Efficient Hessian Free Optimization of Deep Neural Networks

Niklas Brunn, Noël E. Kury, Clemens A. Schächter

Project presentation at the Albert Ludwigs University of Freiburg
Numerical Optimization
Prof. Dr. Moritz Diehl
M.Sc. Florian Messerer

6. Februar 2022

Unconstrained NLP

$$\underset{\theta \in \mathbb{R}^d}{\arg\min} \quad f_{\mathrm{M}}(\theta)$$

Unconstrained NLP

$$\underset{\theta \in \mathbb{R}^d}{\arg\min} \quad f_{\mathrm{M}}(\theta)$$

- θ decision variables and d dimension of parameter space

Unconstrained NLP

$$\underset{\theta \in \mathbb{R}^d}{\arg\min} \quad f_{\mathrm{M}}(\theta)$$

- θ decision variables and d dimension of parameter space
- objective function

$$f_{\mathrm{M}}(\theta) := \frac{1}{N} \sum_{(x,y) \in \mathrm{M}} \mathrm{L}_y(\mathsf{akt}_{(.)}(\mathrm{R}_x(\theta)))$$

Unconstrained NLP

$$\underset{\theta\in\mathbb{R}^d}{\arg\min} \quad f_{\mathrm{M}}(\theta)$$

- θ decision variables and d dimension of parameter space
- objective function

$$f_{\mathrm{M}}(\theta) := \frac{1}{N}\sum_{(x,y)\in\mathrm{M}} \mathrm{L}_y(\mathsf{akt}_{(.)}(\mathrm{R}_x(\theta)))$$

- $\mathrm{M} \subset \mathrm{D} := \{(x_i, y_i)_{1\le i\le N}\}$ observation data;
  $\mathrm{R}_x(\theta)$ realisation of a DNN given the observed value x with current parameters θ;
  $\mathsf{akt}_{(.)}(\hat{y}) := (akt(\hat{y}_1), \dots, akt(\hat{y}_1))$ elementwise application of a convex activation function on the vector $\hat{y}$;
  $\mathrm{L}_z := \mathsf{Loss}(z, y)$ non-decreasing, convex loss function in z given the target y

Unconstrained NLP

$$\underset{\theta \in \mathbb{R}^d}{\arg\min} \quad f_{\mathrm{M}}(\theta)$$

Unconstrained NLP

$$\underset{\theta \in \mathbb{R}^d}{\arg\min} \quad f_{\mathrm{M}}(\theta)$$

- θ decision variables and d dimension of parameter space

Unconstrained NLP

$$\underset{\theta \in \mathbb{R}^d}{\arg\min} \quad f_{\mathrm{M}}(\theta)$$

- θ decision variables and d dimension of parameter space
- objective function

$$f_{\mathrm{M}}(\theta) := \frac{1}{N} \sum_{(x,y) \in \mathrm{M}} \mathrm{L}_y(\mathsf{akt}_{(.)}(\mathrm{R}_x(\theta)))$$

Unconstrained NLP

$$\underset{\theta \in \mathbb{R}^d}{\arg\min} \quad f_{\mathrm{M}}(\theta)$$

- θ decision variables and d dimension of parameter space
- objective function

$$f_{\mathrm{M}}(\theta) := \frac{1}{N} \sum_{(x,y) \in \mathrm{M}} \mathrm{L}_y(\mathrm{akt}_{(.)}(\mathrm{R}_x(\theta)))$$

- $\mathrm{M} \subset \mathrm{D} := \{(x_i, y_i)_{1 \leq i \leq N}\}$ observation data;
  $\mathrm{R}_x(\theta)$ realisation of a DNN given the observed value x with current parameters θ;
  $\mathrm{akt}_{(.)}(\hat{y}) := (akt(\hat{y}_1), \ldots, akt(\hat{y}_1))$ elementwise application of a convex activation function on the vector $\hat{y}$;
  $\mathrm{L}_z := \mathrm{Loss}(z, y)$ non-decreasing, convex loss function in z given the target y

Unconstrained NLP

$$\underset{\theta \in \mathbb{R}^d}{\arg\min} \quad f_{\mathrm{M}}(\theta)$$

Unconstrained NLP

$$\arg\min_{\theta \in \mathbb{R}^d} \quad f_{\mathrm{M}}(\theta)$$

- θ decision variables and d dimension of parameter space

Unconstrained NLP

$$\operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \quad f_{\mathrm{M}}(\theta)$$

- θ decision variables and d dimension of parameter space
- objective function

$$f_{\mathrm{M}}(\theta) := \frac{1}{N} \sum_{(x,y)\in \mathrm{M}} \mathrm{L}_y(\mathsf{akt}_{(.)}(\mathrm{R}_x(\theta)))$$

Unconstrained NLP

$$\underset{\theta \in \mathbb{R}^d}{\arg\min} \quad f_{\mathrm{M}}(\theta)$$

- θ decision variables and d dimension of parameter space
- objective function

$$f_{\mathrm{M}}(\theta) := \frac{1}{N} \sum_{(x,y) \in \mathrm{M}} \mathrm{L}_y(\mathrm{akt}_{(.)}(\mathrm{R}_x(\theta)))$$

- $\mathrm{M} \subset \mathrm{D} := \{(x_i, y_i)_{1 \leq i \leq N}\}$ observation data;
  $\mathrm{R}_x(\theta)$ realisation of a DNN given the observed value x with current parameters θ;
  $\mathrm{akt}_{(.)}(\hat{y}) := (akt(\hat{y}_1), \ldots, akt(\hat{y}_1))$ elementwise application of a convex activation function on the vector $\hat{y}$;
  $\mathrm{L}_z := \mathrm{Loss}(z, y)$ non-decreasing, convex loss function in z given the target y

Sources

[1] Moritz Diehl, "Lecture Notes on Numerical Optimization (Preliminary Draft)", Albert Ludwigs University of Freiburg, September 29, 2017

[2] James Martens, "Deep learning via Hessian-free optimization", University of Toronto, Ontario, M5S 1A1, Canada, 2010

[3] James Martens, "New Insights and Perspectives on the Natural Gradient Method", Jurnal of Machine Learning Research 21, arXiv:1412.1193v11 [cs.LG], September 19, 2020

[4] Nicol N. Schraudolph, "Fast Curvature Matrix-Vector Products for Second-Order Gradient Descent", Neural Computation, August 2002