

# Hessian Chain Bracketing

UWE NAUMANN AND SHUBHADITYA BURELA

Second derivatives of mathematical models for real-world phenomena are fundamental ingredients of a wide range of numerical simulation methods including parameter sensitivity analysis, uncertainty quantification, nonlinear optimization and model calibration. The evaluation of such Hessians often dominates the overall computational effort. The combinatorial HESSIAN ACCUMULATION problem aiming to minimize the number of floating-point operations required for the computation of a Hessian turns out to be NP-complete. We propose a dynamic programming formulation for the solution of HESSIAN ACCUMULATION over a sub-search space. This approach yields improvements by factors of ten and higher over the state of the art based on second-order tangent and adjoint algorithmic differentiation.

AMS 2000 SUBJECT CLASSIFICATIONS: Primary 90C39, 65D25; secondary 90C60.

KEYWORDS AND PHRASES: Chain rule of differentiation, algorithmic differentiation, Hessian, NP-completeness, dynamic programming.

## 1. Motivation and introduction

We consider twice differentiable multivariate vector functions

$$(1) \quad F : \mathbb{R}^n \rightarrow \mathbb{R}^m : \mathbf{x} \mapsto \mathbf{y} = F(\mathbf{x})$$

implemented as computer programs evaluating sequences of  $q > 0$  *elemental functions*

$$F_i : \mathbb{R}^{n_{i-1}} \rightarrow \mathbb{R}^{n_i} : \mathbf{v}_{i-1} \mapsto \mathbf{v}_i = F_i(\mathbf{v}_{i-1})$$

for  $i = 1, \dots, q$ ,  $\mathbf{v}_0 = \mathbf{x}$  and  $\mathbf{y} = \mathbf{v}_q$ . This layered structure of

$$(2) \quad F = F_q \circ F_{q-1} \circ F_{q-2} \circ \dots \circ F_1$$

is typical for many numerical simulations. Even if it is not explicit in the given source program finding suitable vertex separators representing the

---

arXiv: [2103.09480](https://arxiv.org/abs/2103.09480)

$\mathbf{v}_i$  in the directed acyclic data dependence graph is straightforward. We set  $F_{[i,j]} \equiv F_i \circ \dots \circ F_{j+1}$  implying  $F_i = F_{[i,i-1]}$  and  $F = F_{[q,0]}$ . We use  $=$  to denote mathematical equality and  $\equiv$  in the sense of “is defined as.”

Elemental Jacobians

$$F'_i = F'_i(\mathbf{v}_{i-1}) \equiv \frac{dF_i}{d\mathbf{v}_{i-1}}(\mathbf{v}_{i-1}) \in \mathbb{R}^{n_i \times n_{i-1}}$$

and Hessians

$$F''_i = F''_i(\mathbf{v}_{i-1}) \equiv \frac{d^2F_i}{d\mathbf{v}_{i-1}^2}(\mathbf{v}_{i-1}) \in \mathbb{R}^{n_i \times n_{i-1} \times n_{i-1}}$$

are assumed to be given. For example, they can be computed by application of Algorithmic Differentiation (AD) [19, 26] to a given implementation of the  $F_i$  as a differentiable subprogram.

The chain rule of differential calculus yields

$$(3) \quad F' = F'(\mathbf{x}) = F'_q \cdot F'_{q-1} \cdot \dots \cdot F'_1$$

Differentiation of Equation (3) with respect to  $\mathbf{x}$  yields

$$(4) \quad \begin{aligned} [F'']_{\delta, \alpha_1, \alpha_2} &= \sum_{j=1}^q [F'_{[q,j]}]_{\delta, \gamma} \cdot [F''_j]_{\gamma, \beta_1, \beta_2} \cdot [F'_{[j-1,0]}]_{\beta_1, \alpha_1} \cdot [F'_{[j-1,0]}]_{\beta_2, \alpha_2} \\ &= \sum_{j=1}^q \left[ \prod_{i=j+1}^q F'_i \right]_{\delta, \gamma} \cdot [F''_j]_{\gamma, \beta_1, \beta_2} \cdot \left[ \prod_{k=1}^{j-1} F'_k \right]_{\beta_1, \alpha_1} \cdot \left[ \prod_{k=1}^{j-1} F'_k \right]_{\beta_2, \alpha_2} \cdot \end{aligned}$$

We use index notation for tensor products. Tensors are enclosed in square brackets and summation runs over the common index. Jacobians and Hessians of subchains of Equation (2) are denoted as

$$F'_{[i,j]} \equiv \frac{dF_{[i,j]}}{d\mathbf{v}_j} \in \mathbb{R}^{n_i \times n_j} \quad \text{and} \quad F''_{[i,j]} \equiv \frac{d^2F_{[i,j]}}{d\mathbf{v}_j^2} \in \mathbb{R}^{n_i \times n_j \times n_j} .$$

In the following, we use the simplified notation

$$F'' = \sum_{j=1}^q \left( F'_{[q,j]} \cdot F''_j \cdot F'_{[j-1,0]} \otimes F'_{[j-1,0]} \right) ,$$

where  $\otimes$  denotes the outer product of two matrices as defined in Equation (4).

Equations (3) and (4) imply a range of combinatorial optimization problems due to the associativity of multiplication over the given elemental Jacobians and Hessians. Different orders of application of the chain rule during the evaluation of Equation (4) yield varying computational complexities in terms of the number of scalar *fused multiply-add* (fma) operations required. The minimization of this cost can be stated formally as the following decision problem.

**Problem 1.1** (HESSIAN ACCUMULATION). *Given are a layered twice differentiable function  $F$  as in Equation (2) together with elemental Jacobians  $F'_i$  and Hessians  $F''_i$  for  $i = 1, \dots, q$  and a non-negative integer  $K$ .*

*Is there a sequence of fma operations of length less than or equal to  $K$  over the individual entries of the given elemental Jacobians and Hessians which results in the evaluation of the Hessian  $F''$  of  $F$ ?*

**Theorem 1.** HESSIAN ACCUMULATION is NP-complete.

The proof of Theorem 1 builds on the same fundamental ideas as similar arguments presented in [25]. It uses reduction from ENSEMBLE COMPUTATION (EC) which was shown to be NP-complete in [9]. EC can be stated as follows:

Given is a collection  $C = \{C_\nu \subseteq A : \nu = 1, \dots, |C|\}$  of subsets  $C_\nu = \{c_i^\nu : i = 1, \dots, |C_\nu|\}$  of a finite set  $A$  and a non-negative integer  $K$ . Is there a sequence  $u_i = s_i \cup t_i$  for  $i = 1, \dots, k$  of  $k \leq K$  union operations, where each  $s_i$  and  $t_i$  is either  $\{a\}$  for some  $a \in A$  or  $u_j$  for some  $j < i$ , such that  $s_i$  and  $t_i$  are disjoint for  $i = 1, \dots, k$  and such that for every subset  $C_\nu \in C$ ,  $\nu = 1, \dots, |C|$ , there is some  $u_i$ ,  $1 \leq i \leq k$ , that is identical to  $C_\nu$ ?

**Example 1.1** (from [25]). *Consider an instance  $(A, C, K)$  of EC with  $A = \{a_1, a_2, a_3, a_4\}$ ,  $C = \{\{a_1, a_2\}, \{a_2, a_3, a_4\}, \{a_1, a_3, a_4\}\}$  and  $K = 4$ . It yields a positive answer to the decision problem as  $C_1 = u_1 = \{a_1\} \cup \{a_2\}$ ,  $u_2 = \{a_3\} \cup \{a_4\}$ ,  $C_2 = u_3 = \{a_2\} \cup u_2$  and  $C_3 = u_4 = \{a_1\} \cup u_2$ .*

*Proof.* (Theorem 1) Consider an arbitrary instance  $(A, C, K)$  of EC and a bijection  $A \leftrightarrow \tilde{A}$ , where  $\tilde{A}$  consists of  $|A|$  mutually distinct primes. A corresponding bijection  $C \leftrightarrow \tilde{C} \equiv \{\tilde{C}_\nu\}_{\nu=1, \dots, |C|}$  is implied. Create an extension  $(\tilde{A} \cup \tilde{B}, \tilde{C}, K + |\tilde{B}|)$  by adding unique entries from a sufficiently large set  $\tilde{B}$  of primes not in  $\tilde{A}$  to the  $\tilde{C}_\nu$  such that they all have the same cardinality  $|\tilde{C}_\nu| = q$ ,  $\nu = 1, \dots, |\tilde{C}|$ . Note that a solution for this extended instance of EC implies a solution of the original instance of EC as each entry of  $\tilde{B}$  appears exactly once.

Fix the order of the elements of the  $\tilde{C}_\nu$  arbitrarily yielding  $\tilde{C}_\nu = \{\tilde{c}_i^\nu\}_{i=1,\dots,q}$  for  $\nu = 1, \dots, |\tilde{C}|$ . Let

$$F : \mathbb{R} \rightarrow \mathbb{R}^{|\tilde{C}|} : \mathbf{y} = \mathbf{z}_q = F(x)$$

with  $F = F_q \circ F_{q-1} \circ \dots \circ F_1$  defined as

$$\begin{aligned} F_1 : \mathbb{R} \rightarrow \mathbb{R}^{|\tilde{C}|} : \mathbf{z}_1 = F_1(x) : z_j^1 &= \frac{\tilde{c}_1^j}{2} \cdot x^2 \\ F_i : \mathbb{R}^{|\tilde{C}|} \rightarrow \mathbb{R}^{|\tilde{C}|} : \mathbf{z}_i = F_i(\mathbf{z}_{i-1}) : z_j^i &= \tilde{c}_i^j \cdot z_j^{i-1} \end{aligned}$$

yielding

$$F_1' = \left( \tilde{c}_1^j \cdot x \right) \in \mathbb{R}^{|\tilde{C}|} = \mathbb{R}^{|\tilde{C}| \times 1} \quad \text{and} \quad F_1'' = \left( \tilde{c}_1^j \right) \in \mathbb{R}^{|\tilde{C}|} = \mathbb{R}^{|\tilde{C}| \times 1 \times 1}$$

as well as diagonal Jacobians

$$F_i' = (d_{j,k}^i) \in \mathbb{R}^{|\tilde{C}| \times |\tilde{C}|},$$

where

$$d_{j,k}^i = \begin{cases} \tilde{c}_i^j & \text{if } j = k \\ 0 & \text{otherwise,} \end{cases}$$

and vanishing Hessians  $F_i'' = 0$  for  $j = 1, \dots, |\tilde{C}|$  and  $i = 2, \dots, q$ . Equation (4) simplifies to

$$F'' = \prod_{i=2}^q F_i' \cdot F_1''.$$

According to the fundamental theorem of arithmetic [10] the elements of  $\tilde{C}$  correspond to unique (up to commutativity of scalar multiplication) factorizations of the  $|\tilde{C}|$  nonzero entries of  $F'' \in \mathbb{R}^{|\tilde{C}|} = \mathbb{R}^{|\tilde{C}| \times 1 \times 1}$ . This uniqueness property extends to arbitrary subsets of the  $\tilde{C}_j$  considered during the exploration of the search space of the HESSIAN ACCUMULATION problem. A solution implies a solution of the associated extended instance of EC and, hence, of the original instance of EC.

A proposed solution for HESSIAN ACCUMULATION is easily validated by counting the at most  $|\tilde{C}| \cdot q$  scalar multiplications performed.  $\square$

### Example 1.2.

$$A = \{a_1, a_2, a_3, a_4\} \Rightarrow \tilde{A} = \{2, 3, 5, 7\}$$

$$\tilde{B} = \{11\}$$

$$C = \{\{a_1, a_2\}, \{a_2, a_3, a_4\}, \{a_1, a_3, a_4\}\} \Rightarrow \tilde{C} = \{\{2, 3, 11\}, \{3, 5, 7\}, \{2, 5, 7\}\}$$

$$K + |\tilde{B}| = K + 1 = 5 .$$

The three nonzero entries of

$$F'' = F'_3 \cdot F'_2 \cdot F''_1 = \begin{pmatrix} 11 & & \\ & 7 & \\ & & 7 \end{pmatrix} \cdot \begin{pmatrix} 3 & & \\ & 5 & \\ & & 5 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 3 \\ 2 \end{pmatrix}$$

are computed as

$$\begin{aligned} \frac{dz_1^2}{dx} &= \frac{dz_1^2}{dz_1^1} \cdot \frac{d^2z_1^1}{dx^2} = 3 \cdot 2 = 6 \\ \frac{dz_2^3}{dz_2^1} &= \frac{dz_3^3}{dz_3^1} = \frac{dz_3^3}{dz_3^2} \cdot \frac{dz_3^2}{dz_3^1} = 7 \cdot 5 = 35 \\ F''_1 &= \tilde{b}_1 \cdot \left( \frac{dz_1^2}{dz_1^1} \cdot \frac{d^2z_1^1}{dx^2} \right) = 11 \cdot 6 = 66 \\ F''_2 &= \frac{dz_2^3}{dz_2^1} \cdot \frac{d^2z_2^1}{dx^2} = 35 \cdot 3 = 105 \\ F''_3 &= \frac{dz_3^3}{dz_3^1} \cdot \frac{d^2z_3^1}{dx^2} = 35 \cdot 2 = 70 . \end{aligned}$$

at the expense of five `fma` (no additions involved) yielding a positive answer to this instance of the decision version of HESSIAN ACCUMULATION. A corresponding answer to the decision version of EC is implied.

The proof of Theorem 1 exploits potential algebraic dependences among the entries of the elemental Hessians (equality in particular). The following heuristic assumes these entries to be mutually independent (distinct). Insight gained from the proof does not enter the heuristic presented in this paper. Moreover, it remains open if the purely structural version of HESSIAN ACCUMULATION, which ignores algebraic dependences among the entries of the elemental Hessians, is also NP-complete. We conjecture this to be case.

The dynamic programming [3, 14] method to be proposed targets HESSIAN CHAIN BRACKETING formally defined as a combinatorial optimization problem as follows:

**Problem 1.2** (HESSIAN CHAIN BRACKETING). *Given a layered twice differentiable function as in Equation (2) together with elemental Jacobians*

and Hessians, determine a bracketing of Equation (2) such that the number of `fma` operations required by Equation (4) becomes minimal.

**Example 1.3.** To illustrate the potential for optimized instances of HESSIAN CHAIN BRACKETING consider  $F = F_3 \circ F_2 \circ F_1$  such that  $F_3, F_1 \in \mathbb{R}^n \rightarrow \mathbb{R}^m$  and  $F_2 \in \mathbb{R}^m \rightarrow \mathbb{R}^n$ . Hence,  $F'_3, F'_1 \in \mathbb{R}^{m \times n}$ ,  $F_2 \in \mathbb{R}^{n \times m}$  and  $F''_3, F''_1 \in \mathbb{R}^{m \times n \times n}$ ,  $F_2 \in \mathbb{R}^{n \times m \times m}$ . Without loss of generality, all elemental Jacobians and Hessians are assumed to be dense. Tracking of highly likely sparsity would complicate the presentation of the example while not offering any further conceptual insight.

There are two ways to split  $F$  yielding the following `fma` costs

- $F = F_3 \circ (F_2 \circ F_1)$ : From  $F'' = F''_{[3,0]} = F'_3 \cdot F''_{[2,0]} + F''_3 \cdot F'_{[2,0]} \otimes F'_{[2,0]}$  with  $F'_{[2,0]} = F'_2 \cdot F'_1$  and  $F''_{[2,0]} = F''_2 \cdot F'_1 + F'_2 \cdot F''_1 \otimes F'_1$  it follows that  $\text{fma}(F'_{[2,0]}) = mn^2$  and  $\text{fma}(F''_{[2,0]}) = 2mn^3 + m^2n^2$  and hence  $\text{fma}(F'') = \text{fma}(F''_{[3,0]}) = 5mn^3 + m^2n^2 + mn^2$ .
- $F = (F_3 \circ F_2) \circ F_1$ : From  $F'' = F''_{[3,0]} = F''_{[3,1]} \cdot F'_1 + F'_{[3,1]} \cdot F'_1 \otimes F'_1$  with  $F'_{[3,1]} = F'_3 \cdot F'_2$  and  $F''_{[3,1]} = F''_3 \cdot F'_2 + F'_3 \cdot F''_2 \otimes F'_2$  it follows that  $\text{fma}(F'_{[3,1]}) = m^2n$  and  $\text{fma}(F''_{[3,1]}) = 2m^3n + m^2n^2$  and hence  $\text{fma}(F'') = \text{fma}(F''_{[3,0]}) = 3m^3n + 3m^2n^2 + m^2n$ .

The cost of bracketing from the right grows as  $n^3$  and  $m^2$ . The opposite holds for the cost of bracketing from the left growing as  $m^3$  and  $n^2$ . Linear growth of the discrepancy suggests significant potential for further analysis of HESSIAN CHAIN BRACKETING. For example,  $n = 2$  and  $m = 1$  yield costs of `48fma` and `20fma` when bracketing from right and left. Further results presented in Section 4 suggest that the theoretical savings also yield corresponding speedups when evaluating the Hessian chains numerically.

The efficient evaluation of Hessians has been investigated actively in the context of AD since the 1970s [33]. Particular focus has been set on the detection [4] and exploitation of structure [15, 20] and sparsity [35]. More recent contributions include [16] and [31].

This paper draws its motivation from our ongoing research into uncertainty quantification for pruned (e.g., deep feed-forward) neural networks. Such surrogate models are trained to replace a given computationally expensive implementation of a multivariate vector function  $F$  as in Equation (1) such that test data generated by sampling the relevant subdomain of  $F$  is reproduced sufficiently well. The argument  $\mathbf{x}$  of the at least twice continuously differentiable surrogate model  $\tilde{F}$  is assumed to follow a random distribution with known moments up to fourth order (mean  $\mu_{\mathbf{x}}$ , variance  $\sigma_{\mathbf{x}}$ , skewness  $S_{\mathbf{x}}$

and kurtosis  $K_{\mathbf{x}}$ ). We are interested in the mean  $\mu_{\mathbf{y}}$  a the variance  $\sigma_{\mathbf{y}}$  of the (randomly distributed) result  $\mathbf{y} = \tilde{F}(\mathbf{x})$  assuming stochastic independence of the entries of  $\mathbf{x}$ .

The first- and second-(and higher-)order moments can hence be approximated by a second-order Taylor series expansion of  $\tilde{F}$ , that is,

$$\begin{aligned} \tilde{F}_{\nu}(\mathbf{x}) \approx & \tilde{F}_{\nu}(\mu_{\mathbf{x}}) + \sum_{i=0}^{n-1} \frac{\partial \tilde{F}_{\nu}(\mu_{\mathbf{x}})}{\partial x_i} \cdot (x_i - \mu_{\mathbf{x}_i}) \\ & + \frac{1}{2} \cdot \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \frac{\partial^2 \tilde{F}_{\nu}(\mu_{\mathbf{x}})}{\partial x_i \partial x_j} \cdot (x_i - \mu_{\mathbf{x}_i}) \cdot (x_j - \mu_{\mathbf{x}_j}) \end{aligned}$$

for  $\nu = 0, \dots, m - 1$ . It follows that

$$\mu_{\mathbf{y}_{\nu}} \approx \tilde{F}_{\nu}(\mu_{\mathbf{x}}) + \frac{1}{2} \cdot \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \frac{\partial^2 \tilde{F}_{\nu}(\mu_{\mathbf{x}})}{\partial x_i^2} \cdot \sigma_{\mathbf{x}_i}^2$$

and

$$\begin{aligned} \sigma_{\mathbf{y}_{\nu, \kappa}}^2 \approx & \sum_{i=0}^{n-1} \frac{\partial \tilde{F}_{\nu}(\mu_{\mathbf{x}})}{\partial x_i} \cdot \frac{\partial \tilde{F}_{\kappa}(\mu_{\mathbf{x}})}{\partial x_i} \cdot \sigma_{\mathbf{x}_i}^2 \\ & + \frac{1}{2} \cdot \sum_{i=0}^{n-1} \left( \frac{\partial \tilde{F}_{\nu}(\mu_{\mathbf{x}})}{\partial x_i} \cdot \frac{\partial^2 \tilde{F}_{\kappa}(\mu_{\mathbf{x}})}{\partial x_i^2} + \frac{\partial^2 \tilde{F}_{\nu}(\mu_{\mathbf{x}})}{\partial x_i^2} \cdot \frac{\partial \tilde{F}_{\kappa}(\mu_{\mathbf{x}})}{\partial x_i} \right) \cdot S_{\mathbf{x}_i} \cdot \sigma_{\mathbf{x}_i}^3 \\ & + \frac{1}{2} \cdot \sum_{i=0}^{n-1} \sum_{j=0, j \neq i}^{n-1} \frac{\partial^2 \tilde{F}_{\nu}(\mu_{\mathbf{x}})}{\partial x_i \partial x_j} \cdot \frac{\partial^2 \tilde{F}_{\kappa}(\mu_{\mathbf{x}})}{\partial x_i \partial x_j} \cdot \sigma_{\mathbf{x}_i}^2 \cdot \sigma_{\mathbf{x}_j}^2 \\ & + \frac{1}{4} \cdot \sum_{i=0}^{n-1} \frac{\partial^2 \tilde{F}_{\nu}(\mu_{\mathbf{x}})}{\partial x_i^2} \cdot \frac{\partial^2 \tilde{F}_{\kappa}(\mu_{\mathbf{x}})}{\partial x_i^2} \cdot (K_{\mathbf{x}_i} - 1) \cdot \sigma_{\mathbf{x}_i}^4 ; \end{aligned}$$

see [2, 24] and references therein, including [8, 11, 30]. The full Hessian tensor of  $\tilde{F}$  is required. Alternatively, the computationally expensive terms are sometimes omitted in practice, which is what we would like to avoid as much as possible.

Over-parameterized neural networks are known to exhibit improved robustness with regard to the prevention of over-fitting [22, 36]. Subsequent pruning [1, 6] results in a post-training reduction of the dimension of the parameter space. Better run time efficiency of the surrogate model  $\tilde{F}$  can be achieved. Moreover,  $\tilde{F}$  becomes a layered multivariate vector function as in

Equation (2) with rectangular dense local Jacobians of varying dimensions. The HESSIAN CHAIN BRACKETING problem follows immediately.

The upcoming material is organized as follows: A dynamic programming algorithm for HESSIAN CHAIN BRACKETING is proposed in Section 2 including a detailed illustration of the individual steps performed by the algorithm for the simple example introduced above. A case study from computational finance is discussed in Section 3. Numerical results presented in Section 4 show potential reductions of the operations count over the obvious approaches (bracketing from left or right) by factors of ten and more on a set of sample problems of growing size. The savings are shown to translate into actual improvements in runtime. All results can be reproduced with the open-source reference implementation presented in the appendix. Conclusions drawn in Section 5 are complemented with remarks on ongoing and future research and development. A sample session of our proof-of-concept implementation of the dynamic programming algorithm from Section 2 is provided as an appendix.

## 2. Dynamic programming

The number of bracketings of  $F = F_{[q,0]}$  is known to be equal to  $\frac{1}{q} \binom{2(q-1)}{q-1} = \frac{(2q-2)!}{q!(q-1)!}$  [7], which grows exponentially with  $q$ . Subproblems are defined by recursive bisection as

$$F_{[i,k]} = F_{[i,j]} \circ F_{[j,k]} = (F_i \circ \dots \circ F_{j+1}) \circ (F_j \circ \dots \circ F_k)$$

for  $i = 1, \dots, q$ ,  $i - k = 1, \dots, q$  and  $k < j < i$ . An fma-optimal bracketing of the Jacobian chain product in Equation (3) can be computed by dynamic programming. Solutions to subproblems of growing length  $i - k$  are tabulated as

$$\mathbf{fma}(F'_{[i,k]}) = \min_{k < j < i} (\mathbf{fma}(F'_{[i,j]}) + \mathbf{fma}(F'_{[j,k]}) + \mathbf{fma}(F'_{[i,j]} \cdot F'_{[j,k]})) .$$

The tabulated costs are used for the minimization of the numbers of fma operations required for the computations of the Hessians  $F''_{[i,k]} \in \mathbb{R}^{n_i \times n_k \times n_k}$  as follows:

$$\begin{aligned} \mathbf{fma}(F''_{[i,k]}) = \min_{k < j < i} & (\mathbf{fma}(F'_{[i,j]}) + \mathbf{fma}(F''_{[j,k]}) + \mathbf{fma}(F'_{[i,j]} \cdot F''_{[j,k]}) \\ & + \mathbf{fma}(F''_{[i,j]}) + \mathbf{fma}(F'_{[j,k]}) + \mathbf{fma}(F''_{[i,j]} \cdot F'_{[j,k]} \otimes F_{[j,k]})) . \end{aligned}$$



Correctness of the algorithm follows immediately from the *optimal substructure* and *overlapping subproblems* properties [3] exhibited by both JACOBIAN and HESSIAN CHAIN BRACKETING.

**Example 2.1.** We use the same example as in Section 1 with  $n = 2$  and  $m = 1$  for illustration of the individual steps of the dynamic programming algorithm, that is,  $F = F_3 \circ F_2 \circ F_1$  such that  $F_1, F_3 : \mathbb{R}^2 \rightarrow \mathbb{R}$  and  $F_2 : \mathbb{R} \rightarrow \mathbb{R}^2$ . Again, and without loss of generality we assume elemental Jacobians and Hessians to be dense. The number of **fma** required for the product of matrix with a vector is invariant with respect to potential symmetry of the matrix. Hence, the exploitation of likely symmetry of the Hessians does not lead to a reduction in the **fma**-cost.

For a function composition of length three there are only two choices corresponding to bracketing from the left at the computational cost of  $20\mathbf{fma}$  or bracketing from the right at  $48\mathbf{fma}$ . The algorithm favors the former as the result of performing the following steps:

The optimal bracketings of all Jacobian subchains are computed as

$$\begin{aligned} \mathbf{fma}(F'_{[2,0]}) &= n \cdot m \cdot n = m \cdot n^2 = 4; \quad \mathbf{fma}(F'_{[3,1]}) = m \cdot n \cdot m = m^2 \cdot n = 2 \\ \mathbf{fma}(F'_{[3,0]}) &= \min(\mathbf{fma}(F'_{[2,0]}) + \mathbf{fma}(F'_3 \cdot F'_{[2,0]}), \mathbf{fma}(F'_{[3,1]}) + \mathbf{fma}(F'_{[3,1]} \cdot F'_1)) \\ &= \min(m \cdot n^2 + m \cdot n \cdot n, m^2 \cdot n + m \cdot m \cdot n) \\ &= \min(2 \cdot m \cdot n^2, 2 \cdot m^2 \cdot n) = \min(8, 4) = 4. \end{aligned}$$

The whole chain for is evaluated with minimal **fma** cost of four as  $F' = (F'_3 \cdot F'_2) \cdot F'_1$ .

Dynamic programming for HESSIAN CHAIN BRACKETING yields costs for the two subchains of length two as

$$\begin{aligned} \mathbf{fma}(F''_{[2,0]}) &= \mathbf{fma}(F'_2 \cdot F''_1) + \mathbf{fma}(F''_2 \cdot F'_1 \otimes F'_1) \\ &= n \cdot m \cdot n \cdot n + n \cdot m \cdot n \cdot (m + n) \\ &= m \cdot n^3 + m \cdot n^2 \cdot (m + n) = 8 + 12 = 20 \\ \mathbf{fma}(F''_{[3,1]}) &= \mathbf{fma}(F'_3 \cdot F''_2) + \mathbf{fma}(F''_3 \cdot F'_2 \otimes F'_2) \\ &= m \cdot n \cdot m \cdot m + m \cdot n \cdot m \cdot (m + n) \\ &= m^3 \cdot n + m^2 \cdot n \cdot (m + n) = 2 + 6 = 8 \end{aligned}$$

which are looked up during the optimization of  $\mathbf{fma}(F''_{[3,0]})$  as

$$\mathbf{fma}(F''_{[3,0]}) = \min_{0 < j < 3} (\mathbf{fma}(F'_{[3,j]} \cdot F''_{[j,0]}) + \mathbf{fma}(F''_{[3,j]} \cdot F'_{[j,0]} \otimes F'_{[j,0]}))$$

$$\begin{aligned}
&= \min( \\
&\quad \mathbf{fma}(F'_{[3,1]} \cdot F''_1) + \mathbf{fma}(F''_{[3,1]} \cdot F'_1 \otimes F'_1), \\
&\quad \mathbf{fma}(F'_3 \cdot F''_{[2,0]}) + \mathbf{fma}(F''_3 \cdot F'_{[2,0]} \otimes F'_{[2,0]}) \\
&\quad ) \\
&= \min( \\
&\quad \mathbf{fma}(F'_{[3,1]}) + \mathbf{fma}(F''_1) + m \cdot m \cdot n^2 \\
&\quad + \mathbf{fma}(F''_{[3,1]}) + \mathbf{fma}(F'_1) + m \cdot m \cdot n \cdot (m + n), \\
&\quad \mathbf{fma}(F'_3) + \mathbf{fma}(F''_{[2,0]}) + m \cdot n \cdot n^2 \\
&\quad + \mathbf{fma}(F''_3) + \mathbf{fma}(F'_{[2,0]}) + m \cdot n \cdot n \cdot (n + n) \\
&\quad ) \\
&= \min( \\
&\quad (m^2 \cdot n) + 0 + (m^2 \cdot n^2) \\
&\quad + (m^3 \cdot n + m^2 \cdot n \cdot (m + n)) \\
&\quad + 0 + (m^2 \cdot n \cdot (m + n)), \\
&\quad 0 + (m \cdot n^3 + m \cdot n^2 \cdot (m + n)) + m \cdot n^3 \\
&\quad + 0 + n^2 + 2 \cdot m \cdot n^3 \\
&\quad ) \\
&= \min(20, 48) = 20 .
\end{aligned}$$

*This result validates the observations made in Section 1.*

A detailed illustration of the dynamic programming algorithm for the composite function  $F = F_4 \circ F_3 \circ F_2 \circ F_1$  can be found in Figure 1 with further comments provided in the corresponding caption. The solution to HESSIAN CHAIN BRACKETING for  $F_1 : \mathbb{R}^2 \rightarrow \mathbb{R}^5$ ,  $F_2 : \mathbb{R}^5 \rightarrow \mathbb{R}$ ,  $F_3 : \mathbb{R} \rightarrow \mathbb{R}^3$ ,  $F_4 : \mathbb{R}^3 \rightarrow \mathbb{R}^2$  is computed based on  $F = (F_4 \circ F_3) \circ (F_2 \circ F_1)$  with a total cost of 156fma required for the accumulation of the Hessian  $F''$ . Again and without loss of generality, all elemental Jacobians and Hessians are regarded as dense.

### 3. Case study

As a real-world case study we consider the LIBOR<sup>1</sup> market model introduced in [5] and used in [12] as illustration of the benefits of adjoint AD for

---

<sup>1</sup>London Interbank Offered Rate

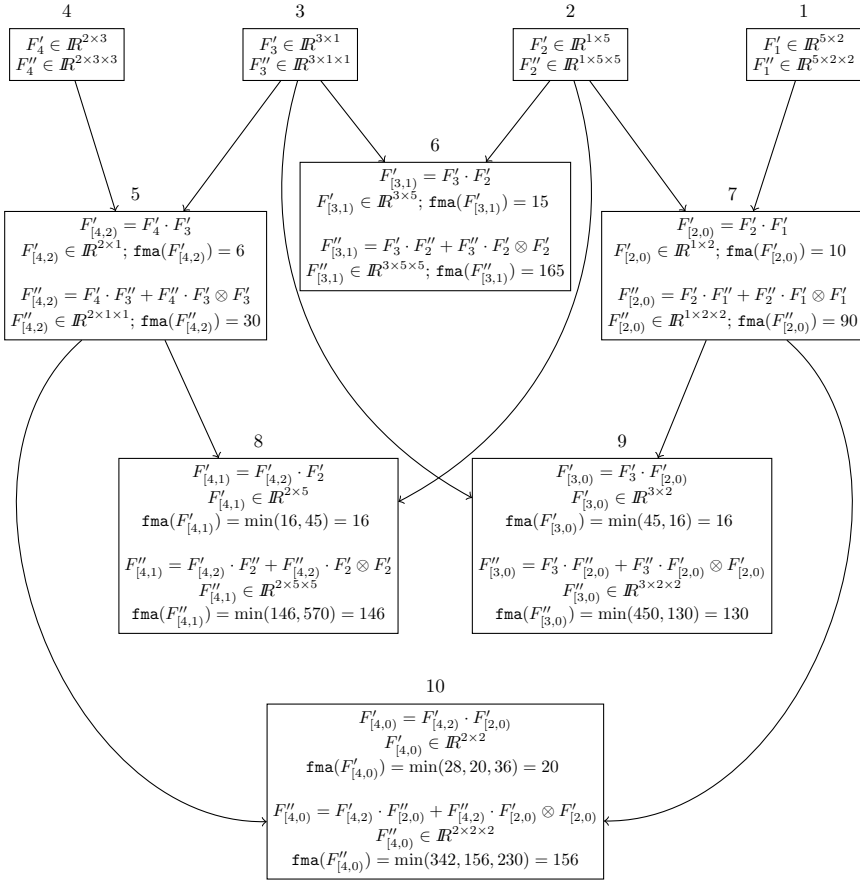


Figure 1: Dynamic programming for HESSIAN CHAIN BRACKETING: The algorithm is visualized as a directed acyclic graph for  $F = F_4 \circ F_3 \circ F_2 \circ F_1$  with  $F_4 : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ ,  $F_3 : \mathbb{R} \rightarrow \mathbb{R}^3$ ,  $F_2 : \mathbb{R}^5 \rightarrow \mathbb{R}$ ,  $F_1 : \mathbb{R}^2 \rightarrow \mathbb{R}^5$ . Vertices 1 to 4 (5 to 7 | 8 to 9) correspond to subchains of length one (two | three). The optimal bracketing is represented by vertex 10. Vertices contain information on the computation of Jacobians and Hessians of the respective subchains corresponding to the optimal bracketing. The dimensions of the resulting Jacobians and Hessian are stated as well as the numbers of **fma** required for their computation. Edges visualize split positions by linking a chain to its two subchains according to an optimal bracketing. For example,  $F''_{[4,0]}$  is computed optimally based on the bracketing  $(F_4 \circ F_3) \circ (F_2 \circ F_1)$  at the cost of  $\text{fma}(F''_{[4,2]}) + \text{fma}(F'_{[4,2]}) + \text{fma}(F''_{[2,0]}) + \text{fma}(F'_{[2,0]}) + n_4 n_2 n_0^2 + n_4 n_2^2 n_0 + n_4 n_2 n_0^2 = 30 + 6 + 90 + 10 + 8 + 4 + 8 = 156 \text{fma}$ .

simulations in finance. Over recent years adjoint AD has gained significant importance in computational finance driven mainly by increasing gradient sizes in the context of XVA calculations and documented by a large number of related publications, e.g, [32, 23]. Considerable effort has been going into the training of surrogate models based on artificial neural networks (ANN) [21].

The LIBOR sample code simulates the evolution of the LIBOR rates for a portfolio of swaptions with given swap rates and maturities. As in [12], swaps of the floating forward rate  $L \in \mathbb{R}^n$  and a given fixed swap rate are considered for  $n = 80$ . Monte Carlo simulation with a normally distributed random variable  $Z \in \mathbb{R}^{p \times m}$  performs  $p$  path calculations evolving  $L = L(t)$  for  $m$  time steps to the target time  $t = T$  and starting from a given initial state  $L(0)$ . Refer to [13] for further discussion of the mathematical details behind the LIBOR market model. All numerical results obtained by our implementation were validated against the implementation used in [12].

On the given computer the run time of  $p = 10^4$  primal Monte Carlo path simulations is 1.9s. We consider the accumulation of the Hessian  $\frac{d^2 L(T)}{dL(0)^2} \in \mathbb{R}^{80 \times 80 \times 80}$  based on a surrogate model in form of an ANN with 11 layers and 80 nodes per layer trained to 99% accuracy in terms of mean squared error. Subsequent pruning eliminates insignificant nodes from hidden layers as described in [1] and based on the results an interval adjoint significance analysis introduced in [34]. A layered function is generated as in Equation (2) with  $q = 11$  and  $n_0 = 80$ ,  $n_1 = 32$ ,  $n_2 = 65$ ,  $n_3 = 64$ ,  $n_4 = 55$ ,  $n_5 = 46$ ,  $n_6 = n_7 = 49$ ,  $n_8 = 53$ ,  $n_9 = 62$ ,  $n_{10} = 48$ ,  $n_{11} = 80$ . The pruned ANN preserves the 99% target accuracy on the given test set.

Based on the measured primal runtime of 1.9s the accumulation of the Hessian in second-order tangent mode of AD is estimated to take approximately  $1.5 \cdot 80^2 \cdot 1.9 = 18,240s$  or 5 hours. The factor 1.5 is due to the overhead of a tangent (directional derivative) propagation induced by our AD library dco/c++ [29]. A total of  $80^2$  tangents need to be evaluated.

The runtime of the surrogate is negligible (a few milliseconds; *ms*). So is the cost of evaluation of the elemental Hessians (a few seconds). Our runtime measurements assume the latter to be given. Different bracketing of Equation (2) are compared. Bracketing from the left [right] performs 388,844,400fma [517,283,120fma] in 855ms [1,125ms]. A greedy heuristic based on locally optimal decisions results in 298,631,368fma taking 638ms. Dynamic programming yields an optimal bracketing with 149,061,728fma performed in 311ms. The reduction in the number of fma by a factor of almost three carries over to the runtime. The optimal bracketing evaluates

Table 1: Random Test Cases

$q$	optimized bracketing	bracketing from left	... from right	$\mathbf{fma}_{\text{rel}}$
3	20	20	48	<b>1</b>
4	156	342	230	<b>1.47</b>
5	1218	2210	1860	<b>1.52</b>
10	11,220	53,118	20,952	<b>1.86</b>
15	10,200	217,555	51,152	<b>5.01</b>
20	56,830	1,057,665	1,053,117	<b>18.53</b>
25	286,366	3,953,376	1,345,312	<b>4.69</b>
30	195,620	1,655,596	8,615,838	<b>8.46</b>
35	614,499	23,461,452	4,440,564	<b>7.22</b>
40	2,254,794	24,388,365	66,718,064	<b>10.81</b>
45	1,787,606	28,170,189	139,760,800	<b>15.75</b>
50	8,271,082	170,383,616	65,760,913	<b>7.95</b>

Table 2: Run times for Larger Random Test Cases

$q$	optimized bracketing	best unidirectional bracketing	$\mathbf{fma}_{\text{rel}}$	$t_{\text{rel}}$
60	37,989,141	331,166,304	8.71	<b>5.54</b>
70	9,241,074	554,021,568	59.95	<b>14.05</b>
80	33,796,544	200,622,919	5.93	<b>2.44</b>
90	19,058,174	251,632,865	13.20	<b>2.36</b>
100	22,951,156	2,313,419,043	100.79	<b>16.20</b>
150	502,245,226	16,928,546,112	33.70	<b>13.53</b>
200	816,938,109	4,620,627,490	5.65	<b>1.42</b>

the Hessian based on

$$F = (F_{11} \circ (F_{10} \circ (F_9 \circ (F_8 \circ (F_7 \circ (F_6 \circ (F_5 \circ (F_4 \circ (F_3 \circ F_2)))))))))) \circ F_1 .$$

All results can be reproduced (run times qualitatively) using the reference implementation described in the appendix.

#### 4. Further numerical results

Table 1 lists the results obtained by applying the dynamic programming heuristic for HESSIAN CHAIN BRACKETING to chains of elemental functions of growing length  $q$ . The latter also serves as an upper bound for the randomly generated dimensions of domains and images of the individual elemental functions. We compare the numbers of  $\mathbf{fma}$  required for the accumulation of the Hessian when bracketing from the left or from the right with the numbers resulting from optimized bracketing. The factor quantifying the

$q$	$\mathbf{fma}_{\text{rel}}$	$t_{\text{DP}} (s)$
500	83	<b>0.6</b>
1000	1085	<b>7</b>
1500	2555	<b>15</b>
2000	498	<b>70</b>
2500	291	<b>150</b>
3000	74	<b>285</b>

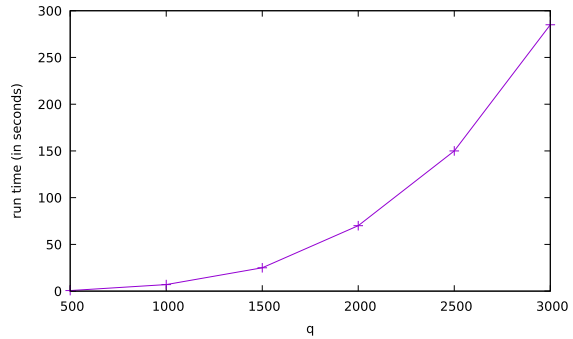


Figure 2: Run Times of the Dynamic Programming Algorithm.

improvement due to optimized bracketing over the better out of the uniform bracketings is shown in the last column. Relative savings in the  $\mathbf{fma}$  count of up to eighteen can be observed.

Savings in the number of  $\mathbf{fma}$  required for the accumulation of the Hessian can be expected to yield adequate reductions in runtime. A set of larger problem instances is presented for this purpose in Table 2. Relative savings in the  $\mathbf{fma}$  count of up to one hundred result in speedups of up to sixteen as shown in the last column. Our reference implementation is not tuned for speed. It uses Eigen<sup>2</sup> for the matrix products. While we consider this approach to be a realistic scenario further optimization is likely to yield even better efficiency. For example, the use of GPGPU has been shown to be beneficial [17, 18].

The computational complexity of the dynamic programming algorithm is cubic in  $q$ . For illustration we report and plot run times<sup>3</sup> for even larger random instances in Figure 2. Relative savings in the  $\mathbf{fma}$  count become even more pronounced. Again, these results can be reproduced with the help of the reference implementation described in the appendix.

## 5. Conclusion and outlook

The results presented in this paper are promising. Reductions in the number of  $\mathbf{fma}$  required for the accumulation of Hessian tensors yield corresponding speedups. Nevertheless, significant effort is required to bridge the present gap

<sup>2</sup><https://eigen.tuxfamily.org>

<sup>3</sup>All numerical experiments were performed on a standard PC with Intel<sup>®</sup> CORE i7 processor.

to seamless integration into software tools for AD. A matrix-free formulation in particular is necessary to handle computationally complex elemental functions similar to the first-order scenario investigated in [27]. The assumption about elemental Hessians being given turns out to be infeasible in many practical applications. ANN represent an exception as differentiation of the individual layers often turns out to be relatively straightforward.

Dynamic programming for JACOBIAN and HESSIAN CHAIN BRACKETING generalizes to arbitrary order. So does the proof of NP-completeness of JACOBIAN and HESSIAN ACCUMULATION as shown in [28]. The obvious discrepancies between the respective formulations give rise to further ongoing investigations into the combinatorics induced by the chain rule of differentiation.

## Appendix A. Implementation

An open-source reference implementation is provided for easy reproduction of our computational results; see

`git@github.com:un110076/HessianChainBracketing.git` .

The software consists of three separate executables resulting from implementations given as three C++ source files. Problem instances are generated randomly by `generate.exe` for given length of the chain and upper bound on the dimensions of domains and images of the elemental functions. The resulting text file serves as input for `solve.exe` which computes a solution for the corresponding instance of (dense) HESSIAN CHAIN BRACKETING. Both the problem formulation and the solution can be passed to `run.exe` to perform the numerical evaluation of the Hessian chain product for given randomly initialized elemental Jacobians and Hessians. Eigen is expected to be installed in `./Eigen`. The code has been tested with the GNU C++ compiler under Linux. A `Makefile` is provided. Essential information on how to build and run the software is given in `README.md`.

A sample session could proceed as follows:

### 1. Running

```
generate.exe 4 4
```

might yield the output

```
4
5 2
```

```

1 5
3 1
2 3

```

corresponding to the example from Figure 1. The chain  $F_4 \circ F_3 \circ F_2 \circ F_1$  of length four (first line) consists of elemental functions  $F_1 : \mathbb{R}^2 \rightarrow \mathbb{R}^5$  (line two),  $F_2 : \mathbb{R}^5 \rightarrow \mathbb{R}$  (line three),  $F_3 : \mathbb{R} \rightarrow \mathbb{R}^3$  (line four),  $F_4 : \mathbb{R}^3 \rightarrow \mathbb{R}^2$  (line five). Let this output be stored in `problem.txt`.

2. The dynamic programming algorithm is executed as illustrated in Figure 1 by running

```
solve.exe problem.txt
```

Diagnostic output is generated.

```

left bracketing fma = 342
right bracketing fma = 230
heuristic bracketing fma = 156
optimized bracketing fma = 156

```

Dynamic Programming Table:

```

fma(F''(1,0))=90; split before 1; dim(F''(1,0))=1x2x2
fma(F''(2,1))=165; split before 2; dim(F''(2,1))=3x5x5
fma(F''(2,0))=130; split before 2; dim(F''(2,0))=3x2x2
fma(F''(3,2))=30; split before 3; dim(F''(3,2))=2x1x1
fma(F''(3,1))=146; split before 2; dim(F''(3,1))=2x5x5
fma(F''(3,0))=156; split before 2; dim(F''(3,0))=2x2x2

```

The number of `fma` required by the optimized bracketing is compared with the numbers resulting from uniform bracketing from the left and from the right as well as with the result of the greedy heuristic. Moreover, the optimized bracketing is stored in a text file `solution.txt` as follows:

```

3 3 2
1 1 0
3 2 0

```

Visiting the lines in reverse order we find that the first split position is set before  $F_3$  yielding  $(F_4 \circ F_3) \circ (F_2 \circ F_1)$ . The remaining two lines indicate (unique) split positions before  $F_2$  and before  $F_4$  within the two subchains (of length two).

3. Passing both `problem.txt` and `solution.txt` as command line arguments to `run.exe` as



```
run.exe problem.txt solution.txt heuristic_solution.txt
```

run times for the numerical evaluation of the uniform bracketings are compared with the run time of computing the Hessian based on the optimized bracketing yielding, for example,

```
Elapsed time (in microseconds):  
left bracketing: 69  
right bracketing: 52  
heuristic bracketing: 48  
optimized bracketing: 48
```

Obviously, the numbers become more reliable for larger problems.

## References

- [1] S. Afghan and U. Naumann. Interval adjoint significance analysis for neural networks. In V. et al. Krzhizhanovskaya, editor, *Computational Science – ICCS 2020*, pages 365–378. Springer International Publishing, 2020. [MR4152510](#)
- [2] M. Beckers. *Toward global robust optimization*. PhD thesis, RWTH Aachen University, 2014.
- [3] R. Bellman. *Dynamic Programming*. Dover Publications, 1957. [MR0090477](#)
- [4] S. Bhowmick and P. Hovland. A polynomial-time algorithm for detecting directed axial symmetry in Hessian computational graphs. In C. Bischof, M. Bücker, P. Hovland, U. Naumann, and J. Utke, editors, *Advances in Automatic Differentiation*, pages 91–102. Springer, 2008. [MR2531681](#)
- [5] A. Brace, D. Gatarek, and M. Musiela. The market model of interest rate dynamics. *Mathematical Finance*, 7:127–147, 1997. [MR1446645](#)
- [6] G. Castellano, A. Fanelli, and M. Pelillo. An iterative pruning algorithm for feed-forward neural networks. *IEEE Trans. Neural Networks*, 3(8):519–531, 1997. [MR2696096](#)
- [7] E. Catalan. Note sur une équation aux différences finies. *J. Math. Pures Appl.*, 3:508–516, 1838.
- [8] B. Christianson and M. Cox. Automatic propagation of uncertainties. In M. Bücker, G. Corliss, P. Hovland, U. Naumann, and B. Norris, editors,

- Automatic Differentiation: Applications, Theory, and Implementations*, Lecture Notes in Computational Science and Engineering, pages 47–58. Springer, 2005. [MR2194171](#)
- [9] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences)*. W. H. Freeman, first edition edition, 1979. [MR0519066](#)
- [10] C. Gauss and tr. A. Clarke. *Disquisitiones Arithmeticae*. Yale University Press, 1965. [MR0197380](#)
- [11] D. Ghate and M. Giles. Inexpensive Monte Carlo uncertainty analysis. In B. Uthup, S. Koruthu, R. K. Sharma, and P. Priyadarshi, editors, *Recent Trends in Aerospace Design and Optimization*, pages 203–210. Tata McGraw-Hill, New Delhi, 2006.
- [12] M. Giles and P. Glasserman. Smoking adjoints: Fast Monte Carlo Greeks. *Risk*, pages 88–92, January 2006.
- [13] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, 2003. [MR1999614](#)
- [14] S. Godbole. On efficient computation of matrix chain products. *IEEE Transactions on Computers*, C-22(9):864–866, Sep. 1973. [MR0351547](#)
- [15] D. Goldfarb and P. Toint. Optimal estimation of Jacobian and Hessian matrices that arise in finite difference calculations. *Mathematics of Computation*, 43(167):69–88, 1984. [MR0744925](#)
- [16] R. Gower and M. Mello. A new framework for the computation of Hessians. *Optimization Methods and Software*, 27(2):251–273, 2012. [MR2901960](#)
- [17] F. Gremse, A. Hoefter, L. Ole Schwen, F. Kiessling, and U. Naumann. GPU-accelerated sparse matrix-matrix multiplication by row merging. *SIAM J. Sci. Comp.*, 37(1):C54–C71, 2015. [MR3302599](#)
- [18] F. Gremse, A. Höfter, L. Razik, F. Kiessling, and U. Naumann. Gpu-accelerated adjoint algorithmic differentiation. *Computer Physics Communications*, 200:300–311, 2016.
- [19] A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Number 105 in Other Titles in Applied Mathematics. SIAM, Philadelphia, PA, 2nd edition, 2008. [MR2454953](#)

- [20] P. Hovland, C. Bischof, D. Spiegelman, and M. Casella. Efficient derivative codes through automatic differentiation and interface contraction: An application in biostatistics. *SIAM Journal on Scientific Computing*, 18(4):1056–1066, 1997. [MR1453557](#)
- [21] B. Huge and A. Savine. Differential machine learning: the shape of things to come. *Risk.net*, September 2020.
- [22] A. Lewkowycz and G. Gur-Ari. On the training dynamics of deep networks with  $l_2$  regularization. In *34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
- [23] D. Lu. *The XVA of Financial Derivatives: CVA, DVA and FVA Explained*. Springer, 2016.
- [24] M. Menshikova. *Uncertainty Estimation Using the Moments Method Facilitated by Automatic Differentiation in Matlab*. PhD thesis, Defence College of Management and Technology, Cranfield, United Kingdom, 2010.
- [25] U. Naumann. Optimal Jacobian accumulation is NP-complete. *Mathematical Programming, Ser. A*, 112(2):427–441, 2008. [MR2361931](#)
- [26] U. Naumann. *The Art of Differentiating Computer Programs: An Introduction to Algorithmic Differentiation*. Number 24 in Software, Environments, and Tools. SIAM, Philadelphia, PA, 2012. [MR2894688](#)
- [27] U. Naumann. Optimization of generalized Jacobian chain products without memory constraints. *arXiv:2003.05755*, 2020.
- [28] U. Naumann. On the computational complexity of the chain rule of differential calculus. *arXiv:2107.05355*, 2021.
- [29] U. Naumann, K. Leppkes, and J. Lotz. Derivative Code by Overloading in C++ (dco/c++): Introduction and Summary of Features. Technical Report AIB-2016-08, RWTH Aachen University, September 2016.
- [30] M. Padulo, S. Forth, and M. Guenov. Robust aircraft conceptual design using automatic differentiation in Matlab. In C. Bischof, M. Bücker, P. Hovland, U. Naumann, and J. Utke, editors, *Advances in Automatic Differentiation*, pages 271–280. Springer, 2008. [MR2531674](#)
- [31] C. Petra, F. Qiang, M. Lubin, and J. Huchette. On efficient Hessian computation using the edge pushing algorithm in Julia. *Optimization Methods & Software*, 33(4–6):1010–1029, 2018. [MR3853226](#)

- [32] A. Pfadler. Computing sensitivities of CVA using adjoint algorithmic differentiation. Master’s thesis, University of Oxford, 2015.
- [33] M. Powell and P. Toint. On the estimation of sparse Hessian matrices. *SIAM J. Numer. Anal.*, 16:1060–1074, 1979. [MR0551326](#)
- [34] V. Vassiliadis, J. Riehme, J. Deussen, K. Parasyris, C. Antonopoulos, N. Bellas, S. Lalisa, and U. Naumann. Towards automatic significance analysis for approximate computing. In *International Symposium on Code Generation and Optimization*, pages 182–193. IEEE/ACM, 2016.
- [35] A. Walther. Computing sparse Hessians with automatic differentiation. *ACM Transaction on Mathematical Software*, 34(1):3:1–3:15, 2008. [MR2404443](#)
- [36] G. Zhang, J. Martens, and R. Grosse. Fast convergence of natural gradient descent for overparameterized neural networks. In *33th Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

UWE NAUMANN  
DEPARTMENT OF COMPUTER SCIENCE  
RWTH AACHEN UNIVERSITY  
GERMANY  
*E-mail address:* [naumann@stce.rwth-aachen.de](mailto:naumann@stce.rwth-aachen.de)

SHUBHADITYA BURELA  
DEPARTMENT OF COMPUTER SCIENCE  
RWTH AACHEN UNIVERSITY  
GERMANY  
*E-mail address:* [shubhaditya.burela@rwth-aachen.de](mailto:shubhaditya.burela@rwth-aachen.de)

RECEIVED OCTOBER 26, 2021