Identifying and attacking the saddle point problem in high-dimensional non-convex optimization

Dauphin et al.

Zhikai Guo
Seo Taek Kong
Nitin Tangellamudi
Main Problem: Saddle points are the true problem in high-dimensional non-convex optimization as opposed to common intuition that descent algorithms are stuck near local minima.
Review

All critical points have a zero vector as the gradient.

Local **maxima**: Hessian is **negative** definite
Local **minima**: Hessian matrix is **positive** definite

Saddle point: Hessian is **indefinite**. It contains both positive and negative eigenvalues
Motivation

Define:

\[ x^* \] is a critical point on a gaussian random field

\[ e(x^*) = f(x^*) - f(x_{gm}) \]

Where \( x_{gm} \) is the global minimum

\[ \lambda \] is an eigenvalue of the Hessian matrix evaluated at a critical point
ε vs α plane (Bray and Dean 2007)

- **ε** - $E[e(x^*)]$ (the expected error attained at a critical point)
- **α** - \[
\frac{\# \text{ negative eigenvalues}}{\text{dimension of matrix}}
\]
- Monotonically increasing α with respect to ε
- At critical point with ε >> global minimum, exponentially likely to be saddle points
MNIST

(a) Train error $\varepsilon$ (%)

Index of critical point $\alpha$

(b) $p(\lambda)$

- Error 0.32%
- Error 23.49%
- Error 28.23%

Eigenvalue $\lambda$
Wigner’s Semicircular Law

Consider a symmetric real matrix \( A \in \mathbb{R}^{N \times N} \)

Where \( N >> 1 \) and each entry in the upper triangular part of the matrix is iid \( N(0,1) \).

Wigner’s Semicircular Law states that the distribution of the eigenvalues of \( A \) is given by

Becomes exponentially unlikely to randomly pick all eigenvalues to be positive or negative, therefore most critical points are saddle points
Dynamics of optimization problem near saddle point

Gradient descent

Newton Method

Trust region method
Gradient Descent

for gradient descent:

\[ f(\theta^* + \Delta \theta) = f(\theta^*) + \frac{1}{2} \sum_{i=1}^{d} \lambda_i \cdot \Delta v_i^2 \]

\[ \Delta V = \Lambda \Delta \theta \]

where \( \Lambda \) is matrix with columns being eigenvectors, and \( \lambda_i \) is the \( i \)th eigenvalue of \( H \)

\[ \nabla f(\Delta \theta + \theta^*)_i = \frac{1}{2} \lambda_i (\nabla \theta \Delta v_i^2) \]

if \( |\lambda_i| \) is small in magnitude, then gradient is nearly 0
Newton’s Method

for Newton method: \( x^{t+1} = x^t - \gamma_t H^{-1} \nabla f(x) \), where \( \gamma_t \) is the step size

\[
f(x, y) = 5x^2 - y^2, (x, y) = (0, \pm \infty)
\]

is optimal min of this function. Here we argue that a point near saddle point will be attracted. Suppose initial point is \((1, 1)\), then \( \nabla f(x, y) = [10, -2] \) and Hessian is

\[
\begin{bmatrix}
10 & 0 \\
0 & -2
\end{bmatrix}
\]

and H inverse is

\[
\begin{bmatrix}
0.1 & 0 \\
0 & -0.5
\end{bmatrix}
\]

In this case, \( x^{t+1} = x^t - \gamma_t [1, 1] \)
Trust region method

for trust region method: the idea is to make every eigenvalue positive. but the problem is it will incur small eigenvalues and result in slow convergence rate. 

\[ x^{t+1} = x^t - \gamma_t B^{-1} \nabla f, \text{ where } B = H + \alpha I, \text{ and } \alpha > |\min_i \lambda_i| \]
Main Contribution:

Formulate each iteration as a first-order minimization problem:

$$\min_{\Delta \theta} \mathcal{T}_1(\theta, \Delta \theta, f)$$

subject to  \(d(\theta, \theta + \Delta \theta) \leq \alpha\)

Possible distance metric choices:

1) \(d(\theta, \theta + \Delta \theta) = \frac{1}{2} \|\theta + \Delta \theta - \theta\|_2^2 = \frac{1}{2} \|\Delta \theta\|_2^2 : \quad \Delta \theta^* = -\frac{\nabla f(\theta)}{\|\nabla f(\theta)\|_2^2} 2\alpha\)

2) \(d(\theta, \theta + \Delta \theta) = |\mathcal{T}_2(\theta, \Delta \theta, f) - \mathcal{T}_1(\theta, \Delta \theta, f)| = |\Delta \theta^T H \Delta \theta|\)
Lemma 1: \[ |\Delta \theta^T H \Delta \theta| \leq \Delta \theta^T |H| \Delta \theta \]

Proof:

Suppose H is diagonalizable with orthonormal eigenvectors. Using Triangle Inequality:

\[
|\Delta \theta^T H \Delta \theta| = |\Delta \theta \Lambda^T H \Lambda \Delta \theta| = \left| \sum_{i=1}^{n} \lambda_i \Delta v_i^2 \right| \\
\leq \sum_{i=1}^{n} |\lambda_i \Delta v_i^2| = \sum_{i=1}^{n} |\lambda_i| \Delta v_i^2 = \Delta \theta^T |H| \Delta \theta
\]
To leverage curvature information, we can solve the first order Taylor expansion subject to the second distance constraint.

This lemma suggests the following problem:

\[
\begin{align*}
\min_{\Delta \theta} & \quad \mathcal{T}_1(\theta, \Delta \theta, f) \\
\text{S.T.} & \quad |\Delta \theta^T H \Delta \theta| \leq \Delta \theta^T |H| \Delta \theta \leq \alpha
\end{align*}
\]

This can be analytically solved using Lagrange multipliers to obtain the optimal step which is proportional to:

\[
\Delta \theta = -|H|^{-1} \nabla f
\]
Experimental Results

MNIST

Train error $\epsilon$ (%)

$10^1$

$10^{-1}$

# hidden units

5

25

50

Deep Autoencoder

Train error $\epsilon$ (%)

$10^2$

$10^1$

# epochs

0

10

20

30

40

50

Recurrent Neural Network

Most negative $\lambda$

$10^0$

$10^{-1}$

$10^{-2}$

$10^{-3}$

$10^{-4}$

# epochs

0

10

20

30

40

50

(a)

(b)

(c)

(d)