frac{1}{k!} \left[(k+1)! f^{(k+1)}(a + \theta(x-a)) + g(x) \right],$$

즉

$$f^{(k+1)}(x) = (k+1)\theta f^{(k+1)}(a + \theta(x-a)) + \theta \frac{1}{k!} g(x).$$

o) 식을 간단히 정리하여

$$f^{(k+1)}(x) = (k+1)\theta f^{(k+1)}(a + \theta(x-a)) + \theta \frac{1}{k!} g(x), \quad g(x) := \sum_{j=1}^{k+1} \binom{k+1}{j} G^{(j)}(x) \Psi^{(k+1-j)}(x).$$

(Step 3) $x = a$ 에서의 평가 및 θ 결정

위 식을 $x = a$ 에 대입하면,

$$f^{(k+1)}(a) = (k+1)\theta f^{(k+1)}(a + \theta(a-a)) + \theta \frac{1}{k!} g(a) = (k+1)\theta f^{(k+1)}(a) + 0 \quad (\because g(a) = 0).$$

따라서 모든 C^{k+1} 함수 f 에 대해 다음 식이 항상 성립해야 한다:

$$f^{(k+1)}(a) - (k+1)\theta f^{(k+1)}(a) = 0 \iff f^{(k+1)}(a)[1 - (k+1)\theta] = 0.$$

여기서 중요한 점은 “어떤 함수에서는 $f^{(k+1)}(a) = 0$ 이고, 다른 함수에서는 $f^{(k+1)}(a) \neq 0$ ” 인 경우를 모두 고려하더라도 위 등식이 항상 참이 되어야 한다는 것이다. 그러므로 곱해진 계수

$$1 - (k+1)\theta$$

o) 반드시 0이어야만 하고, 그 결과

$$\theta = \frac{1}{k+1}$$

임을 유도할 수 있다. 따라서 최종 잔여항

$$\begin{aligned} R_{k+1}(x) &= \theta \frac{f^{(k+1)}(a + \theta(x-a))}{k!} (x-a)^{k+1} \\ &= \frac{1}{k+1} \frac{f^{(k+1)}\left(a + \frac{1}{k+1}(x-a)\right)}{k!} (x-a)^{k+1} \\ &= \frac{f^{(k+1)}(c)}{(k+1)!} (x-a)^{k+1}, \quad c = a + \frac{1}{k+1}(x-a) \in (a, x). \end{aligned}$$

결국

$$f(x) = \sum_{i=0}^k \frac{f^{(i)}(a)}{i!} (x-a)^i + \frac{f^{(k+1)}(c)}{(k+1)!} (x-a)^{k+1}, \quad c = a + \frac{1}{k+1}(x-a) \in (a, x),$$

가 되어 $n = k$ 인 경우에도 테일러-라그랑주 정리가 성립함을 보였다. 이로써 수학적 귀납법에 의해 임의의 정수 $n \geq 0$ 에 대해 원래 정리가 모두 성립함이 증명되었다.

$$\begin{aligned} f(x) &= \sum_{i=0}^k \frac{f^{(i)}(a)}{i!} (x-a)^i \\ &\quad + \frac{f^{(k+1)}(c)}{(k+1)!} (x-a)^{k+1}, \\ c &= a + \frac{1}{k+1}(x-a) \in (a, x). \end{aligned}$$

□

D 데이터셋 출처

Table 7: 데이터셋 출처 (URL · 저자 · 연도 · 라이선스)

데이터셋	출처
MNIST	LeCun et al. (1998), http://yann.lecun.com/exdb/mnist , CC BY-SA 3.0
CIFAR-10/100	Krizhevsky et al. (2009), https://www.cs.toronto.edu/~kriz/cifar.html , MIT License
20 Newsgroups	Lang (1995), http://qwone.com/~jason/20Newsgroups/ , Public Domain
Imbalance	Scikit-learn 예제 데이터, https://scikit-learn.org/ , BSD License
WineQuality-Red	Cortez et al. (2009), UCI ML Repo, https://archive.ics.uci.edu/ml/datasets/Wine+Quality , Public Domain
Fashion-MNIST	Xiao et al. (2017), https://github.com/zalandoresearch/fashion-mnist , MIT License
HAR	Anguita et al. (2013), UCI ML Repo, https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones , Public Domain
합성 Gaussian	본 연구 생성 (scikit-learn BSD License)

E 실험 결과 (250 epochs) 부록

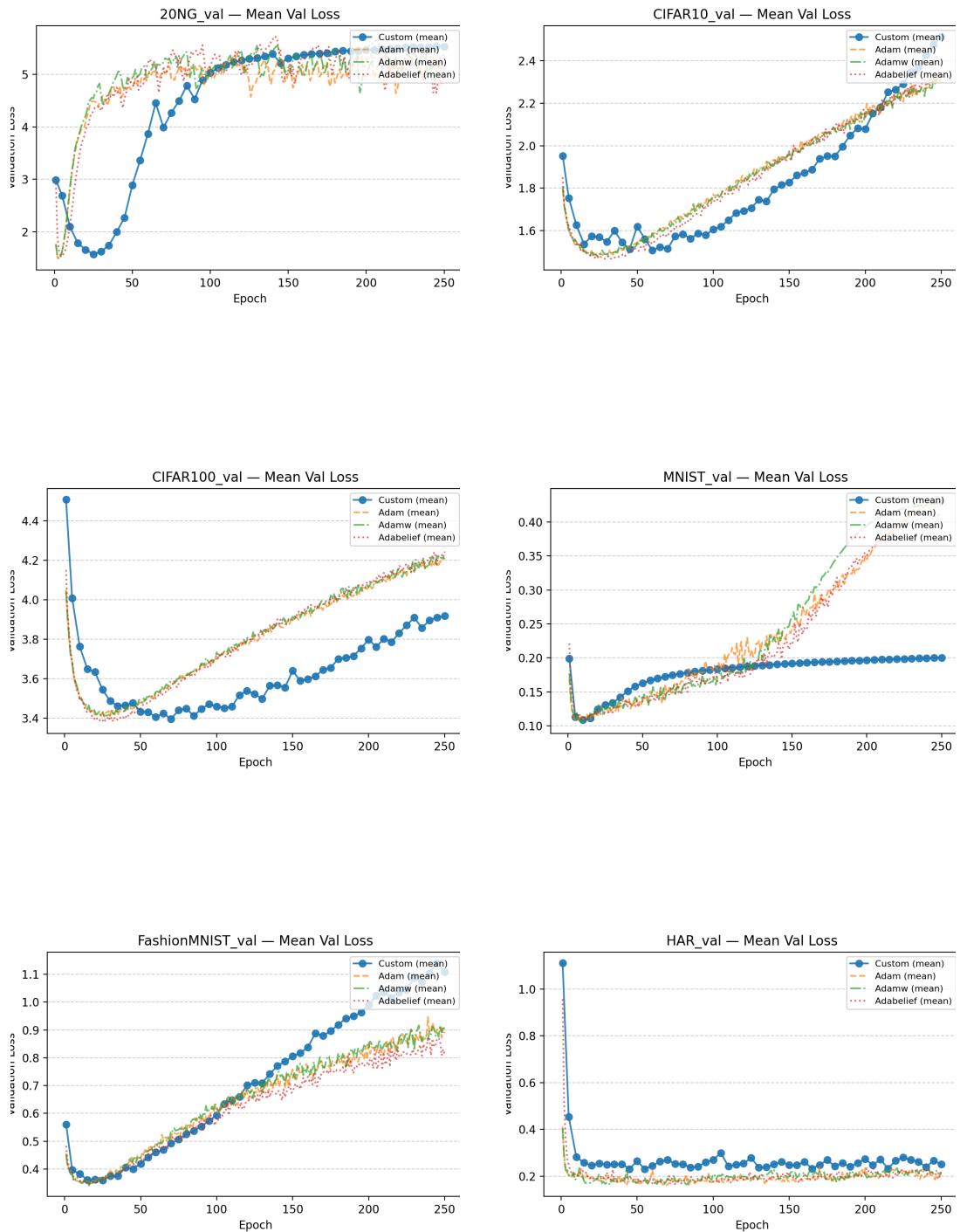
E.1 Hidden7 No Regulation

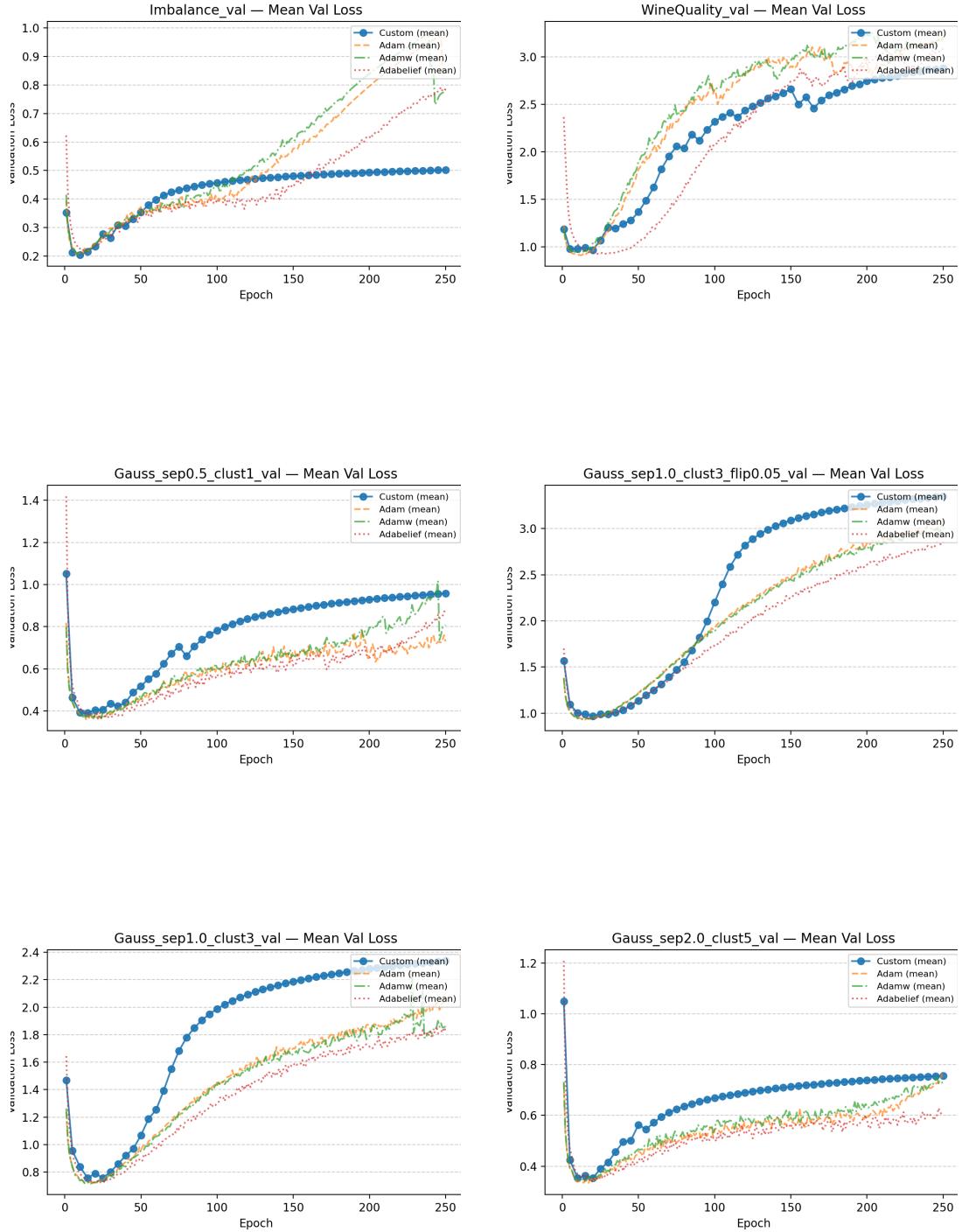
TRAIN Results								VAL Results							
	Custom	Adam	Adamw	Adabelief	p_c_vs_a	p_c_vs_aw	p_c_vs_ab		Custom	Adam	Adamw	Adabelief	p_c_vs_a	p_c_vs_aw	p_c_vs_ab
MNIST	0.0	0.0	0.0	0.0	0.0	0.0	0.0	MNIST	0.2003	0.4329	0.428	0.4111	0.0	0.0	0.0
CIFAR10	0.7277	0.7849	0.7933	0.7213	0.0373	0.0083	0.7472	CIFAR10	2.5121	2.327	2.334	2.3341	0.0882	0.0776	0.0403
CIFAR100	2.45	2.448	2.446	2.3281	0.9612	0.901	0.0149	CIFAR100	3.917	4.2137	4.2053	4.2443	0.0	0.0009	0.0
20NG	0.0905	0.0937	0.0908	0.0939	0.1326	0.0031	0.2458	20NG	5.5281	5.1607	5.3914	4.964	0.1316	0.428	0.0319
Imbalance	0.0	0.0063	0.0055	0.0013	0.3577	0.1806	0.3648	Imbalance	0.5022	0.8549	0.7853	0.7864	0.0168	0.0504	0.0112
WineQuality	0.0029	0.0045	0.004	0.0052	0.0781	0.2337	0.0928	WineQuality	2.8786	3.1925	3.2007	3.0861	0.1774	0.0616	0.1311
FashionMNIST	0.0207	0.0411	0.0401	0.0407	0.0028	0.0028	0.0009	FashionMNIST	1.1088	0.8888	0.9017	0.8121	0.0011	0.0068	0.0005
HAR	0.0511	0.0184	0.0207	0.0203	0.0004	0.001	0.0002	HAR	0.2507	0.2094	0.2063	0.219	0.0039	0.0124	0.0715
Gauss_sep0.5_clust1	0.0001	0.0368	0.0249	0.0161	0.0252	0.0239	0.1893	Gauss_sep0.5_clust1	0.9584	0.7298	0.8071	0.8728	0.0002	0.0964	0.2043
Gauss_sep1.0_clust3	0.0002	0.0581	0.0519	0.0691	0.012	0.0003	0.0012	Gauss_sep1.0_clust3	2.3376	2.0799	1.8611	1.8493	0.0219	0.0003	0.0
Gauss_sep2.0_clust5	0.0001	0.0085	0.0088	0.0213	0.1743	0.1017	0.0463	Gauss_sep2.0_clust5	0.7562	0.7503	0.7352	0.5986	0.8987	0.807	0.0004
Gauss_sep1.0_clust3_flip0.05	0.0003	0.101	0.1165	0.1292	0.0	0.0027	0.0	Gauss_sep1.0_clust3_flip0.05	3.3487	3.1043	3.0213	2.86	0.0029	0.0002	0.0

ACC Results								F1 Results							
	Custom	Adam	Adamw	Adabelief	p_c_vs_a	p_c_vs_aw	p_c_vs_ab		Custom	Adam	Adamw	Adabelief	p_c_vs_a	p_c_vs_aw	p_c_vs_ab
MNIST	0.9775	0.9803	0.9801	0.9794	0.0038	0.0029	0.0324	MNIST	0.9773	0.9801	0.9799	0.9792	0.0039	0.0026	0.0293
CIFAR10	0.4819	0.4529	0.4518	0.463	0.0002	0.0002	0.0001	CIFAR10	0.4814	0.4514	0.4513	0.4614	0.0	0.0002	0.0001
CIFAR100	0.1917	0.1821	0.1826	0.1908	0.0094	0.0014	0.7564	CIFAR100	0.1839	0.1745	0.176	0.1847	0.0104	0.0064	0.7922
20NG	0.5482	0.5515	0.5528	0.5473	0.1866	0.1749	0.8173	20NG	0.5427	0.5464	0.5481	0.5425	0.1414	0.1035	0.9572
Imbalance	0.9525	0.958	0.9537	0.9585	0.0023	0.6124	0.0001	Imbalance	0.9017	0.9139	0.9046	0.9152	0.0013	0.5781	0.0003
WineQuality	0.6576	0.6571	0.6683	0.6437	0.9775	0.4457	0.2389	WineQuality	0.4352	0.4241	0.437	0.4082	0.4657	0.9263	0.1567
FashionMNIST	0.8784	0.8831	0.8836	0.8841	0.0084	0.0049	0.0008	FashionMNIST	0.8783	0.8831	0.8837	0.8842	0.0052	0.004	0.001
HAR	0.9397	0.9573	0.9564	0.956	0.0003	0.0018	0.0	HAR	0.9395	0.958	0.9571	0.9569	0.0003	0.0009	0.0
Gauss_sep0.5_clust1	0.9005	0.8939	0.8942	0.901	0.157	0.2875	0.9212	Gauss_sep0.5_clust1	0.9005	0.8938	0.8941	0.9009	0.1547	0.2853	0.935
Gauss_sep1.0_clust3	0.7852	0.7686	0.7711	0.7665	0.0216	0.0034	0.0003	Gauss_sep1.0_clust3	0.7852	0.7684	0.771	0.7665	0.0213	0.003	0.0004
Gauss_sep2.0_clust5	0.9161	0.9216	0.9189	0.9163	0.1705	0.4596	0.9508	Gauss_sep2.0_clust5	0.9161	0.9216	0.9189	0.9163	0.1723	0.4578	0.9546
Gauss_sep1.0_clust3_flip0.05	0.7158	0.6872	0.6896	0.6869	0.0001	0.0	0.0	Gauss_sep1.0_clust3_flip0.05	0.7157	0.6868	0.6893	0.6867	0.0001	0.0	0.0

TIME Results							
	Custom	Adam	Adamw	Adabelief	p_c_vs_a	p_c_vs_aw	p_c_vs_ab
MNIST	144.1717	247.7555	254.9061	259.5063			
CIFAR10	169.7086	256.3487	261.5442	265.9932			
CIFAR100	197.1079	276.4862	287.2436	289.3341			
20NG	50.1049	86.0445	91.1562	93.4816			
Imbalance	55.3221	112.9044	115.5945	118.6108			
WineQuality	7.3027	11.9501	13.1443	14.1618			
FashionMNIST	150.2349	270.2277	276.7918	280.9411			
HAR	25.5033	49.6695	51.5042	50.9249			
Gauss_sep0.5_clust1	54.0626	109.2435	113.3247	114.6174			
Gauss_sep1.0_clust3	53.0335	109.8698	113.3813	114.7416			
Gauss_sep2.0_clust5	52.844	110.2151	112.9742	114.4973			
Gauss_sep1.0_clust3_flip0.05	52.8154	109.7917	112.6658	113.6371			

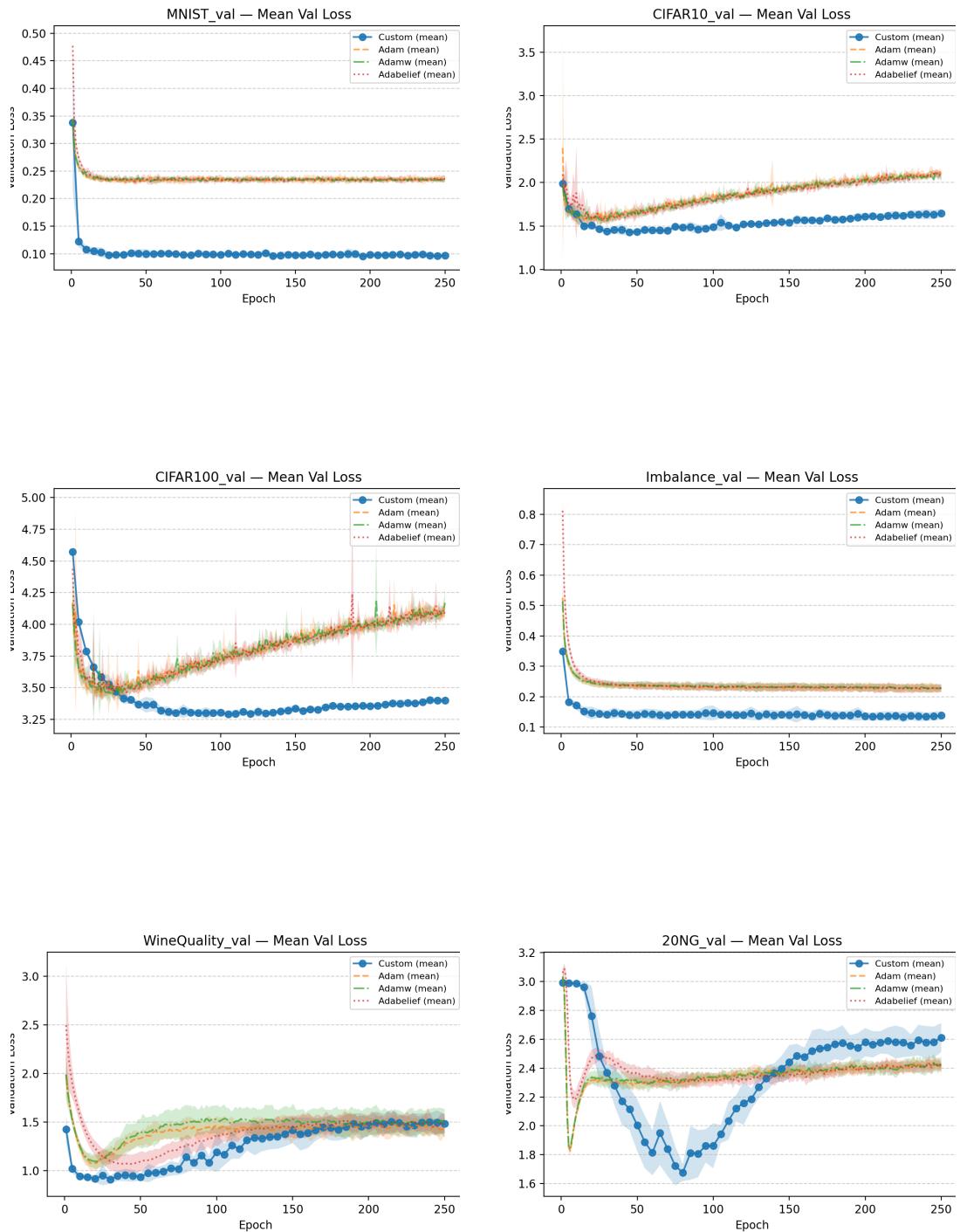
Figure 1: Hidden7 No Reg 실험 결과 요약

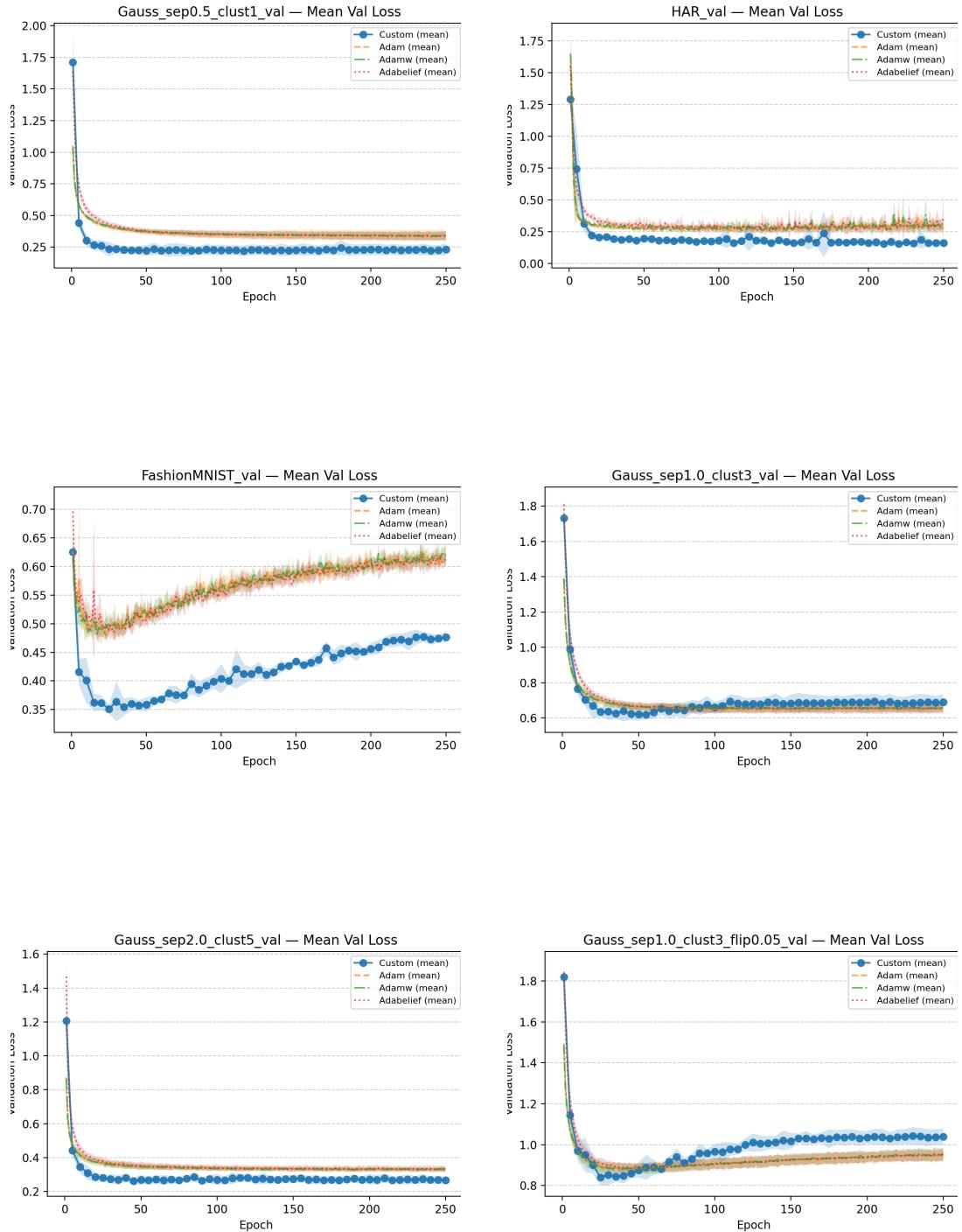




E.2 Hidden7 Regulation

TRAIN Results													
dataset	custom (\pm sd)	adam (\pm sd)	adamw (\pm sd)	adabelief (\pm sd)	d_c_vs_a	d_c_vs_aw	d_c_vs_ab	p_ttest_c_vs_a	p_ttest_c_vs_aw	p_ttest_c_vs_ab	p_ttest_a_vs_a	p_ttest_a_vs_ab	p_ttest_aw_vs_ab
MNIST	0.0247 \pm 0.0013	0.1612 \pm 0.0005	0.1609 \pm 0.0004	0.1608 \pm 0.0003	-13.4832	-142.3431	-145.0940	0.0000	0.0000	0.0000	0.1267	0.2365	0.9261
CIFAR10	0.7429 \pm 0.0350	0.6283 \pm 0.0058	0.6327 \pm 0.0043	0.6257 \pm 0.0051	4.4669	4.3190	4.5811	0.0001	0.0003	0.0001	0.2332	0.3800	0.0054
CIFAR100	2.5413 \pm 0.0339	2.1177 \pm 0.0093	2.1199 \pm 0.0090	2.1065 \pm 0.0045	17.0171	16.9728	17.9634	0.0000	0.0000	0.0000	0.6930	0.0521	0.0117
20NG	0.1198 \pm 0.0044	0.2794 \pm 0.0022	0.2786 \pm 0.0018	0.2798 \pm 0.0019	-45.7579	-47.0248	-46.7261	0.0000	0.0000	0.0000	0.2816	0.7298	0.3081
Imbalance	0.0237 \pm 0.0026	0.1272 \pm 0.0024	0.1274 \pm 0.0015	0.1285 \pm 0.0017	-41.5153	-48.7101	-47.2325	0.0000	0.0000	0.0000	0.7157	0.0892	0.1075
WineQuality	0.0454 \pm 0.0167	0.1936 \pm 0.0096	0.1882 \pm 0.0068	0.2046 \pm 0.0079	-10.8753	-11.1981	-12.1948	0.0000	0.0000	0.0000	0.1046	0.0072	0.0018
FashionMNIST	0.0529 \pm 0.0055	0.1969 \pm 0.0016	0.1980 \pm 0.0017	0.1971 \pm 0.0016	-35.2898	-35.3752	-35.3329	0.0000	0.0000	0.0000	0.3131	0.8176	0.3232
HAR	0.0865 \pm 0.0178	0.1746 \pm 0.0043	0.1730 \pm 0.0026	0.1753 \pm 0.0042	-6.8059	-6.8033	-6.8699	0.0000	0.0000	0.0000	0.4865	0.6997	0.2808
Gauss_sep0.5_-clust1	0.0353 \pm 0.0078	0.2123 \pm 0.0150	0.2118 \pm 0.0130	0.2112 \pm 0.0139	-14.4376	-16.4366	-15.6166	0.0000	0.0000	0.0000	0.8796	0.7452	0.7496
Gauss_sep1.0_-clust3	0.0800 \pm 0.0149	0.3568 \pm 0.0140	0.3571 \pm 0.0113	0.3572 \pm 0.0106	-19.1802	-20.9936	-21.4800	0.0000	0.0000	0.0000	0.9113	0.8822	0.9365
Gauss_sep2.0_-clust3	0.0318 \pm 0.0040	0.2043 \pm 0.0063	0.2041 \pm 0.0061	0.2050 \pm 0.0071	-32.7129	-33.2350	-30.1852	0.0000	0.0000	0.0000	0.7092	0.7055	0.4967
Gauss_sep1.0_-clust3_flip0.05	0.1433 \pm 0.0061	0.4851 \pm 0.0109	0.4850 \pm 0.0165	0.4868 \pm 0.0143	-38.8360	-27.5374	-31.3080	0.0000	0.0000	0.0000	0.9689	0.4600	0.3235
VAL Results													
dataset	custom (\pm sd)	adam (\pm sd)	adamw (\pm sd)	adabelief (\pm sd)	d_c_vs_a	d_c_vs_aw	d_c_vs_ab	p_ttest_c_vs_a	p_ttest_c_vs_aw	p_ttest_c_vs_ab	p_ttest_a_vs_a	p_ttest_a_vs_ab	p_ttest_aw_vs_ab
MNIST	0.0971 \pm 0.0054	0.2358 \pm 0.0050	0.2348 \pm 0.0016	0.2359 \pm 0.0057	-26.8527	-34.7578	-25.0900	0.0000	0.0000	0.0000	0.6600	0.9625	0.5541
CIFAR10	1.6472 \pm 0.0500	2.1045 \pm 0.0392	2.0711 \pm 0.0620	2.0737 \pm 0.0442	-10.1742	-7.5224	-9.0391	0.0000	0.0000	0.0000	0.3043	0.0672	0.9298
CIFAR100	3.3990 \pm 0.0244	4.1294 \pm 0.0386	4.1691 \pm 0.1282	4.0858 \pm 0.0203	-22.6393	-8.3457	-30.6258	0.0000	0.0000	0.0000	0.5099	0.0059	0.1731
20NG	2.6117 \pm 0.1015	2.4017 \pm 0.0340	2.4243 \pm 0.0555	2.4122 \pm 0.0331	2.7748	2.2916	2.6434	0.0022	0.0007	0.0006	0.3401	0.6446	0.4984
Imbalance	0.1378 \pm 0.0141	0.2283 \pm 0.0126	0.2258 \pm 0.0108	0.2268 \pm 0.0119	-6.7576	-7.0102	-6.8124	0.0000	0.0000	0.0000	0.3081	0.6342	0.7454
WineQuality	1.4845 \pm 0.0949	1.4565 \pm 0.0139	1.4934 \pm 0.1438	1.4861 \pm 0.0967	0.2821	-0.0729	-0.0161	0.2752	0.8270	0.9523	0.3108	0.4156	0.8398
FashionMNIST	0.4764 \pm 0.0098	0.6230 \pm 0.0163	0.6192 \pm 0.0207	0.6153 \pm 0.0118	-10.8971	-8.8080	-12.8013	0.0000	0.0000	0.0000	0.7679	0.1703	0.7195
HAR	0.1630 \pm 0.0281	0.3083 \pm 0.0459	0.2765 \pm 0.0187	0.3631 \pm 0.1665	-3.8163	-4.7600	-1.6762	0.0012	0.0001	0.0180	0.1420	0.4071	0.1792
Gauss_sep0.5_-clust1	0.2300 \pm 0.0471	0.3437 \pm 0.0392	0.3373 \pm 0.0315	0.3378 \pm 0.0375	-2.6257	-2.6811	-2.5333	0.0000	0.0000	0.0000	0.1405	0.3320	0.9221
Gauss_sep0.5_-clust3	0.6899 \pm 0.0451	0.6601 \pm 0.0205	0.6595 \pm 0.0245	0.6531 \pm 0.0292	0.8491	0.8374	0.9691	0.0347	0.0549	0.0142	0.8849	0.3124	0.2992
Gauss_sep2.0_-clust3	0.2669 \pm 0.0209	0.3318 \pm 0.0120	0.3329 \pm 0.0125	0.3335 \pm 0.0159	-3.8098	-3.8302	-3.4846	0.0000	0.0000	0.0000	0.5368	0.9346	0.6696
Gauss_sep1.0_-clust3_flip0.05	1.0385 \pm 0.0389	0.9441 \pm 0.0334	0.9517 \pm 0.0294	0.9501 \pm 0.0287	2.6036	2.5167	2.5823	0.0000	0.0001	0.0001	0.0632	0.3756	0.6657
ACC Results													
dataset	custom (\pm sd)	adam (\pm sd)	adamw (\pm sd)	adabelief (\pm sd)	d_c_vs_a	d_c_vs_aw	d_c_vs_ab	p_ttest_c_vs_a	p_ttest_c_vs_aw	p_ttest_c_vs_ab	p_ttest_a_vs_a	p_ttest_a_vs_ab	p_ttest_aw_vs_ab
MNIST	0.9800 \pm 0.0008	0.9802 \pm 0.0012	0.9799 \pm 0.0006	0.9798 \pm 0.0011	-0.2809	0.0396	0.1160	0.6667	0.9180	0.8254	0.5330	0.5506	0.8419
CIFAR10	0.5027 \pm 0.0071	0.4641 \pm 0.0065	0.4656 \pm 0.0123	0.4698 \pm 0.0081	5.6702	3.6800	4.3022	0.0000	0.0003	0.0000	0.7964	0.1962	0.5142
CIFAR100	0.2105 \pm 0.0062	0.1898 \pm 0.0049	0.1868 \pm 0.0137	0.1961 \pm 0.0041	3.7047	2.2222	2.7240	0.0004	0.0055	0.0008	0.6435	0.0203	0.2033
20NG	0.4988 \pm 0.0173	0.5453 \pm 0.0043	0.5427 \pm 0.0096	0.5410 \pm 0.0042	-3.7390	-3.1387	-3.3618	0.0004	0.0004	0.0003	0.4168	0.1056	0.6736
Imbalance	0.9659 \pm 0.0039	0.9676 \pm 0.0036	0.9683 \pm 0.0030	0.9678 \pm 0.0033	-0.4466	-0.6804	-0.5254	0.0633	0.0362	0.1646	0.4015	0.7944	0.6586
WineQuality	0.6518 \pm 0.0202	0.6638 \pm 0.0207	0.6594 \pm 0.0250	0.6545 \pm 0.0313	-0.5897	-0.3342	-0.1018	0.0814	0.4705	0.8137	0.4371	0.1887	0.4941
FashionMNIST	0.8862 \pm 0.0014	0.8795 \pm 0.0044	0.8803 \pm 0.0040	0.8806 \pm 0.0029	2.1623	1.9623	2.4680	0.0048	0.0042	0.0013	0.7690	0.3426	0.8941
HAR	0.9440 \pm 0.0087	0.9458 \pm 0.0160	0.9499 \pm 0.0074	0.9337 \pm 0.0258	-0.1345	-0.7293	0.5372	0.8507	0.3070	0.3802	0.4251	0.2090	0.0802
Gauss_sep0.5_-clust1	0.9354 \pm 0.0114	0.9327 \pm 0.0130	0.9353 \pm 0.0127	0.9347 \pm 0.0122	2.1188	0.0079	0.0543	0.2287	0.9397	0.7185	0.1334	0.3229	0.6960
Gauss_sep1.0_-clust3	0.8039 \pm 0.0088	0.8289 \pm 0.0075	0.8289 \pm 0.0081	0.8295 \pm 0.0097	-3.0487	-2.9510	-2.7561	0.0000	0.0001	0.0000	0.9848	0.7817	0.8195
Gauss_sep2.0_-clust3	0.9243 \pm 0.0073	0.9411 \pm 0.0048	0.9403 \pm 0.0041	0.9406 \pm 0.0062	-2.7464	-2.7074	-2.4205	0.0000	0.0000	0.0003	0.0890	0.6992	0.8422
Gauss_sep1.0_-clust3_flip0.05	0.7459 \pm 0.0068	0.7688 \pm 0.0097	0.7706 \pm 0.0084	0.7690 \pm 0.0080	-2.7621	-3.2432	-3.1391	0.0002	0.0001	0.0003	0.3008	0.9314	0.4158
F1 Results													
dataset	custom (\pm sd)	adam (\pm sd)	adamw (\pm sd)	adabelief (\pm sd)	d_c_vs_a	d_c_vs_aw	d_c_vs_ab	p_ttest_c_vs_a	p_ttest_c_vs_aw	p_ttest_c_vs_ab	p_ttest_a_vs_a	p_ttest_a_vs_ab	p_ttest_aw_vs_ab
MNIST	0.9798 \pm 0.0006	0.9801 \pm 0.0012	0.9798 \pm 0.0007	0.9797 \pm 0.0012	-0.2812	0.0072	0.1210	0.6661	0.9849	0.8175	0.5659	0.5503	0.7949
CIFAR10	0.5027 \pm 0.0072	0.4636 \pm 0.0066	0.4662 \pm 0.0124	0.4688 \pm 0.0080	5.8177	3.6111	4.4457	0.0000	0.0003	0.0000	0.6564	0.1964	0.6633
CIFAR100	0.2012 \pm 0.0062	0.1877 \pm 0.0049	0.1829 \pm 0.0122	0.1913 \pm 0.0049	2.6143	1.9432	1.9139	0.0031	0.0097	0.0026	0.3878	0.2205	0.2144
20NG	0.4952 \pm 0.0171	0.5391 \pm 0.0039	0.5362 \pm 0.0091	0.5346 \pm 0.0046	-3.5498	-2.9965	-3.1486	0.0005	0.0005	0.0003	0.2744	0.0684	0.6014
Imbalance	0.9302 \pm 0.0064	0.9317 \pm 0.0086	0.9335 \pm 0.0065	0.9319 \pm 0.0078	-0.1987	-0.5098	-0.2345	0.4343	0.1147	0.4673	0.3808	0.8701	0.4318
WineQuality	0.4146 \pm 0.0384	0.4265 \pm 0.0336	0.4345 \pm 0.0660	0.4107 \pm 0.0404	-0.3314	-0.3701	0.0989	0.2425	0.3061	0.6163	0.2983	0.3978	
FashionMNIST	0.8862 \pm 0.0013	0.8795 \pm 0.0036	0.8800 \pm 0.0044	0.8804 \pm 0.0026	2.5040	1.9178	2.8281	0.0018	0.0058	0.0005	0.8434	0.3579	0.8664
HAR	0.9446 \pm 0.0087	0.9474 \pm 0.0150	0.9516 \pm 0.0070	0.9345 \pm 0.0278	-0.2340	-0.8933	0.4910	0.7457	0.2172	0.4162	0.4003	0.2085	0.0955
Gauss_sep0.5_-clust1	0.9353 \pm 0.0114	0.9325 \pm 0.0131	0.9351 \pm 0.0127	0.9346 \pm 0.0123	0.2247	0.0151	0.0616	0.2194	0.8847	0.6836	0.1366	0.3282	0.6982





E.3 Hidden10 No Regulation

TRAIN Results

	Custom	Adam	Adamw	Adabelief	p_c_vs_a	p_c_vs_aw	p_c_vs_ab
MNIST	0.0	0.0015	0.0008	0.0039	0.1503	0.1905	0.0119
CIFAR10	0.7543	0.8545	0.8521	0.7894	0.0157	0.0402	0.3134
CIFAR100	2.5776	2.4628	2.4762	2.4193	0.1339	0.1674	0.0245
20NG	0.0904	0.0934	0.0942	0.1018	0.0335	0.0258	0.1874
Imbalance	0.0	0.0042	0.0067	0.0053	0.0573	0.0405	0.0206
WineQuality	0.0025	0.0117	0.002	0.0255	0.2191	0.5287	0.3515
FashionMNIST	0.0244	0.094	0.05	0.0424	0.152	0.0015	0.0014
HAR	0.056	0.0239	0.0281	0.033	0.0008	0.0007	0.0185
Gauss_sep0.5_clust1	0.0001	0.0299	0.0346	0.0262	0.0002	0.0003	0.0004
Gauss_sep1.0_clust3	0.0001	0.0906	0.0706	0.0811	0.0001	0.0	0.0
Gauss_sep2.0_clust5	0.0	0.0279	0.027	0.0259	0.0	0.0	0.0001
Gauss_sep1.0_clust3_flip0.05	0.0002	0.1337	0.1214	0.1429	0.0	0.0	0.0

VAL Results

	Custom	Adam	Adamw	Adabelief	p_c_vs_a	p_c_vs_aw	p_c_vs_ab
MNIST	0.2457	0.3456	0.4691	0.28	0.125	0.027	0.2694
CIFAR10	2.4471	2.0848	2.0848	2.0761	0.014	0.0116	0.0057
CIFAR100	3.9646	4.1963	4.2123	4.0919	0.001	0.0005	0.003
20NG	5.9258	5.0123	4.7149	4.7243	0.0001	0.0001	0.0058
Imbalance	0.5237	0.5612	0.4588	0.5155	0.7219	0.3399	0.9105
WineQuality	3.0536	3.033	3.5869	3.2316	0.9301	0.0028	0.5807
FashionMNIST	1.1178	0.7771	0.7757	0.8454	0.0002	0.0003	0.0004
HAR	0.263	0.2584	0.2181	0.2081	0.8039	0.0503	0.0014
Gauss_sep0.5_clust1	1.1794	0.7212	0.7165	0.7206	0.0	0.0	0.0
Gauss_sep1.0_clust3	2.8179	1.6139	1.6025	1.5534	0.0	0.0	0.0
Gauss_sep2.0_clust5	0.9305	0.545	0.5394	0.5335	0.0	0.0	0.0
Gauss_sep1.0_clust3_flip0.05	3.7196	2.4249	2.4281	2.2781	0.0	0.0	0.0

ACC Results

	Custom	Adam	Adamw	Adabelief	p_c_vs_a	p_c_vs_aw	p_c_vs_ab
MNIST	0.9761	0.9791	0.978	0.9777	0.0024	0.2404	0.0712
CIFAR10	0.4797	0.459	0.4566	0.4701	0.0004	0.0023	0.0443
CIFAR100	0.1793	0.1783	0.1758	0.1827	0.7893	0.1937	0.1322
20NG	0.5365	0.5386	0.5465	0.5377	0.6914	0.045	0.6854
Imbalance	0.9553	0.9564	0.9556	0.9548	0.5305	0.8516	0.8137
WineQuality	0.6491	0.6625	0.6402	0.6554	0.1342	0.3962	0.7043
FashionMNIST	0.8762	0.8755	0.8822	0.8842	0.9146	0.0724	0.0028
HAR	0.9383	0.9535	0.9551	0.9545	0.0025	0.0005	0.0035
Gauss_sep0.5_clust1	0.8995	0.8866	0.8841	0.8867	0.0071	0.0007	0.0059
Gauss_sep1.0_clust3	0.7767	0.7631	0.7685	0.7662	0.0094	0.0312	0.0245
Gauss_sep2.0_clust5	0.9139	0.9119	0.9122	0.9145	0.2571	0.4062	0.7353
Gauss_sep1.0_clust3_flip0.05	0.7183	0.6932	0.6927	0.6936	0.0002	0.0003	0.0001

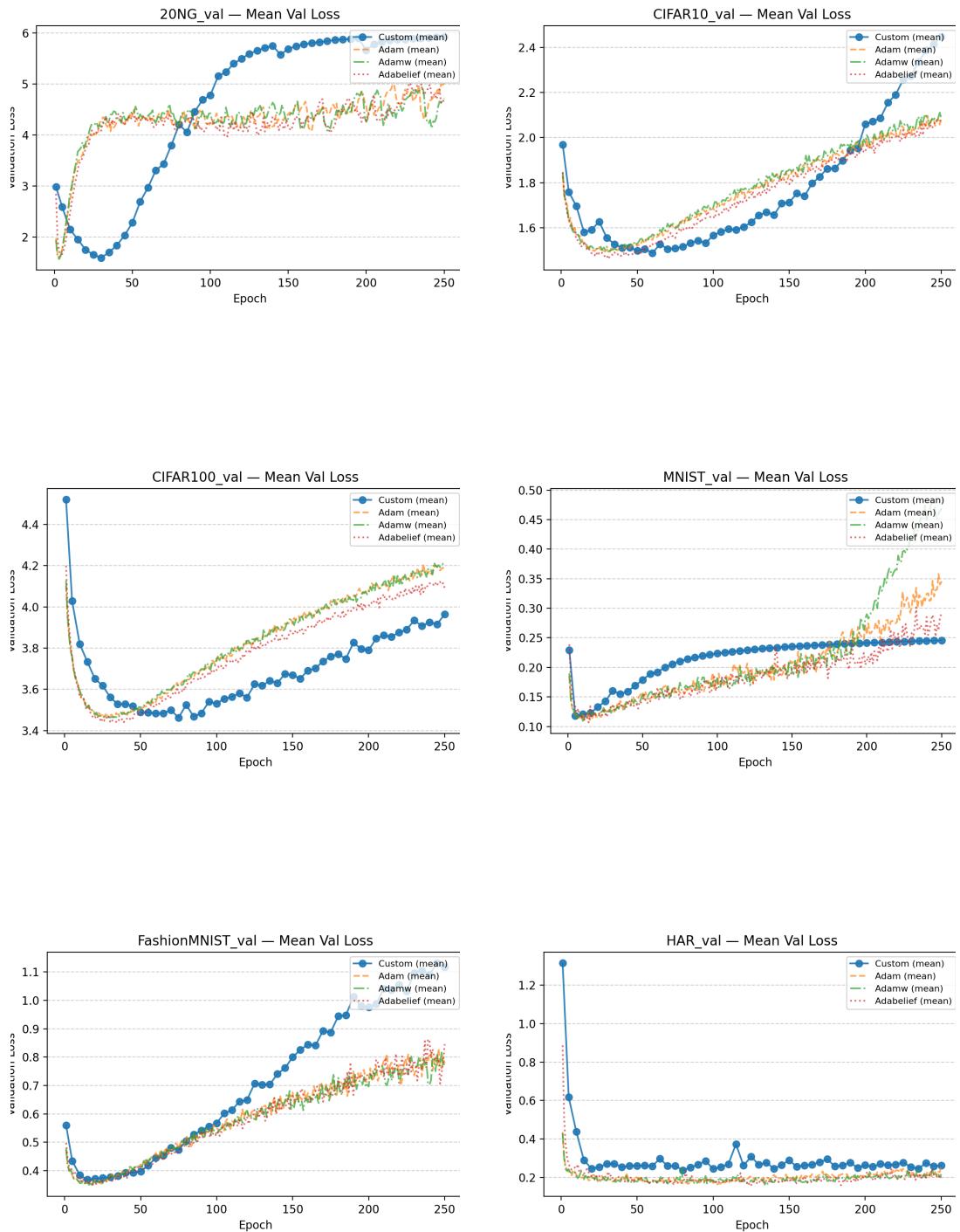
F1 Results

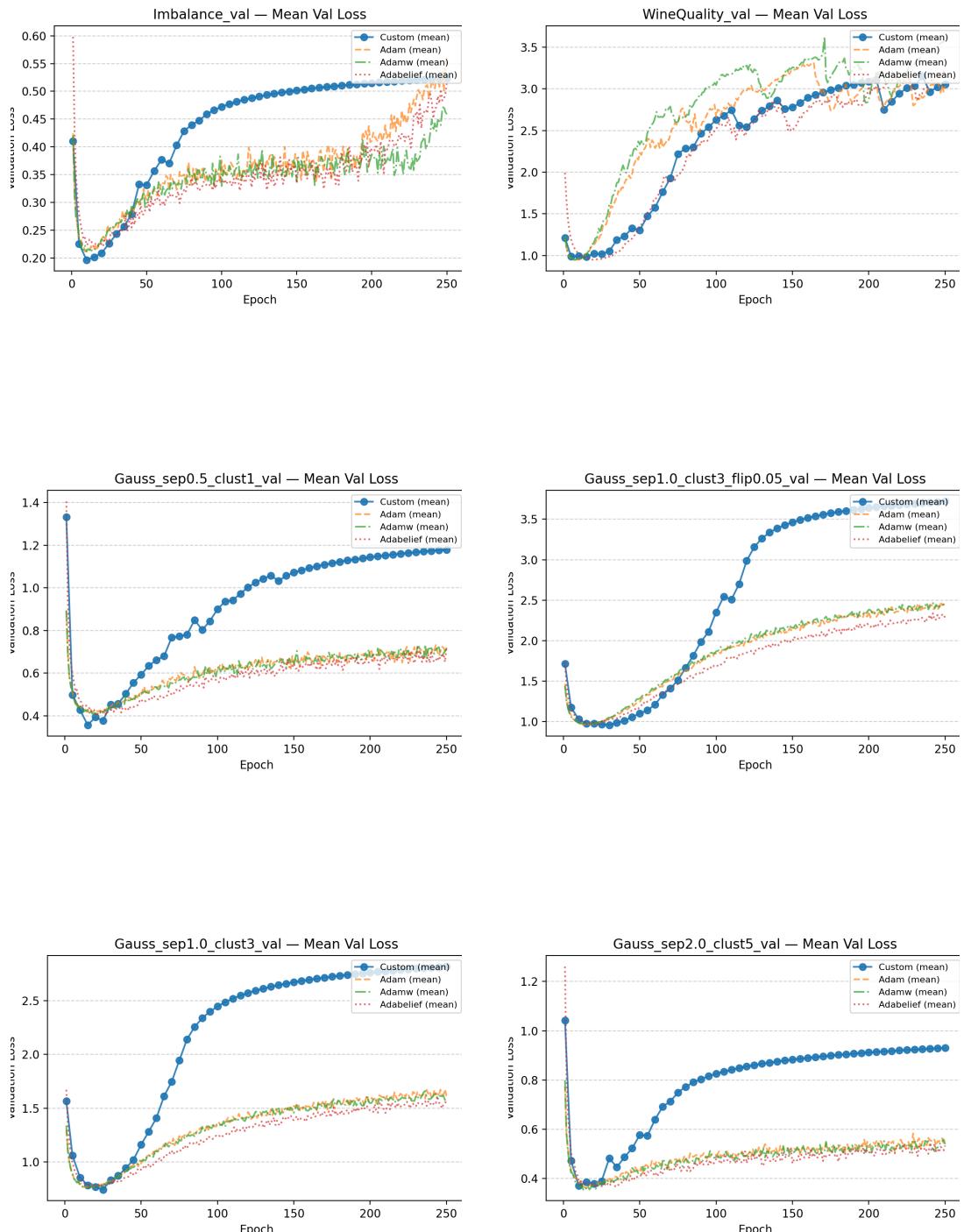
	Custom	Adam	Adamw	Adabelief	p_c_vs_a	p_c_vs_aw	p_c_vs_ab
MNIST	0.9759	0.9789	0.9778	0.9775	0.0028	0.2425	0.0729
CIFAR10	0.479	0.4581	0.4551	0.4687	0.0004	0.0019	0.031
CIFAR100	0.1716	0.1727	0.1699	0.1763	0.7798	0.5497	0.0594
20NG	0.5316	0.5354	0.5427	0.5334	0.4527	0.0283	0.4643
Imbalance	0.9076	0.9121	0.9094	0.9082	0.1834	0.5076	0.8827
WineQuality	0.4209	0.4195	0.4184	0.4236	0.9257	0.8962	0.8686
FashionMNIST	0.8762	0.8754	0.881	0.8844	0.9043	0.087	0.0025
HAR	0.938	0.9538	0.9558	0.9552	0.0026	0.0004	0.0029
Gauss_sep0.5_clust1	0.8995	0.8865	0.8841	0.8867	0.0073	0.0007	0.0061
Gauss_sep1.0_clust3	0.7767	0.7629	0.7684	0.7662	0.0089	0.0269	0.0236
Gauss_sep2.0_clust5	0.9139	0.9119	0.9122	0.9145	0.2643	0.4142	0.7268
Gauss_sep1.0_clust3_flip0.05	0.7181	0.693	0.6924	0.6935	0.0003	0.0002	0.0001

TIME Results

	Custom	Adam	Adamw	Adabelief
MNIST	178.7493	275.1787	284.1111	290.527
CIFAR10	190.9793	277.2164	283.7629	307.2627
CIFAR100	207.4634	282.5955	291.8332	312.1756
20NG	59.46	92.9509	97.325	100.7718
Imbalance	65.5428	118.4874	122.7422	124.601
WineQuality	8.2637	13.4492	16.511	17.7939
FashionMNIST	179.1427	281.9967	292.2724	297.1215
HAR	29.25	51.7072	53.9491	57.4662
Gauss_sep0.5_clust1	64.7157	117.0886	120.0175	122.2473
Gauss_sep1.0_clust3	64.5793	120.7696	122.3737	123.0598
Gauss_sep2.0_clust5	64.151	119.9376	122.3609	123.4184
Gauss_sep1.0_clust3_flip0.05	63.5085	118.705	120.8227	122.9597

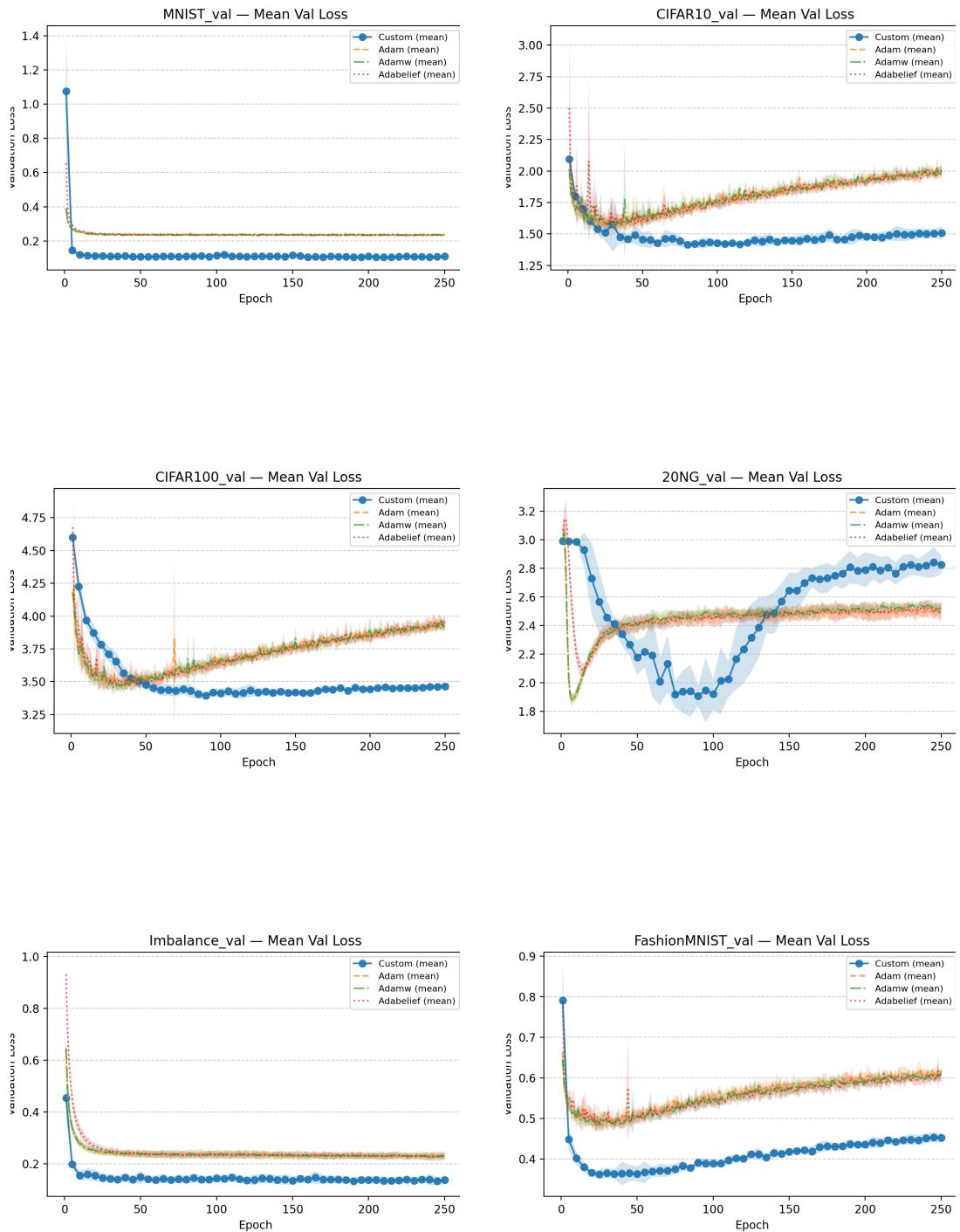
Figure 3: Hidden10 No Reg 실험 결과 요약

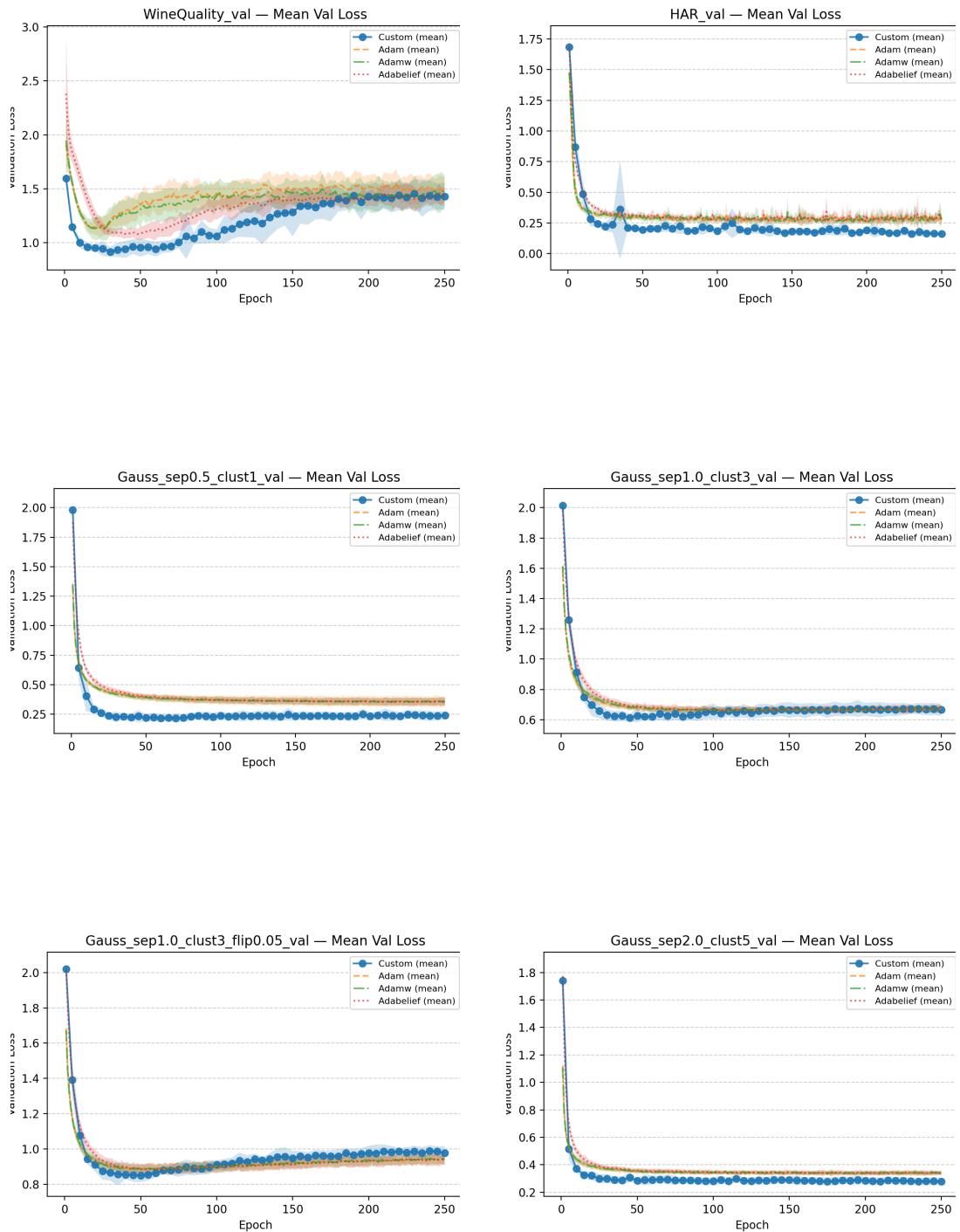




E.4 Hidden10 Regulation

TRAIN Results													
dataset	custom (\pm sd)	adam (\pm sd)	adamw (\pm sd)	adabelief (\pm sd)	d_c_vs_a	d_c_vs_aw	d_c_vs_ab	p_ttest_c_vs_a	p_ttest_c_vs_aw	p_ttest_c_vs_ab	p_ttest_a_vs_aw	p_ttest_a_vs_ab	p_ttest_aw_vs_ab
MNIST	0.0289 \pm 0.0079	0.1616 \pm 0.0007	0.1616 \pm 0.0005	0.1620 \pm 0.0006	-23.8963	-23.9367	-23.9840	0.0000	0.0000	0.0000	0.7908	0.1815	0.2220
CIFAR10	0.8498 \pm 0.0313	0.6565 \pm 0.0078	0.6546 \pm 0.0055	0.6540 \pm 0.0046	8.4787	8.6878	8.7577	0.0000	0.0000	0.0000	0.6072	0.3549	0.6103
CIFAR100	2.8258 \pm 0.0541	2.2222 \pm 0.0058	2.2212 \pm 0.0080	2.2092 \pm 0.0057	15.6940	15.6410	16.0349	0.0000	0.0000	0.0000	0.7897	0.0035	0.0103
20NG	0.1651 \pm 0.0360	0.2858 \pm 0.0015	0.2857 \pm 0.0040	0.2874 \pm 0.0018	-4.7295	-4.7023	-4.7923	0.0001	0.0001	0.0001	0.9779	0.0412	0.3160
Imbalance	0.0264 \pm 0.0071	0.1288 \pm 0.0025	0.1288 \pm 0.0021	0.1295 \pm 0.0026	-19.1598	-19.4729	-19.1675	0.0000	0.0000	0.0000	0.9528	0.2814	0.4535
WineQuality	0.0484 \pm 0.0125	0.2045 \pm 0.0105	0.2019 \pm 0.0103	0.2216 \pm 0.0115	-13.5559	-13.4090	-14.4118	0.0000	0.0000	0.0000	0.6164	0.0207	0.0028
FashionMNIST	0.0737 \pm 0.0090	0.2006 \pm 0.0022	0.2019 \pm 0.0009	0.2008 \pm 0.0019	-19.3428	-19.9826	-19.4849	0.0000	0.0000	0.0000	0.1846	0.8894	0.2842
HAR	0.0825 \pm 0.0128	0.1832 \pm 0.0041	0.1852 \pm 0.0017	0.1856 \pm 0.0040	-10.5530	-11.2100	-10.8457	0.0000	0.0000	0.0000	0.3311	0.0965	0.8099
Gauss_sep0_5_-clust1	0.0419 \pm 0.0106	0.2178 \pm 0.0158	0.2157 \pm 0.0118	0.2151 \pm 0.0115	-13.0821	-15.5027	-15.6837	0.0000	0.0000	0.0000	0.4627	0.4749	0.7862
Gauss_sep1_0_-clust3	0.0901 \pm 0.0063	0.3467 \pm 0.0119	0.3446 \pm 0.0086	0.3482 \pm 0.0112	-27.0071	-33.8030	-28.3326	0.0000	0.0000	0.0000	0.3212	0.5273	0.1318
Gauss_sep2_0_-clust3	0.0323 \pm 0.0033	0.2037 \pm 0.0054	0.2036 \pm 0.0050	0.2047 \pm 0.0050	-38.3166	-40.1472	-40.7110	0.0000	0.0000	0.0000	0.8866	0.3208	0.3994
Gauss_sep1_0_-clust3_flip0.05	0.1423 \pm 0.0103	0.4724 \pm 0.0102	0.4716 \pm 0.0144	0.4706 \pm 0.0130	-32.2868	-26.3526	-28.0454	0.0000	0.0000	0.0000	0.7771	0.2783	0.6642
VAL Results													
dataset	custom (\pm sd)	adam (\pm sd)	adamw (\pm sd)	adabelief (\pm sd)	d_c_vs_a	d_c_vs_aw	d_c_vs_ab	p_ttest_c_vs_a	p_ttest_c_vs_aw	p_ttest_c_vs_ab	p_ttest_a_vs_aw	p_ttest_a_vs_ab	p_ttest_aw_vs_ab
MNIST	0.1092 \pm 0.0057	0.2332 \pm 0.0034	0.2349 \pm 0.0018	0.2352 \pm 0.0049	-26.5773	-29.8988	-23.7693	0.0000	0.0000	0.0000	0.4075	0.2400	0.8819
CIFAR10	1.5079 \pm 0.0502	2.0000 \pm 0.0434	2.0023 \pm 0.0601	2.0078 \pm 0.0515	-10.4885	-8.9285	-9.8345	0.0000	0.0000	0.0000	0.9340	0.8062	0.7941
CIFAR100	3.4660 \pm 0.0216	3.9369 \pm 0.0250	3.9386 \pm 0.0408	3.9587 \pm 0.0494	-20.1333	-14.4769	-12.9123	0.0000	0.0000	0.0000	0.9107	0.3064	0.2538
20NG	2.8253 \pm 0.0668	2.5061 \pm 0.0439	2.5230 \pm 0.0484	2.5019 \pm 0.0630	5.6500	5.1840	4.9818	0.0000	0.0000	0.0000	0.3918	0.8342	0.4048
Imbalance	0.1375 \pm 0.0165	0.2317 \pm 0.0108	0.2321 \pm 0.0177	0.2327 \pm 0.0118	-6.7558	-5.5244	-6.6274	0.0000	0.0000	0.0000	0.9128	0.6461	0.8819
WineQuality	1.4310 \pm 0.1263	1.4731 \pm 0.1315	1.4813 \pm 0.1438	1.4393 \pm 0.0564	-0.3262	-0.3714	-0.0842	0.2406	0.0911	0.8319	0.8336	0.4136	0.3496
FashionMNIST	0.4528 \pm 0.0107	0.6178 \pm 0.0106	0.6108 \pm 0.0170	0.6038 \pm 0.0166	-15.4949	-11.1230	-10.7884	0.0000	0.0000	0.0000	0.4347	0.1516	0.3382
HAR	0.1608 \pm 0.0160	0.2735 \pm 0.0305	0.2785 \pm 0.0409	0.2689 \pm 0.0313	-4.5813	-3.7632	-4.3114	0.0001	0.0005	0.0001	0.8091	0.5319	0.6228
Gauss_sep0_5_-clust1	0.2397 \pm 0.0359	0.3558 \pm 0.0438	0.3569 \pm 0.0310	0.3586 \pm 0.0323	-2.9012	-3.4953	-3.4752	0.0000	0.0000	0.0000	0.8791	0.8034	0.7681
Gauss_sep1_0_-clust3	0.6662 \pm 0.0405	0.6753 \pm 0.0214	0.6672 \pm 0.0204	0.6703 \pm 0.0255	-0.2805	-0.0332	-0.1229	0.4751	0.9107	0.7501	0.1789	0.4973	0.6240
Gauss_sep2_0_-clust3	0.2789 \pm 0.0254	0.3373 \pm 0.0191	0.3418 \pm 0.0156	0.3382 \pm 0.0145	-2.5986	-2.9865	-2.8699	0.0000	0.0002	0.0000	0.2840	0.7476	0.3236
Gauss_sep1_0_-clust3_flip0.05	0.9767 \pm 0.0388	0.9397 \pm 0.0259	0.9416 \pm 0.0286	0.9433 \pm 0.0323	1.1215	1.0298	0.9363	0.0053	0.0086	0.0051	0.7797	0.5218	0.5710
ACC Results													
dataset	custom (\pm sd)	adam (\pm sd)	adamw (\pm sd)	adabelief (\pm sd)	d_c_vs_a	d_c_vs_aw	d_c_vs_ab	p_ttest_c_vs_a	p_ttest_c_vs_aw	p_ttest_c_vs_ab	p_ttest_a_vs_aw	p_ttest_a_vs_ab	p_ttest_aw_vs_ab
MNIST	0.9791 \pm 0.0011	0.9807 \pm 0.0007	0.9801 \pm 0.0006	0.9803 \pm 0.0014	-1.7340	-1.1087	-0.9772	0.0017	0.1114	0.1654	0.1935	0.4562	0.7164
CIFAR10	0.5109 \pm 0.0079	0.4677 \pm 0.0044	0.4686 \pm 0.0147	0.4680 \pm 0.0088	6.5965	5.3812	5.1604	0.0000	0.0000	0.0000	0.8739	0.9573	0.9050
CIFAR100	0.1811 \pm 0.0066	0.1996 \pm 0.0053	0.2025 \pm 0.0061	0.1989 \pm 0.0038	-3.0743	-3.3724	-3.3092	0.0003	0.0002	0.0007	0.2890	0.7867	0.1285
20NG	0.4619 \pm 0.0126	0.5407 \pm 0.0120	0.5401 \pm 0.0208	0.5359 \pm 0.0085	-6.8915	-7.4772	-6.8756	0.0000	0.0000	0.0000	0.8850	0.0582	0.2836
Imbalance	0.9673 \pm 0.0040	0.9665 \pm 0.0031	0.9668 \pm 0.0049	0.9664 \pm 0.0025	0.2119	0.1009	0.2632	0.5061	0.8088	0.4764	0.7630	0.8452	0.7036
WineQuality	0.6589 \pm 0.0279	0.6670 \pm 0.0291	0.6509 \pm 0.0224	0.6634 \pm 0.0216	-0.2822	0.3181	-0.1790	0.3250	0.1755	0.6273	0.0569	0.3932	0.1194
FashionMNIST	0.8845 \pm 0.0024	0.8797 \pm 0.0037	0.8797 \pm 0.0039	0.8815 \pm 0.0043	1.5556	1.5039	0.8754	0.0447	0.0209	0.1213	0.9945	0.4574	0.5382
HAR	0.9460 \pm 0.0078	0.8797 \pm 0.0070	0.8797 \pm 0.0039	0.8815 \pm 0.0090	-0.5099	0.1437	-0.3476	0.3108	0.8134	0.4906	0.4074	0.5687	0.5153
Gauss_sep0_5_-clust1	0.9321 \pm 0.0090	0.9302 \pm 0.0149	0.9305 \pm 0.0103	0.9289 \pm 0.0101	0.1603	0.1719	0.3407	0.5226	0.3828	0.2438	0.9155	0.7655	0.4794
Gauss_sep1_0_-clust3	0.8049 \pm 0.0103	0.8249 \pm 0.0074	0.8282 \pm 0.0062	0.8265 \pm 0.0070	-2.2243	-2.7384	-2.4512	0.0013	0.0000	0.0001	0.2611	0.4892	0.1914
Gauss_sep2_0_-clust3	0.9216 \pm 0.0076	0.9413 \pm 0.0059	0.9385 \pm 0.0063	0.9404 \pm 0.0047	-2.8923	-2.4144	-2.9706	0.0000	0.0001	0.0000	0.1321	0.5534	0.2756
Gauss_sep1_0_-clust3_flip0.05	0.7500 \pm 0.0086	0.7687 \pm 0.0090	0.7683 \pm 0.0108	0.7704 \pm 0.0097	-2.1058	-1.8826	-2.2418	0.0000	0.0002	0.0000	0.8811	0.1718	0.1419
TIME Results													
dataset	custom (\pm sd)			adam (\pm sd)			adabief (\pm sd)						
MNIST	185.4826 \pm 6.4292			376.7281 \pm 11.4109			393.0337 \pm 16.0912			398.2448 \pm 12.3954			
CIFAR10	191.2558 \pm 5.6225			434.8170 \pm 92.2056			503.9228 \pm 54.0965			502.8245 \pm 70.2491			
CIFAR100	204.6625 \pm 6.6011			503.2338 \pm 69.7559			531.9348 \pm 17.1434			539.6793 \pm 18.0731			
20NG	60.3703 \pm 1.1462			165.5293 \pm 29.4061			184.4452 \pm 7.7035			185.8137 \pm 6.6768			
Imbalance	70.8542 \pm 0.7904			265.2795 \pm 1.7394			268.9805 \pm 2.4659			273.2321 \pm 2.7111			
WineQuality	10.0686 \pm 0.6547			28.0815 \pm 0.6277			29.0738 \pm 0.6801			31.7435 \pm 0.4252			
FashionMNIST	192.4882 \pm 4.7716			411.8450 \pm 7.0397			445.4972 \pm 7.8580			458.7871 \pm 56.3374			
HAR	31.9890 \pm 0.6137			101.8450 \pm 2.5930			104.5518 \pm 2.2058			108.2547 \pm 2.1761			
Gauss_sep0.5_-clust1													





E.5 하이퍼파라미터 민감도 분석 히트맵

