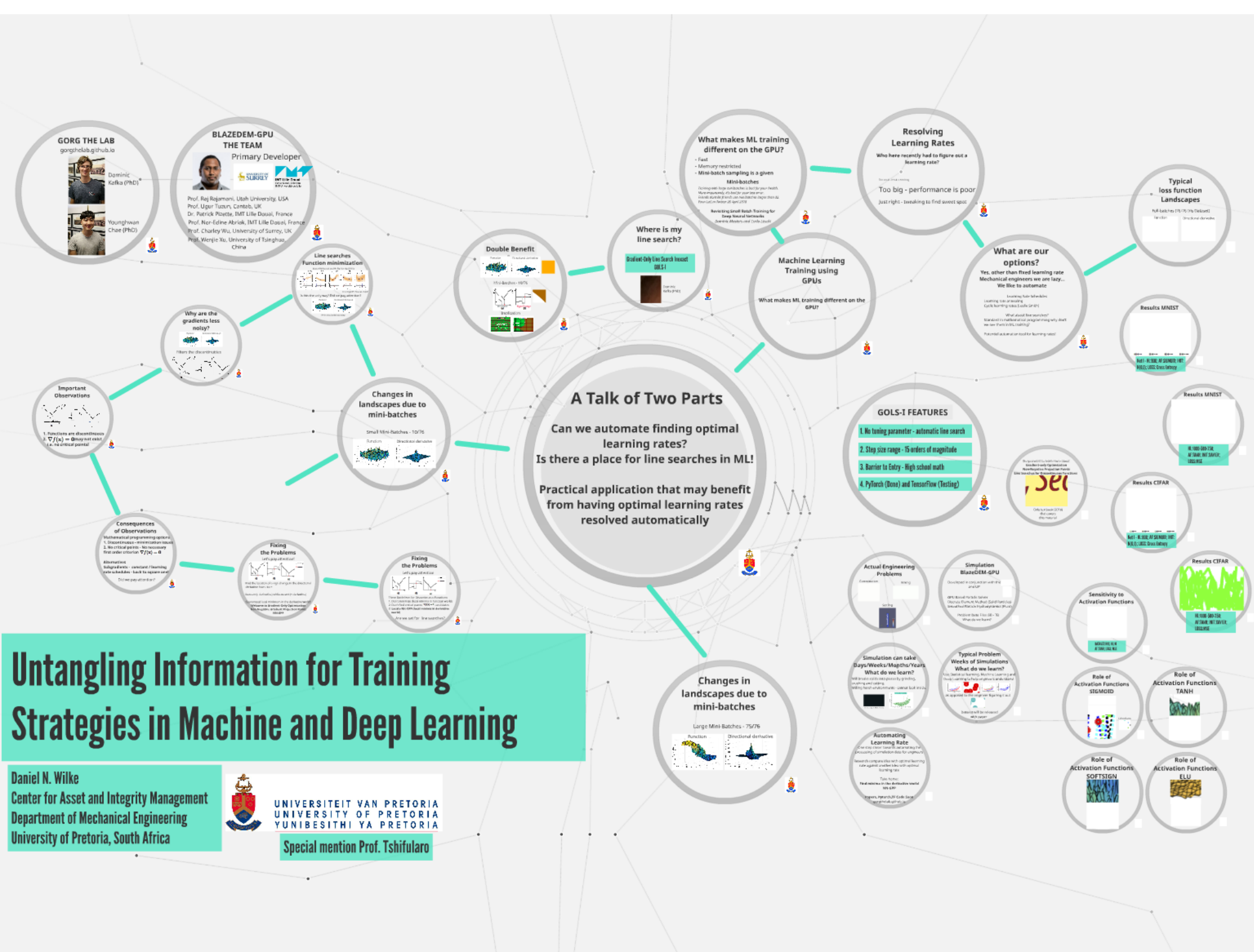
Untangling Information for Training Strategies in Machine and Deep Learning

Daniel N. Wilke
Center for Asset and Integrity Management
Department of Mechanical Engineering
University of Pretoria, South Africa

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Special mention Prof. Tshifularo



Untangling Information for Training Strategies in Machine and Deep Learning

Daniel N. Wilke
Center for Asset and Integrity Management
Department of Mechanical Engineering
University of Pretoria, South Africa

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Special mention Prof. Tshifularo

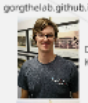
A Talk of Two Parts

Can we automate finding optimal learning rates?
Is there a place for line searches in ML!

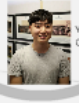
Practical application that may benefit from having optimal learning rates resolved automatically

GORG THE LAB

gorgthelab.github.io



Dominic Kafka (PhD)



Younghwan Chae (PhD)

BLAZEDEM-GPU THE TEAM

Primary Developer

Prof. Raj Rajamani, Utah University, USA
Prof. Ugur Tuzun, Comlab, UK
Dr. Patrick Pizette, IMT Lille Douai, France
Prof. Nor-Edine Abriak, IMT Lille Douai, France
Prof. Charley Wu, University of Surrey, UK
Prof. Wenjie Xu, University of Tsinghua, China

What makes ML training different on the GPU?

- Fast
- Memory restricted
- **Mini-batch sampling is a given Mini-batches**

Training with large mini-batches is bad for your health More importantly, it's bad for your test error. Friends don't let friends use mini-batches larger than 32 Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, Ping Tak Peter Tang

Revisiting Small Batch Training for Deep Neural Networks
Dominic Masters, Carlo Luschi

Resolving Learning Rates

Who here recently had to figure out a learning rate?

Too big - performance is poor

Just right - tweaking to find sweet spot

Typical loss function Landscapes

What are our options?

Yes, other than fixed learning rate
Mechanical engineers we are lazy..
We like to automate

Results MNIST

Results CIFAR

Double Benefit

Where is my line search?

Gradient-Only Line Search Inexact
BWLS-I

Machine Learning Training using GPUs

What makes ML training different on the GPU?

Line searches Function minimization

Why are the gradients less noisy?

Important Observations

Changes in landscapes due to mini-batches

Small Mini-Batches - 10/76

GOLS-I FEATURES

1. No tuning parameter - automatic line search
2. Step size range - 15 orders of magnitude
3. Barrier to Entry - High school math
4. PyTorch (Done) and TensorFlow (Testing)

Consequences of Observations

Fixing the Problems

Fixing the Problems

Changes in landscapes due to mini-batches

Large Mini-Batches - 75/76

Actual Engineering Problems

Simulation BlazeDEM-GPU

Simulation can take Days/Weeks/Months/Years What do we learn?

Typical Problem Weeks of Simulations What do we learn?

Automating Learning Rate

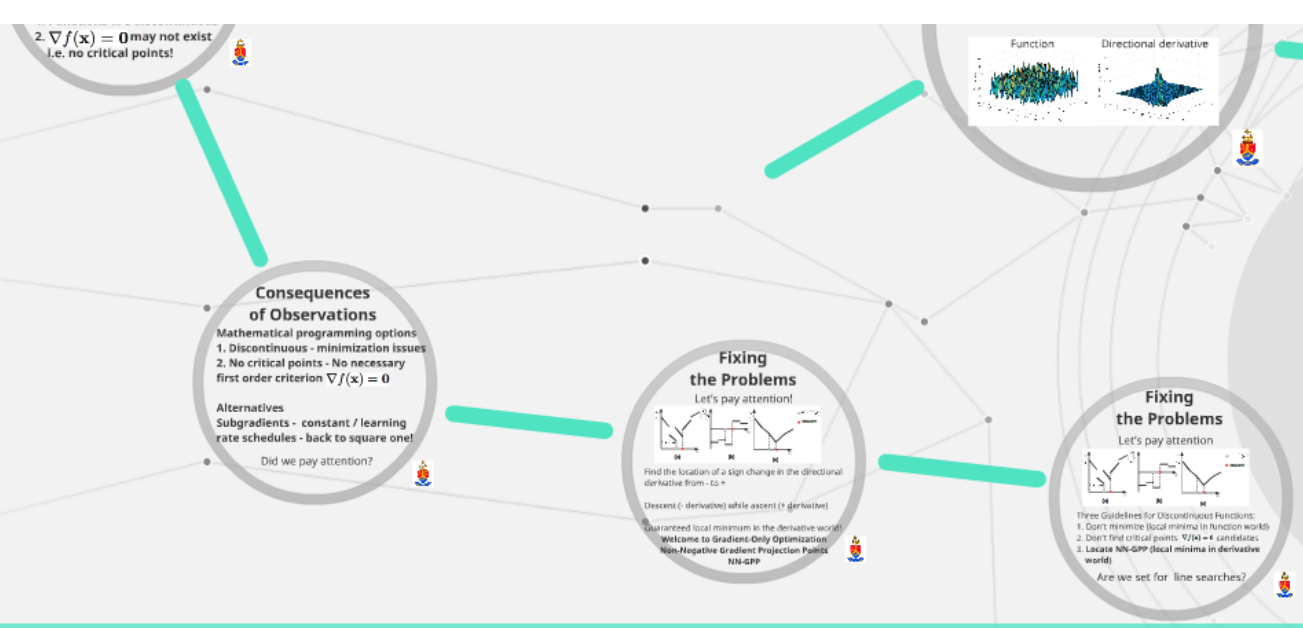
Sensitivity to Activation Functions

Role of Activation Functions SIGMOID

Role of Activation Functions TANH

Role of Activation Functions SOFTSIGN

Role of Activation Functions ELU



Untangling Information for Training Strategies in Machine and Deep Learning

Daniel N. Wilke
Center for Asset and Integrity Management
Department of Mechanical Engineering
University of Pretoria, South Africa

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Special mention Prof. Tshifularo

A Talk of Two Parts

Can we automate finding optimal learning rates?
Is there a place for line searches in ML!

Practical application that may benefit from having optimal learning rates resolved automatically

Machine Learning Training using GPUs

What makes ML training different on the GPU?

What makes ML training different on the GPU?

- Fast
- Memory restricted
- **Mini-batch sampling is a given**

Mini-batches

Training with large minibatches is bad for your health.
More importantly, it's bad for your test error.
Friends dont let friends use minibatches larger than 32.
Yann LeCun Twitter 26 April 2018

Revisiting Small Batch Training for Deep Neural Networks
Dominic Masters and Carlo Luschi

re is my

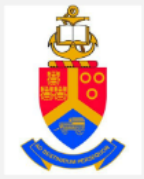
Resolving Learning Rates

Who here recently had to figure out a learning rate?

Too small - it takes too long

Too big - performance is poor

Just right - tweaking to find sweet spot



What are our options?

Yes, other than fixed learning rate Mechanical engineers we are lazy... We like to automate

Learning Rate Schedules

Learning rate annealing

Cyclic learning rates (Leslie Smith)

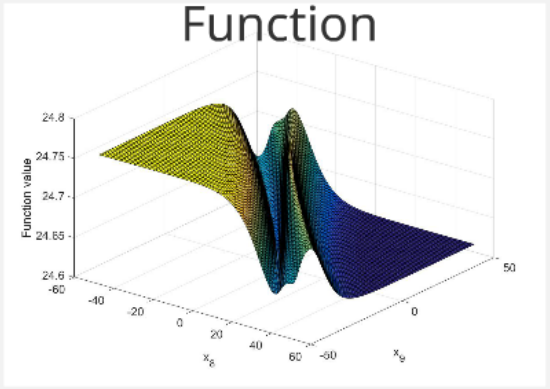
What about line searches?

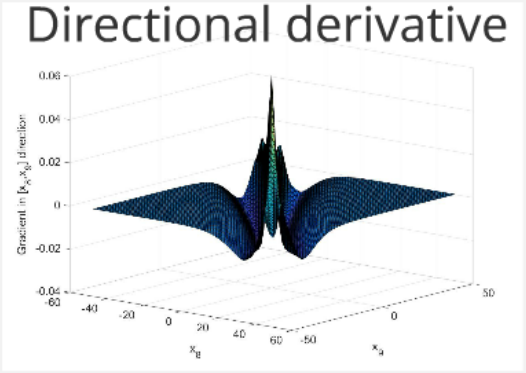
Standard in mathematical programming why don't we see them in ML training?

Potential automation tool for learning rates!

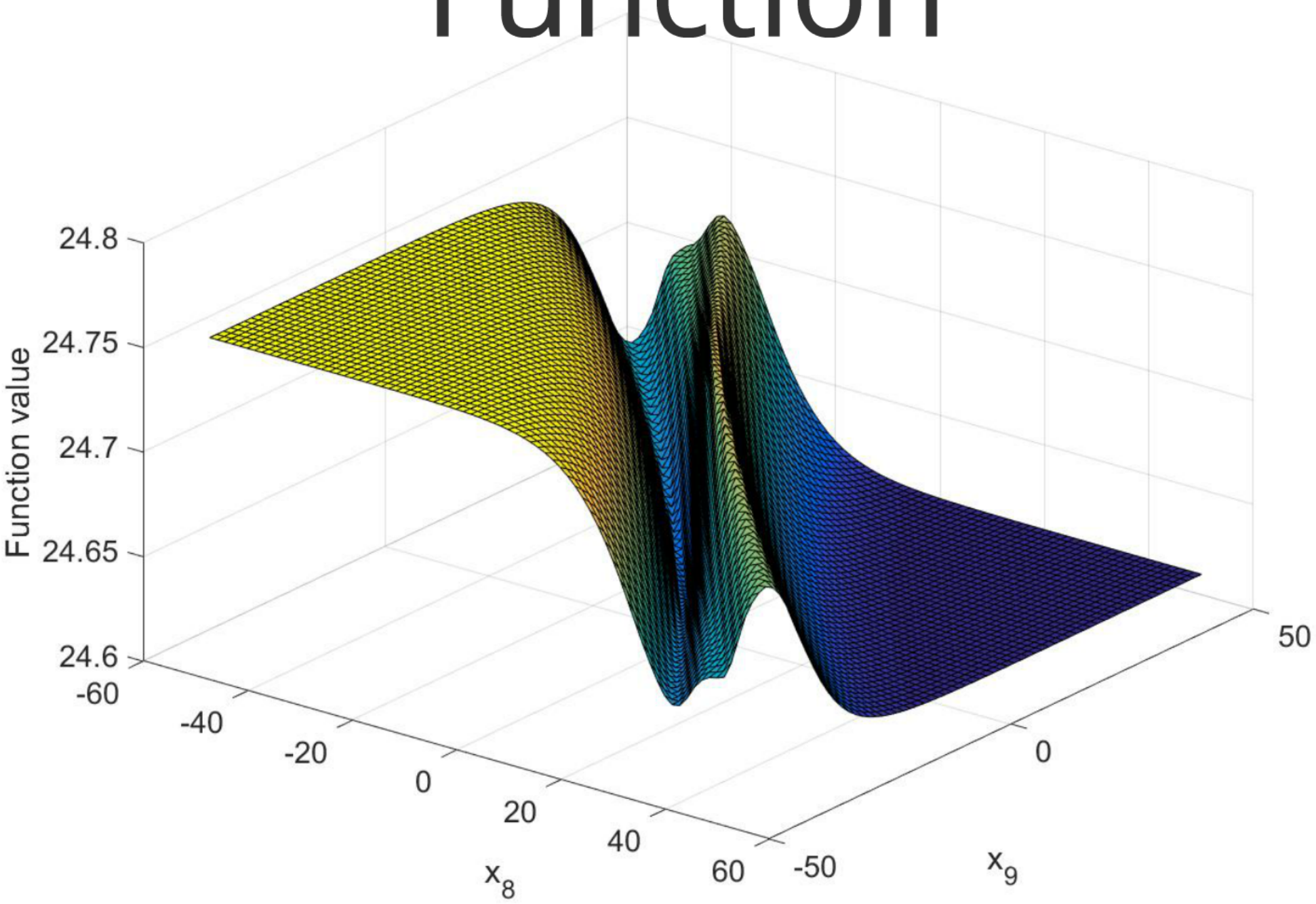
Typical loss function Landscapes

Full-batches (76/76 Iris Dataset)

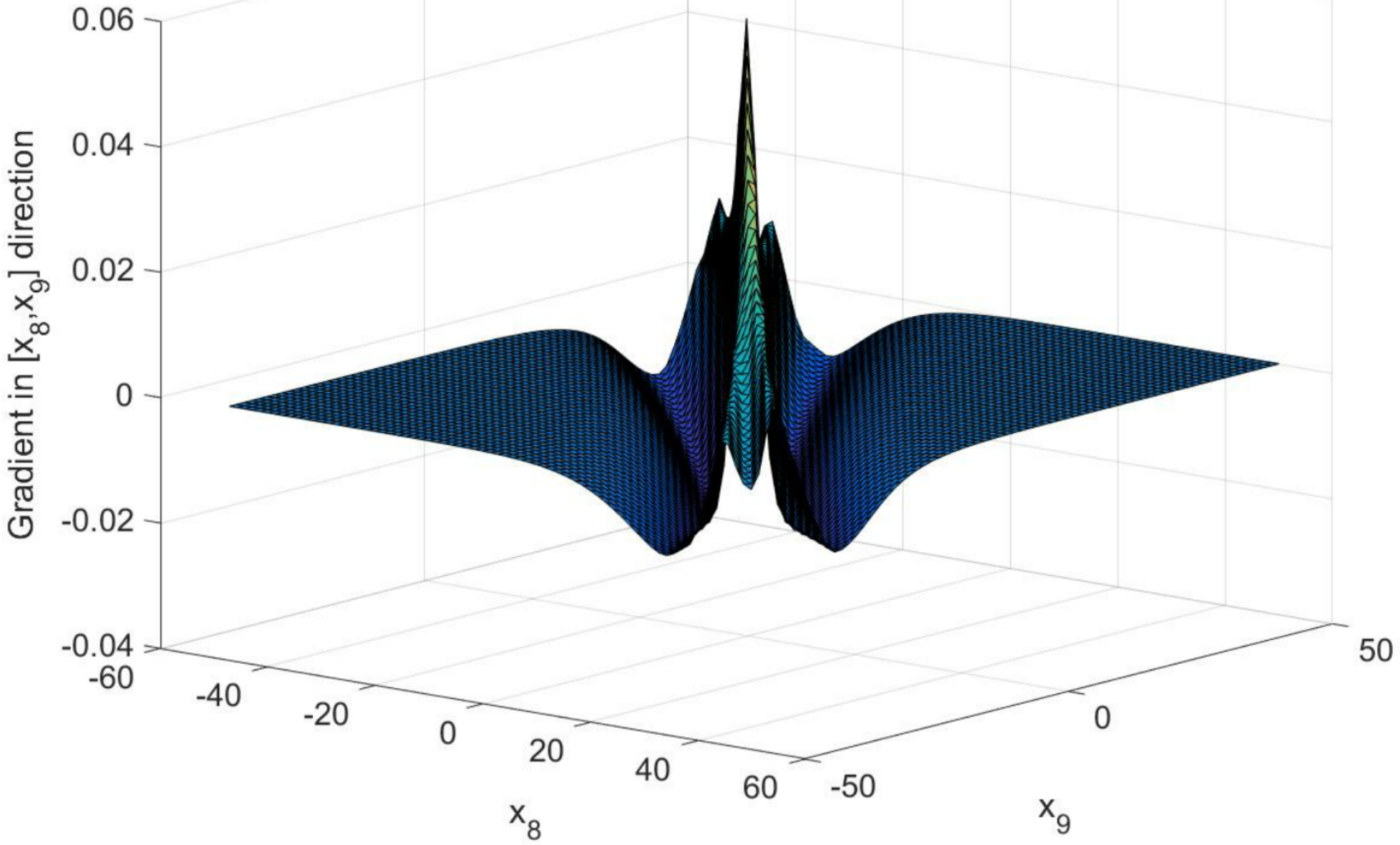




Function

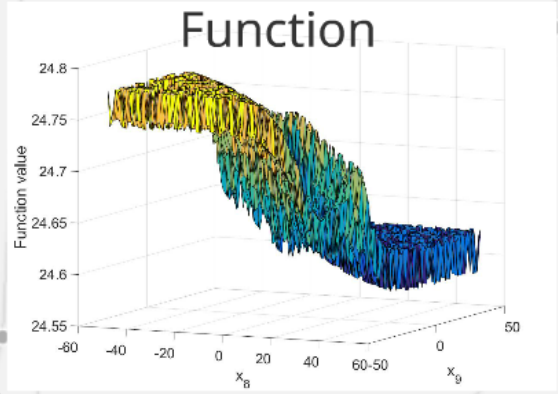

24.8
24.75
24.7
24.65
24.6
Function value
-60
-40
-20
0
20
40
60
x_8
-50
0
50
x_9

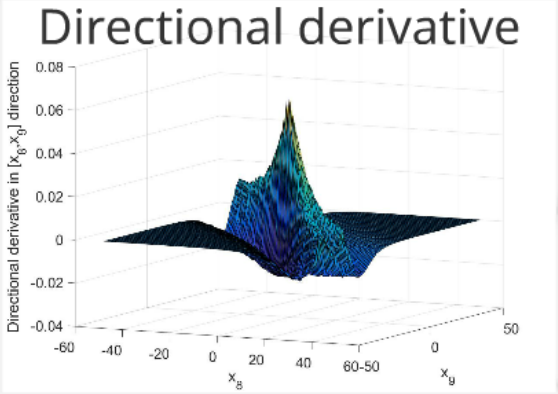
Directional derivative



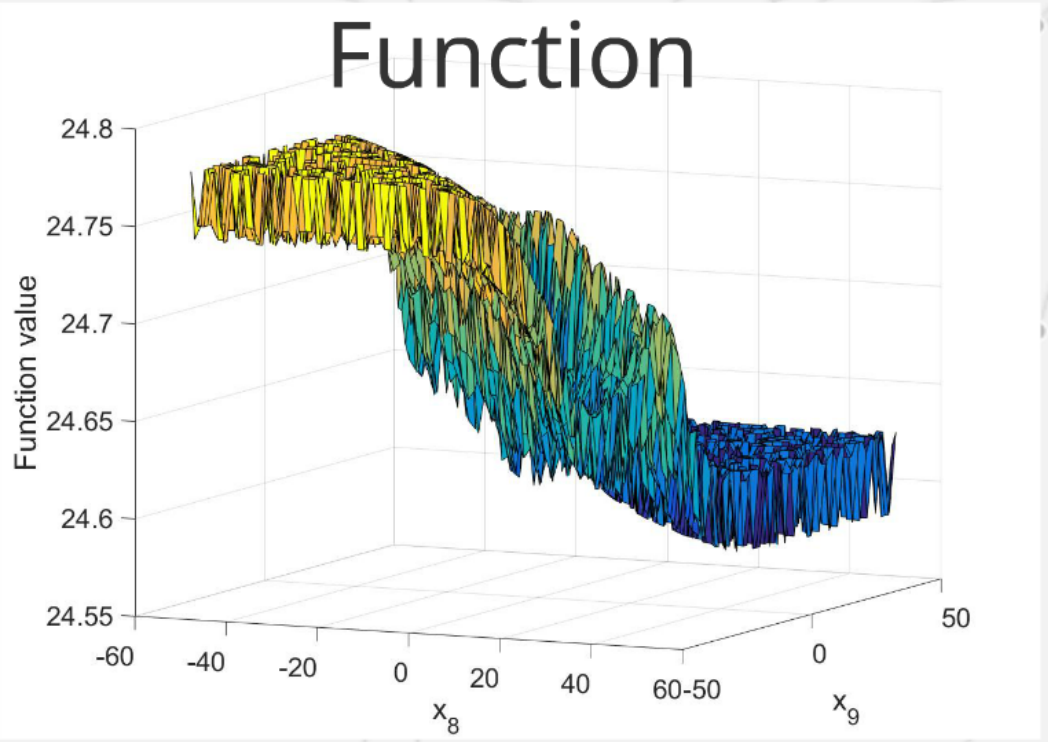
Changes in landscapes due to mini-batches

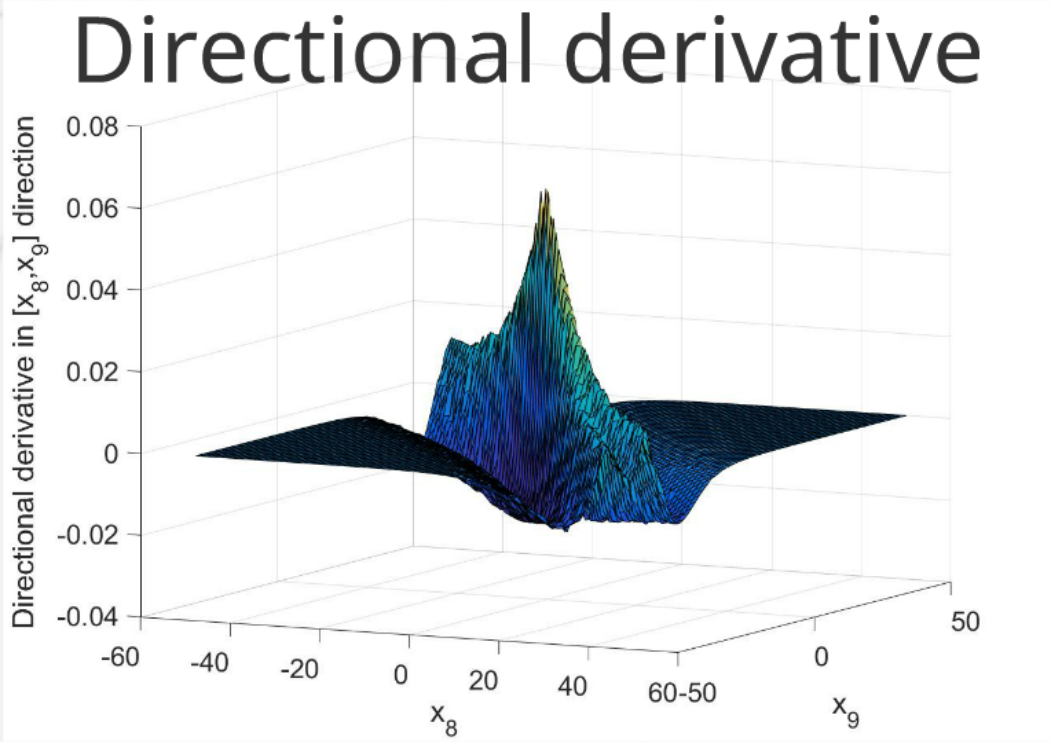
Large Mini-Batches - 75/76



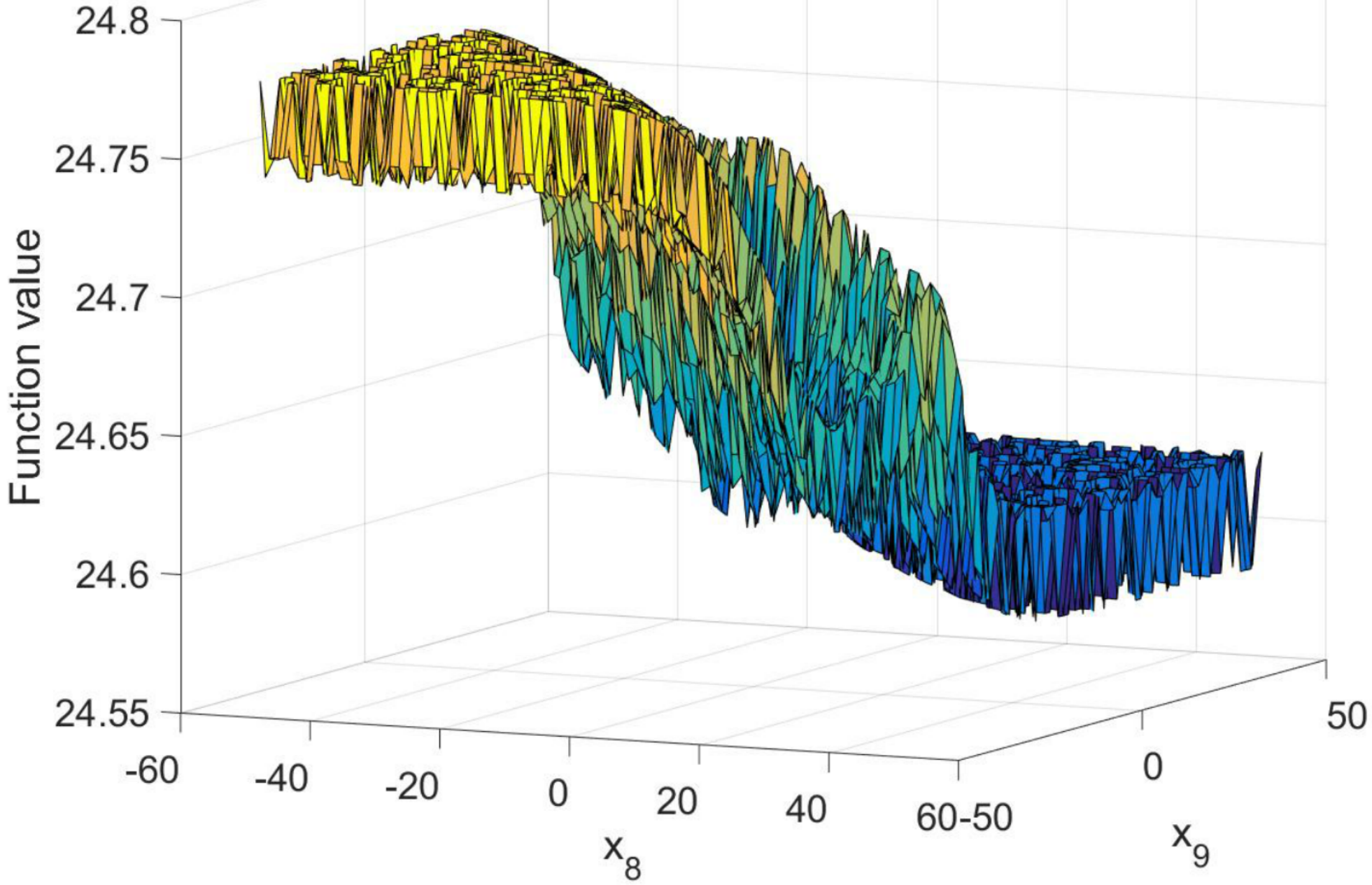


Large Mini-Batches - 75/76

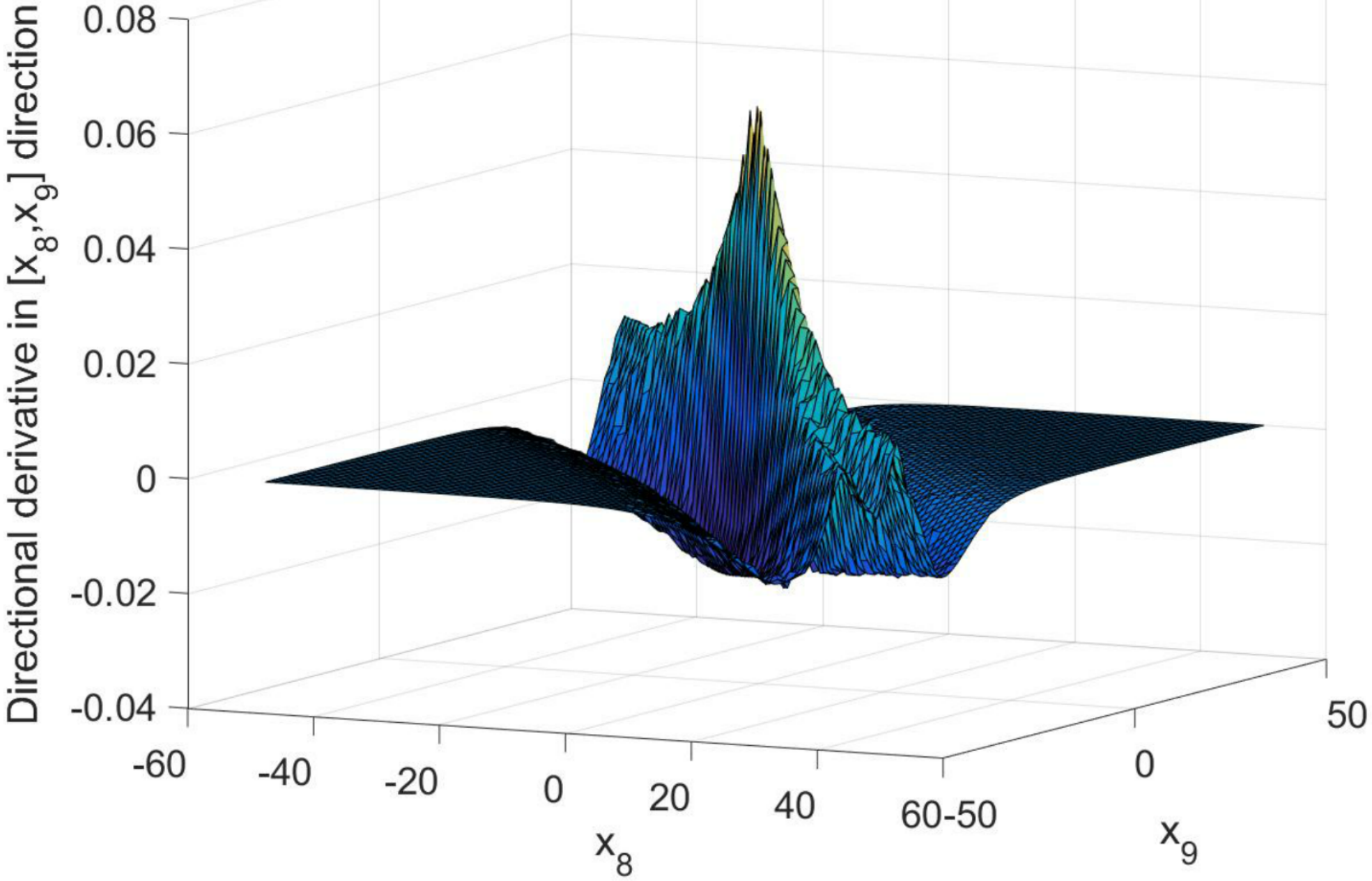




Function

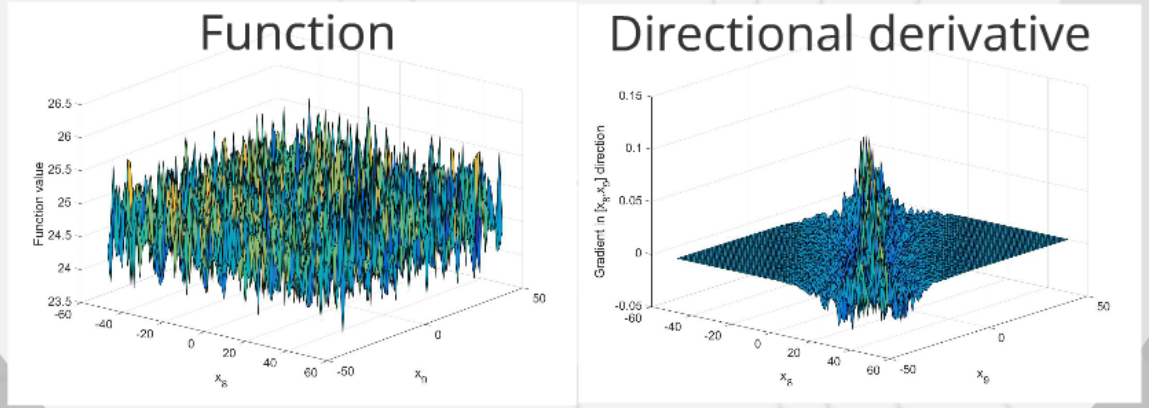

24.8
24.75
24.7
24.65
24.6
24.55
Function value
-60
-40
-20
0
20
40
60
x_8
-50
0
50
x_9

Directional derivative

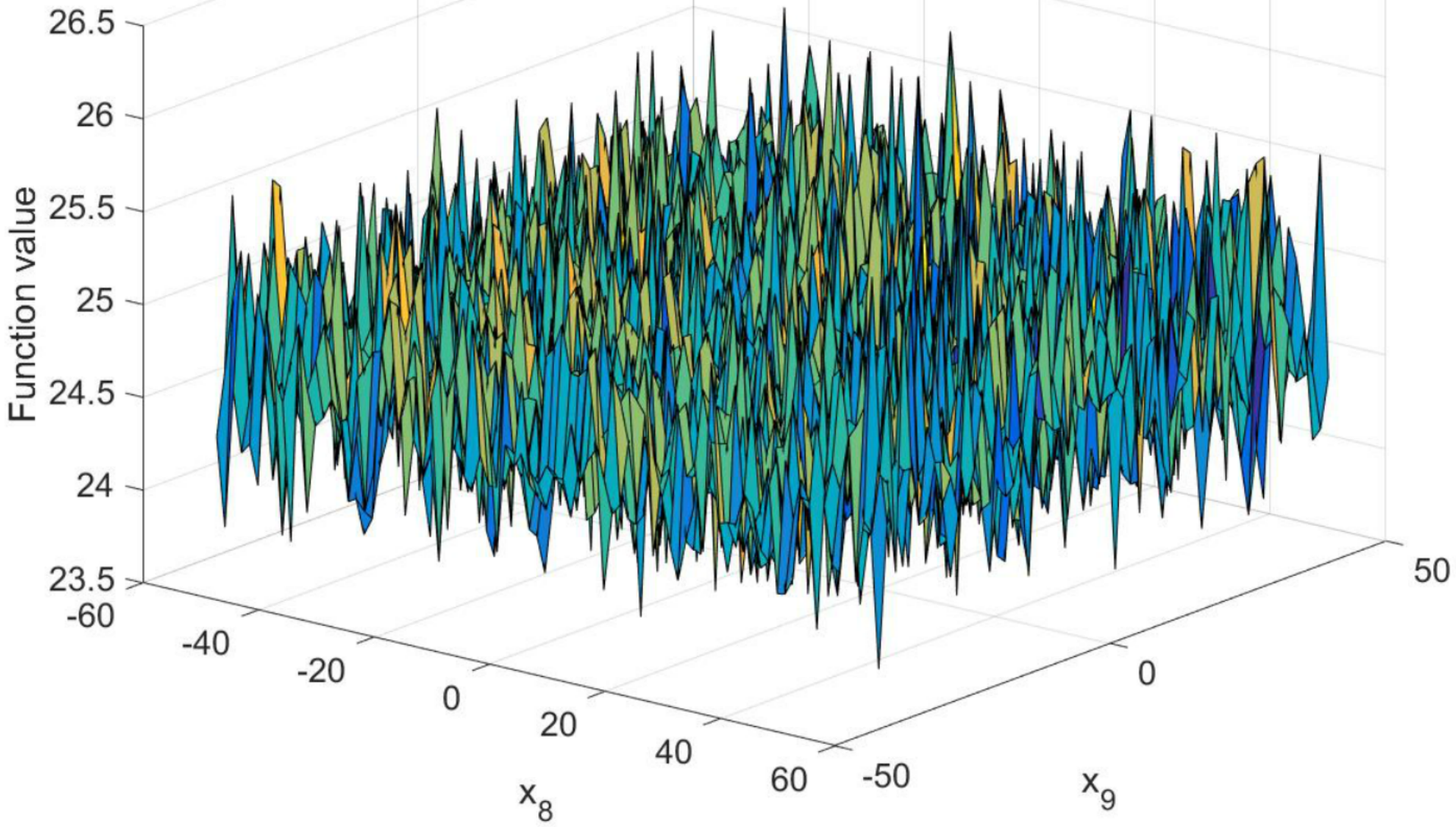


Changes in landscapes due to mini-batches

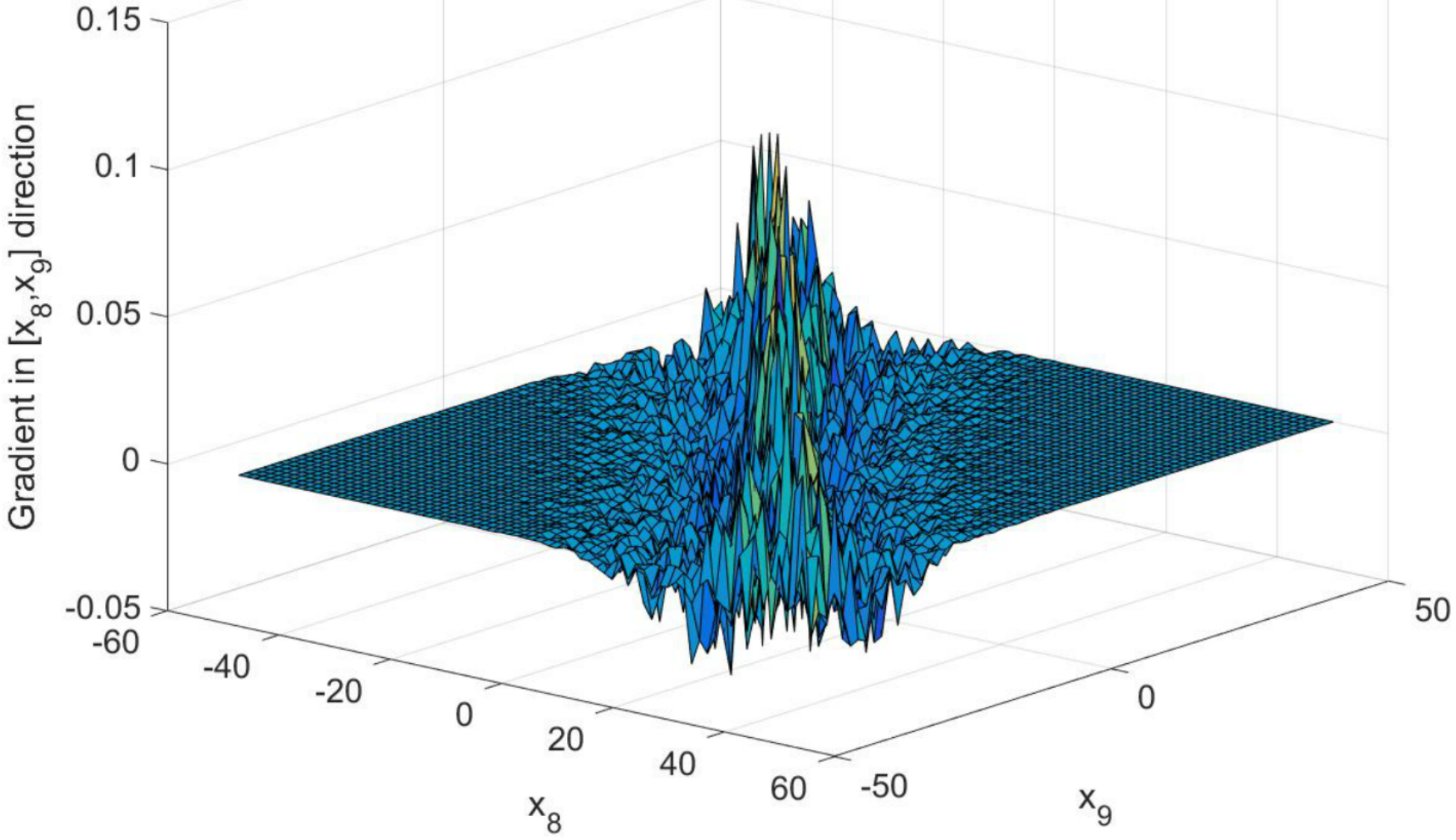
Small Mini-Batches - 10/76



Function

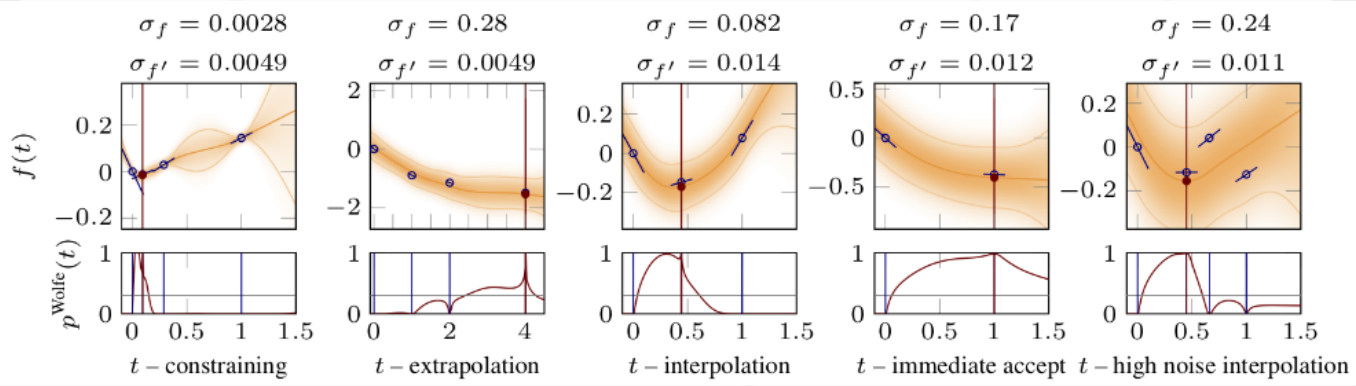

Function value
26.5
26
25.5
25
24.5
24
23.5
-60
-40
-20
0
20
40
60
x_8
-50
0
50
x_9

Directional derivative



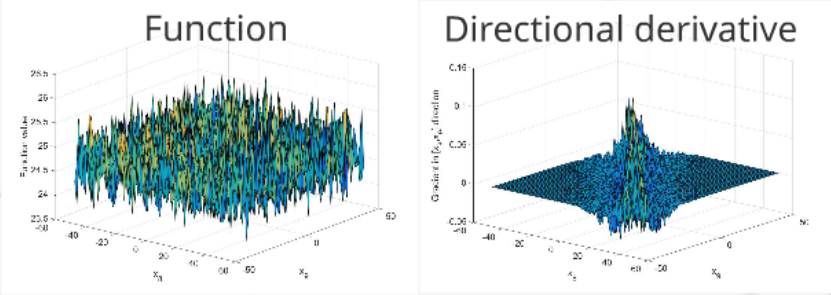
Line searches
Function minimization

Maren Mahsereci and Phillip Hennig (2015)



arxiv.org/pdf/1502.02846.pdf

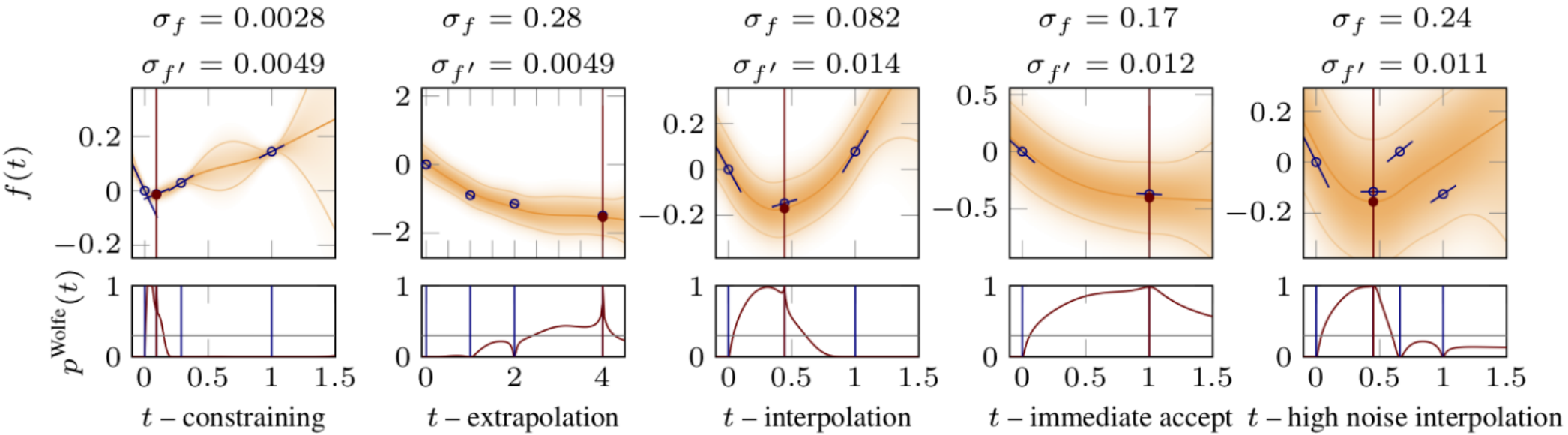
Is this the only way? Did we pay attention?



Which side looks less noisy?

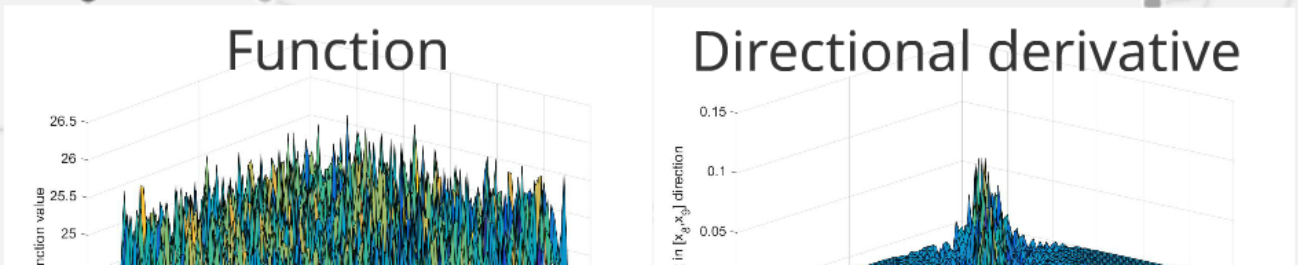
Line searches
Function minimization

Maren Mahsereci and Phillip Hennig (2015)



arxiv.org/pdf/1502.02846.pdf

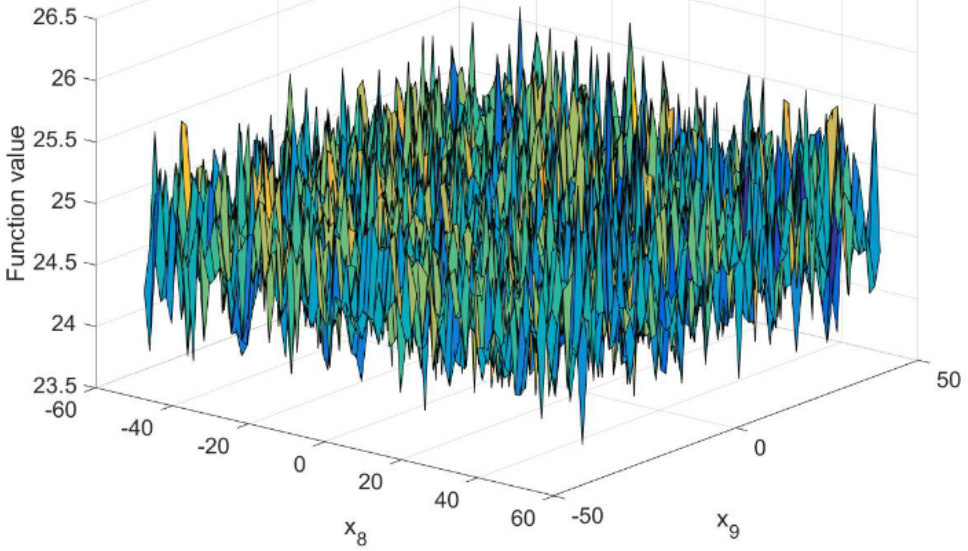
Is this the only way? Did we pay attention?



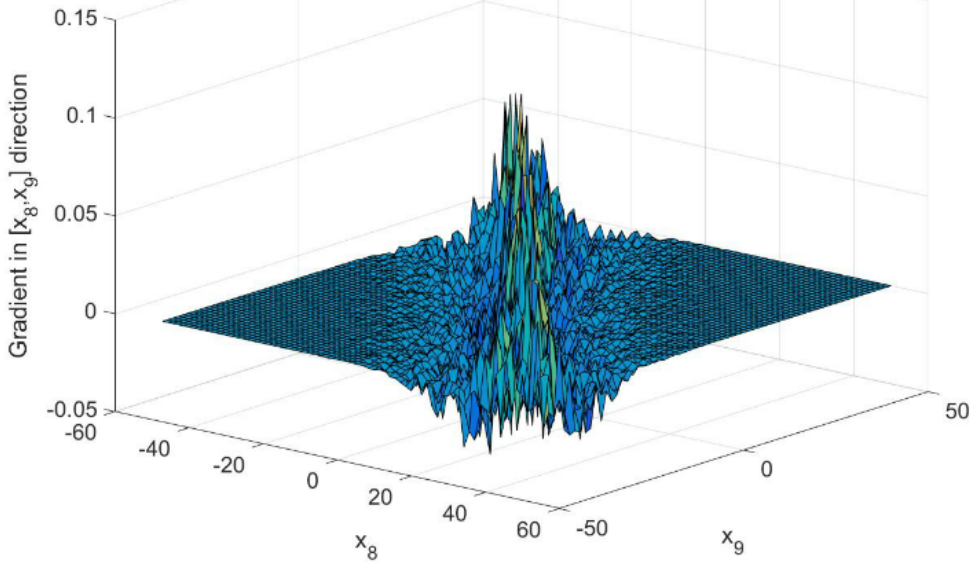
arxiv.org/pdf/1

ne only way? Did we pay attent

Function

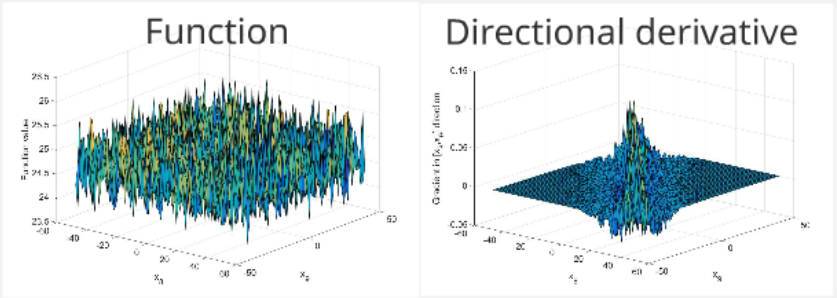


Directional derivative

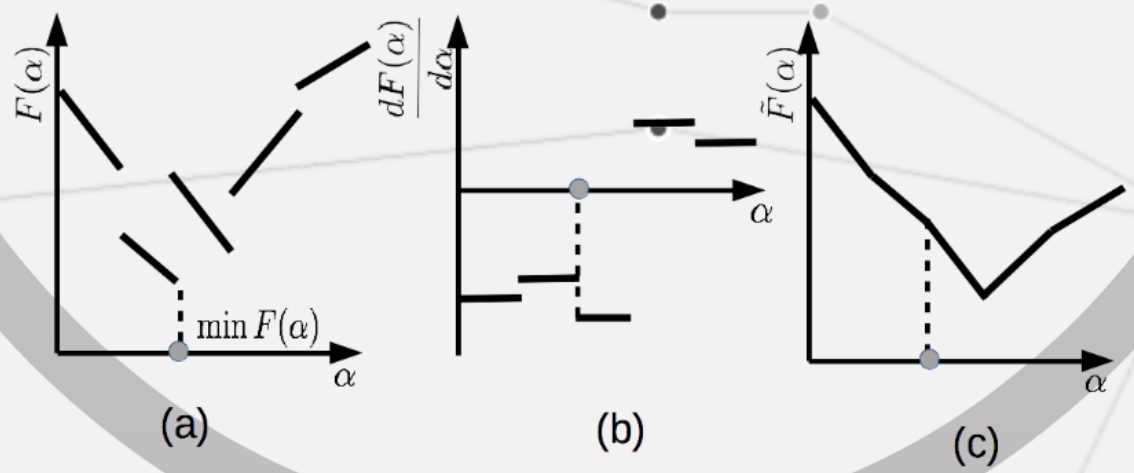


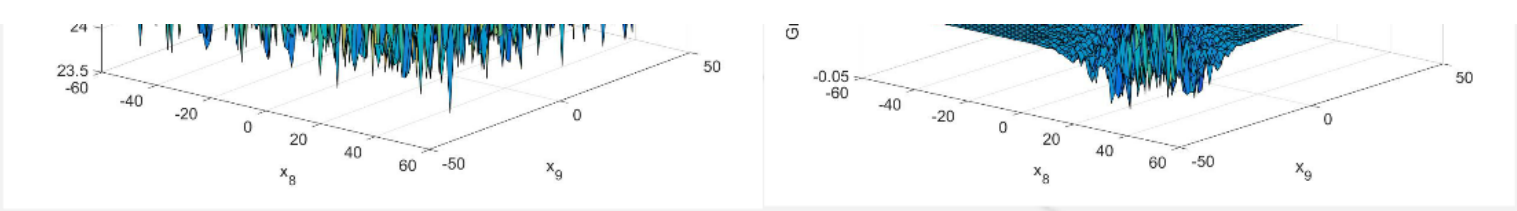
Which side looks less noisy?

Why are the gradients less noisy?

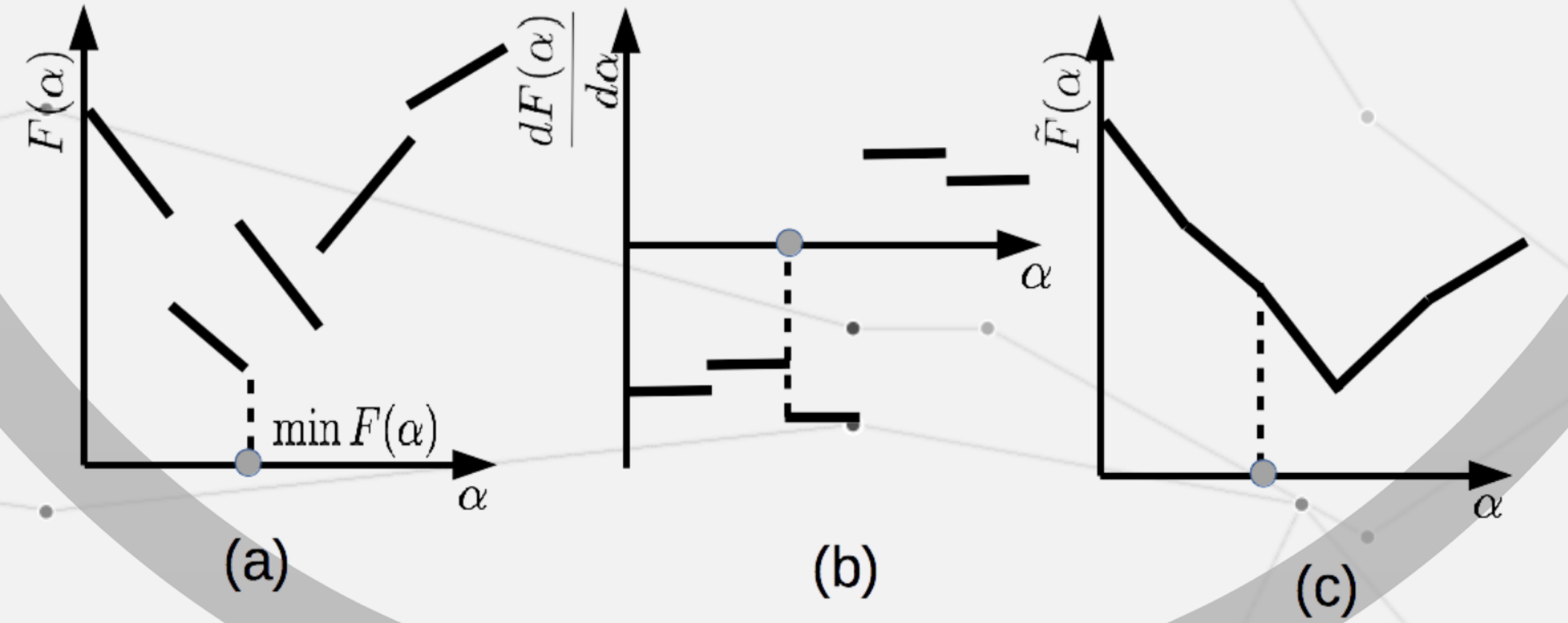


Filters the discontinuities

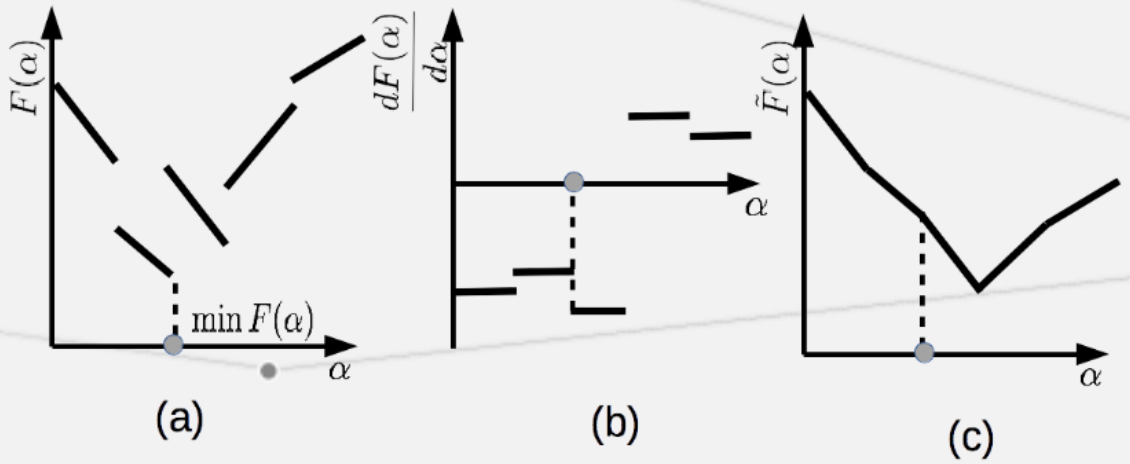




Filters the discontinuities



Important Observations



1. Functions are discontinuous
2. $\nabla f(\mathbf{x}) = \mathbf{0}$ may not exist
   i.e. no critical points!

Consequences of Observations

Mathematical programming options

1. Discontinuous - minimization issues

2. No critical points - No necessary first order criterion $\nabla f(\mathbf{x}) = \mathbf{0}$

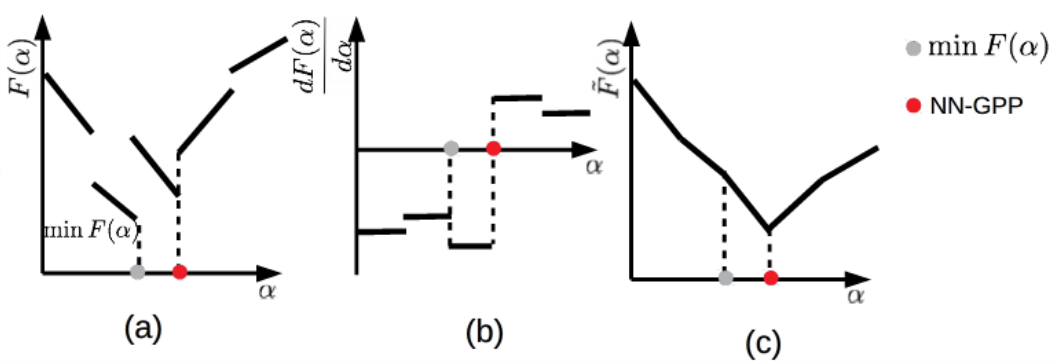
Alternatives

Subgradients - constant / learning rate schedules - back to square one!

Did we pay attention?

Fixing the Problems

Let's pay attention!



Find the location of a sign change in the directional derivative from - to +

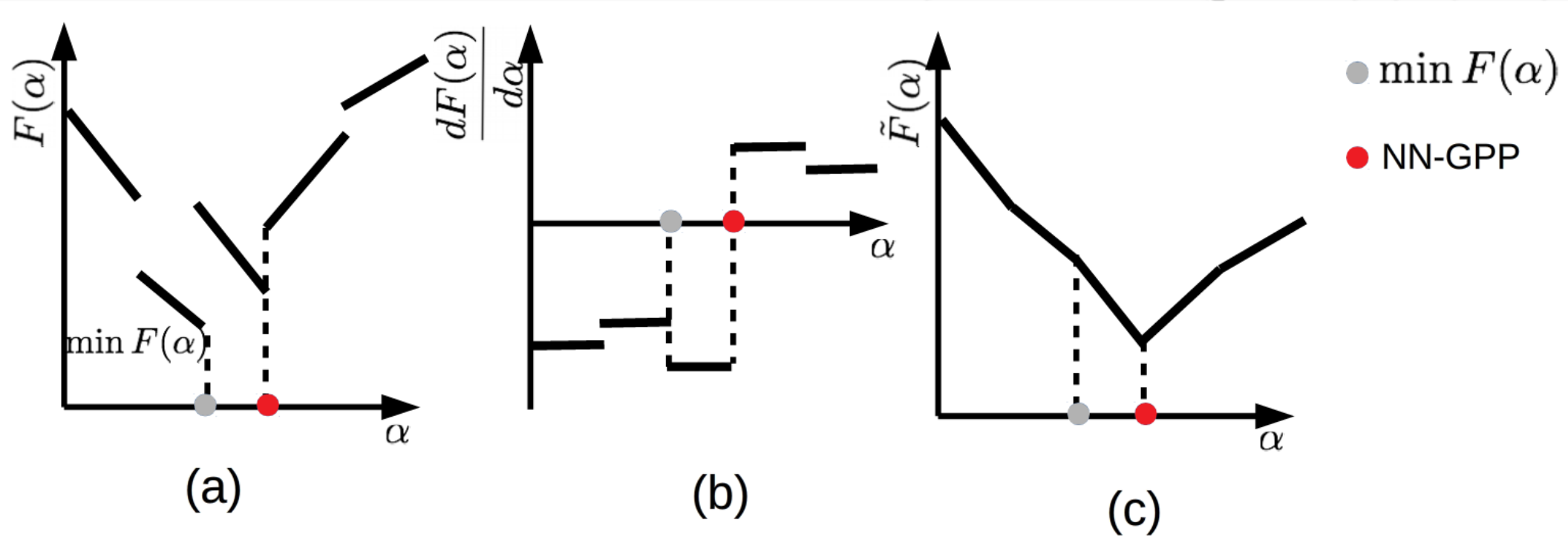
Descent (- derivative) while ascent (+ derivative)

Guaranteed local minimum in the derivative world!

Welcome to Gradient-Only Optimization
Non-Negative Gradient Projection Points
NN-GPP

the Problems

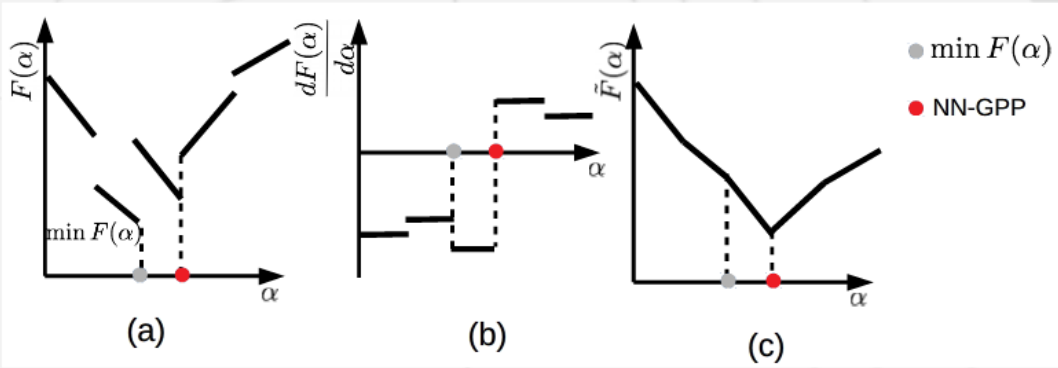
Let's pay attention!



nd the location of a sign change in the direc
rivative from - to +

Fixing the Problems

Let's pay attention

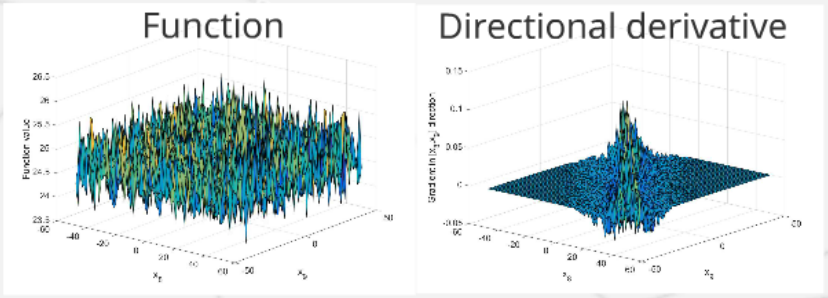


Three Guidelines for Discontinuous Functions:

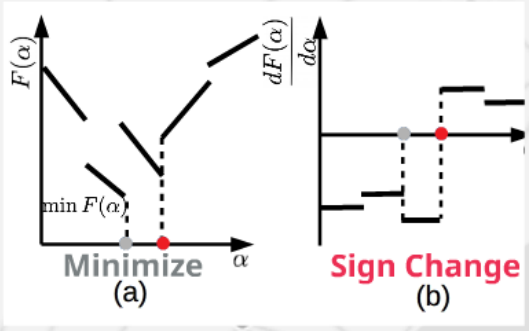
1. Don't minimize (local minima in function world)
2. Don't find critical points $\nabla f(\mathbf{x}) = \mathbf{0}$ candidates
3. **Locate NN-GPP (local minima in derivative world)**

Are we set for line searches?

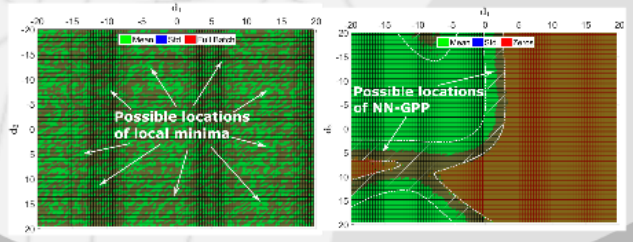
Double Benefit



More Mini-Batches - 10/76



Implication

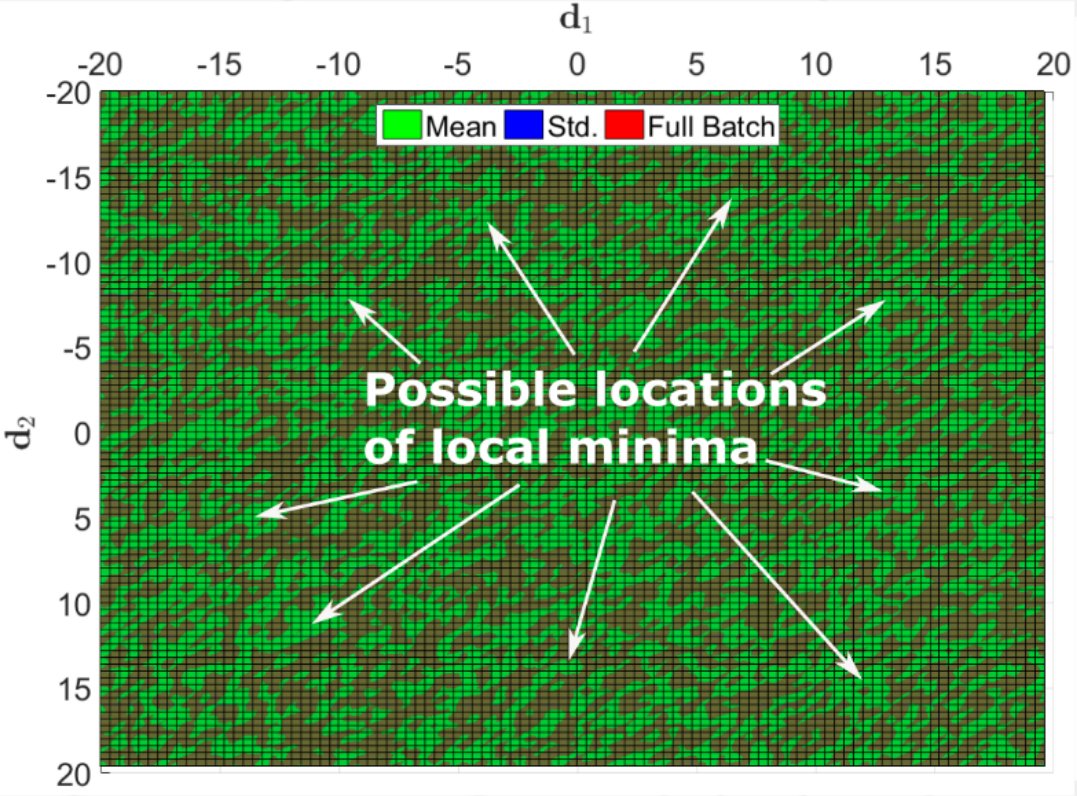


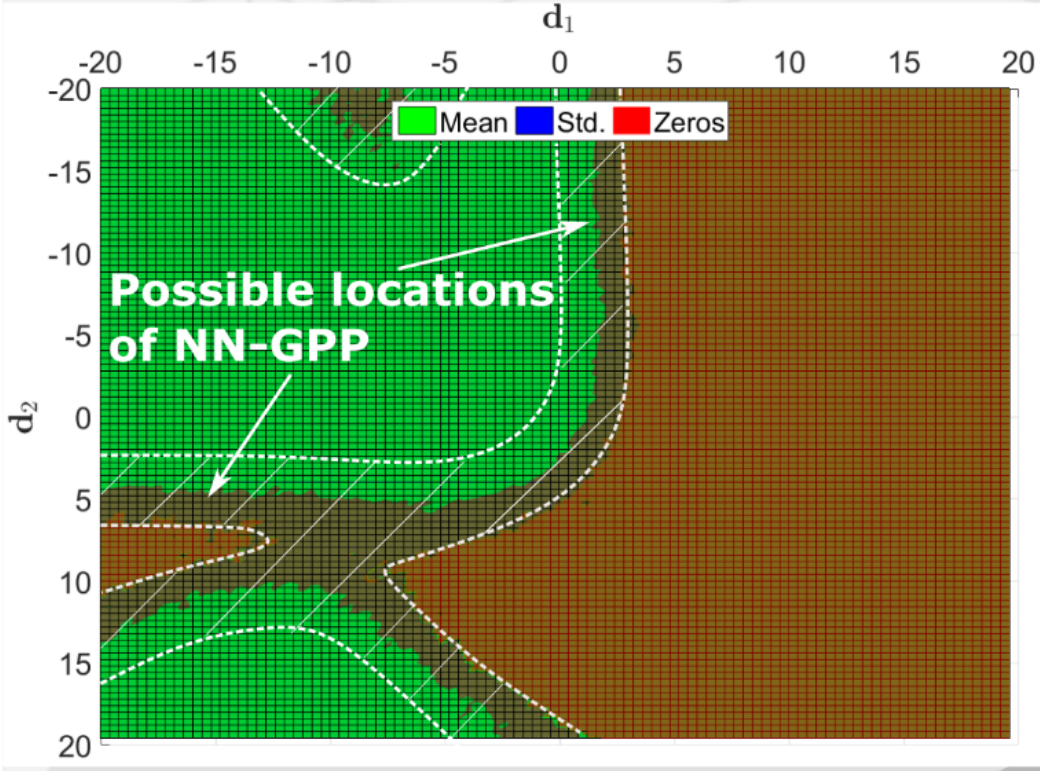
(a)

Sign Change

(b)

Implication





Where is my line search?

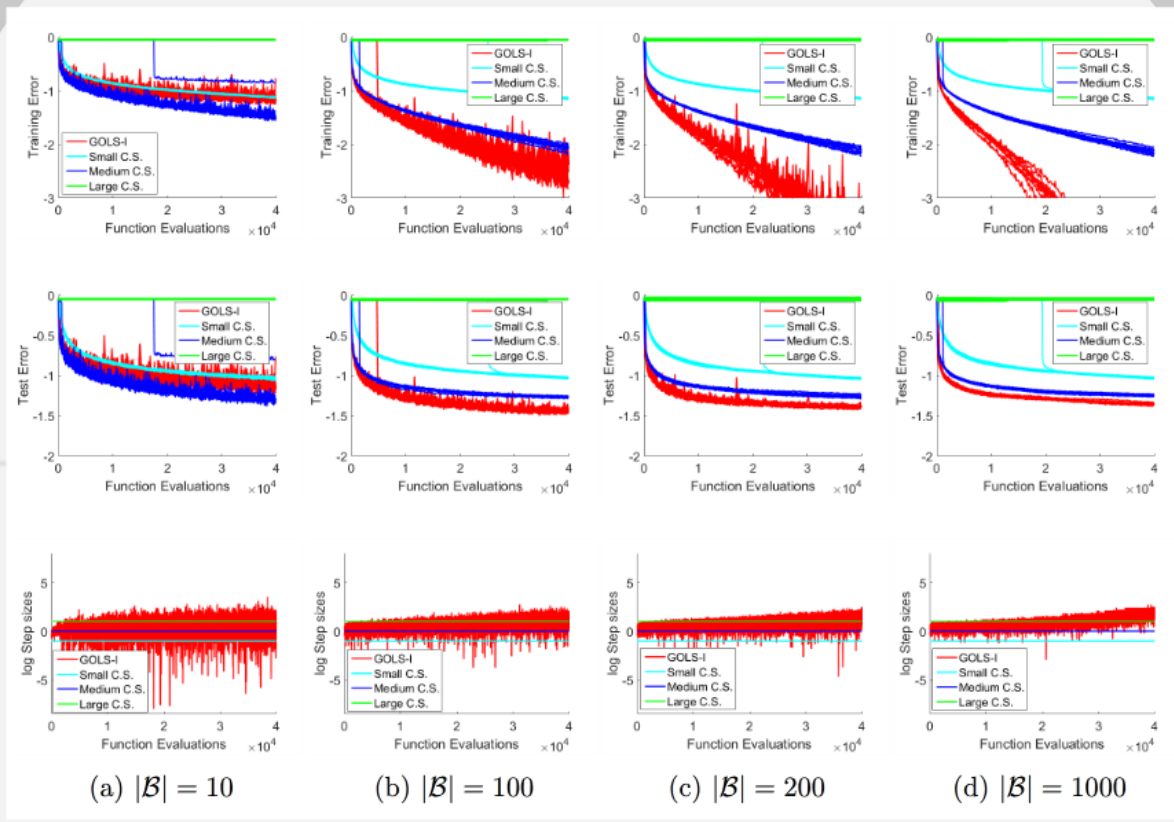
Gradient-Only Line Search Inexact GOLS-I

Dominic
Kafka (PhD)

GOLS-I FEATURES

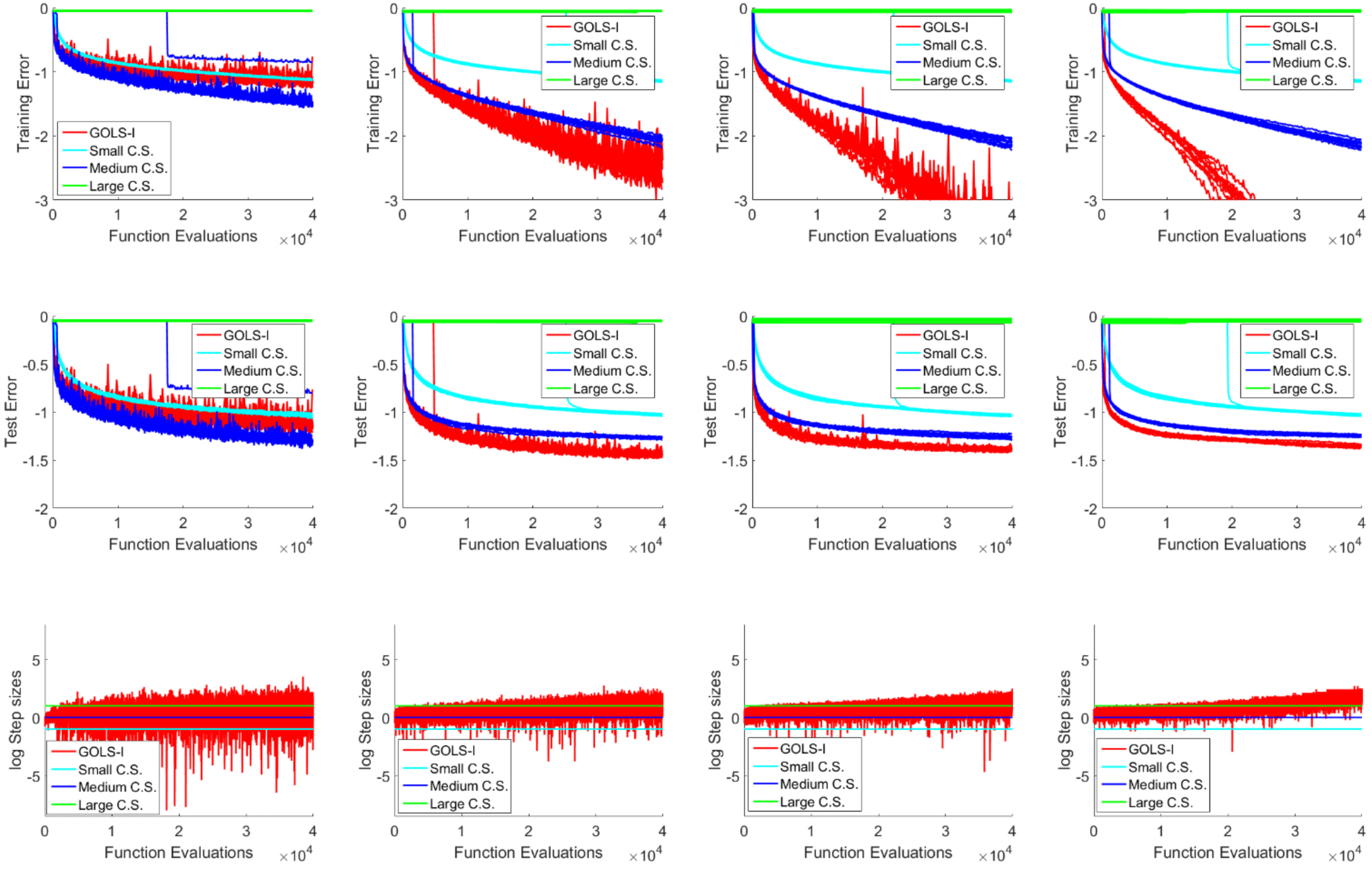
1. No tuning parameter - automatic line search
2. Step size range - 15 orders of magnitude
3. Barrier to Entry - High school math
4. PyTorch (Done) and TensorFlow (Testing)

Results MNIST



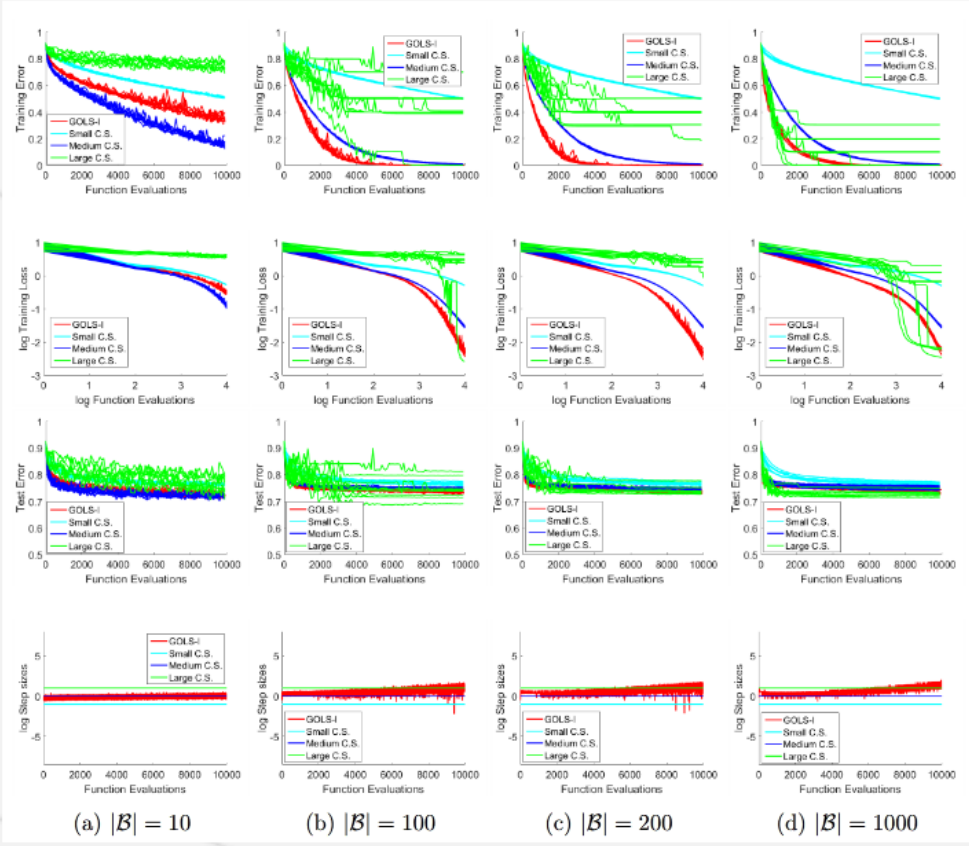
(a) $|\mathcal{B}| = 10$ (b) $|\mathcal{B}| = 100$ (c) $|\mathcal{B}| = 200$ (d) $|\mathcal{B}| = 1000$

Net I - HL:800; AF:SIGMOID; INIT: N(0,1); LOSS: Cross Entropy



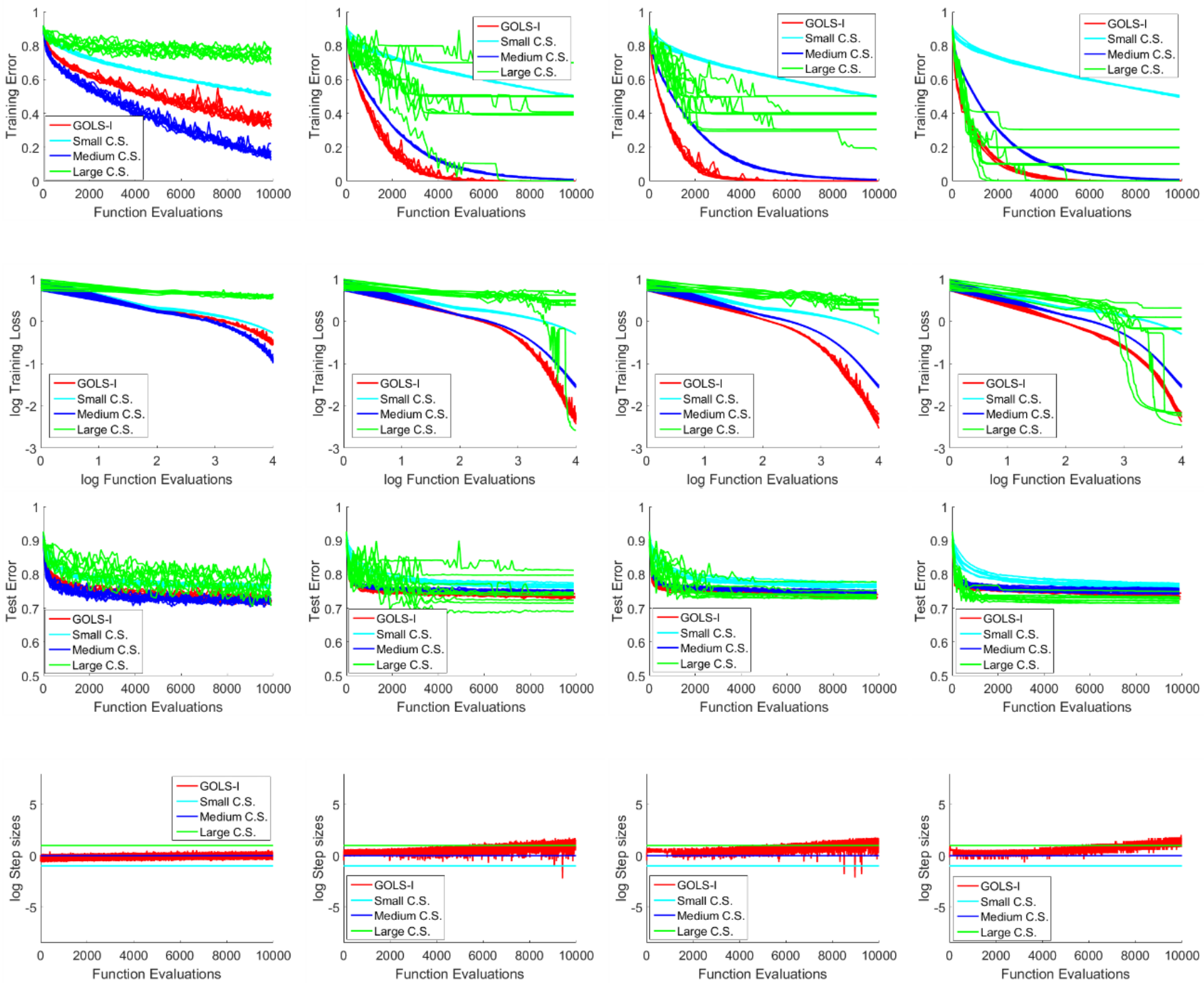
(a) $|\mathcal{B}| = 10$ (b) $|\mathcal{B}| = 100$ (c) $|\mathcal{B}| = 200$ (d) $|\mathcal{B}| = 1000$

Results CIFAR



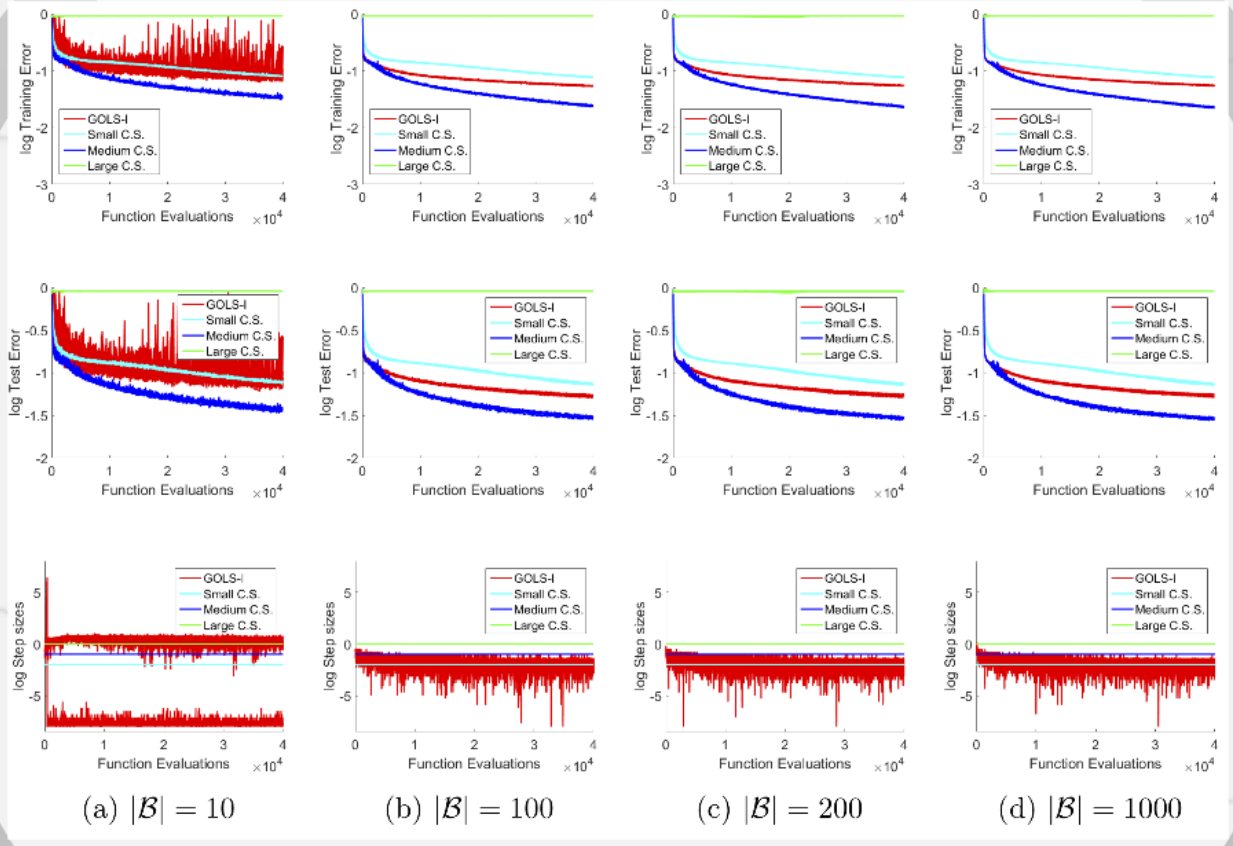
(a) $|\mathcal{B}| = 10$ (b) $|\mathcal{B}| = 100$ (c) $|\mathcal{B}| = 200$ (d) $|\mathcal{B}| = 1000$

Net I - HL:800; AF:SIGMOID; INIT: N(0,1); LOSS: Cross Entropy



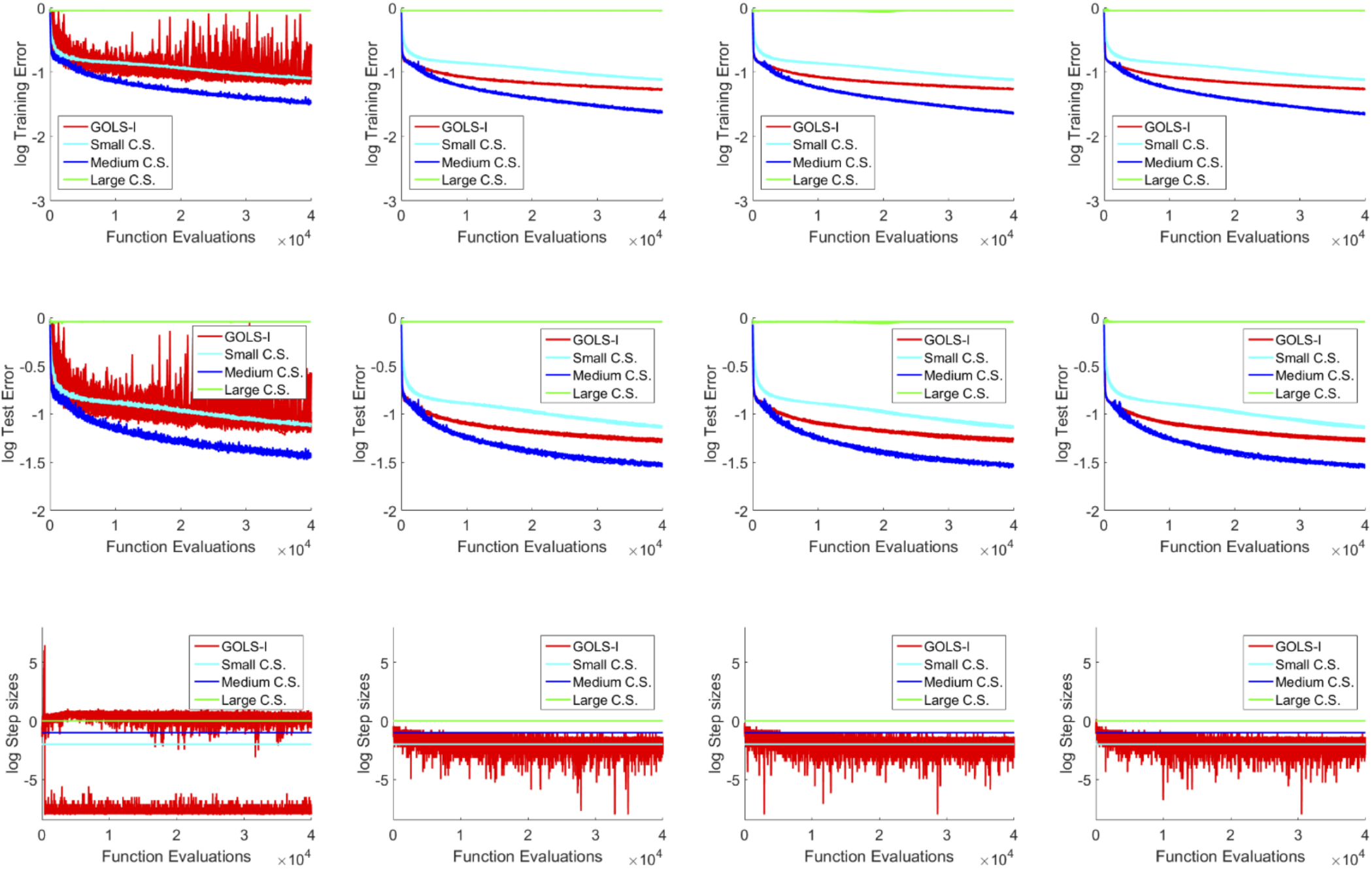
(a) $|\mathcal{B}| = 10$ (b) $|\mathcal{B}| = 100$ (c) $|\mathcal{B}| = 200$ (d) $|\mathcal{B}| = 1000$

Results MNIST



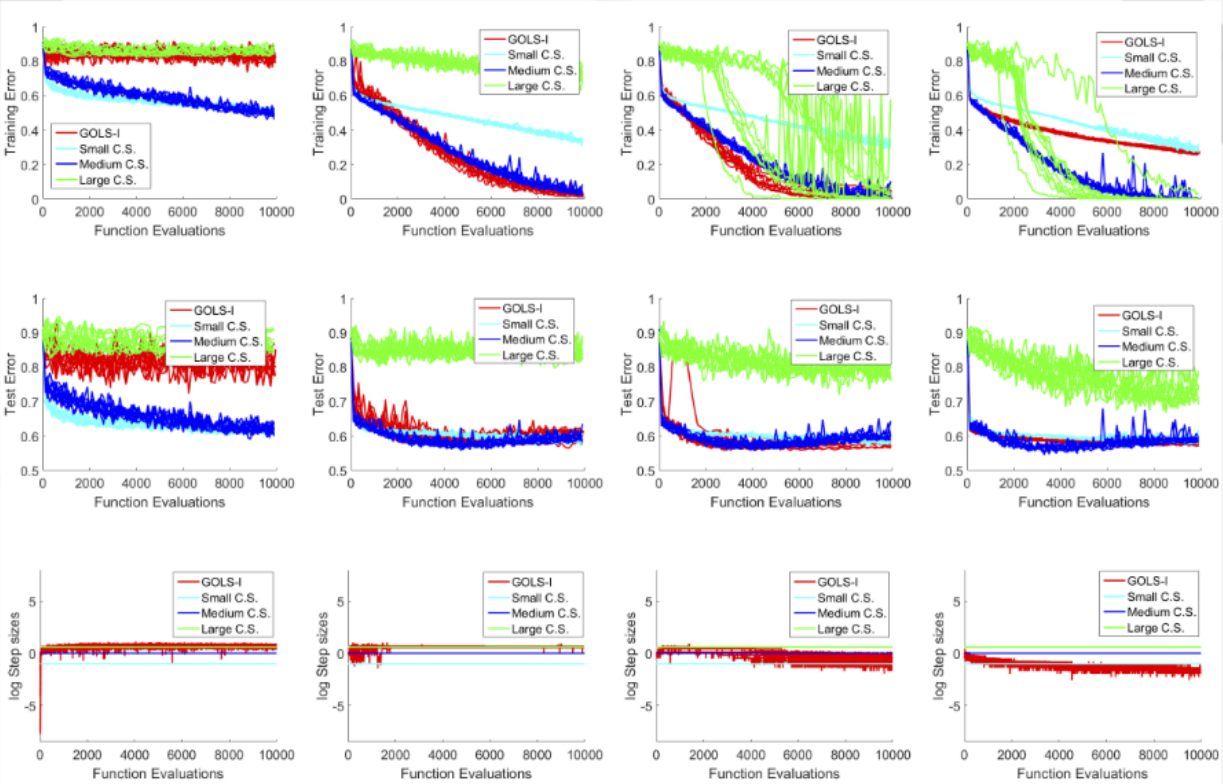
(a) $|\mathcal{B}| = 10$ (b) $|\mathcal{B}| = 100$ (c) $|\mathcal{B}| = 200$ (d) $|\mathcal{B}| = 1000$

HL:1000-500-250; AF:TANH; INIT:XAVIER; LOSS:MSE



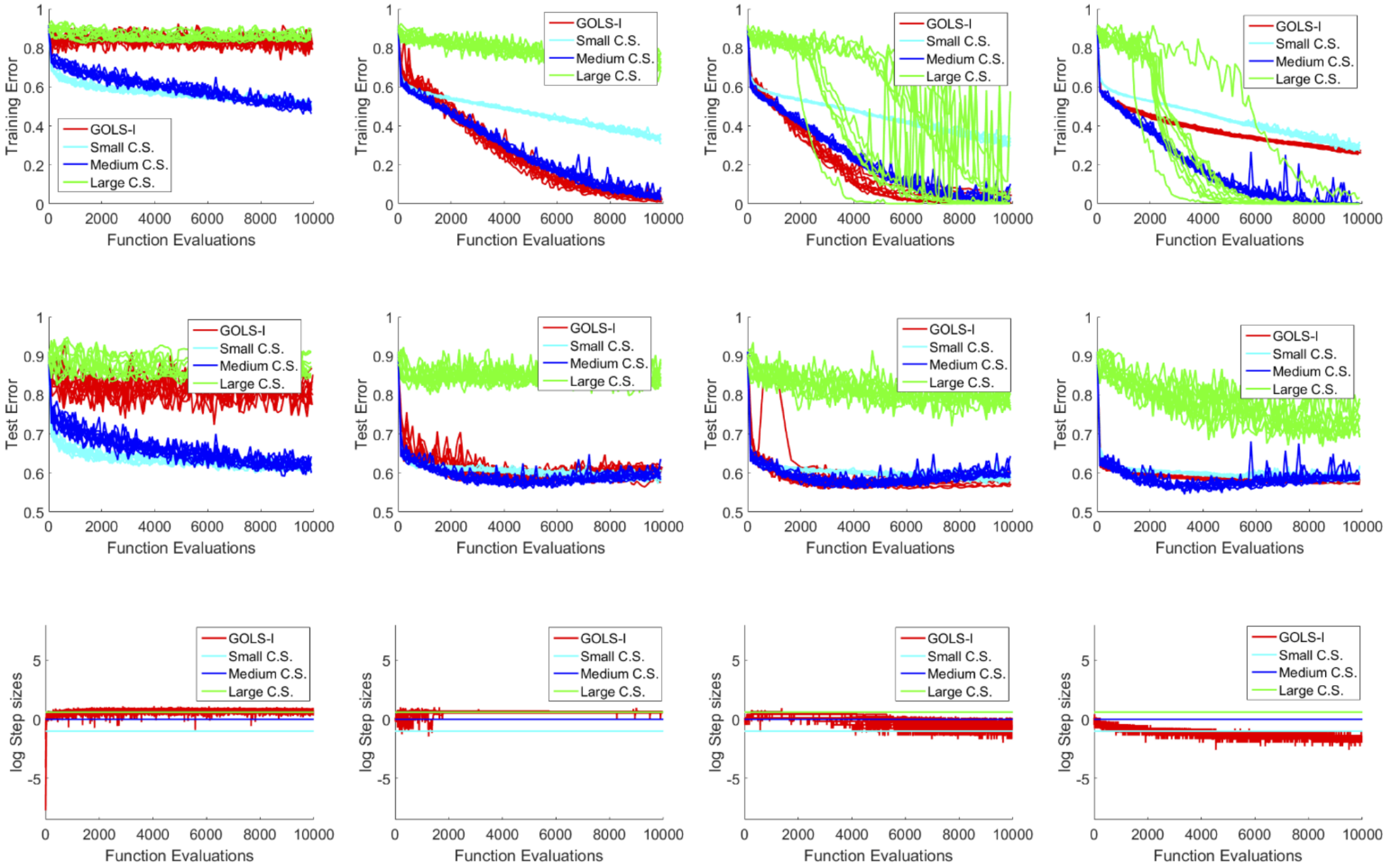
(a) $|\mathcal{B}| = 10$ (b) $|\mathcal{B}| = 100$ (c) $|\mathcal{B}| = 200$ (d) $|\mathcal{B}| = 1000$

Results CIFAR



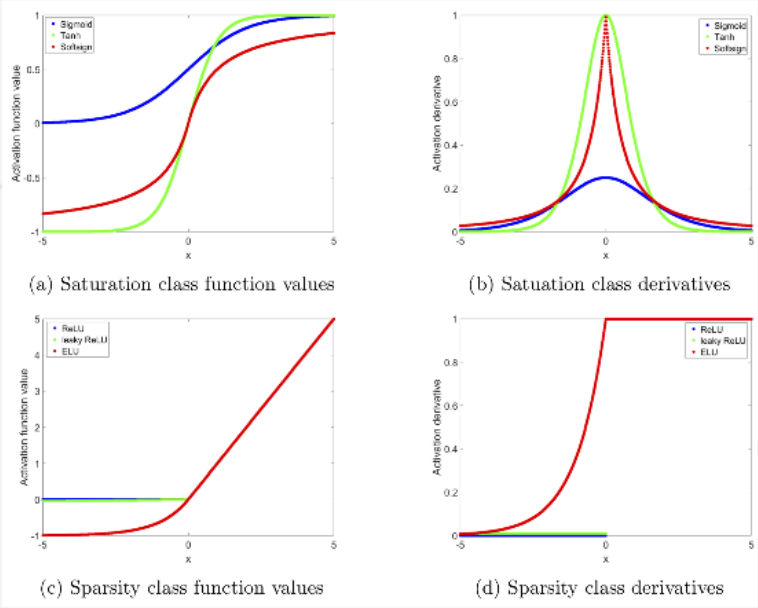
(a) $|\mathcal{B}| = 10$ (b) $|\mathcal{B}| = 100$ (c) $|\mathcal{B}| = 200$ (d) $|\mathcal{B}| = 1000$

HL:1000-500-250;
AF:TANH; INIT:XAVIER;
LOSS:MSE



(a) $|\mathcal{B}| = 10$ (b) $|\mathcal{B}| = 100$ (c) $|\mathcal{B}| = 200$ (d) $|\mathcal{B}| = 1000$

Sensitivity to Activation Functions



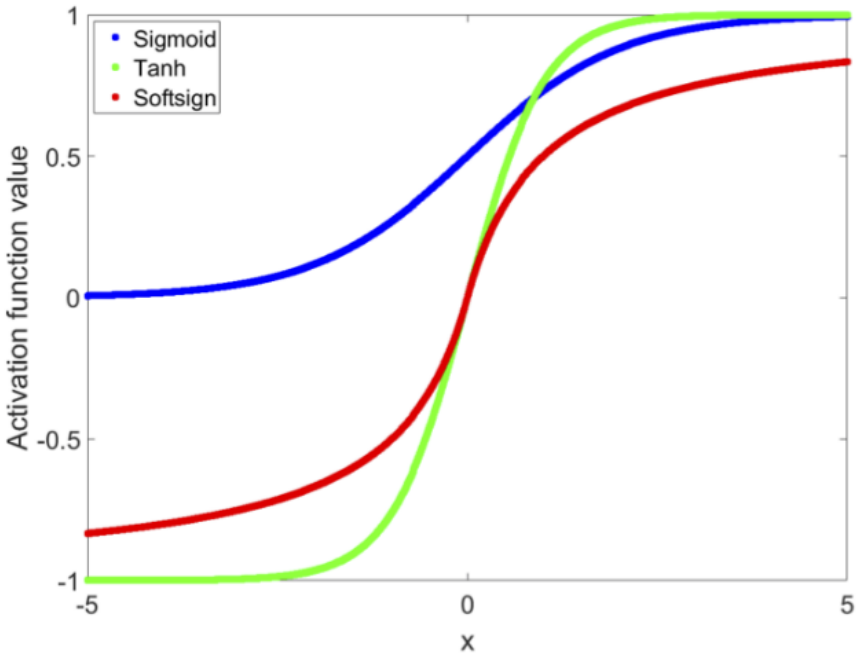
(a) Saturation class function values

(b) Satuation class derivatives

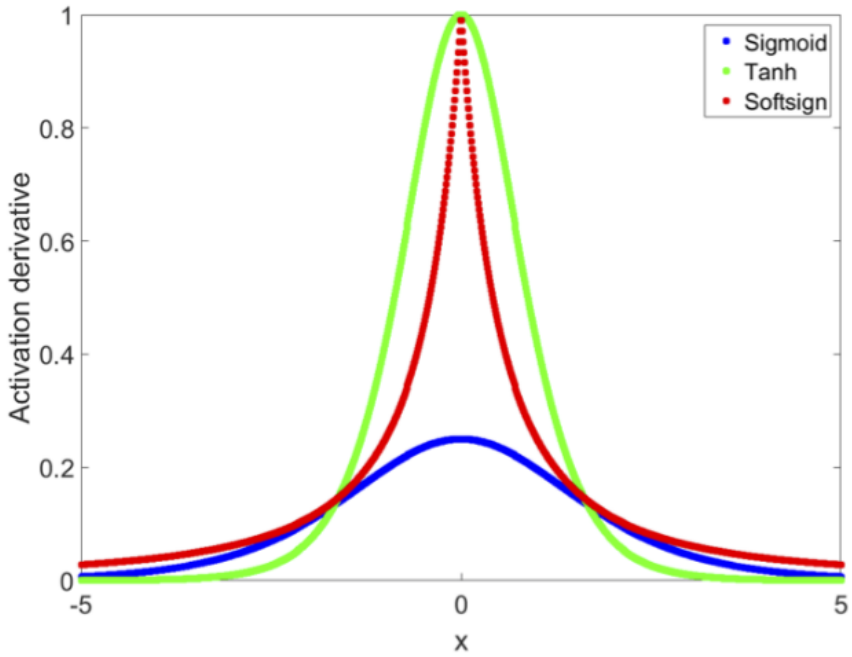
(c) Sparsity class function values

(d) Sparsity class derivatives

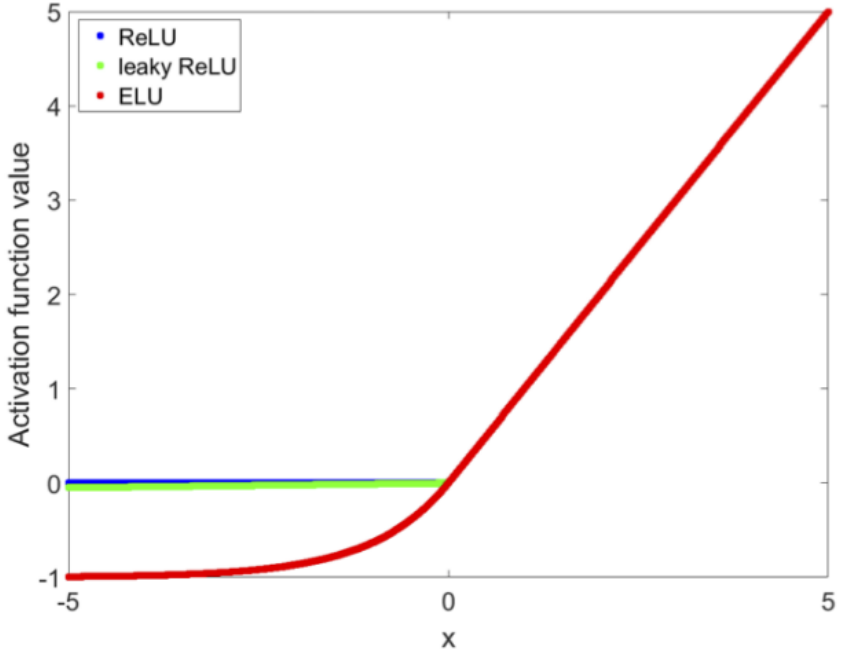
DATASET:IRIS; HL:10
AF:TANH; LOSS: MSE



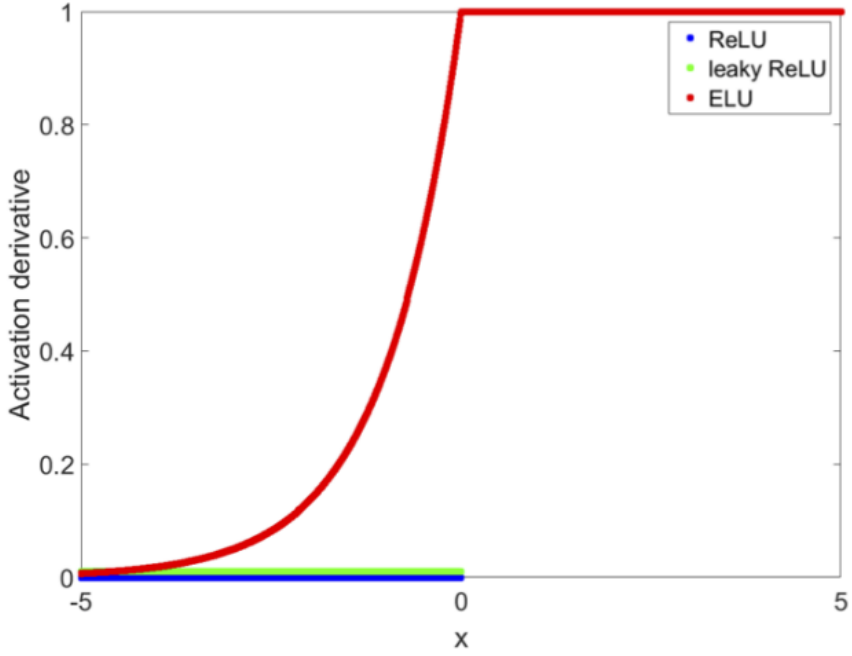
(a) Saturation class function values



(b) Satuation class derivatives

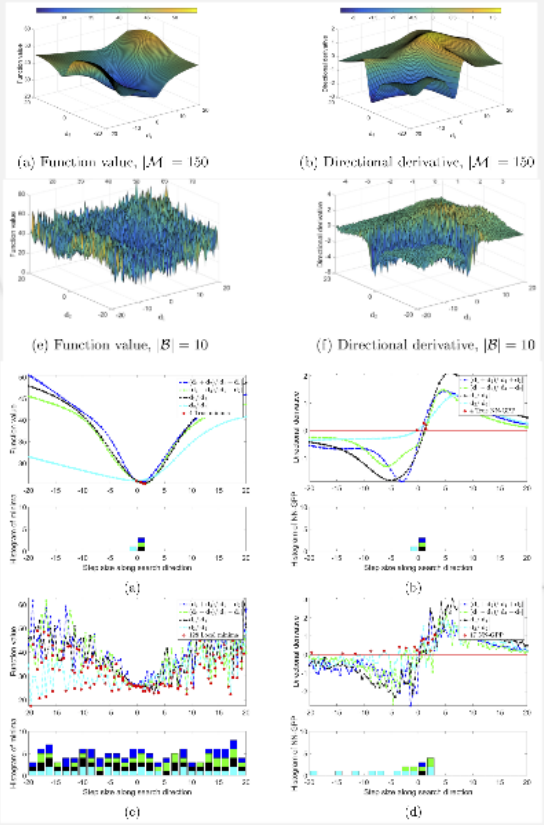


(c) Sparsity class function values



(d) Sparsity class derivatives

Role of Activation Functions SIGMOID



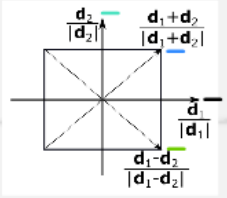
(a) Function value, $|\mathcal{M}| = 150$ (b) Directional derivative, $|\mathcal{M}| = 150$

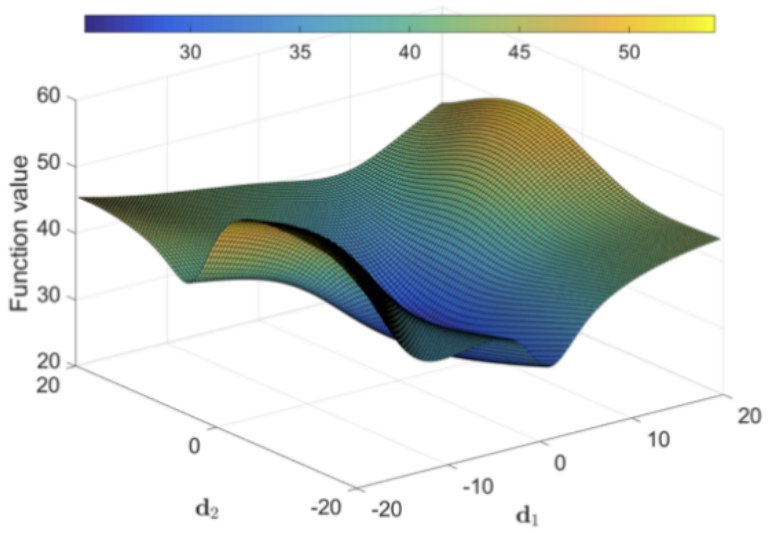
(e) Function value, $|\mathcal{B}| = 10$ (f) Directional derivative, $|\mathcal{B}| = 10$

(a) (b)

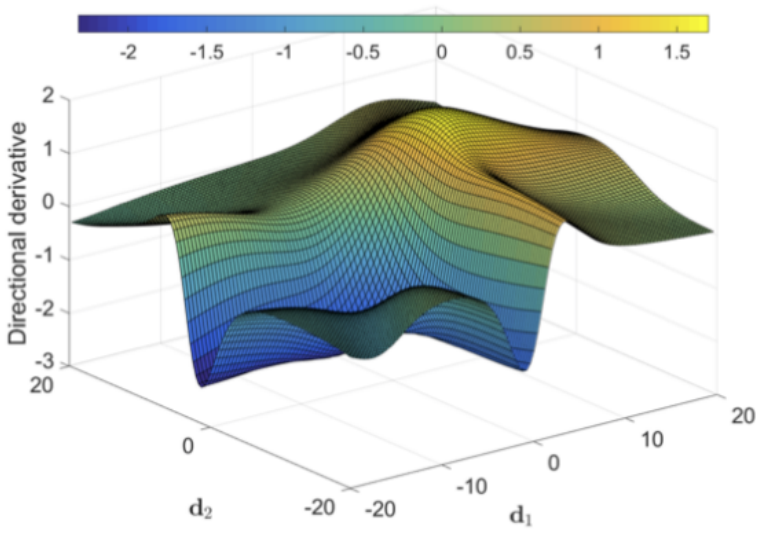
(c) (d)

Directions

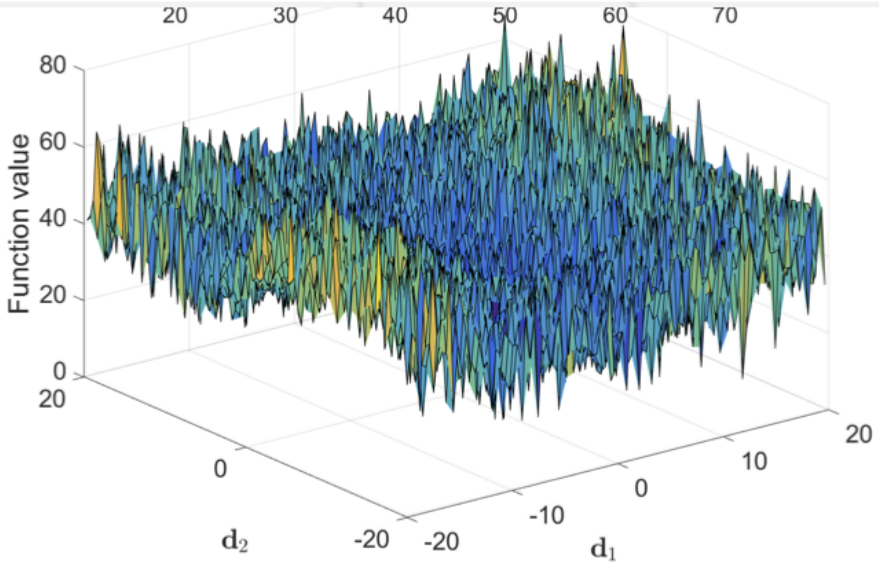




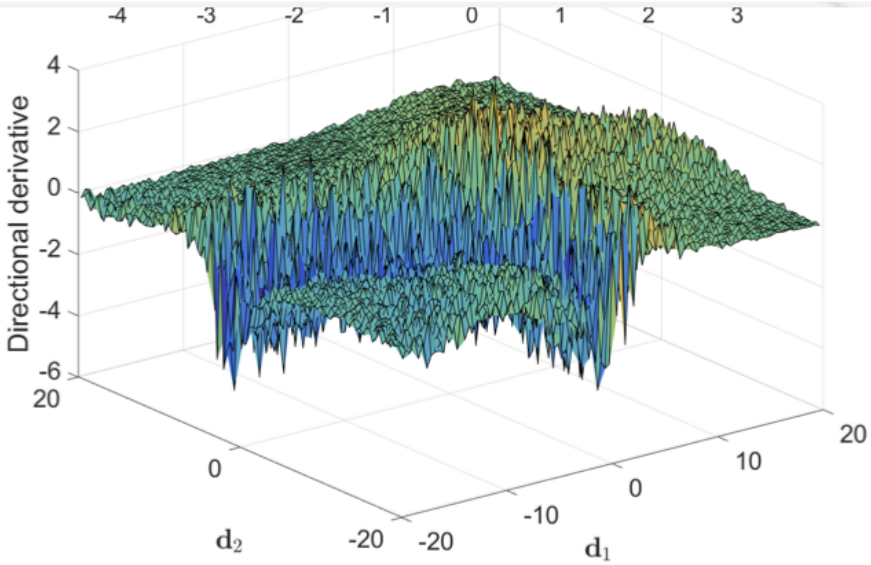
(a) Function value, $|\mathcal{M}| = 150$



(b) Directional derivative, $|\mathcal{M}| = 150$

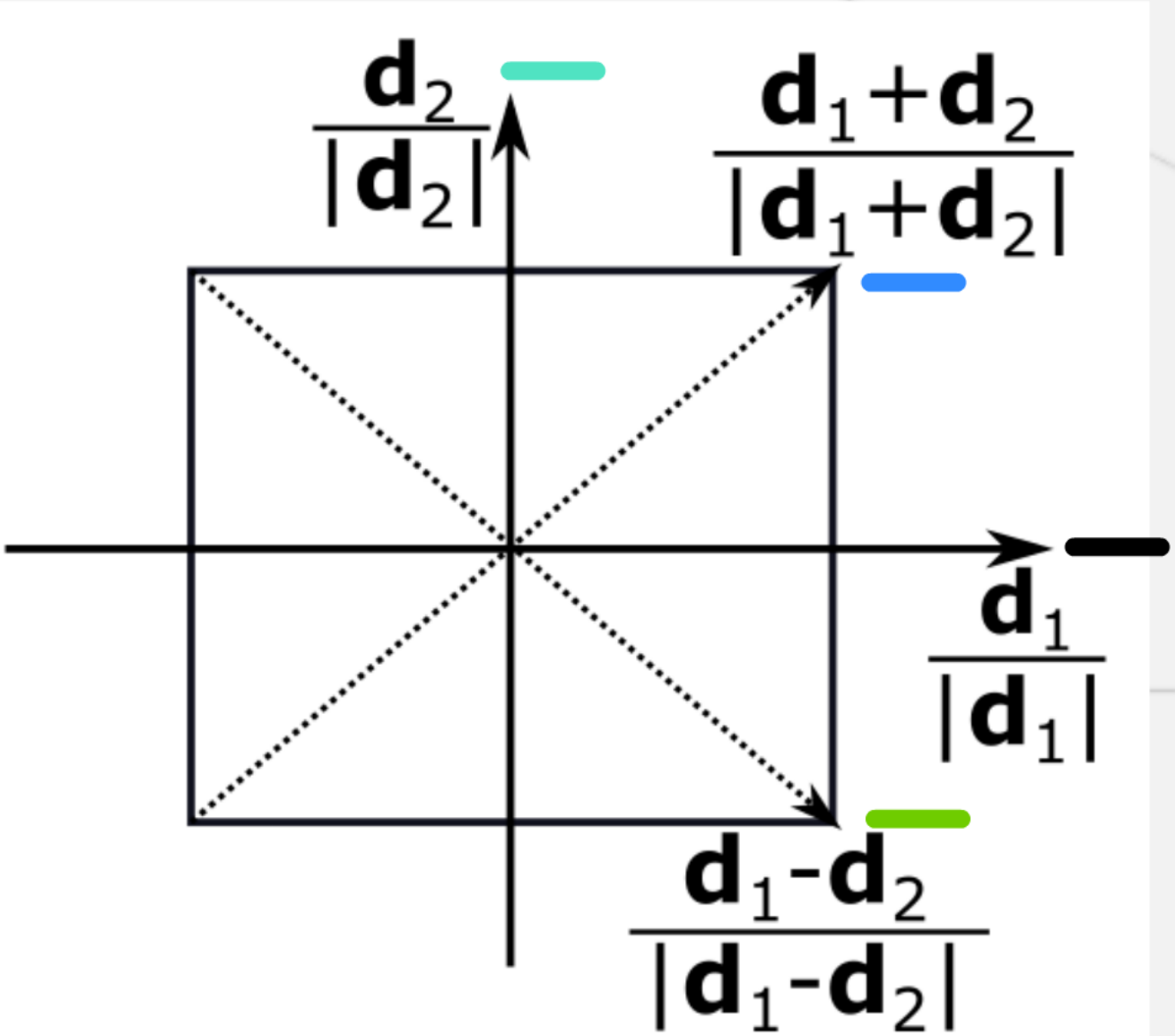


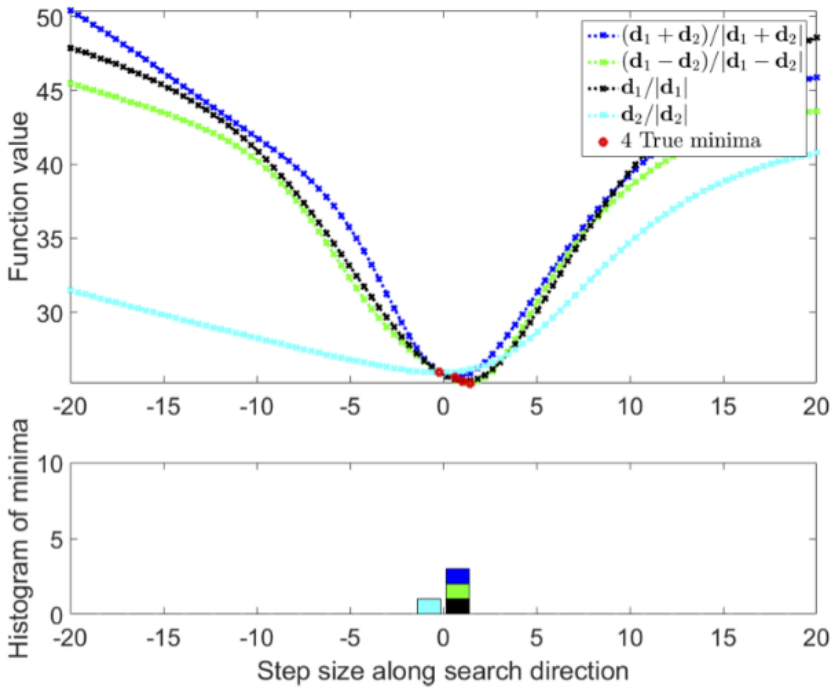
(e) Function value, $|\mathcal{B}| = 10$



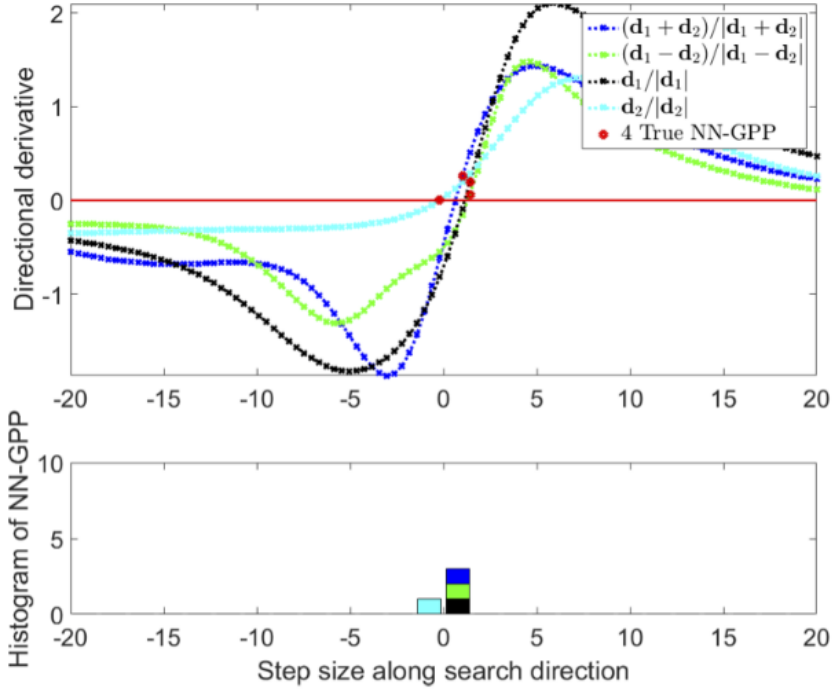
(f) Directional derivative, $|\mathcal{B}| = 10$

Directions

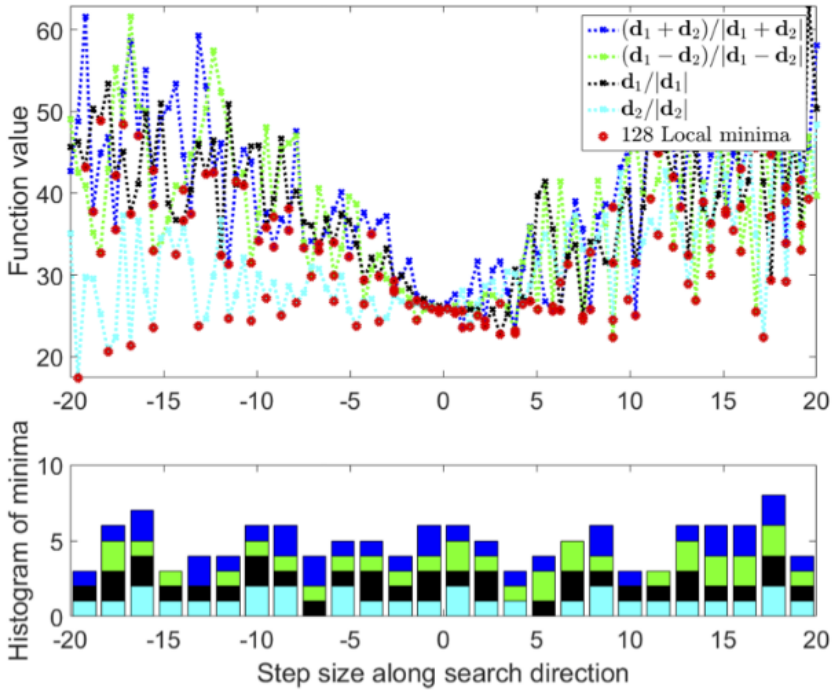

$\frac{\mathbf{d}_2}{|\mathbf{d}_2|}$
$\frac{\mathbf{d}_1+\mathbf{d}_2}{|\mathbf{d}_1+\mathbf{d}_2|}$
$\frac{\mathbf{d}_1}{|\mathbf{d}_1|}$
$\frac{\mathbf{d}_1-\mathbf{d}_2}{|\mathbf{d}_1-\mathbf{d}_2|}$


Function value
$(\mathbf{d}_1+\mathbf{d}_2)/|\mathbf{d}_1+\mathbf{d}_2|$
$(\mathbf{d}_1-\mathbf{d}_2)/|\mathbf{d}_1-\mathbf{d}_2|$
$\mathbf{d}_1/|\mathbf{d}_1|$
$\mathbf{d}_2/|\mathbf{d}_2|$
4 True minima
Histogram of minima
Step size along search direction

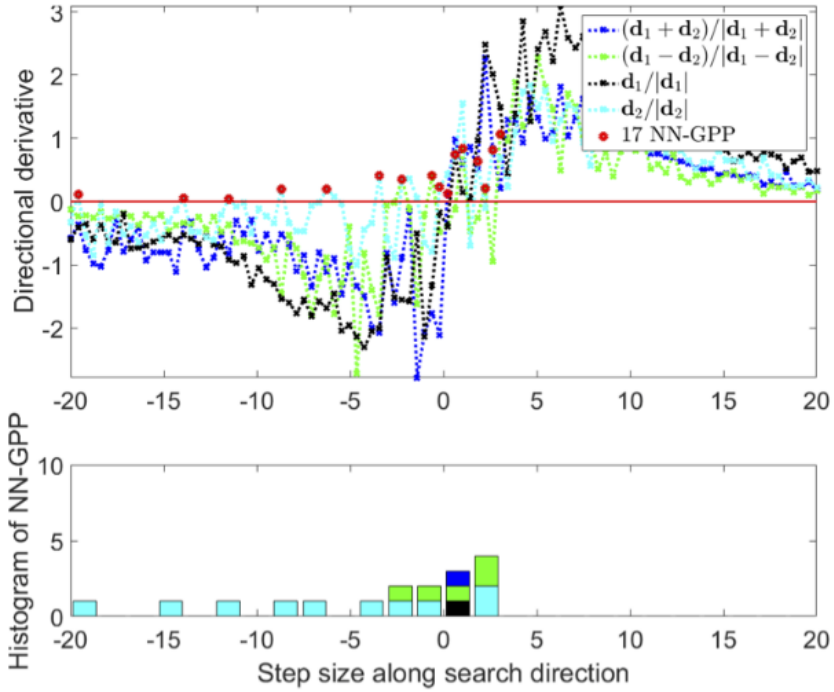
(a)


Directional derivative
$(\mathbf{d}_1+\mathbf{d}_2)/|\mathbf{d}_1+\mathbf{d}_2|$
$(\mathbf{d}_1-\mathbf{d}_2)/|\mathbf{d}_1-\mathbf{d}_2|$
$\mathbf{d}_1/|\mathbf{d}_1|$
$\mathbf{d}_2/|\mathbf{d}_2|$
4 True NN-GPP
Histogram of NN-GPP
Step size along search direction

(b)

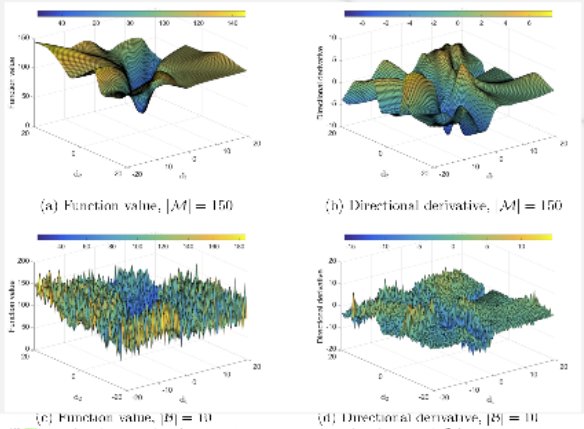

Function value
$(\mathbf{d}_1+\mathbf{d}_2)/|\mathbf{d}_1+\mathbf{d}_2|$
$(\mathbf{d}_1-\mathbf{d}_2)/|\mathbf{d}_1-\mathbf{d}_2|$
$\mathbf{d}_1/|\mathbf{d}_1|$
$\mathbf{d}_2/|\mathbf{d}_2|$
128 Local minima
Histogram of minima
Step size along search direction

(c)


Directional derivative
$(\mathbf{d}_1+\mathbf{d}_2)/|\mathbf{d}_1+\mathbf{d}_2|$
$(\mathbf{d}_1-\mathbf{d}_2)/|\mathbf{d}_1-\mathbf{d}_2|$
$\mathbf{d}_1/|\mathbf{d}_1|$
$\mathbf{d}_2/|\mathbf{d}_2|$
17 NN-GPP
Histogram of NN-GPP
Step size along search direction

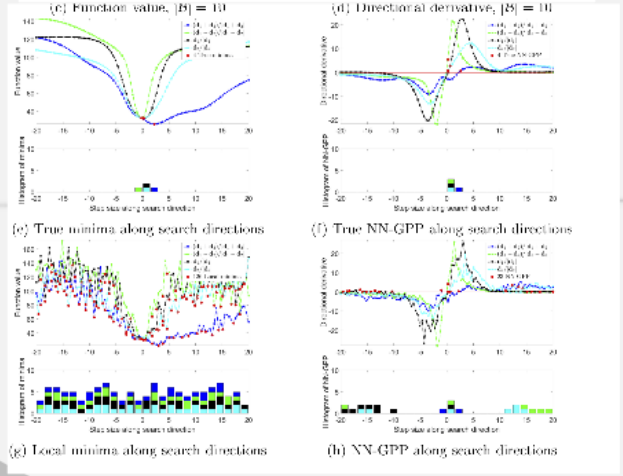
(d)

Role of Activation Functions TANH



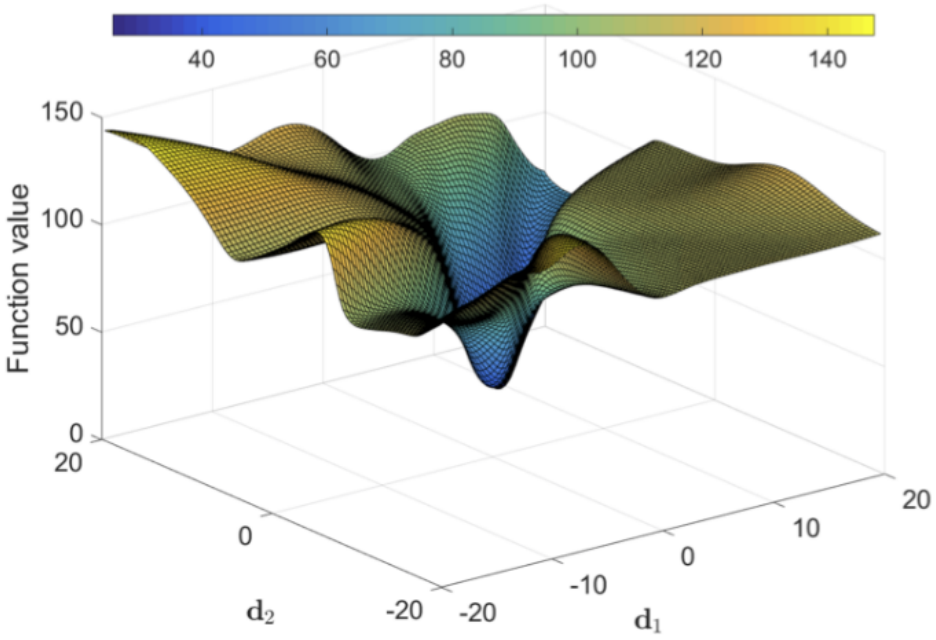
(a) Function value, $|M| = 150$ (b) Directional derivative, $|M| = 150$

(c) Function value, $|S| = 10$ (d) Directional derivative, $|S| = 10$

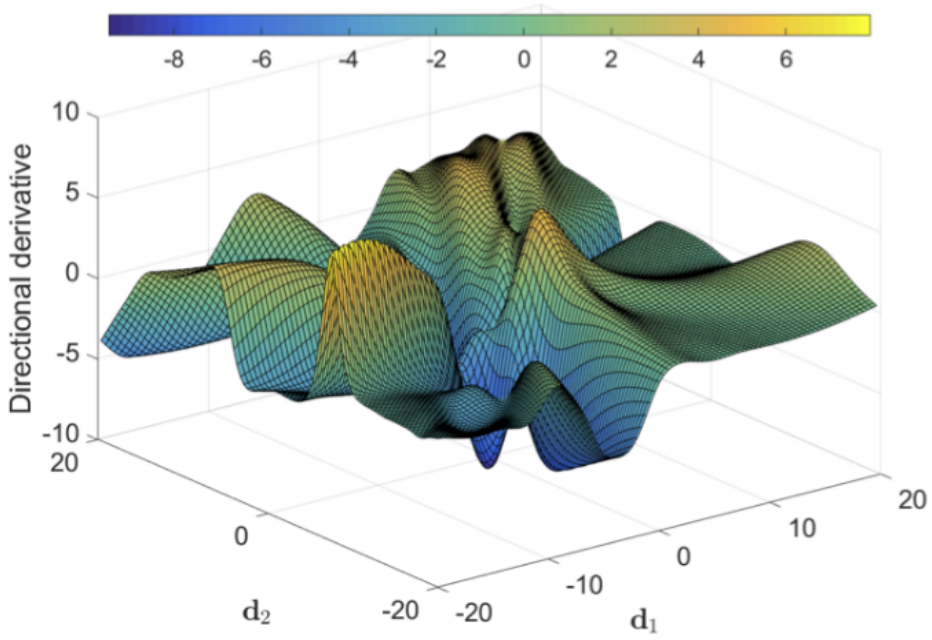


(e) True minima along search directions (f) True NN-GPP along search directions

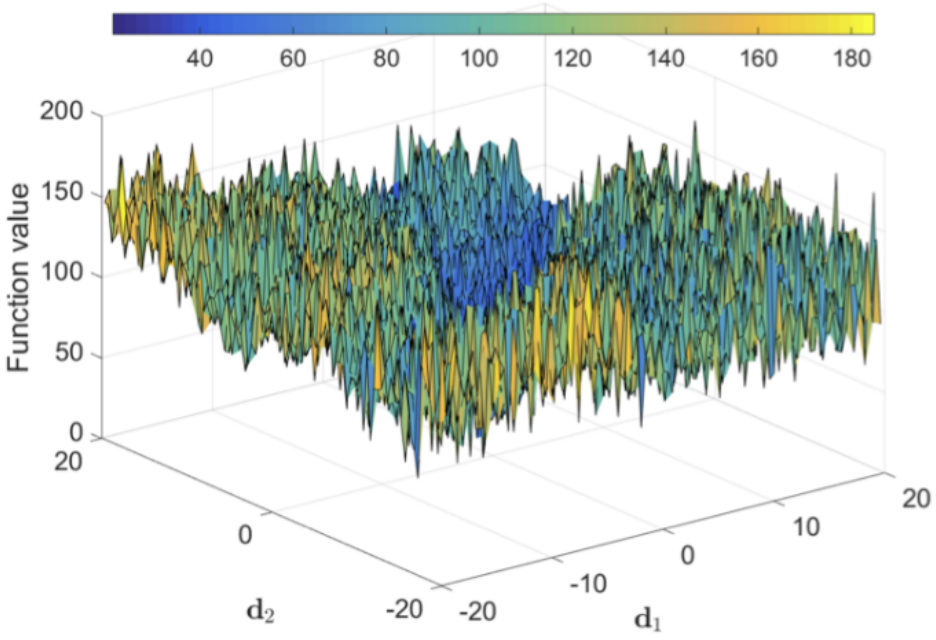
(g) Local minima along search directions (h) NN-GPP along search directions



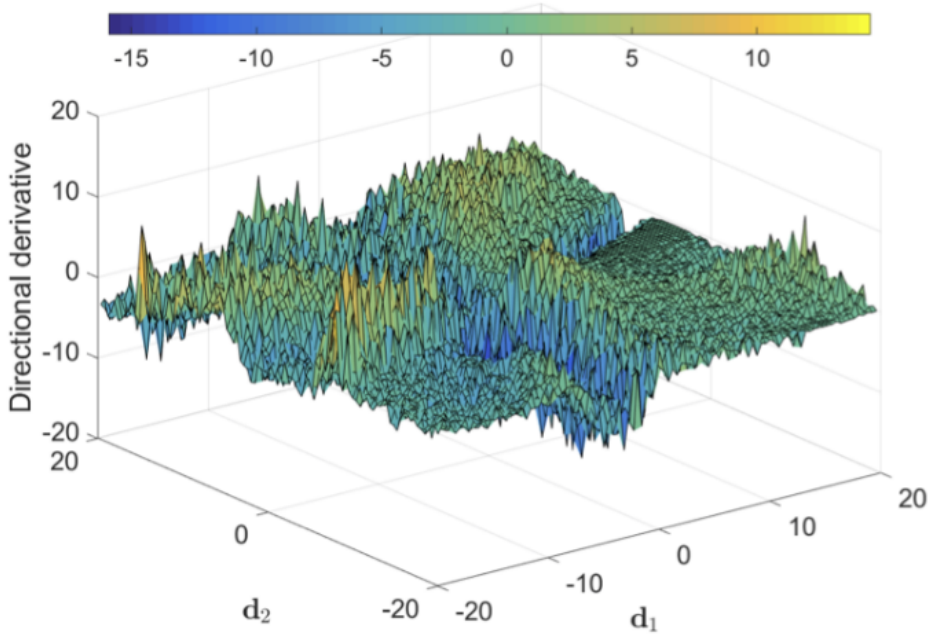
(a) Function value, $|\mathcal{M}| = 150$



(b) Directional derivative, $|\mathcal{M}| = 150$

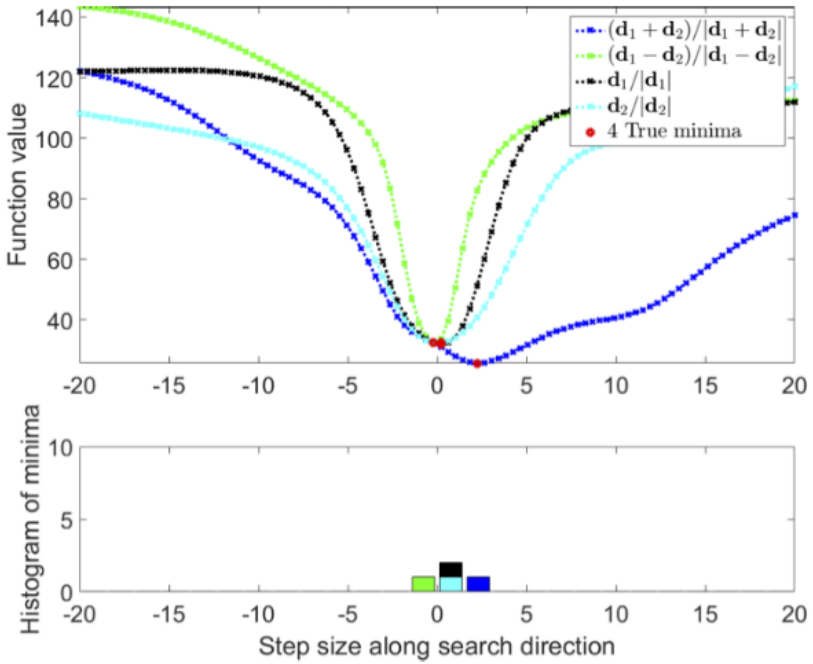


(c) Function value, $|\mathcal{B}| = 10$

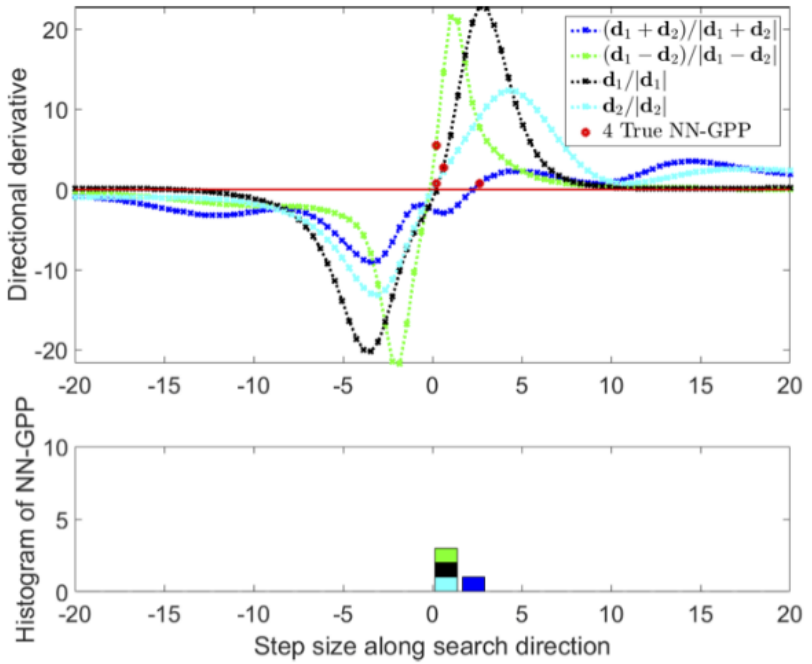


(d) Directional derivative, $|\mathcal{B}| = 10$

(c) Function value, $|\mathcal{B}| = 10$

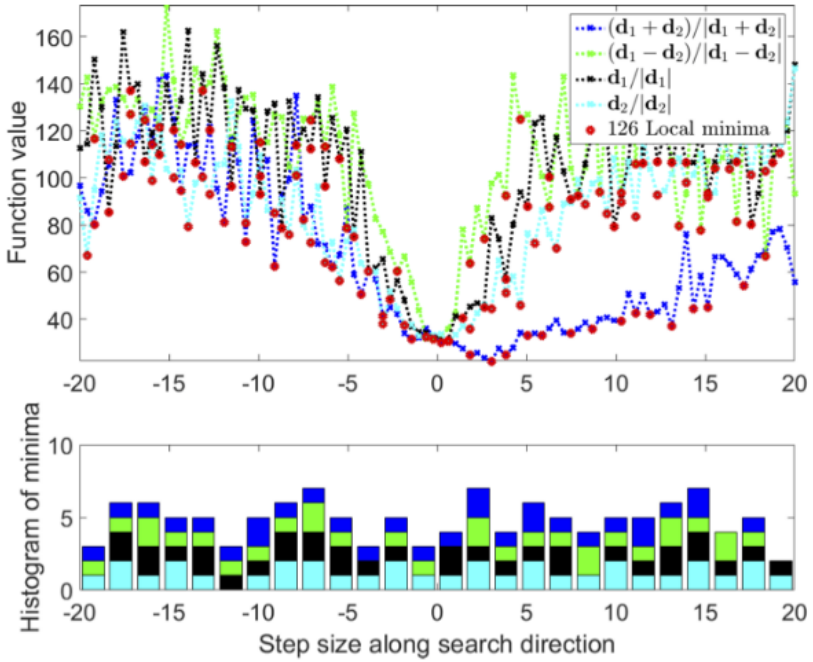


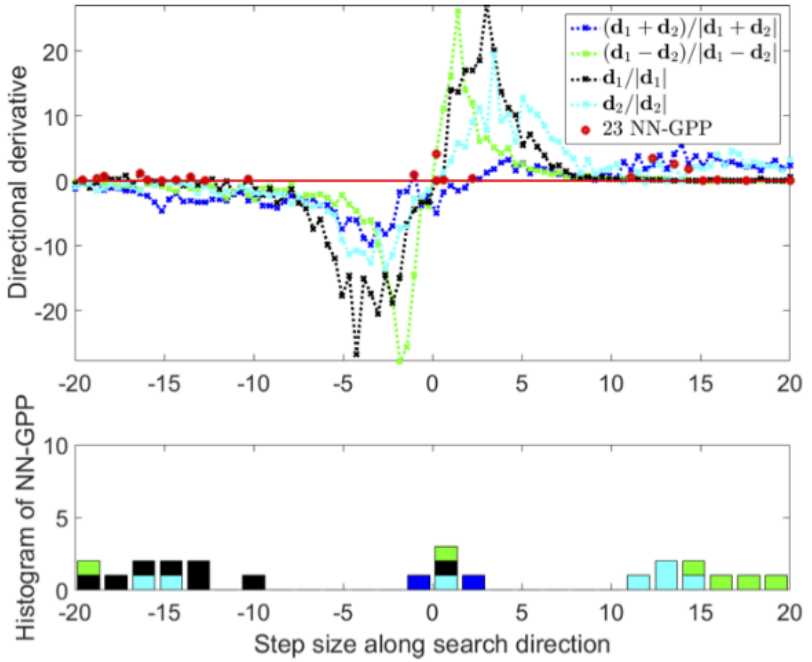
(d) Directional derivative, $|\mathcal{B}| = 10$



(e) True minima along search directions

(f) True NN-GPP along search directions

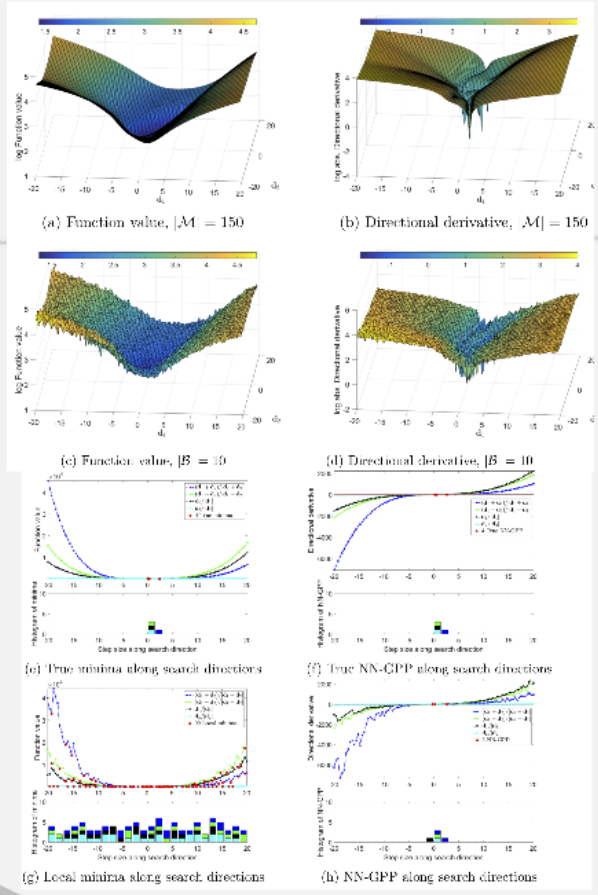




(g) Local minima along search directions

(h) NN-GPP along search directions

Role of Activation Functions ELU



(a) Function value, $|\mathcal{M}| = 150$

(b) Directional derivative, $|\mathcal{M}| = 150$

(c) Function value, $|\mathcal{B}| = 10$

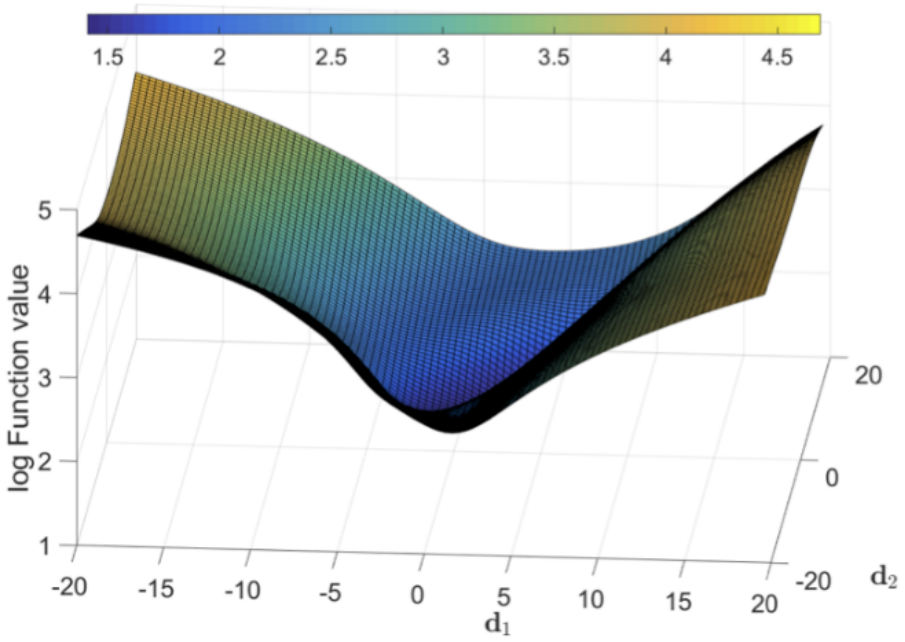
(d) Directional derivative, $|\mathcal{B}| = 10$

(e) True minima along search directions

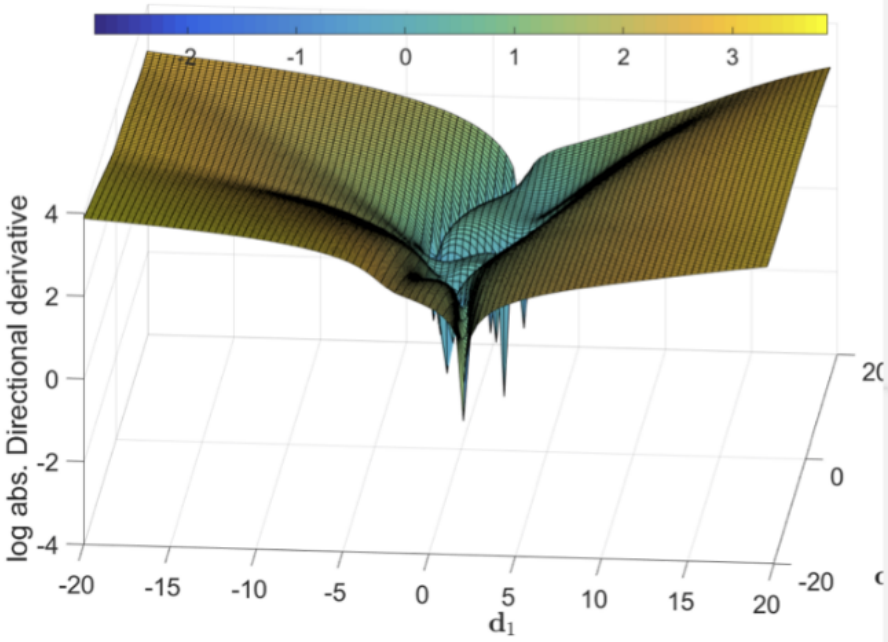
(f) True NN-GPP along search directions

(g) Local minima along search directions

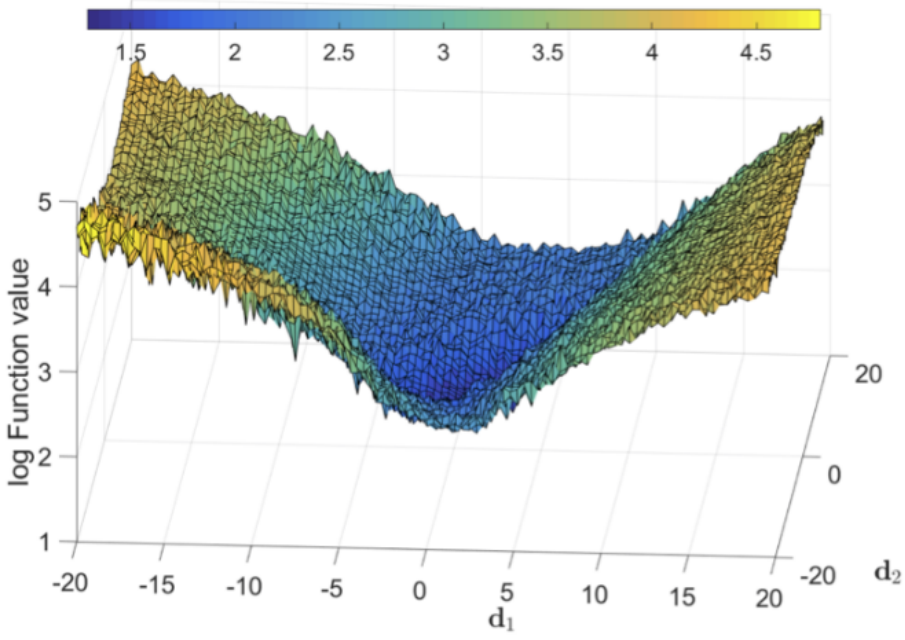
(h) NN-GPP along search directions



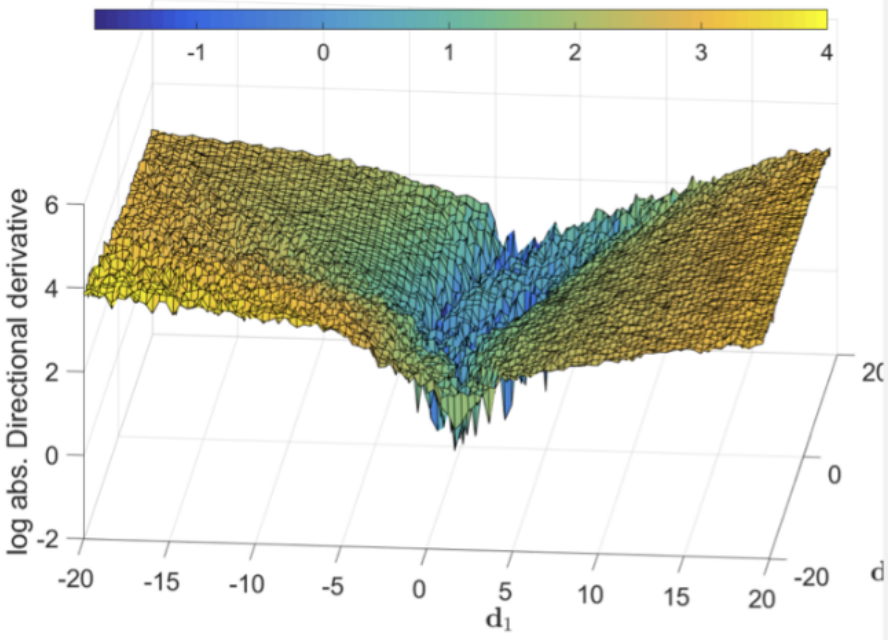
(a) Function value, $|\mathcal{M}| = 150$



(b) Directional derivative, $|\mathcal{M}| = 150$

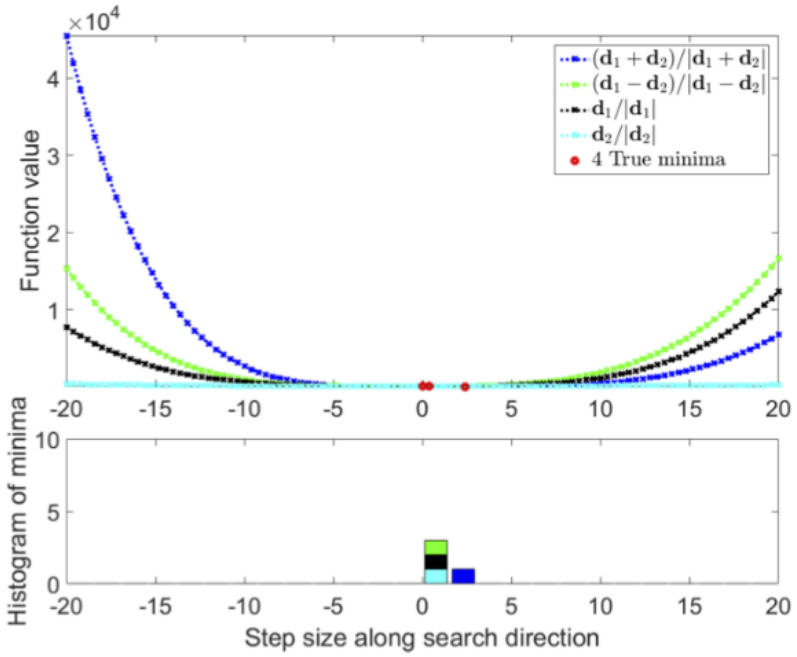


(c) Function value, $|\mathcal{B}| = 10$

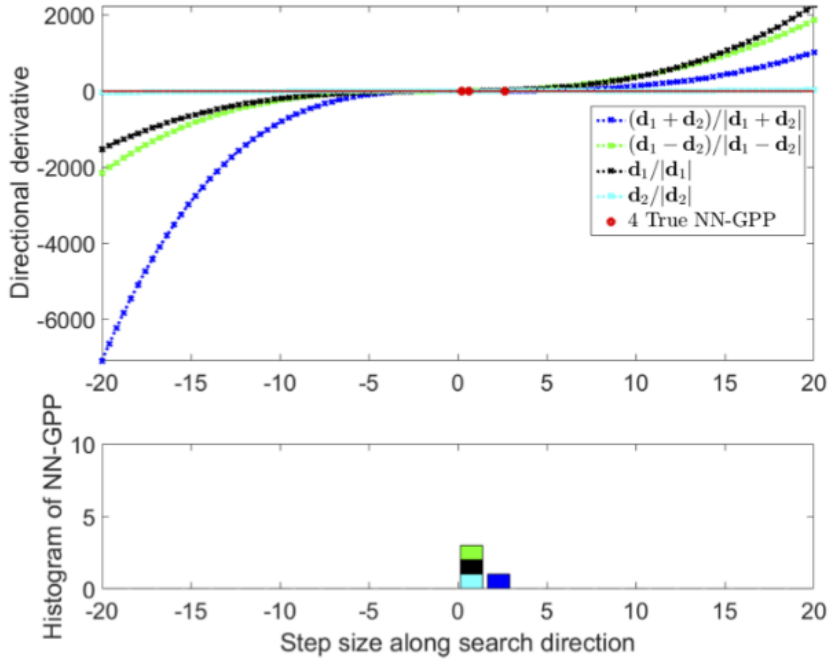


(d) Directional derivative, $|\mathcal{B}| = 10$

(c) Function value, $|\mathcal{B}| = 10$

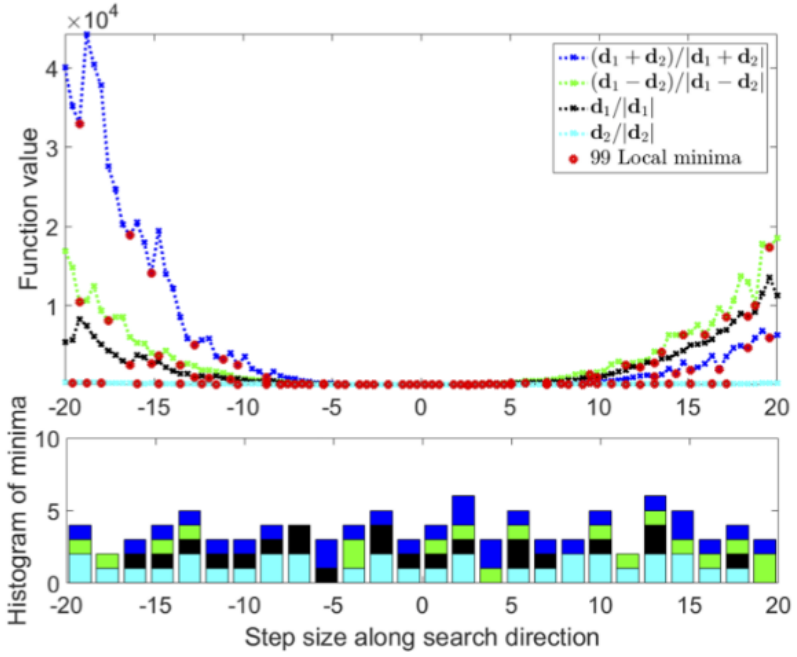


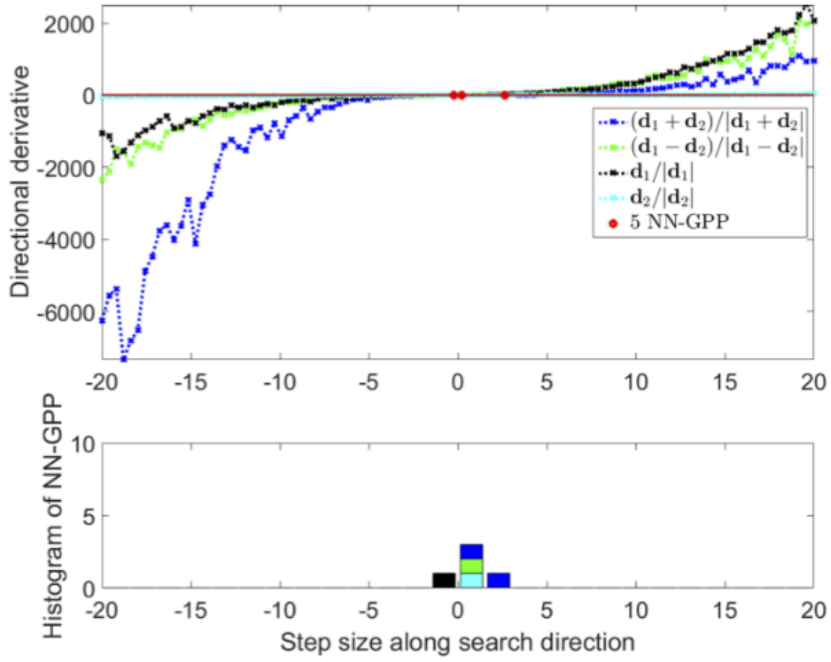
(d) Directional derivative, $|\mathcal{B}| = 10$



(e) True minima along search directions

(f) True NN-GPP along search directions



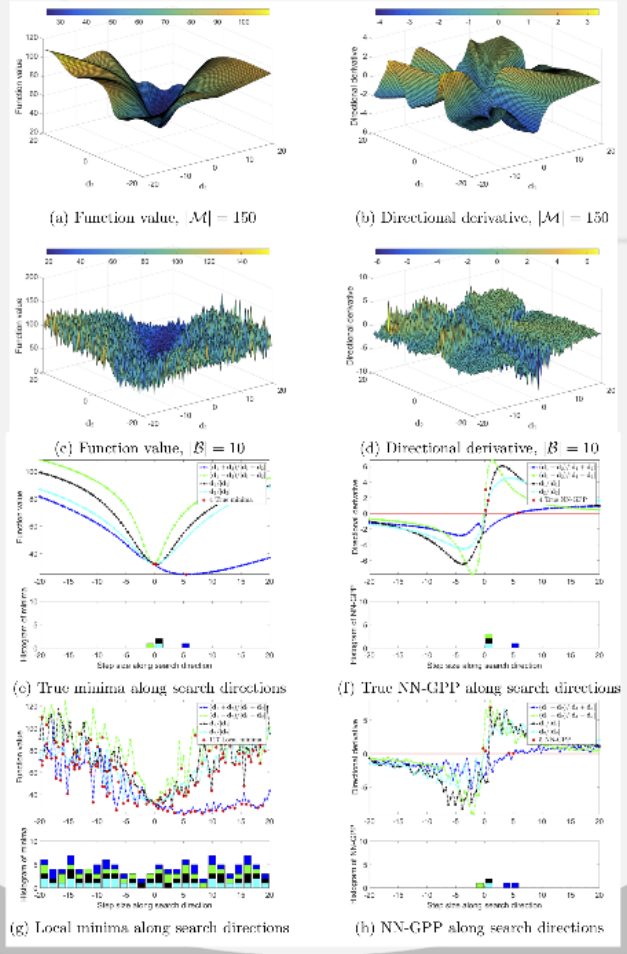


(g) Local minima along search directions

(h) NN-GPP along search directions

Role of Activation Functions

SOFTSIGN



(a) Function value, $|\mathcal{M}| = 150$

(b) Directional derivative, $|\mathcal{M}| = 150$

(c) Function value, $|\mathcal{B}| = 10$

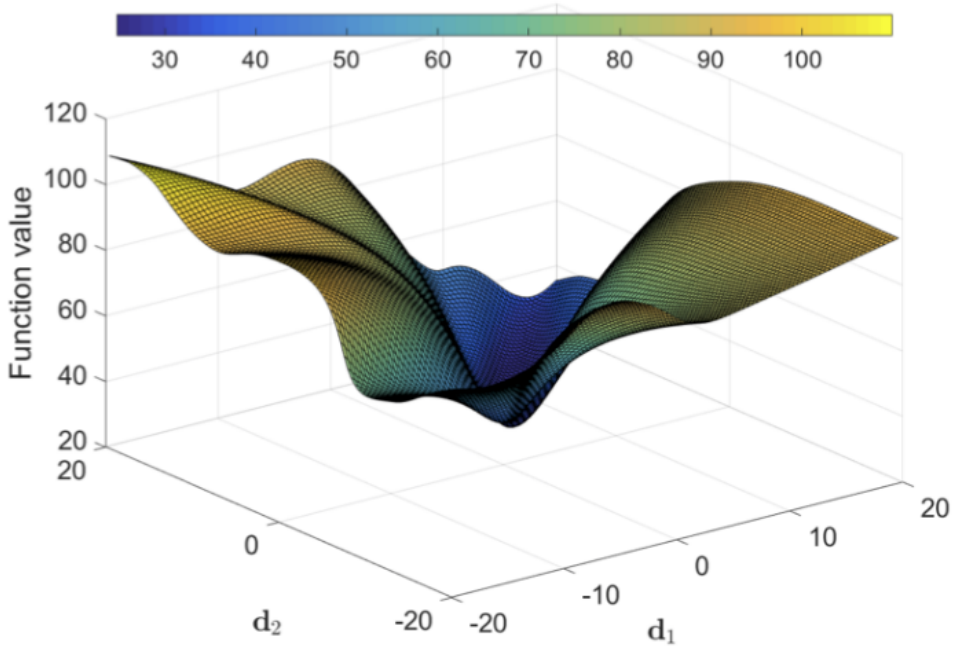
(d) Directional derivative, $|\mathcal{B}| = 10$

(e) True minima along search directions

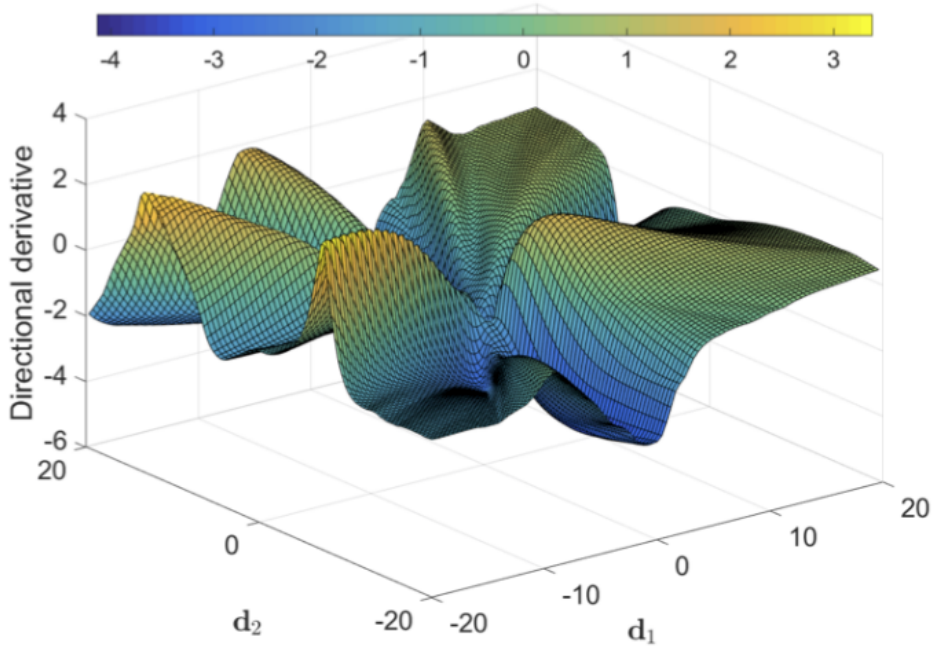
(f) True NN-GPP along search directions

(g) Local minima along search directions

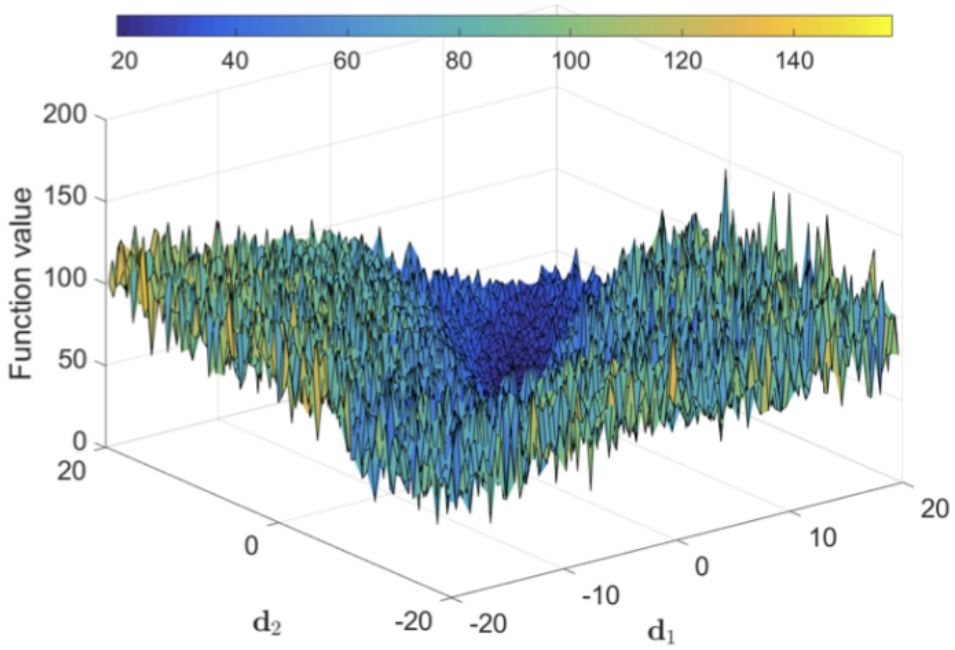
(h) NN-GPP along search directions



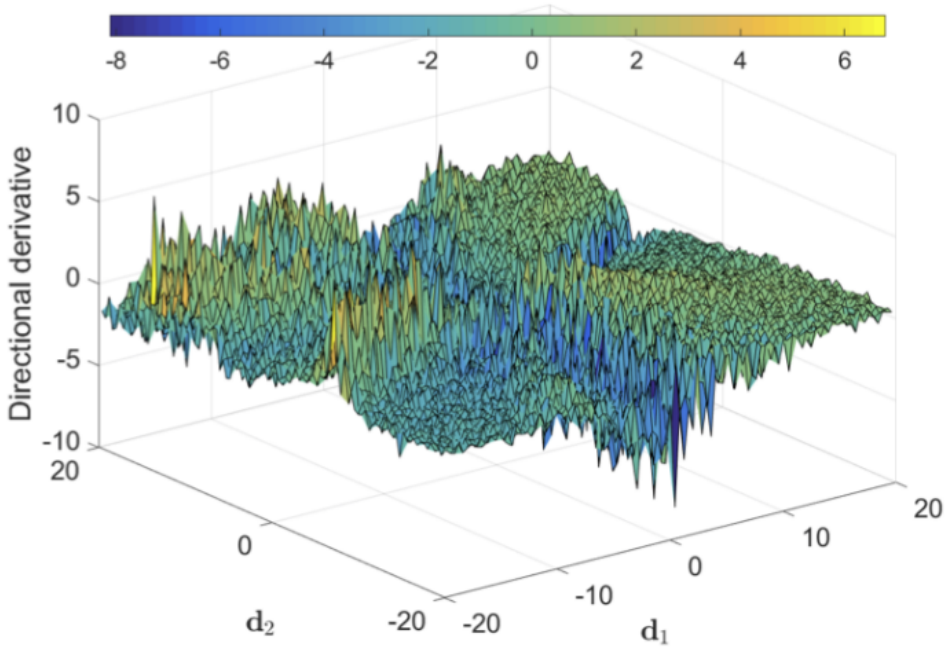
(a) Function value, $|\mathcal{M}| = 150$



(b) Directional derivative, $|\mathcal{M}| = 150$

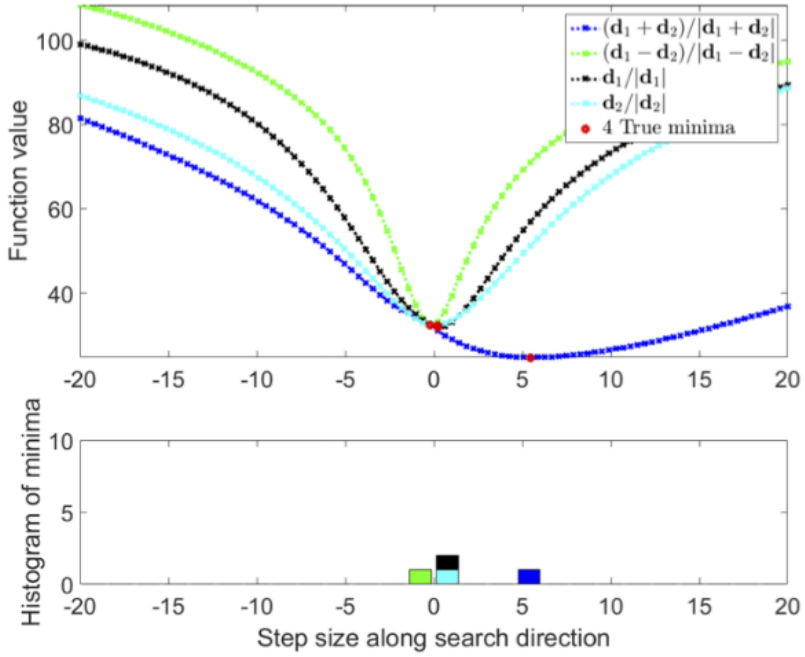


(c) Function value, $|\mathcal{M}| = 10$

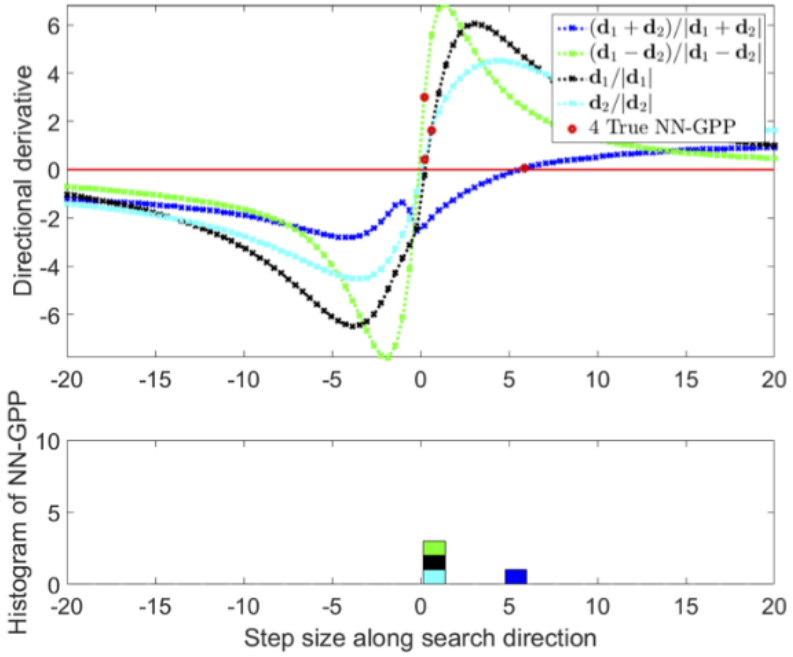


(d) Directional derivative, $|\mathcal{M}| = 10$

(c) Function value, $|\mathcal{B}| = 10$

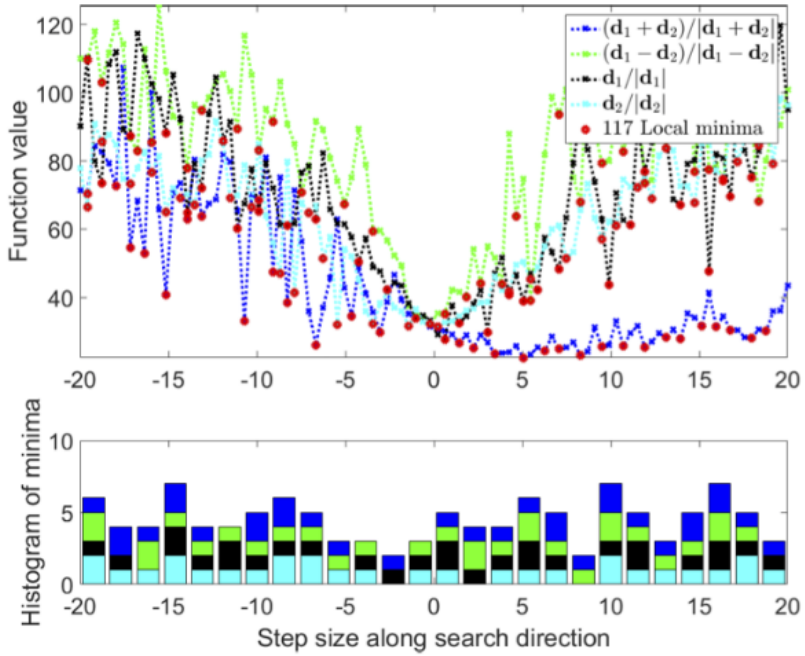


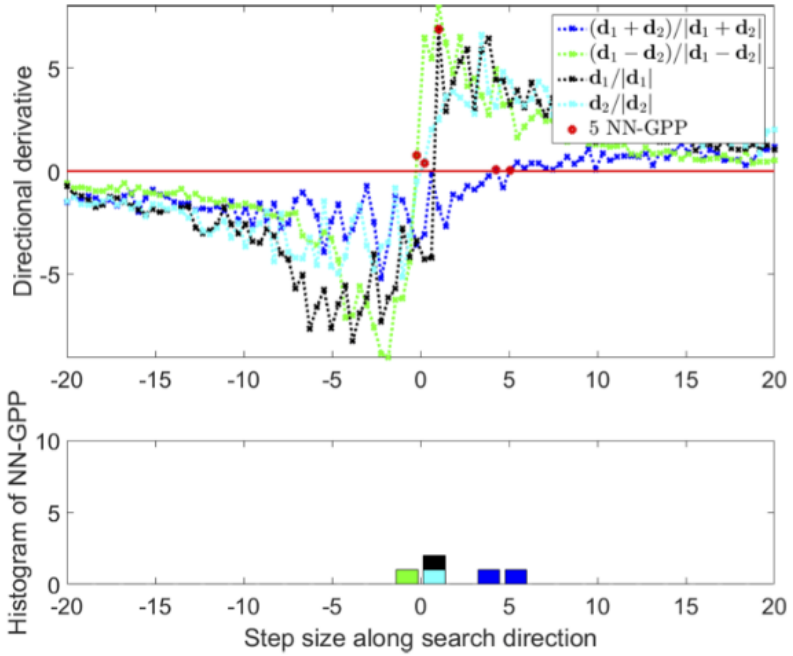
(d) Directional derivative, $|\mathcal{B}| = 10$



(e) True minima along search directions

(f) True NN-GPP along search directions



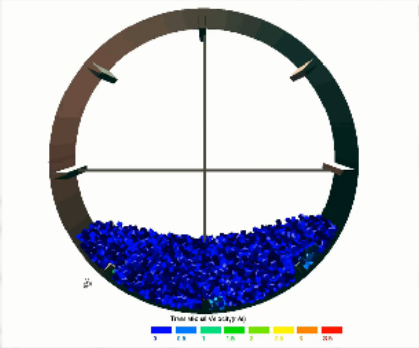


(g) Local minima along search directions

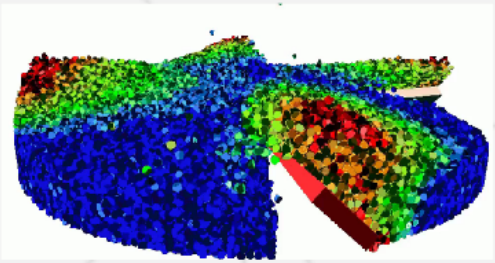
(h) NN-GPP along search directions

Actual Engineering Problems

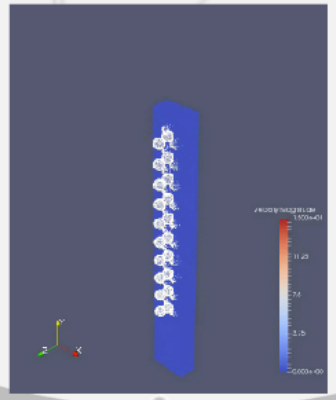
Comminution

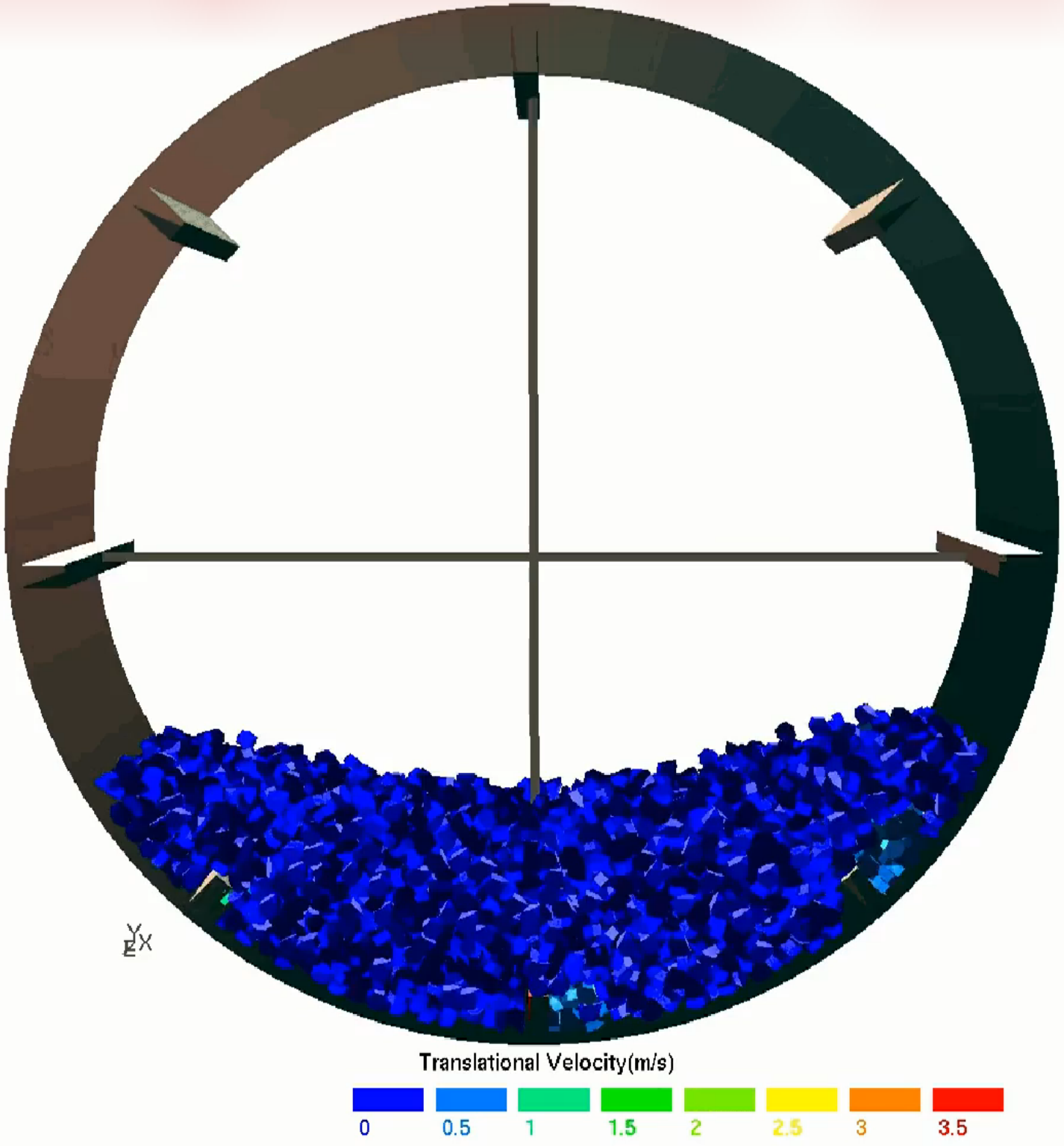


Mixing

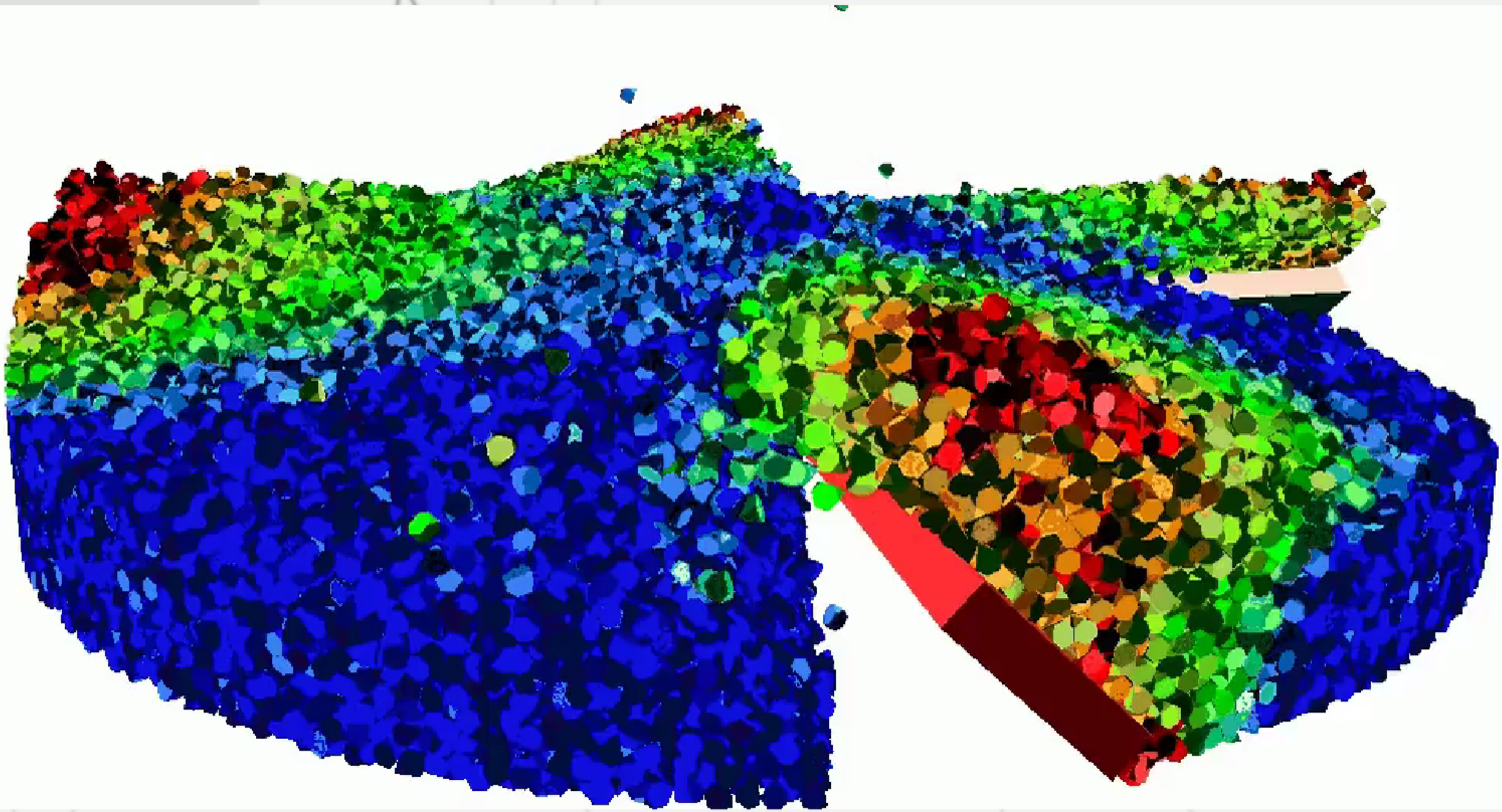


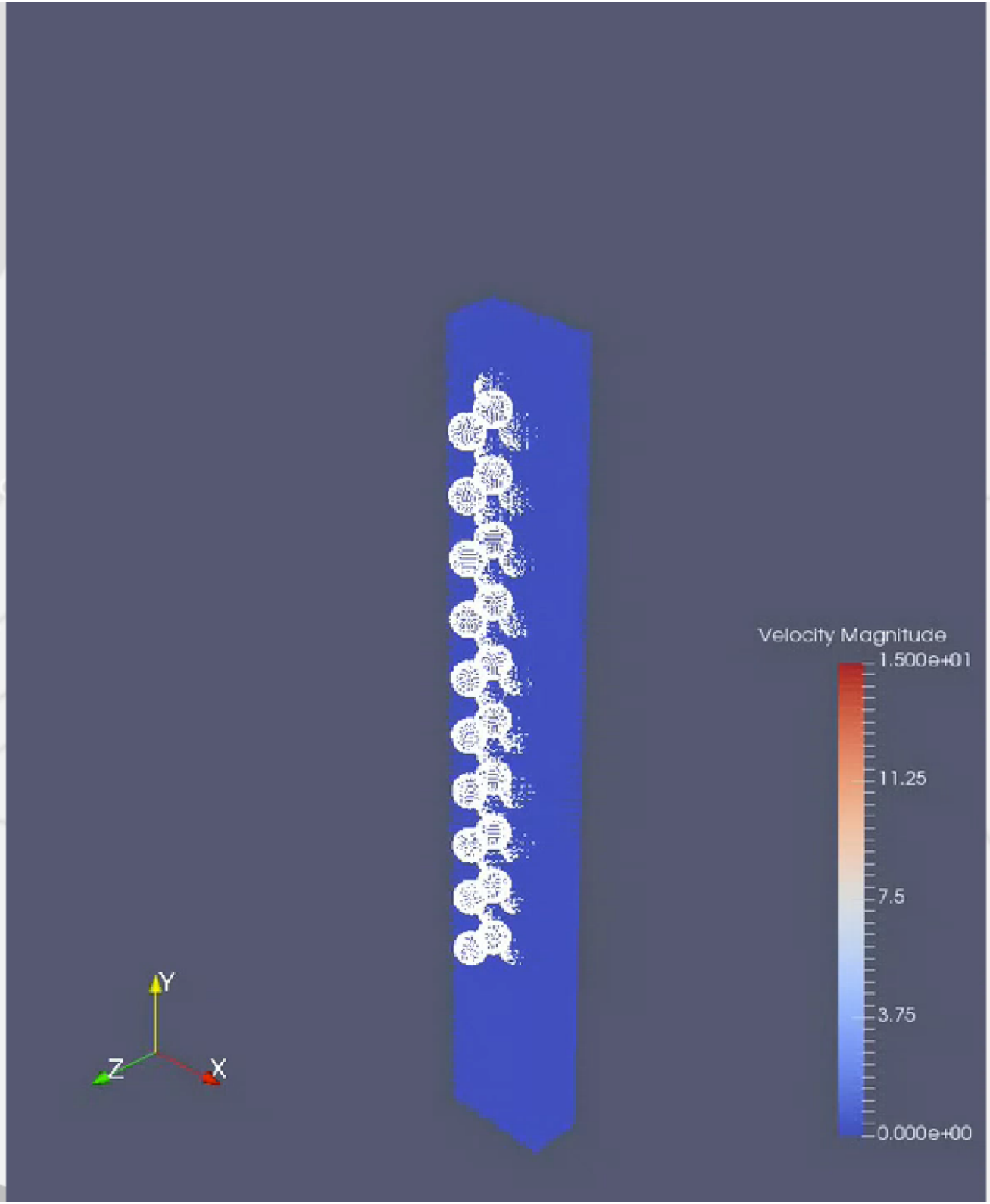
Settling




Y
X
Translational Velocity(m/s)
0
0.5
1
1.5
2
2.5
3
3.5

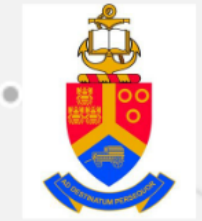
Mixing




Velocity Magnitude
1.500e+01
11.25
7.5
3.75
0.000e+00
Y
Z
X

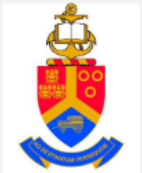
Simulation
BlazeDEM-GPU

Developed in conjunction with the CSIR
and UP



GPU Based Particle Solver
Discrete Element Method (Solid Particles)
Smoothed Particle Hydrodynamics (Fluid)

Problem Data Files GB - TB
What do we learn?

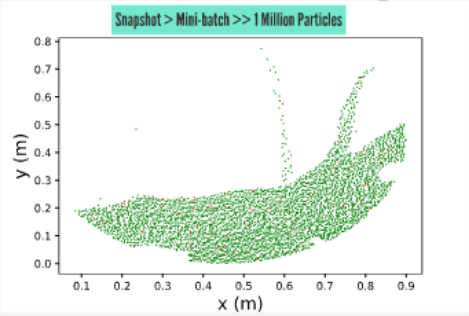


Simulation can take Days/Weeks/Months/Years What do we learn?

Mill breaks solids into pieces by grinding, crushing and cutting.

Milling harsh environments - cannot look inside.

Sentinal Copper Mine, Zambia

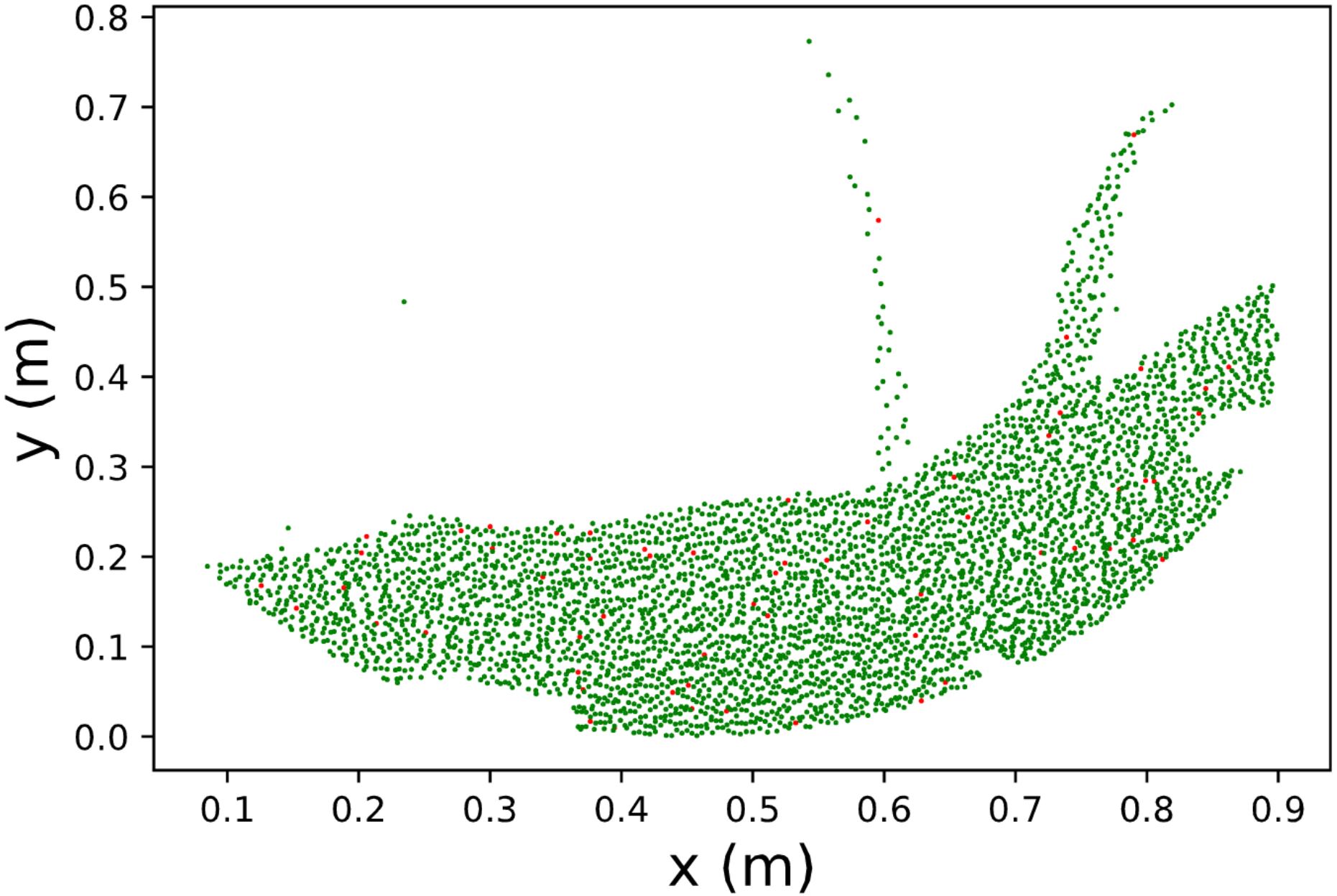


Siemens Gearless Mill Drives

High availability, lower operational costs, best performance.

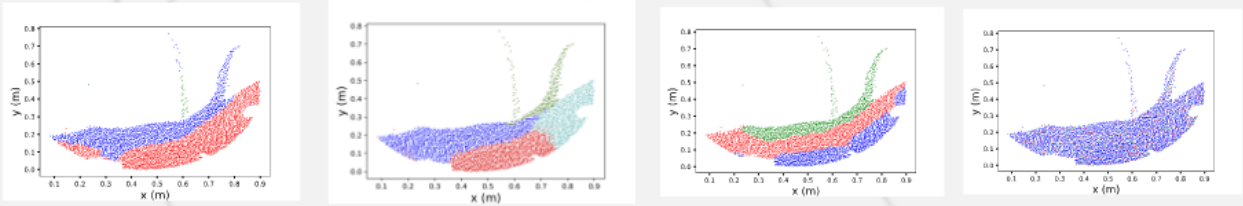
Sentinal Copper Mine, Zambia

Snapshot > Mini-batch >> 1 Million Particles

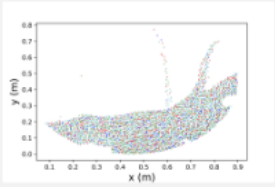


Typical Problem
Weeks of Simulations
What do we learn?

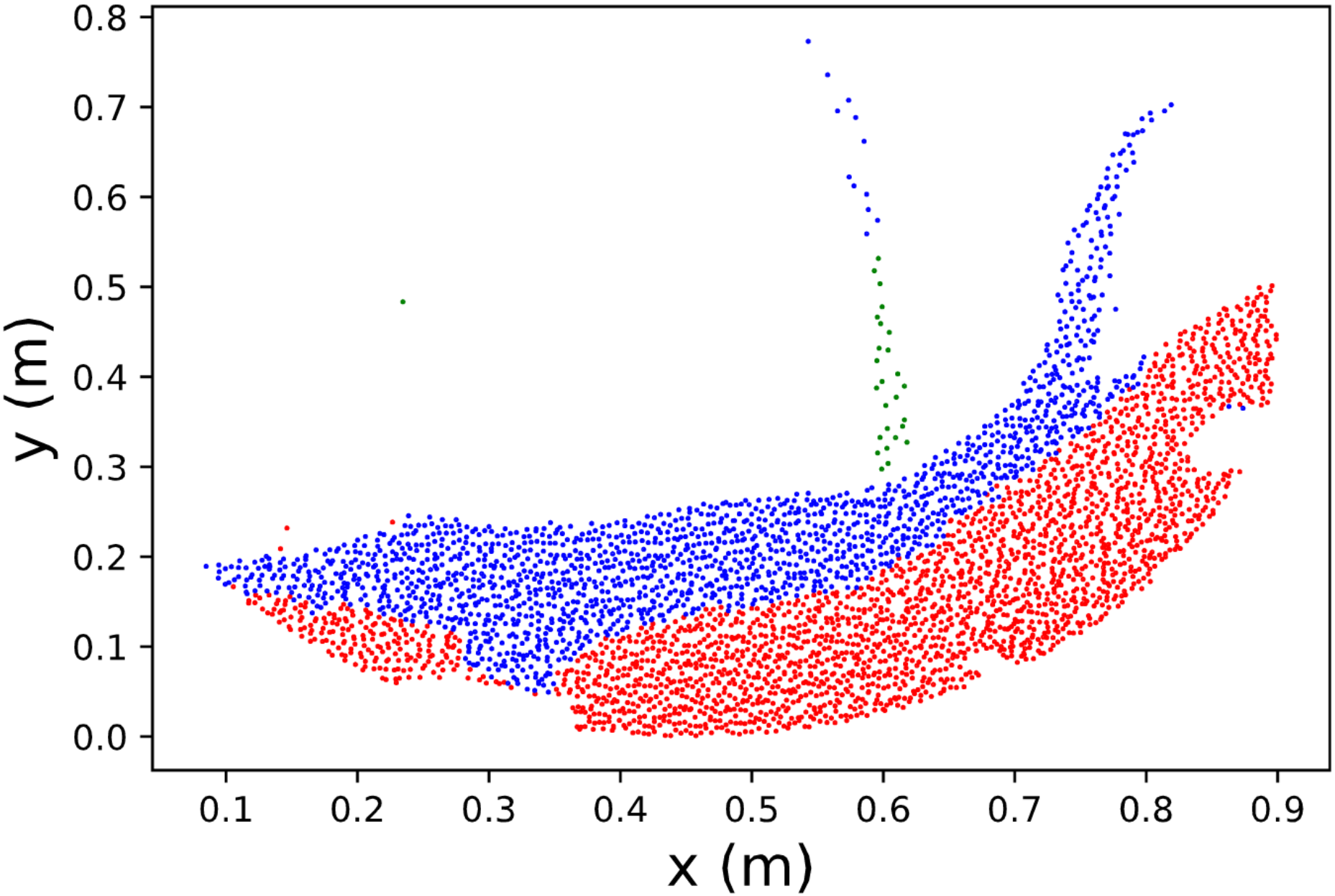
Use Statistical learning, Machine Learning and Deep Learning to help engineers understand

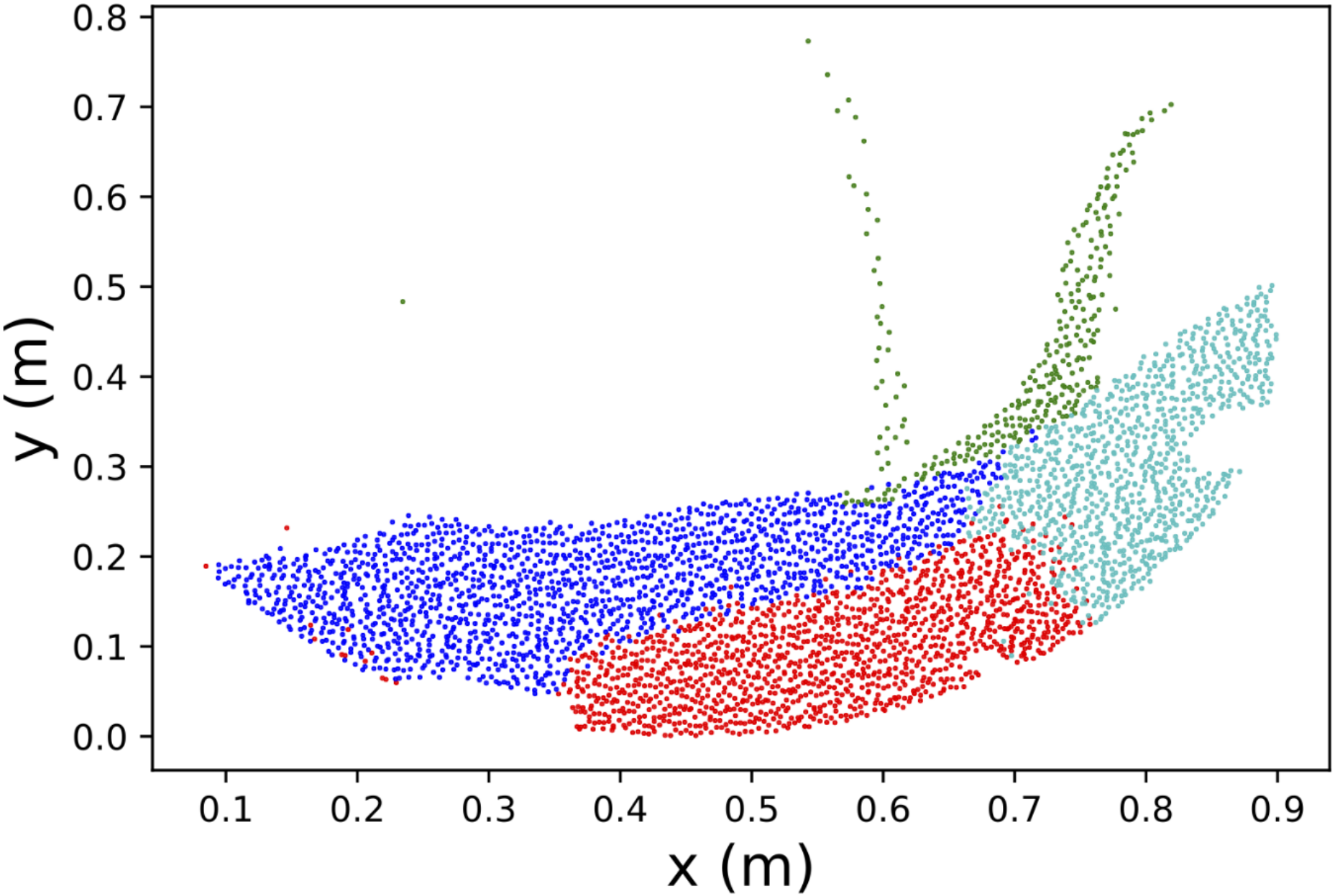


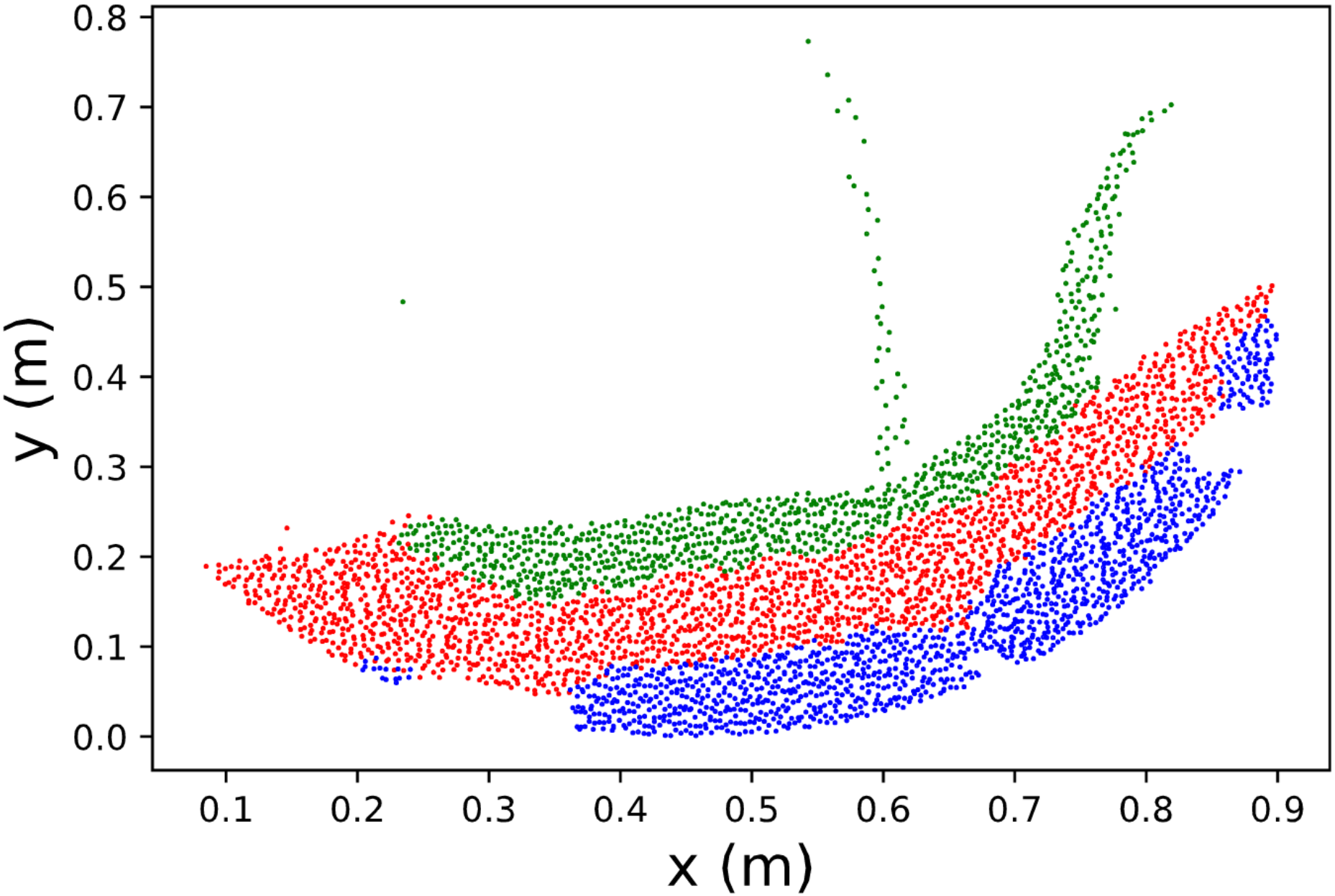
as opposed to the engineer figuring it out

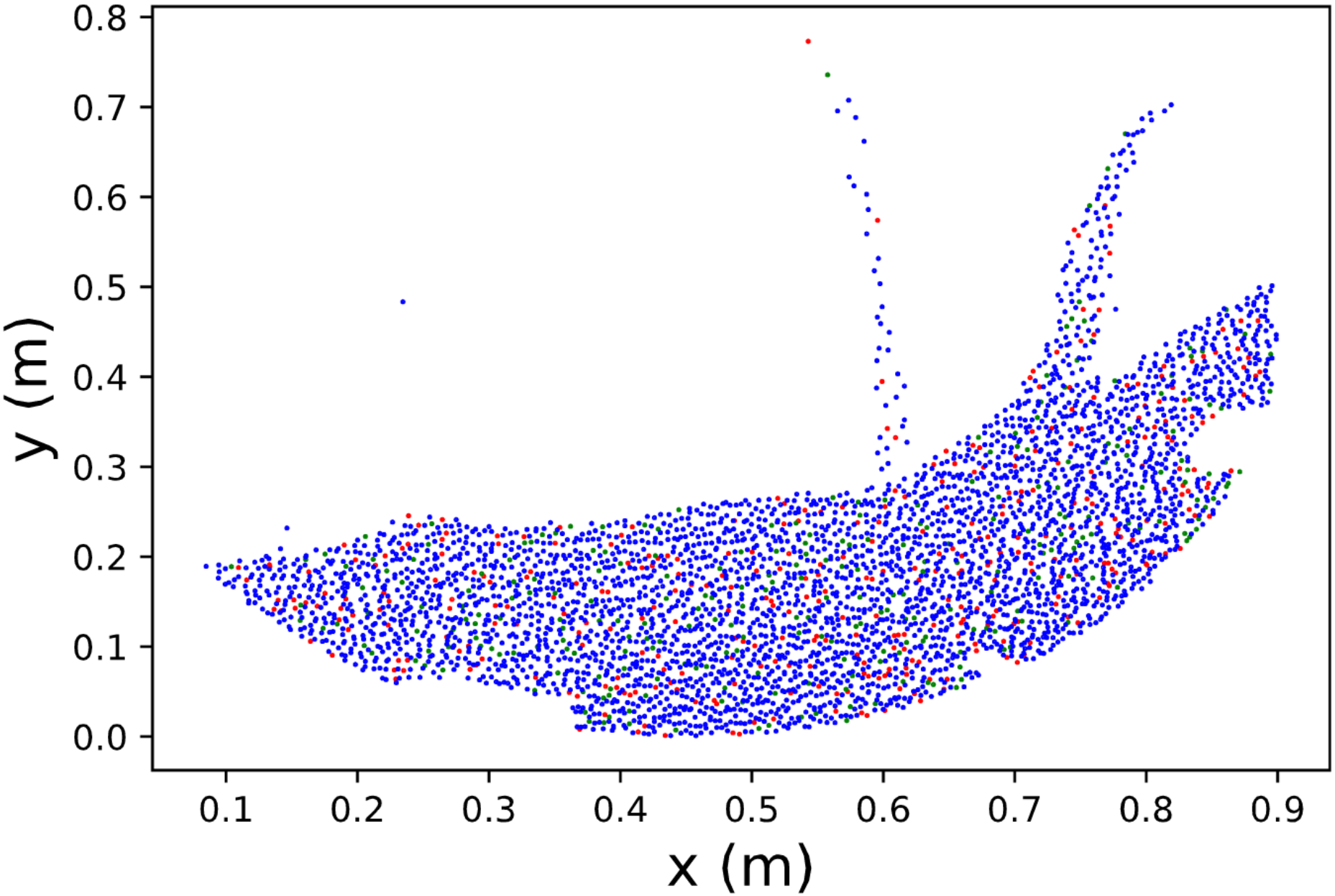


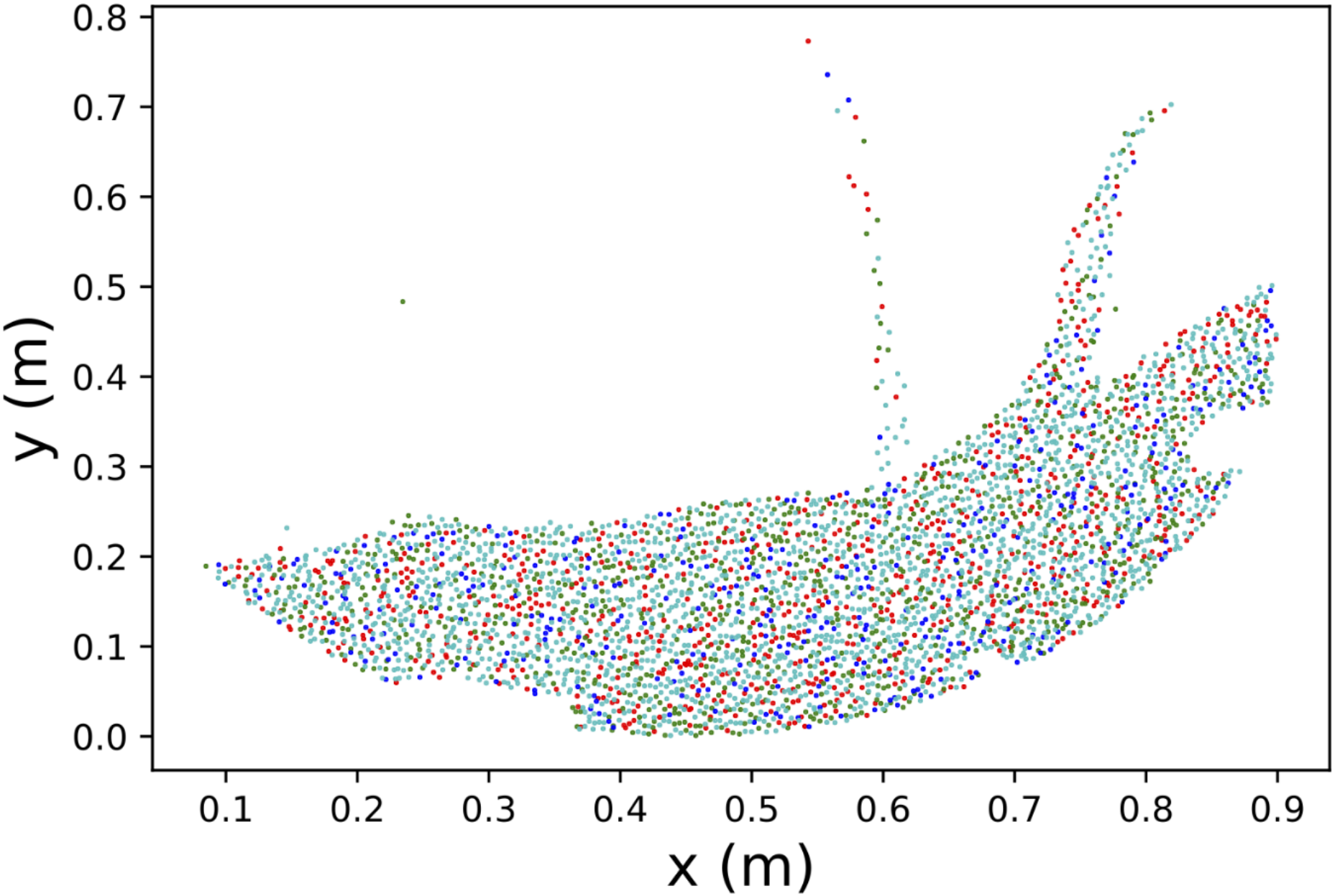
Dataset will be released with paper


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
y (m)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
x (m)


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
y (m)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
x (m)


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
y (m)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
x (m)


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
y (m)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
x (m)


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
y (m)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
x (m)

Automating Learning Rate

One step closer towards automating the processing of simulation data for engineers

Research compare idea with optimal learning rate against another idea with optimal learning rate

Take home:

Find minima in the derivative world

NN-GPP

Papers, Pytorch,TF Code Soon

gorgthelab.github.io

GORG THE LAB

gorgthelab.github.io

Dominic Kafka (PhD)

Younghwan Chae (PhD)

GORG THE LAB

gorgthelab.github.io

Dem

BLAZEDEM-GPU
THE TEAM

Primary Developer

UNIVERSITY OF SURREY

IMT Lille Douai
École Mines-Télécom
IMT-Université de Lille

Prof. Raj Rajamani, Utah University, USA

Prof. Ugur Tuzun, Cantab, UK

Dr. Patrick Pizette, IMT Lille Douai, France

Prof. Nor-Edine Abriak, IMT Lille Douai, France

Prof. Charley Wu, University of Surrey, UK

Prof. Wenjie Xu, University of Tsinghua, China

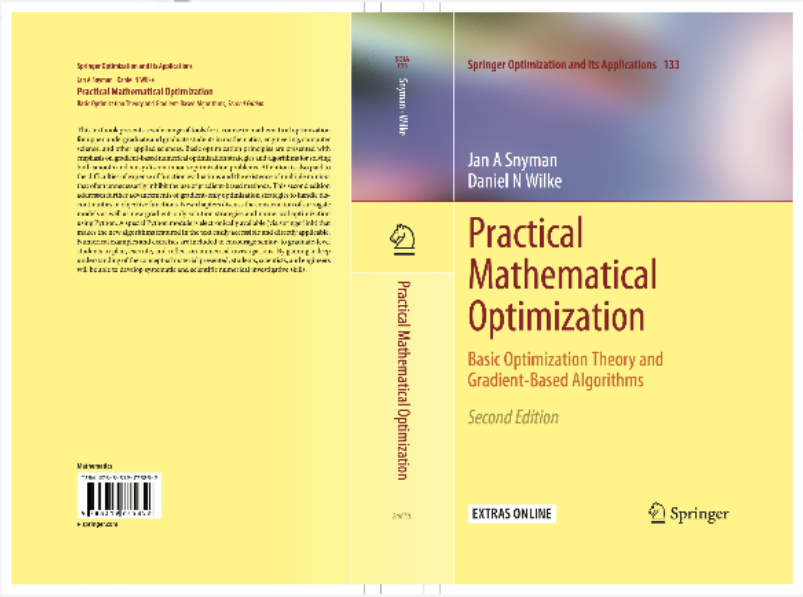
Line search
Function minim

Do you want to learn more about

Gradient-only Optimization

Non-Negative Projection Points

Line Searches for Discontinuous Functions



Only textbook (2018)
that covers
this material

Springer Optimization and Its Applications

Jan A Snyman · Daniel N Wilke

Practical Mathematical Optimization

Basic Optimization Theory and Gradient-Based Algorithms, *Second Edition*

This textbook presents a wide range of tools for a course in mathematical optimization for upper undergraduate and graduate students in mathematics, engineering, computer science, and other applied sciences. Basic optimization principles are presented with emphasis on gradient-based numerical optimization strategies and algorithms for solving both smooth and noisy discontinuous optimization problems. Attention is also paid to the difficulties of expense of function evaluations and the existence of multiple minima that often unnecessarily inhibit the use of gradient-based methods. This second edition addresses further advancements of gradient-only optimization strategies to handle discontinuities in objective functions. New chapters discuss the construction of surrogate models as well as new gradient-only solution strategies and numerical optimization using Python. A special Python module is electronically available (via springerlink) that makes the new algorithms featured in the text easily accessible and directly applicable. Numerical examples and exercises are included to encourage senior- to graduate-level students to plan, execute, and reflect on numerical investigations. By gaining a deep understanding of the conceptual material presented, students, scientists, and engineers will be able to develop systematic and scientific numerical investigative skills.

Mathematics

ISBN 978-3-319-77585-2

▸ springer.com

SOIA 133

Snyman · Wilke

Practical Mathematical Optimization

2nd Ed.

Springer Optimization and Its Applications 133

Jan A Snyman
Daniel N Wilke

Practical Mathematical Optimization

Basic Optimization Theory and Gradient-Based Algorithms

Second Edition

EXTRAS ONLINE

Springer