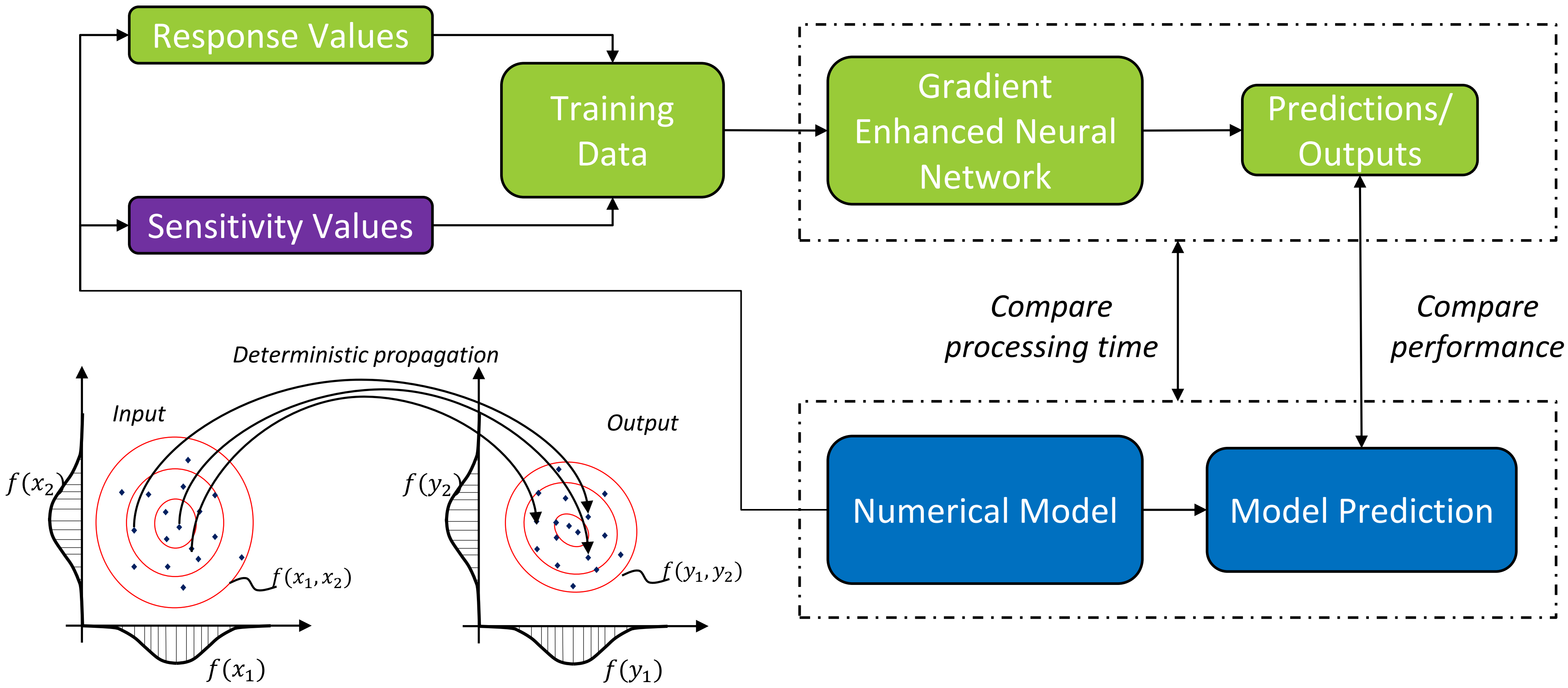
A Sobolev neural network with adaptive residual weighting scheme as a surrogate for computational mechanics

The financial support provided under project-ID 278867966, TRR188 by the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged.

M.G.R. Faes
M. A. Valdebenito

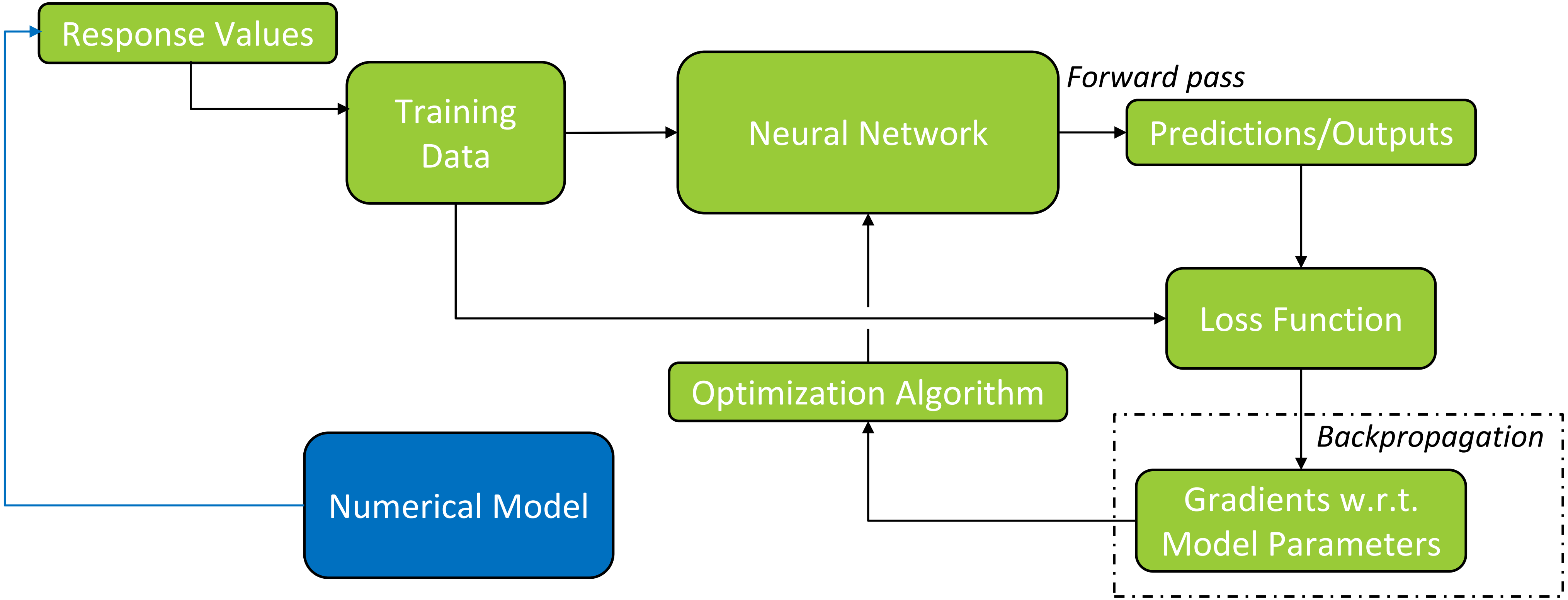
Presented by *Ali O.M. Kilicsoy*,
Chair for Reliability Engineering at TU Dortmund

General process



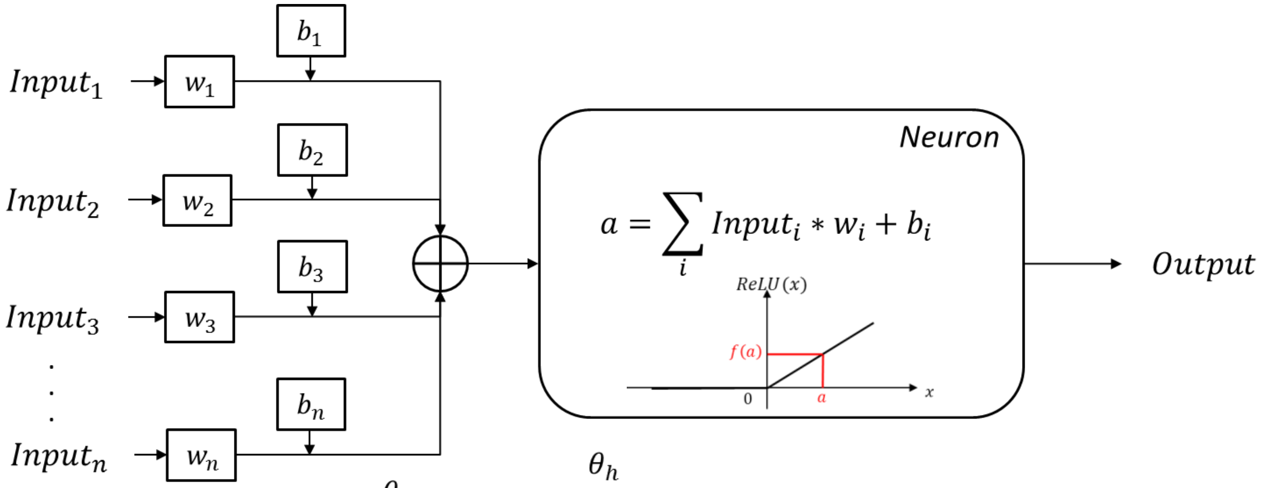
Neural network surrogate

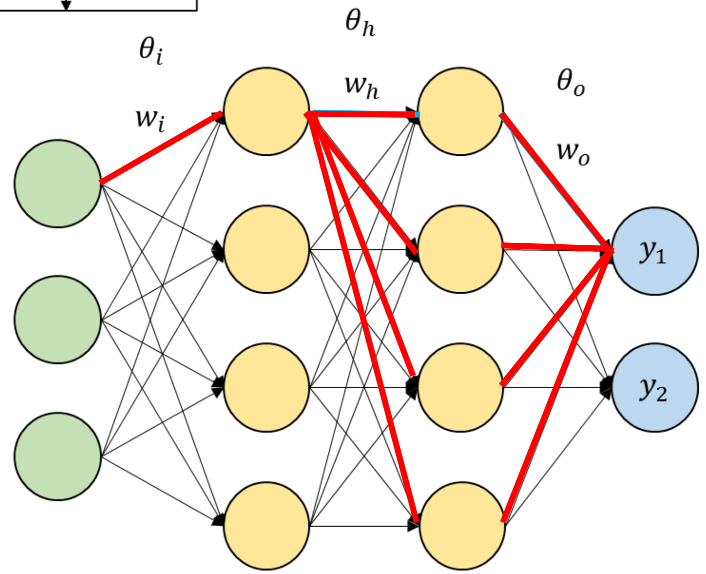
- General NN



Neural network surrogate

- General NN





Model Parameters $\boldsymbol{\theta} = \{\boldsymbol{w}, \boldsymbol{b}\}$

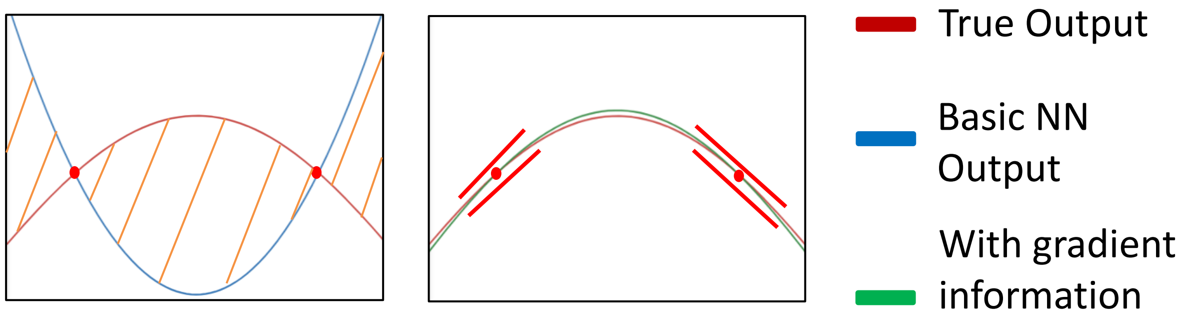
Loss Function $L = \frac{1}{2}\sum(\hat{\boldsymbol{y}} - \boldsymbol{y})^2$

Automatic Differentiation – Chain Rule

$$\frac{\partial y_1}{\partial w_h} = \frac{\partial y_1}{\partial \theta_o}\frac{\partial \theta_o}{\partial w_h}$$

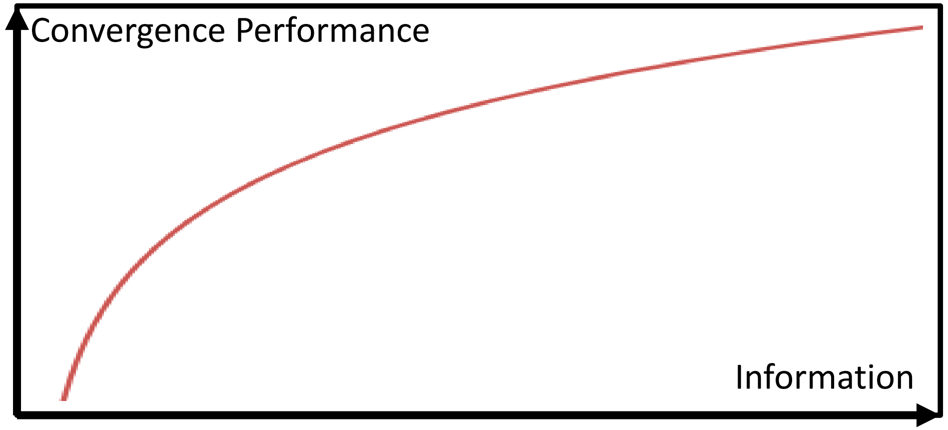
$$\frac{\partial y_1}{\partial w_i} = \frac{\partial y_1}{\partial \theta_o}\frac{\partial \theta_o}{\partial \theta_h}\frac{\partial \theta_h}{\partial w_i}$$

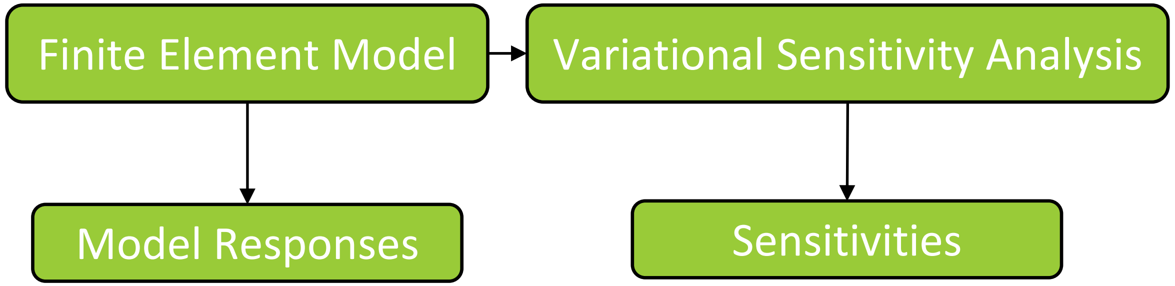
Response and Sensitivity



Basic NN Approximation

With gradient information

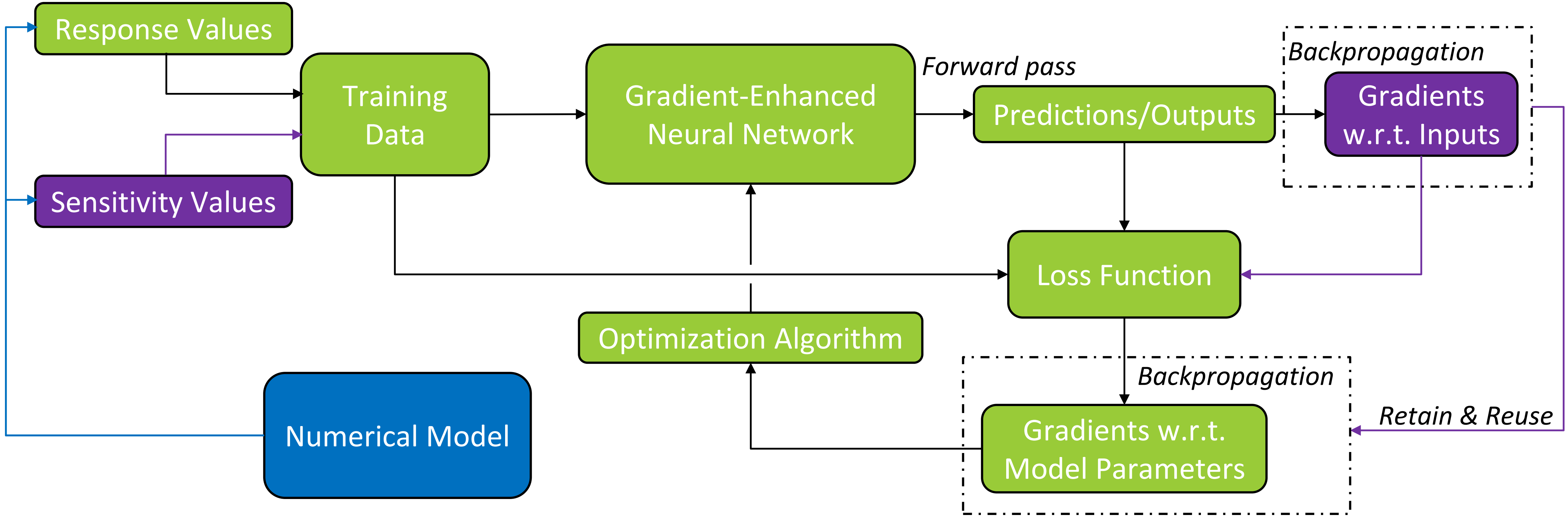




- Sensitivity information decisively influences training convergence
- Sensitivity information is available for low cost (computational, economical, etc.)

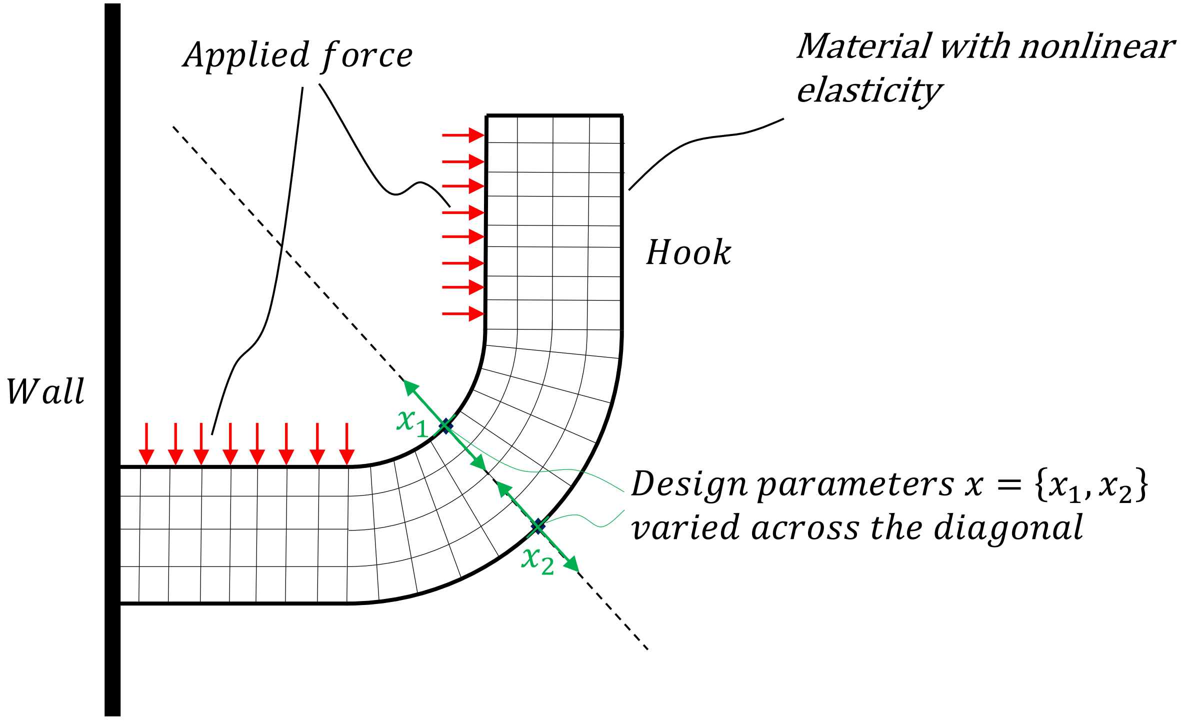
Response and Sensitivity

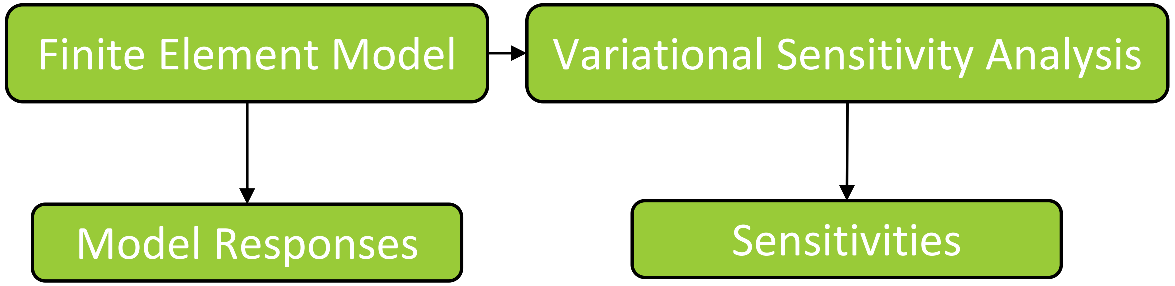
- Sobolev training



Response and Sensitivity

- Finite Element Model – Use Case

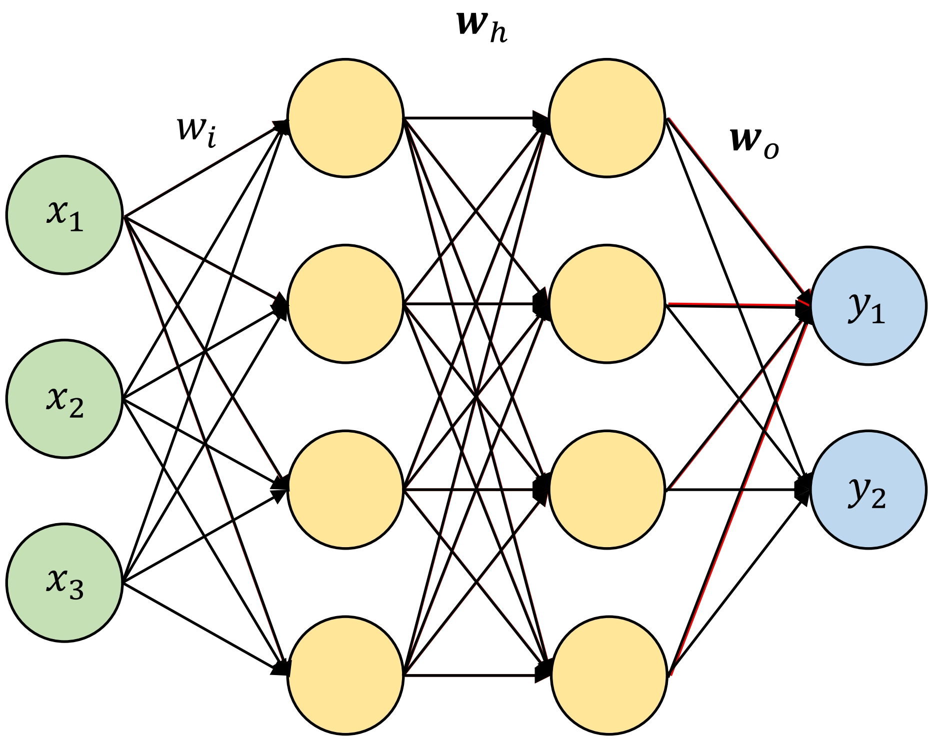




Model Components	Value
Degrees of Freedom per Element	2
Elements	512
Design Parameters	2
Nodes per Element	4
Nonlinear material	St Venant, plane stress
System Behaviour	Static

Response and Sensitivity

- Sobolev training



Model Parameters $\boldsymbol{\theta} = \{\boldsymbol{w}, \boldsymbol{b}\}$

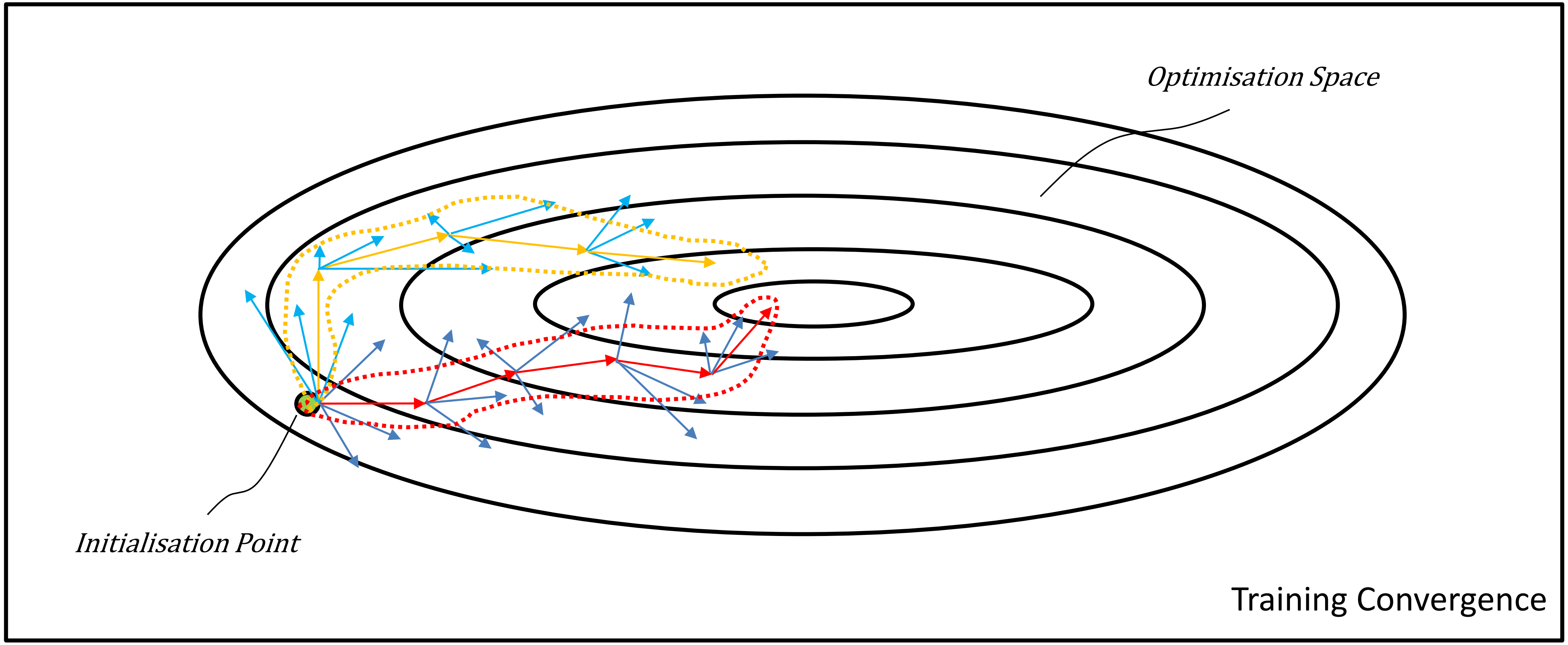
Loss Function $L = \frac{1}{2}\sum(\hat{\boldsymbol{y}} - \boldsymbol{y})^2 + \frac{1}{2}\sum(\frac{\partial\hat{\boldsymbol{y}}}{\partial \boldsymbol{x}} - \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}})^2$

Automatic Differentiation – Chain Rule

$$\frac{\partial y_1}{\partial x_1} = \frac{\partial y_1}{\partial w_o}\frac{\partial w_o}{\partial w_h}\frac{\partial w_h}{\partial w_i}\frac{\partial w_i}{\partial x_1}$$

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial w_i} + \frac{\partial L}{\partial(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{x}})}\frac{\partial(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{x}})}{\partial w_i}$$

Residual Weighting



Residual Weighting

$$L = \frac{1}{2}\sum(\hat{\boldsymbol{y}} - \boldsymbol{y})^2 + \frac{1}{2}\sum\sum(\frac{\partial\hat{\boldsymbol{y}}}{\partial\boldsymbol{x}} - \frac{\partial\boldsymbol{y}}{\partial\boldsymbol{x}})^2$$

- Magnitude $|L_i|$ acts as a weight
- The less accurate L_i, the greater gradient factor $|L_i|$
- Varying difficulty → staggered $|L_i|$ → gradient update intuitive to worst L_i, counterintuitive to best L_i

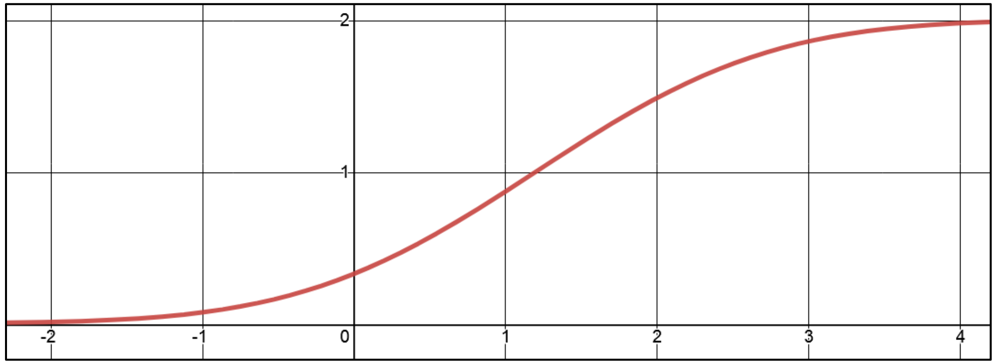
$$\Rightarrow \quad L = \frac{1}{2}\sum\boldsymbol{\lambda_1}(\hat{\boldsymbol{y}} - \boldsymbol{y})^2 + \frac{1}{2}\sum\sum\boldsymbol{\lambda_2}(\frac{\partial\hat{\boldsymbol{y}}}{\partial\boldsymbol{x}} - \frac{\partial\boldsymbol{y}}{\partial\boldsymbol{x}})^2$$

Residual Weighting

- Weighting methods

Abbreviation	Method	Target
L_{max}	$\max L$	Loss Maximisation
CD_{min}	$\min(1 - \frac{\nabla L \cdot \nabla L_i}{\lvert\nabla L\rvert \cdot \lvert\nabla L_i\rvert})$	Gradient Alignment
SNN	Sobolev Trained	-

$$L = \frac{1}{2}\sum \boldsymbol{\lambda_1}(\hat{\boldsymbol{y}} - \boldsymbol{y})^2 + \frac{1}{2}\sum\sum \boldsymbol{\lambda_2}(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{x}} - \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}})^2$$



Error Function or Clipping for numerical stability

$|y| \ll \left|\frac{\partial y}{\partial x}\right|$ | $\left|\sum(\hat{y} - y)^2\right| \ll \left|\sum\left(\frac{\partial \hat{y}}{\partial x} - \frac{\partial y}{\partial x}\right)^2\right|$

OR | OR

$|y| \gg \left|\frac{\partial y}{\partial x}\right|$ | $\left|\sum(\hat{y} - y)^2\right| \gg \left|\sum(\frac{\partial \hat{y}}{\partial x} - \frac{\partial y}{\partial x})^2\right|$

Response Values

Δ

Sensitivity Values

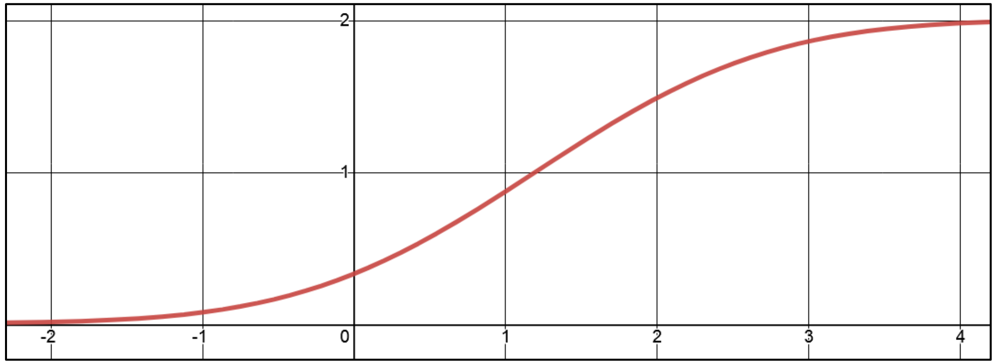
Normalization between (-1,1) but lose sensitivity characteristics

Residual Weighting

- Maximize Loss

Abbreviation	Method	Target
L_{max}	$\max L$	Loss Maximisation
CD_{min}	$\min(1 - \frac{\nabla L \cdot \nabla L_i}{\lvert\nabla L\rvert \cdot \lvert\nabla L_i\rvert})$	Gradient Alignment
SNN	Sobolev Trained	-

$$L = \frac{1}{2}\sum \lambda_1(\hat{\boldsymbol{y}} - \boldsymbol{y})^2 + \frac{1}{2}\sum\sum \lambda_2(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{x}} - \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}})^2$$



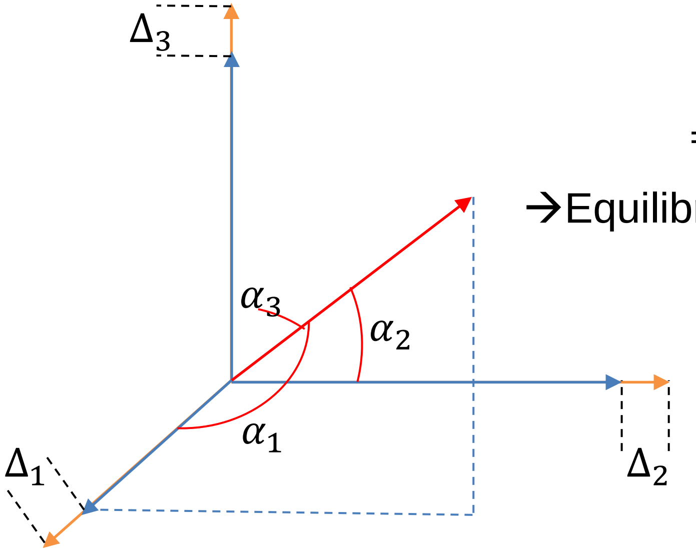
Error Function or Clipping for numerical stability

- Very straightforward and simple
- Essentially a constant gradient step, but -

→$|L_i|$ dicate Δ_i

Residual Weighting

- Minimize Cosine Distance by Magnitude – In $\mathcal{R}^2$



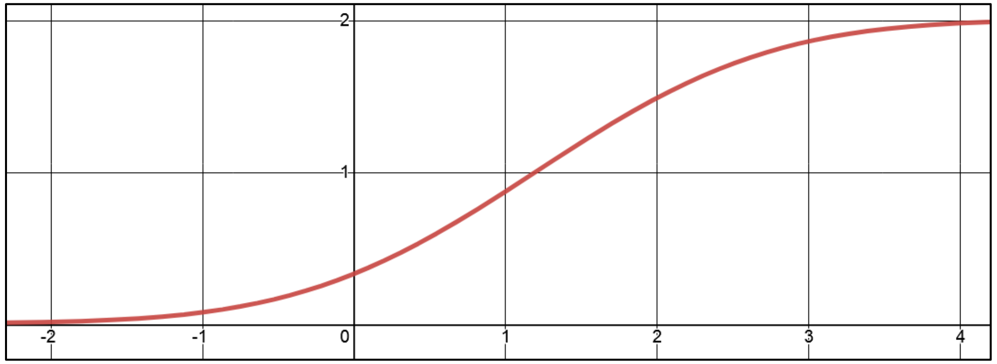
For one training step:

$$\alpha_1 : \alpha_2 : \alpha_3$$

$$\Rightarrow \Delta_1 : \Delta_2 : \Delta_3$$

→Equilibrium of residual weight change

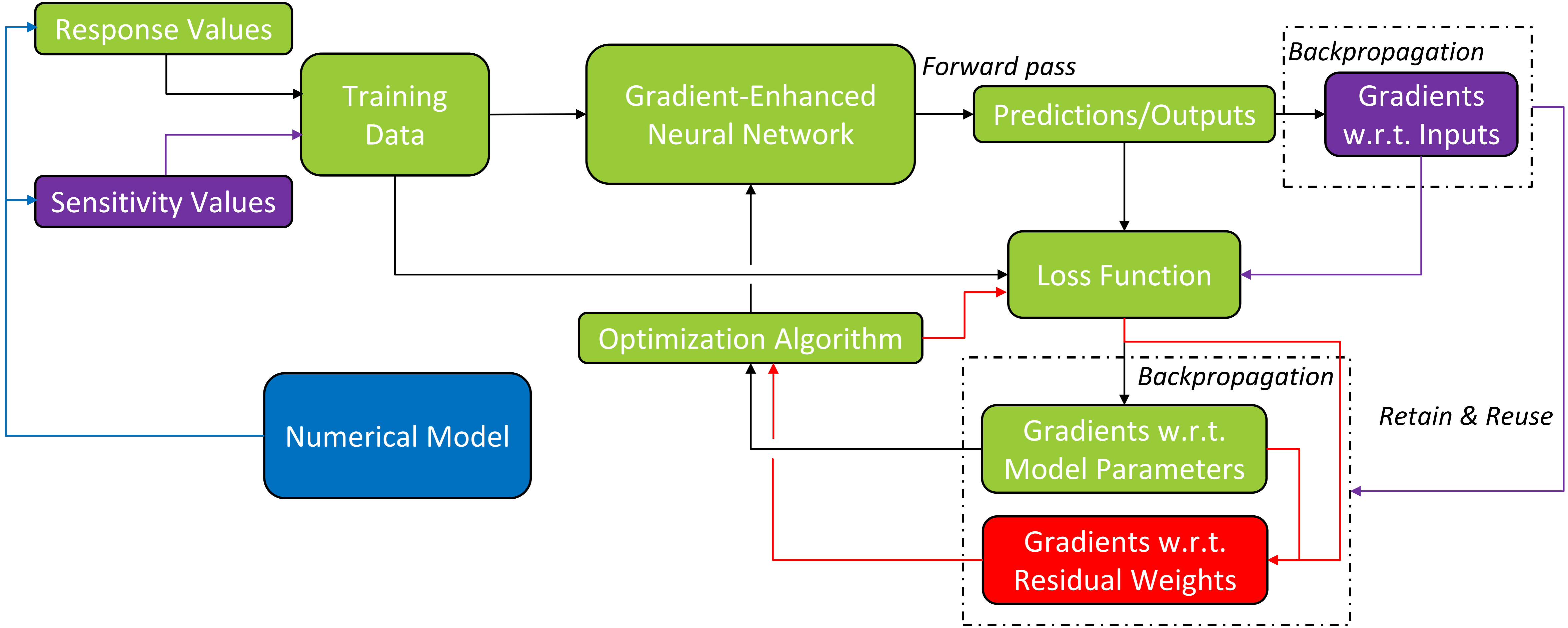
$$L = \frac{1}{2}\sum \boldsymbol{\lambda_1}(\hat{\boldsymbol{y}} - \boldsymbol{y})^2 + \frac{1}{2}\sum\sum \boldsymbol{\lambda_2}(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{x}} - \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}})^2$$



Error Function or Clipping for numerical stability

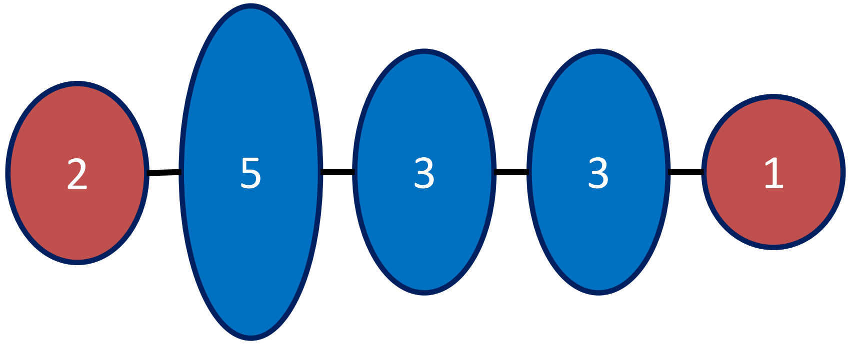
- Ratios $|\nabla L_i|$ dicate α_i, ratios α_i dictate Δ_i
- When a loss target is dominated, its residual weight increases stronger

Residual Weighting

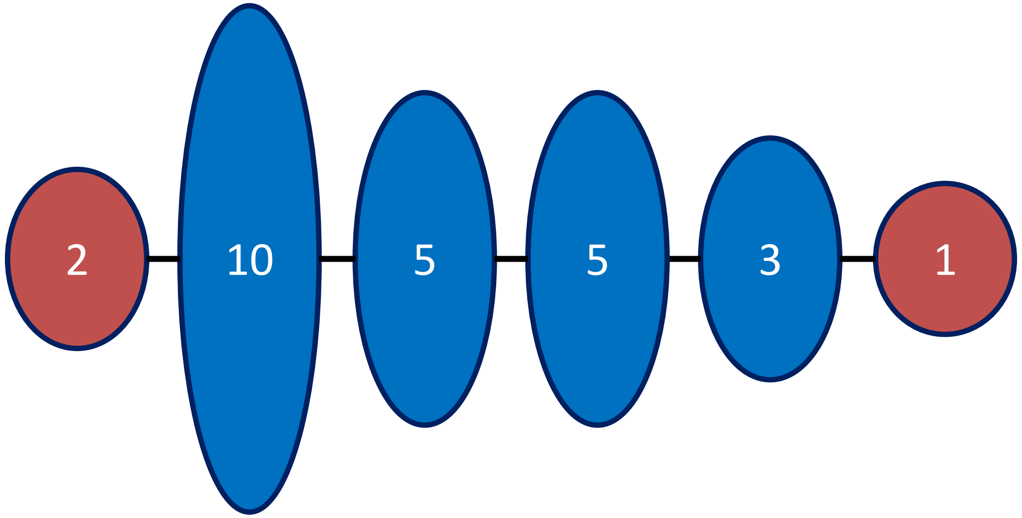


Residual Weighting

- Feedforward structure



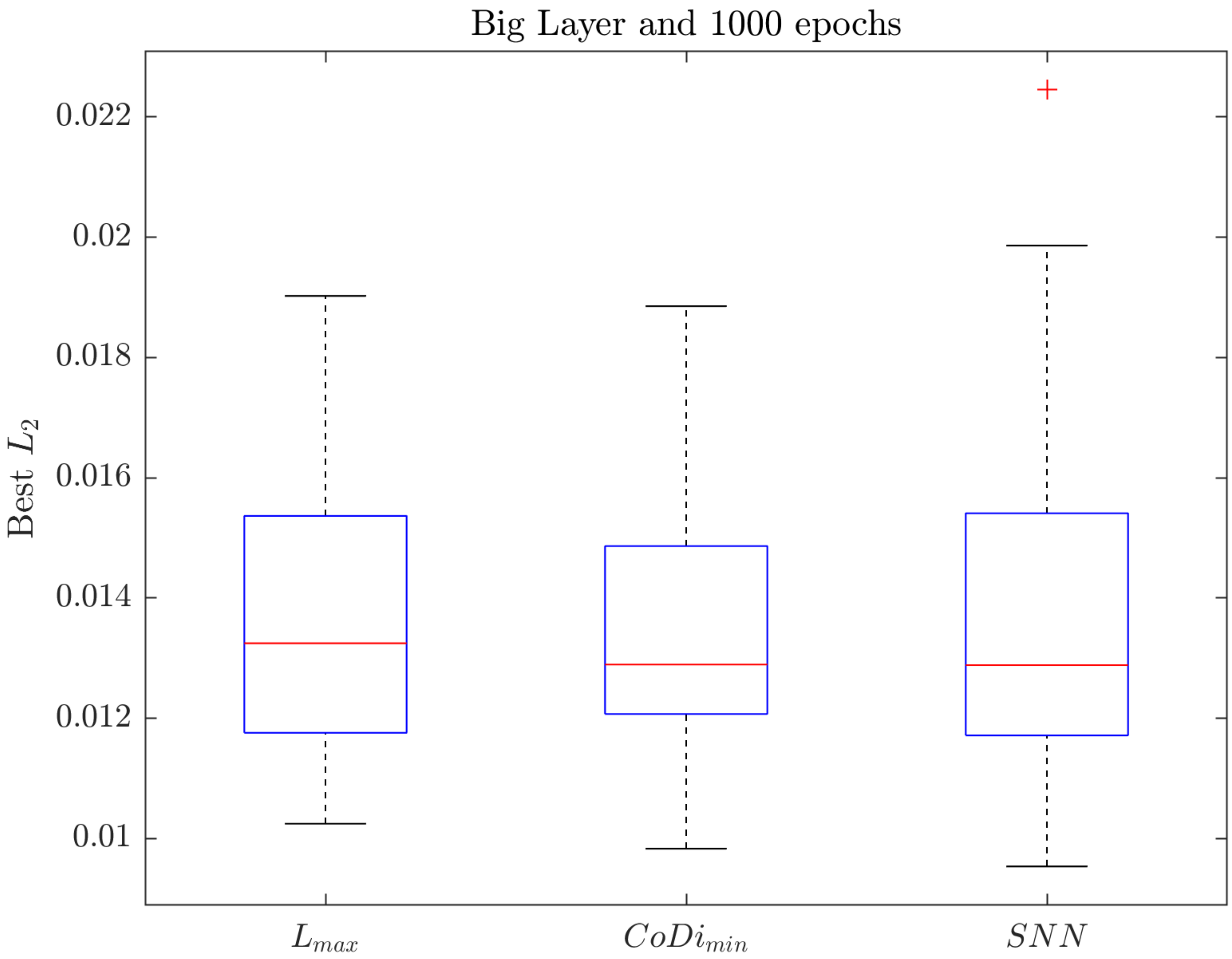
Small Layer – 5,3,3

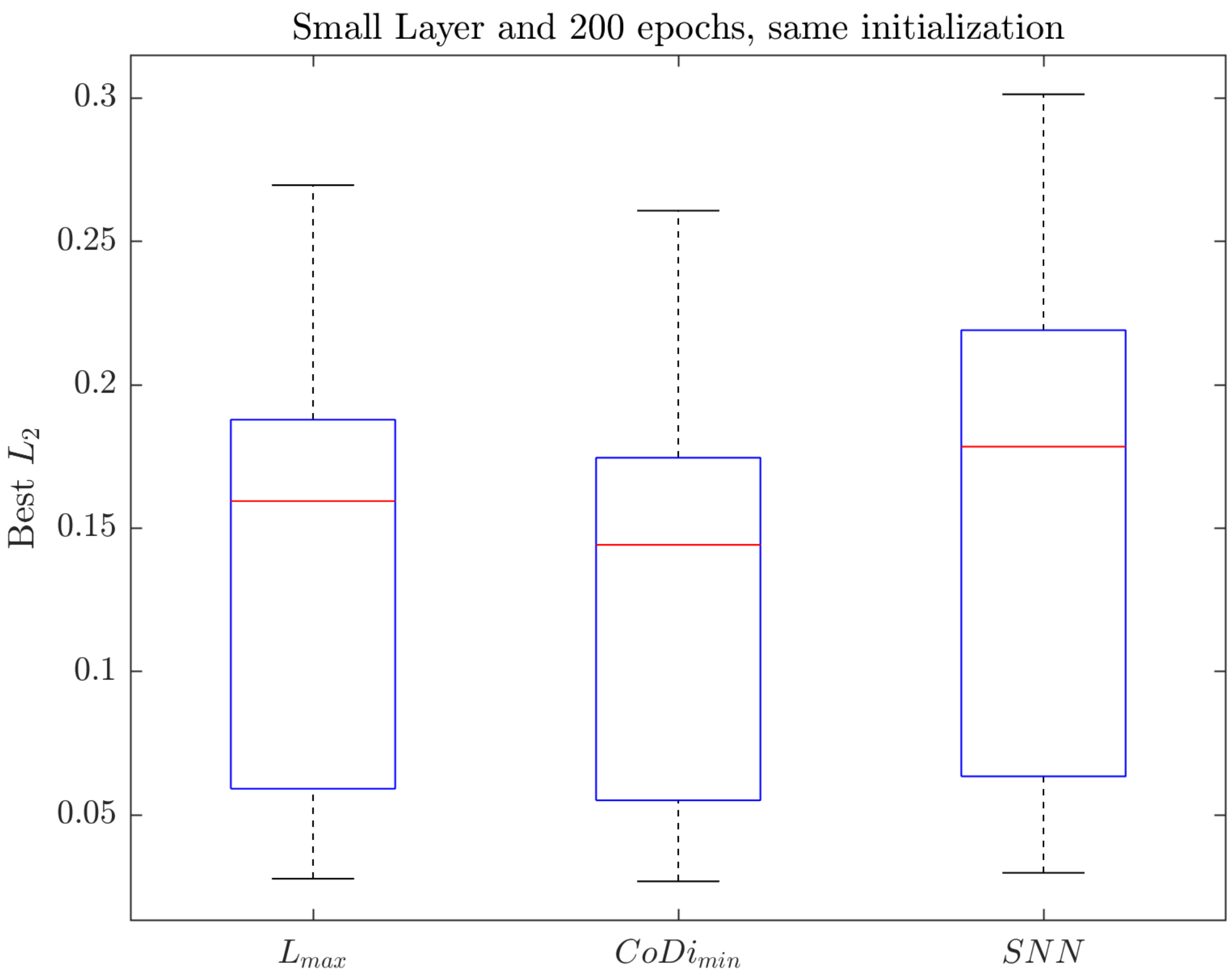


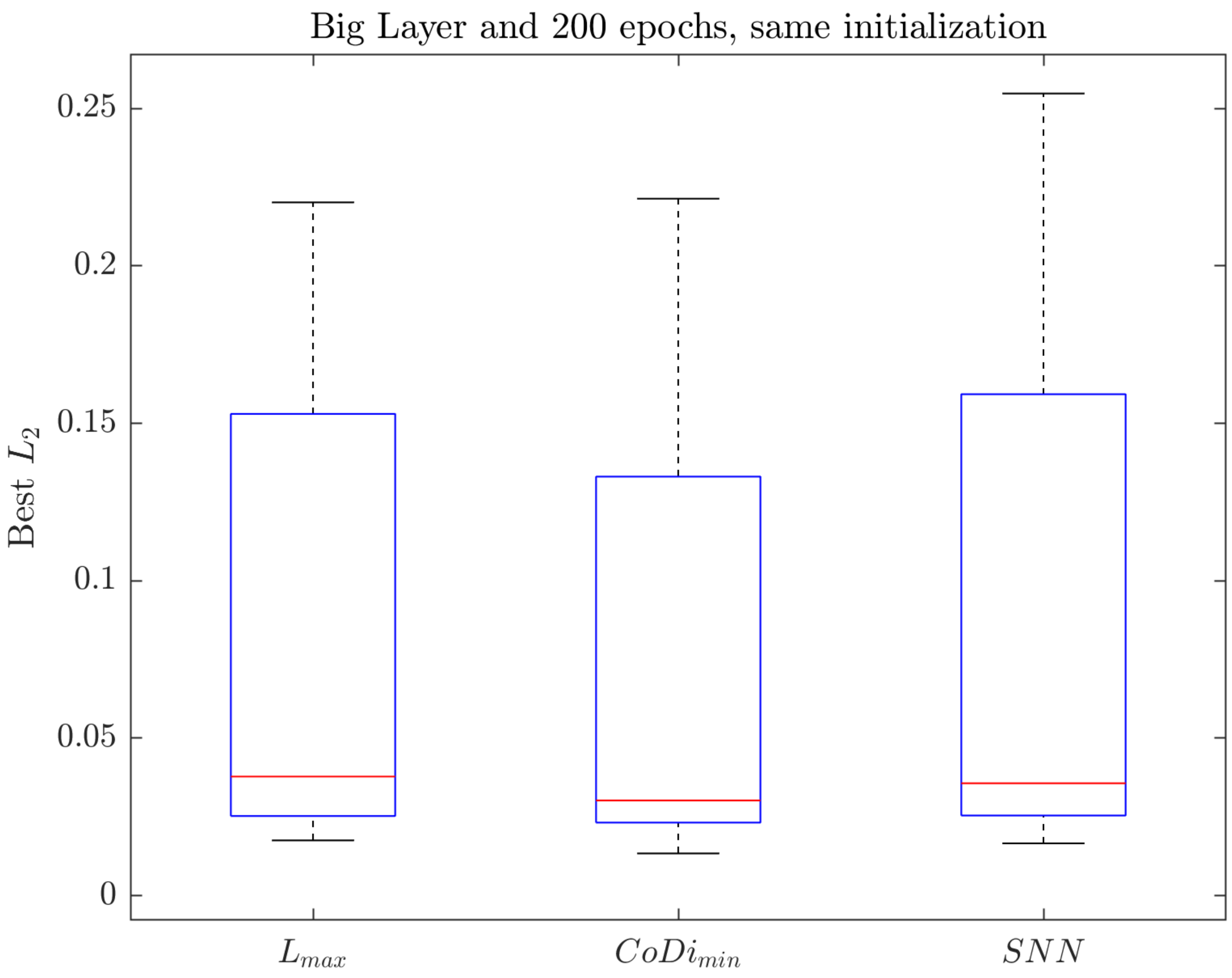
Big Layer– 10,5,5,3

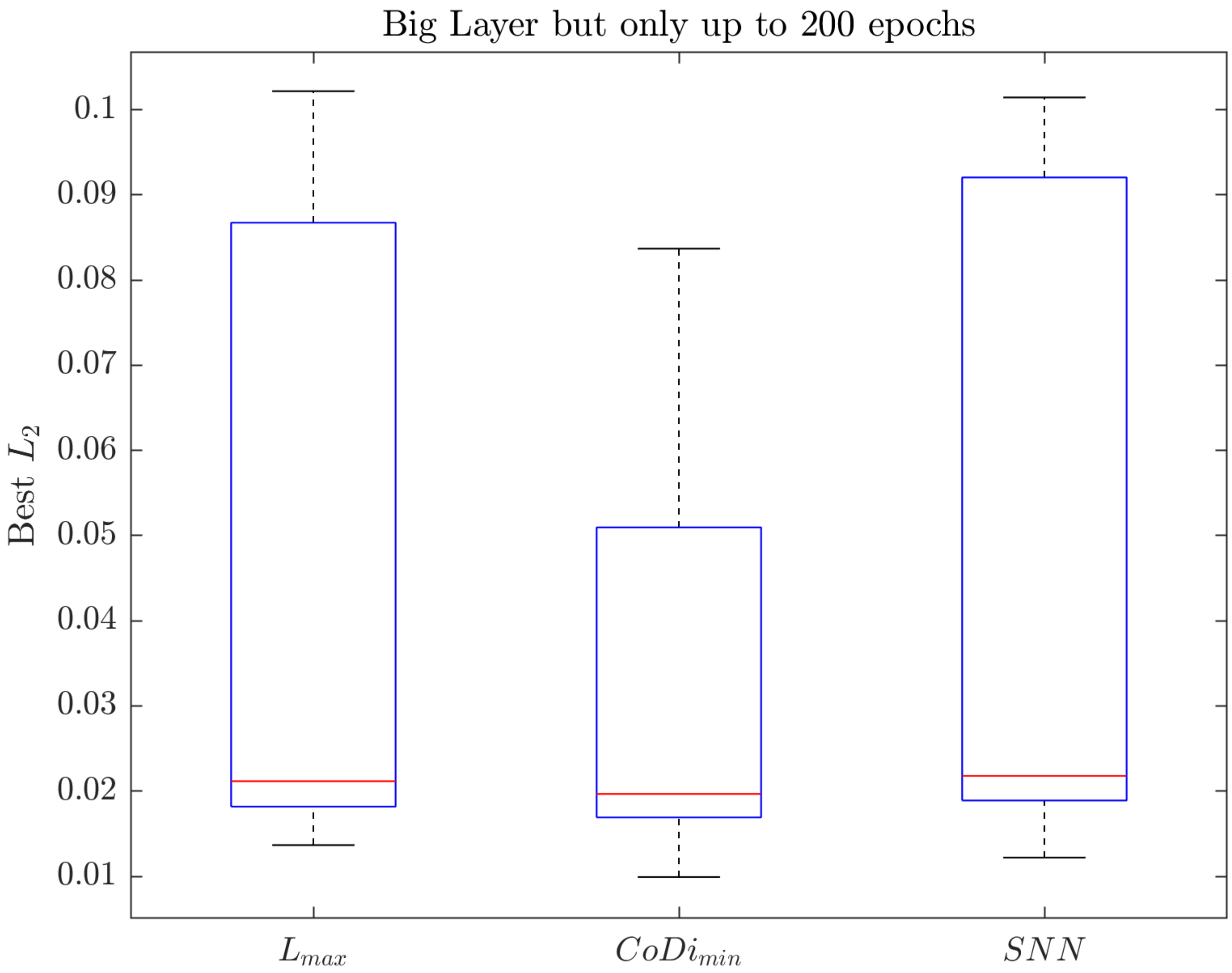
Hyperparameter	Value
Epochs	200/500/1000
Training Data Size	320
Validation Data Size	305

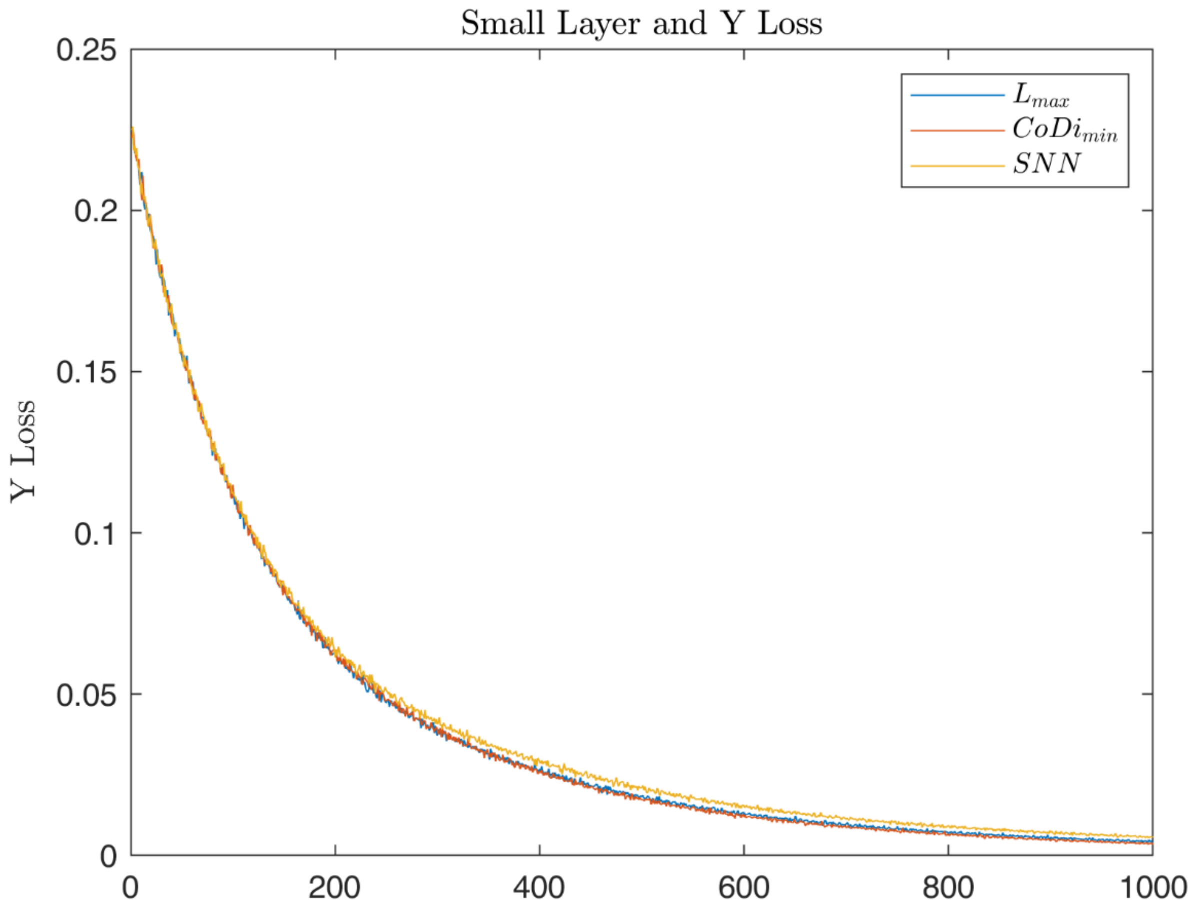
- Data generated via FEM
- Evenly spaced samples from input space
- Validation data inbetween training data

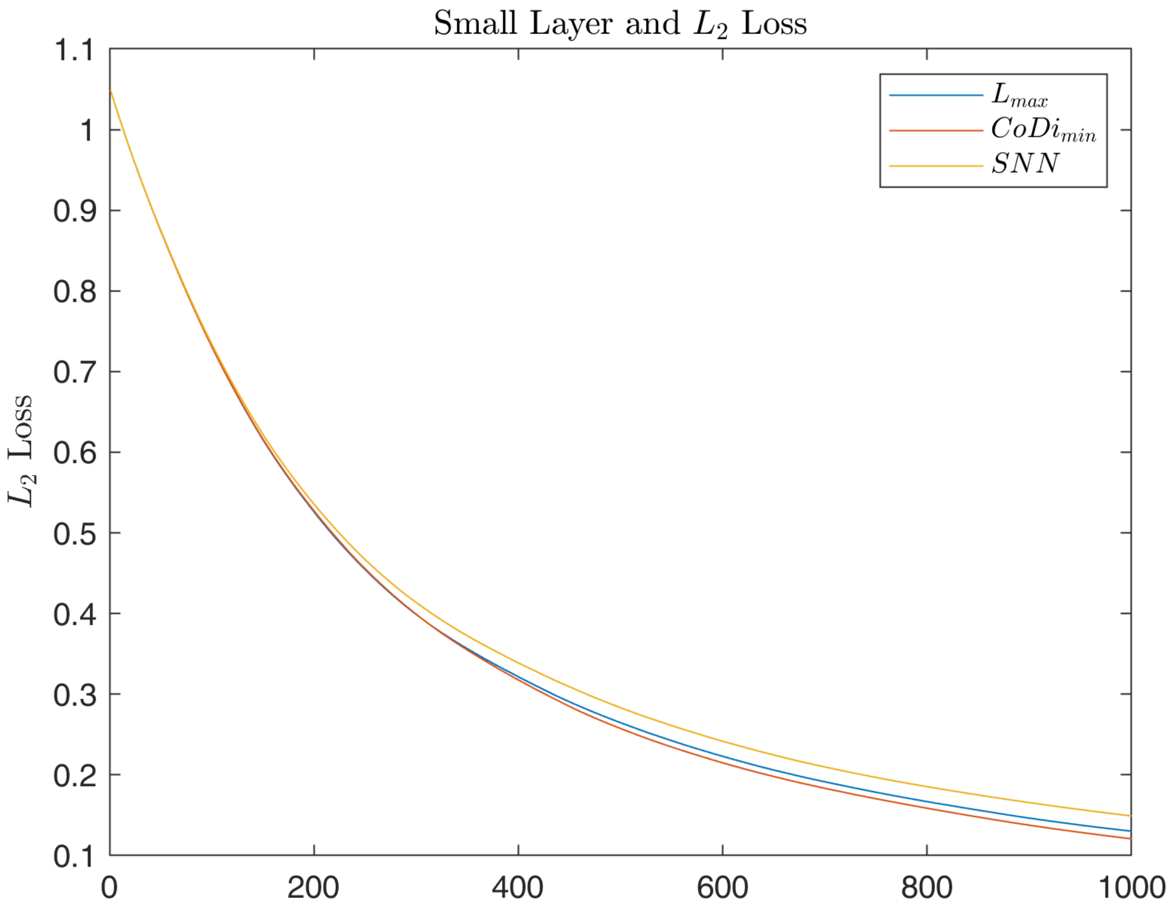

Big Layer and 1000 epochs
Best L_2
0.022
0.02
0.018
0.016
0.014
0.012
0.01
L_{max}
$CoDi_{min}$
SNN

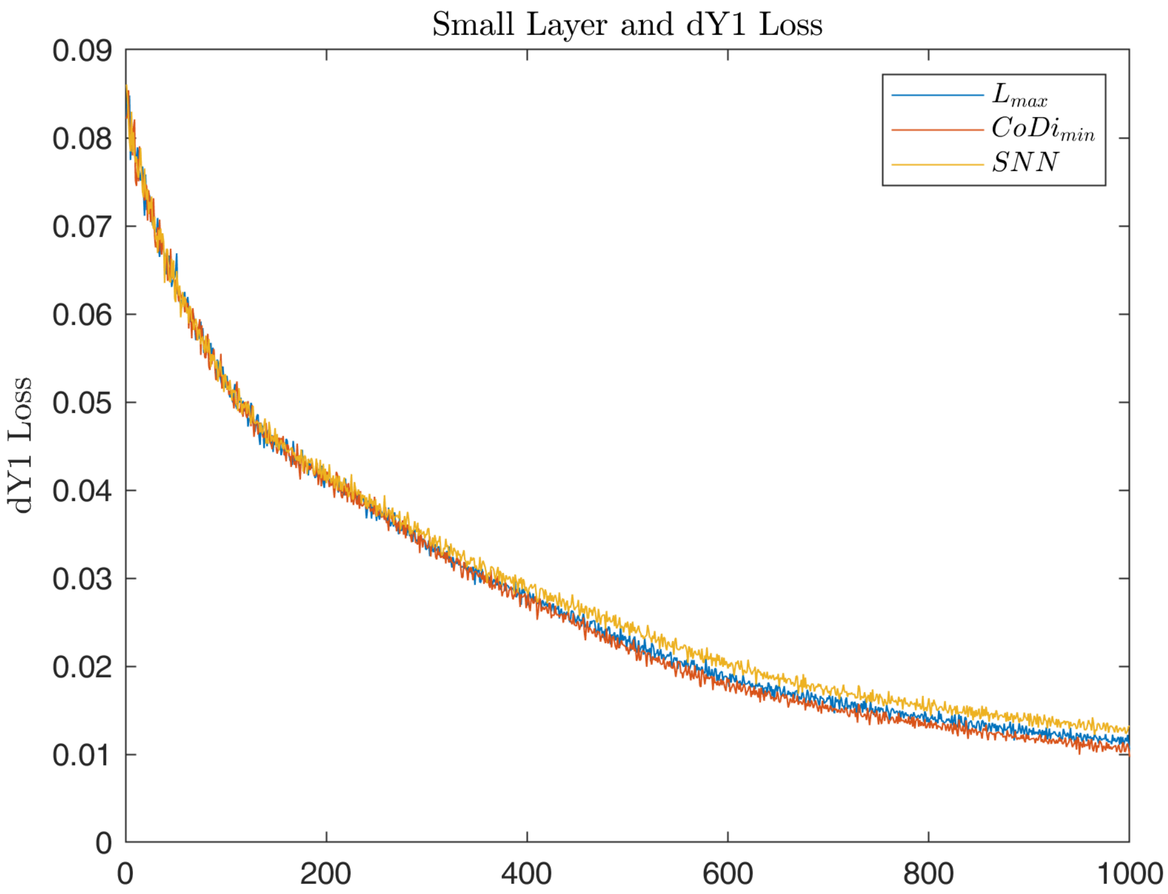

Small Layer and 200 epochs, same initialization
Best L_2
0.3
0.25
0.2
0.15
0.1
0.05
L_{max}
$CoDi_{min}$
SNN

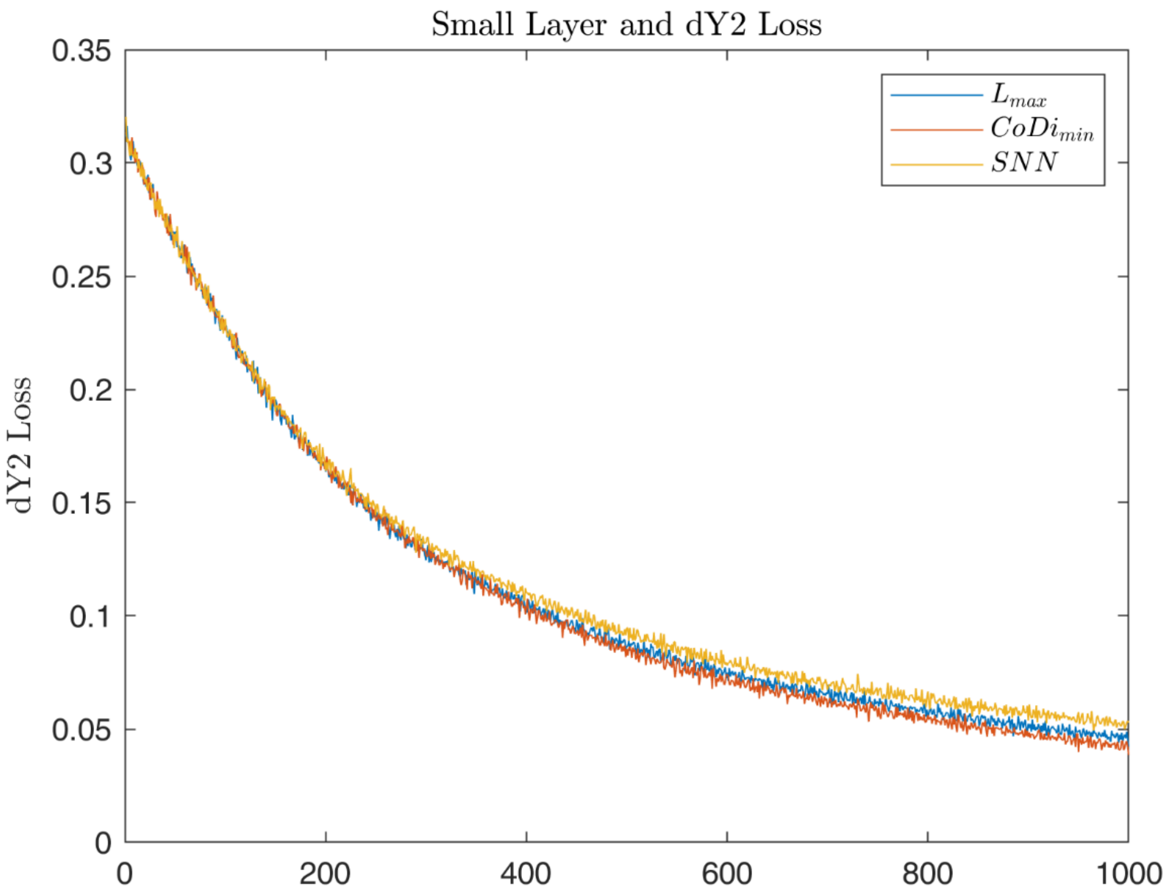

Big Layer and 200 epochs, same initialization
Best L_2
0.25
0.2
0.15
0.1
0.05
0
L_{max}
$CoDi_{min}$
SNN

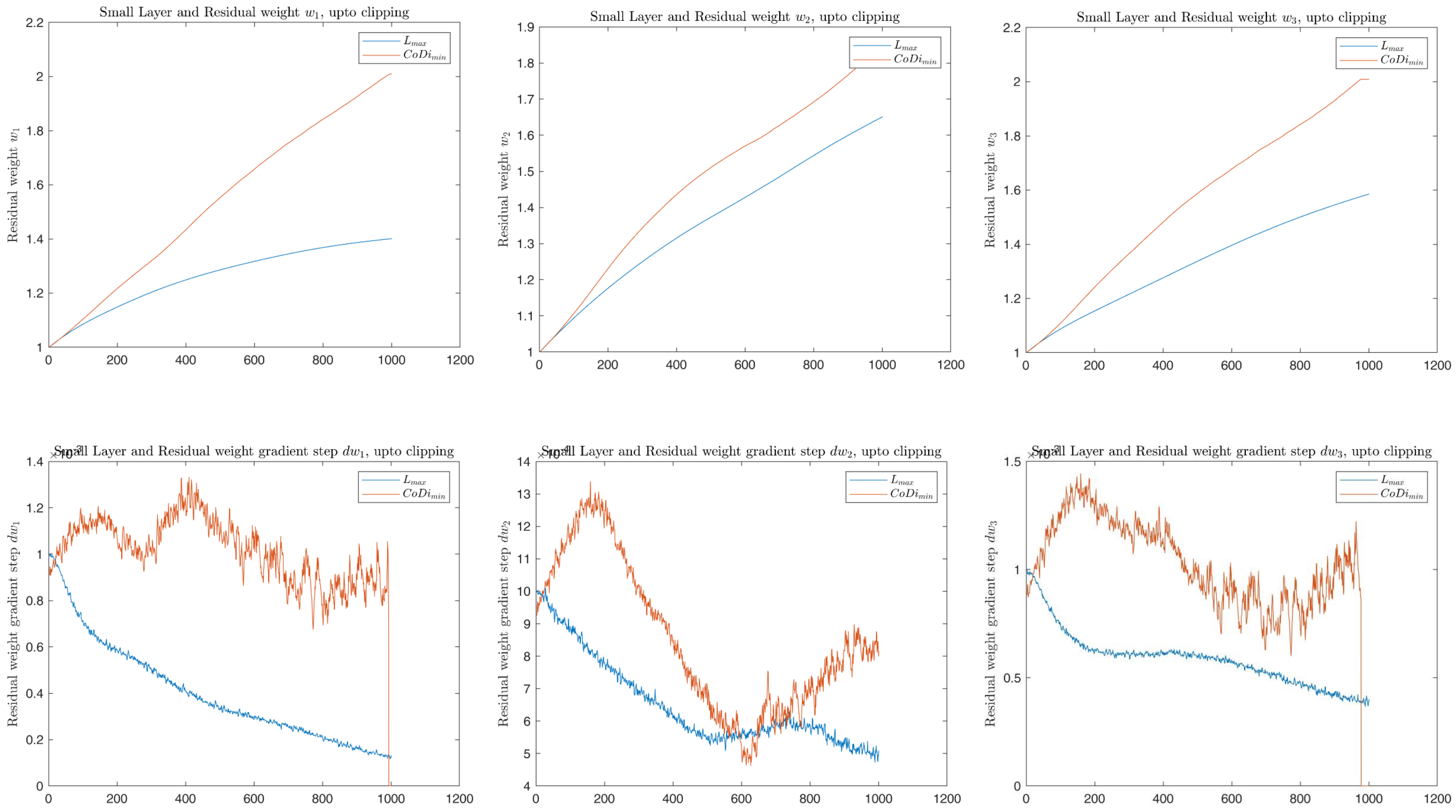

Big Layer but only up to 200 epochs
Best L_2
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
L_{max}
$CoDi_{min}$
SNN


Small Layer and Y Loss
Y Loss
0.25
0.2
0.15
0.1
0.05
0
0
200
400
600
800
1000
L_{max}
$CoDi_{min}$
SNN

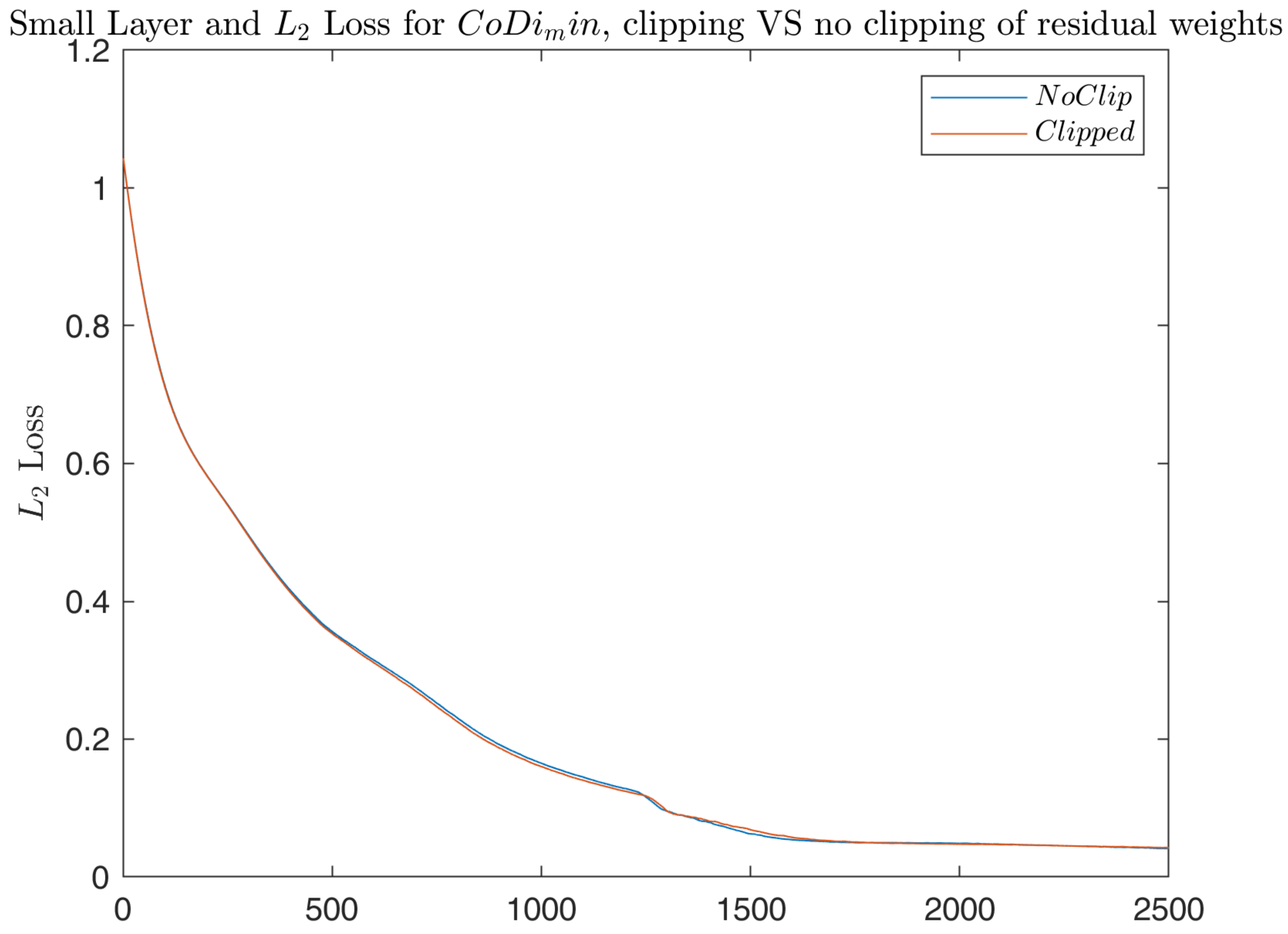

Small Layer and L_2 Loss
L_2 Loss
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
200
400
600
800
1000
L_{max}
$CoDi_{min}$
SNN

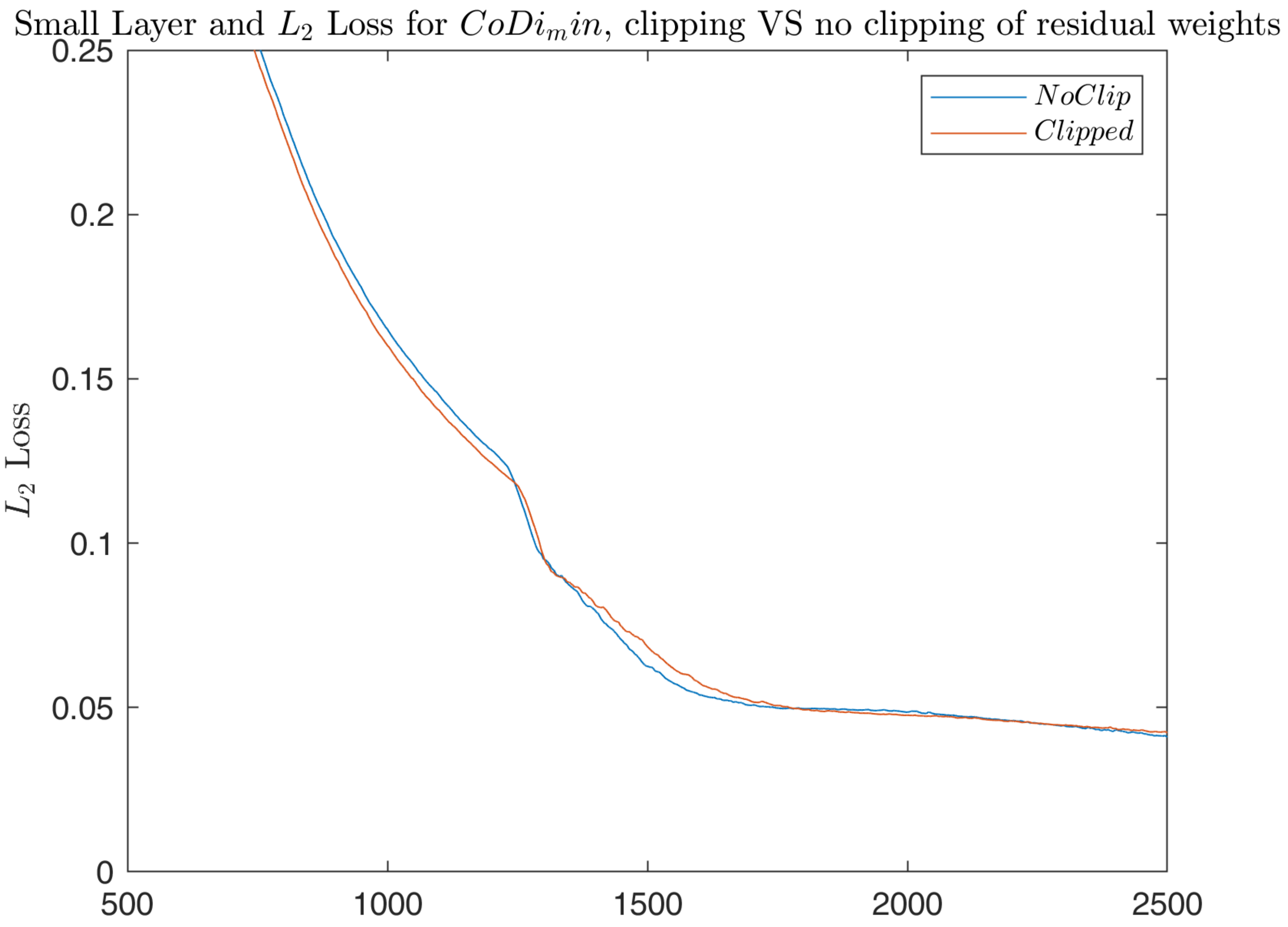

Small Layer and dY1 Loss
dY1 Loss
L_{max}
$CoDi_{min}$
SNN
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
200
400
600
800
1000

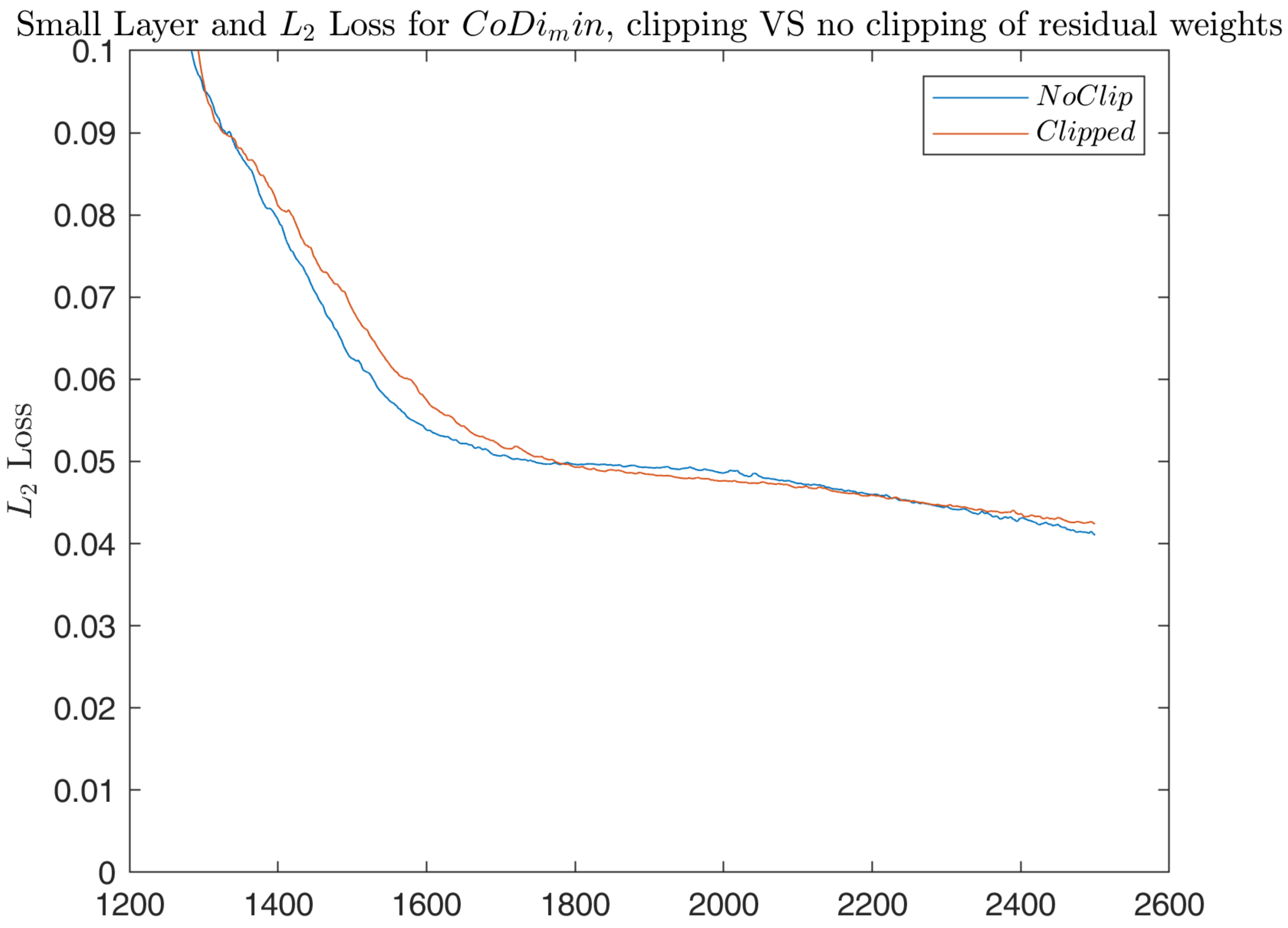

Small Layer and dY2 Loss
dY2 Loss
L_{max}
$CoDi_{min}$
SNN
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
200
400
600
800
1000



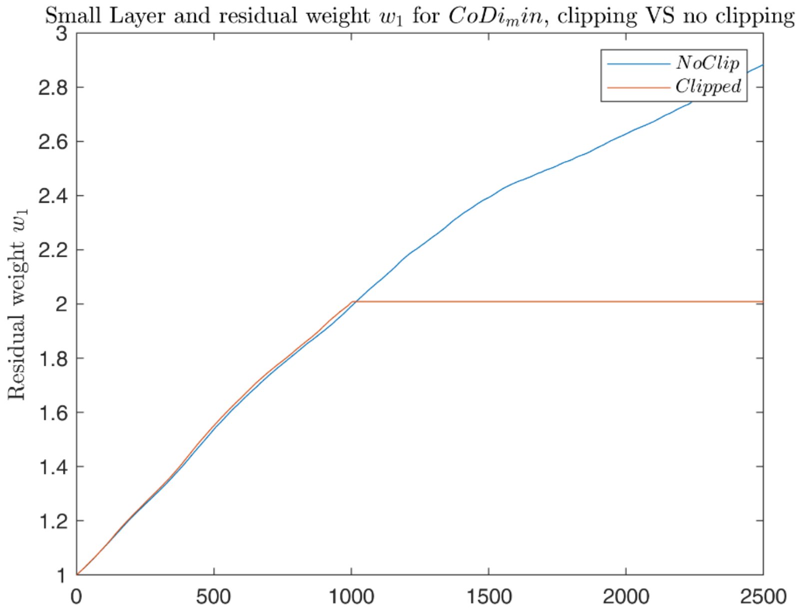
Small Layer and Residual weight w_1, upto clipping
Residual weight w_1
Small Layer and Residual weight w_2, upto clipping
Residual weight w_2
Small Layer and Residual weight w_3, upto clipping
Residual weight w_3
Small Layer and Residual weight gradient step dw_1, upto clipping
Residual weight gradient step dw_1
Small Layer and Residual weight gradient step dw_2, upto clipping
Residual weight gradient step dw_2
Small Layer and Residual weight gradient step dw_3, upto clipping
Residual weight gradient step dw_3
L_{max}
$CoDi_{min}$
0
200
400
600
800
1000
1200

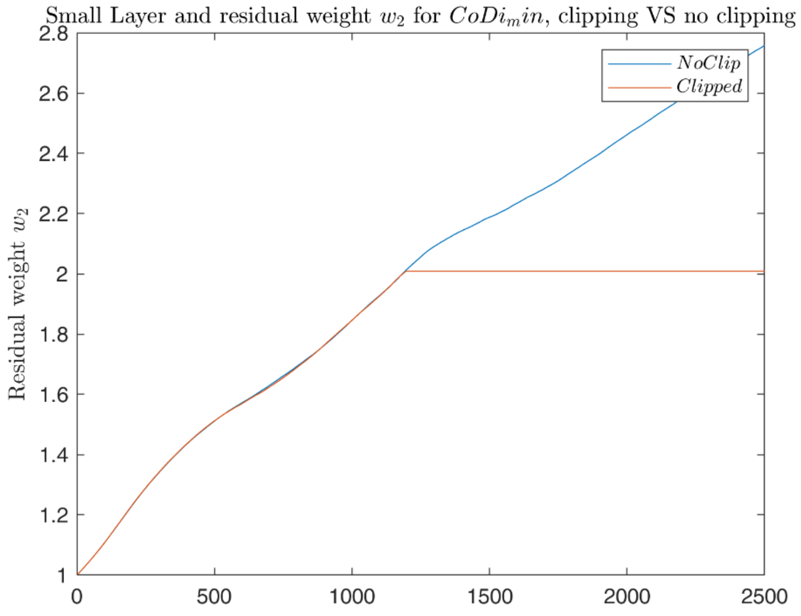

Small Layer and L_2 Loss for $CoDi_m in$, clipping VS no clipping of residual weights
L_2 Loss
NoClip
Clipped
1.2
1
0.8
0.6
0.4
0.2
0
0
500
1000
1500
2000
2500

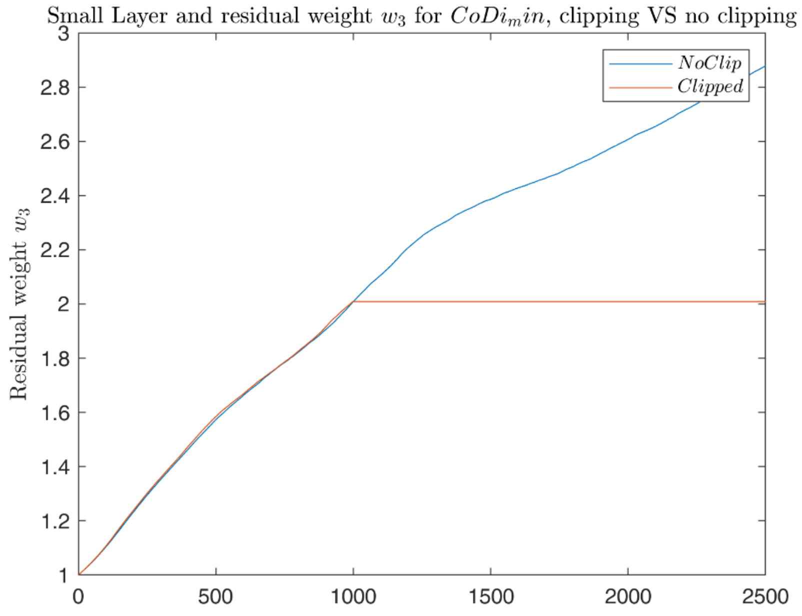

Small Layer and L_2 Loss for $CoDi_min$, clipping VS no clipping of residual weights
$NoClip$
$Clipped$
L_2 Loss
0.25
0.2
0.15
0.1
0.05
0
500
1000
1500
2000
2500

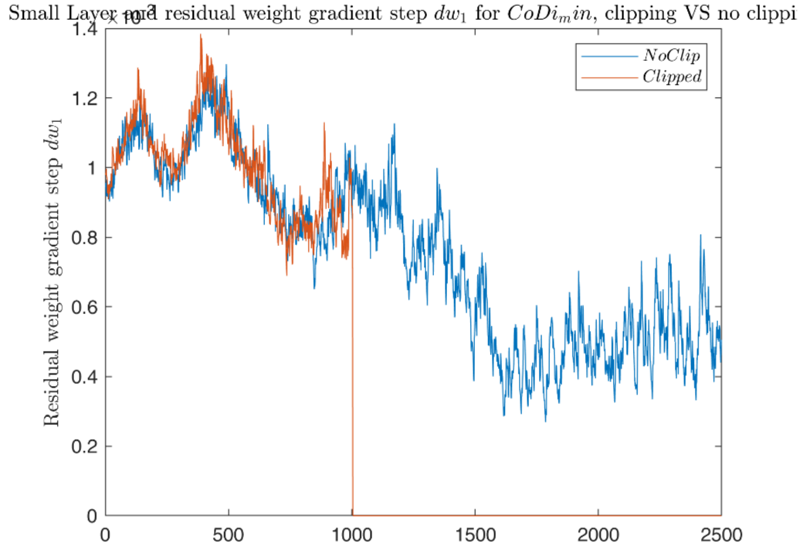


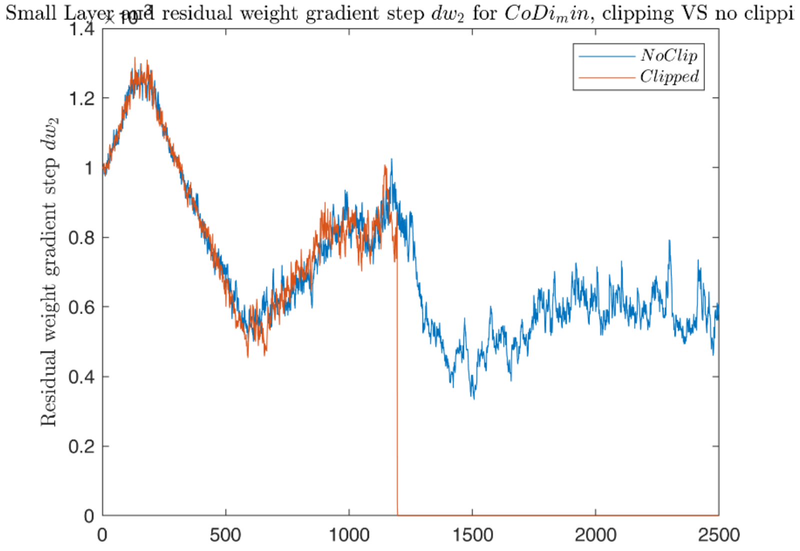
Small Layer and L_2 Loss for $CoDi_min$, clipping VS no clipping of residual weights
$NoClip$
$Clipped$
L_2 Loss
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1200
1400
1600
1800
2000
2200
2400
2600

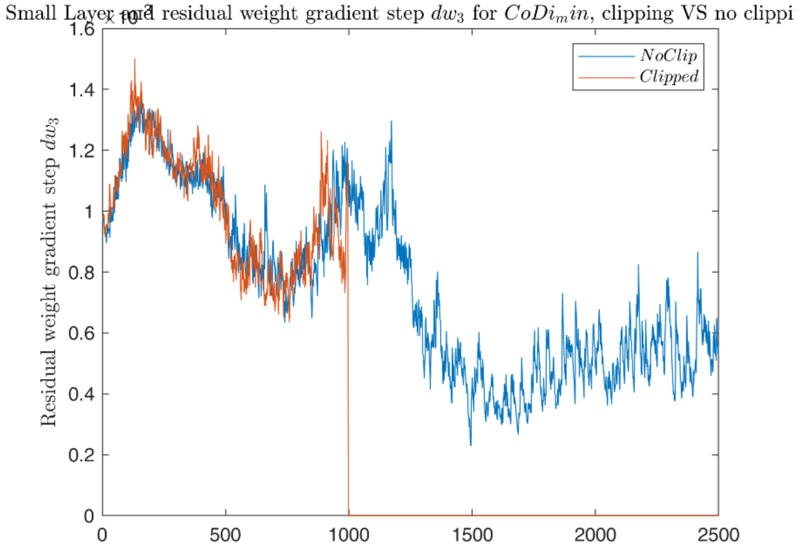

Small Layer and residual weight w_1 for $CoDi_min$, clipping VS no clipping
Residual weight w_1
NoClip
Clipped


Small Layer and residual weight w_2 for $CoDi_min$, clipping VS no clipping
Residual weight w_2
NoClip
Clipped


Small Layer and residual weight w_3 for $CoDi_min$, clipping VS no clipping
Residual weight w_3
NoClip
Clipped


Small Layer and residual weight gradient step dw_1 for $CoDi_min$, clipping VS no clippi
Residual weight gradient step dw_1
NoClip
Clipped


Small Layer and residual weight gradient step dw_2 for $CoDi_min$, clipping VS no clippi
Residual weight gradient step dw_2
NoClip
Clipped


Small Layer and residual weight gradient step dw_3 for $CoDi_min$, clipping VS no clippi
Residual weight gradient step dw_3
NoClip
Clipped

Conclusion

- Multiple loss terms lead to staggered training focus

- Optimal weighting for training convergence
  - dynamic weights to avoid more hyperparameters
  - optimization target/s of residual weights crucial for model quality of interest

  → equilibrium between gradient vectors

- training becomes more robust, not just an average improvement of accuracy

- Unclipped CoDi shows a converging behavior that seems to oscillate

Conclusion

Issues:

- Sensitivities magnitude difference might still introduce bias
- Numerical stability of residual weights for unclipped case
- Optimization target for residual weights: some residual loss terms more important than others, perhaps?

Thank you!

Leonhard-Euler-Straße 5
44227 Dortmund
Campus Nord

https://lre.mb.tu-dortmund.de/

ali.kilicsoy@tu-dortmund.de

https://www.linkedin.com/company/chair-for-reliability-engineering/