ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Prof. Ryan Cotterell

Danilo Dordevic: Assignment 1

ddordevic@student.ethz.ch, 21-254-888.

01/11/2022 - 14:49h

Question 1

(a) The Hessian matrix of a given function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as the matrix of all of the function's second partial derivatives:

$$\boldsymbol{H}_f(\boldsymbol{x}) \triangleq \nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\boldsymbol{x}) \end{bmatrix} \tag{1}$$

The i-th column of the Hessian matrix is:

$$\boldsymbol{H}_f(\boldsymbol{x})_{:,i} \triangleq \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_i}(\boldsymbol{x}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_i}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_i}(\boldsymbol{x}) \end{bmatrix} \tag{2}$$

Noticing that j-th element of the vector 2 is a partial derivative with respect to variable x_j of the derivative function $\frac{\partial f}{\partial x_i}(\boldsymbol{x})$, we come to the conclusion that the vector 2 is nothing but the gradient of the said derivative function:

$$\boldsymbol{H}_f(\boldsymbol{x})_{:,i} = \nabla \left(\frac{\partial f}{\partial x_i} \right) (\boldsymbol{x}) \tag{3}$$

The derivative function $\frac{\partial f}{\partial x_i}(\boldsymbol{x})$ is the i-th element of the gradient of function $f(\boldsymbol{x})$. The i-th element of any vector $\boldsymbol{v} \in \mathbb{R}^n$, v_i, can be extracted by computing the inner product between the i-th standard basis vector $\boldsymbol{e}_i \in \mathbb{R}^n$ and the vector $\boldsymbol{v}$: $\boldsymbol{e}_i^\top \boldsymbol{v} = v_i$. Similarly, the partial derivative with respect to variable x_i of function $f(\boldsymbol{x})$ can be extracted from the derivative as such:

$$\nabla \left(\frac{\partial f}{\partial x_i} \right) (\boldsymbol{x}) = \nabla (\boldsymbol{e}_i^\top \nabla f(\boldsymbol{x})) \tag{4}$$

Finally, plugging 4 into 3 leads to:

$$\boldsymbol{H}_f(\boldsymbol{x})_{:,i} = \nabla \left(\frac{\partial f}{\partial x_i} \right) (\boldsymbol{x}) = \nabla(\boldsymbol{e}_i^\top \nabla f)(\boldsymbol{x}) \tag{5}$$

Equation 5 is a computation of just one column of the Hessian matrix 1. Concatenating all of the columns 5 into a matrix produces the following:

$$\boldsymbol{H}_f(\boldsymbol{x}) = \begin{bmatrix} | & | & & | \\ \nabla(\boldsymbol{e}_1^\top \nabla f)(\boldsymbol{x}) & \nabla(\boldsymbol{e}_2^\top \nabla f)(\boldsymbol{x}) & \cdots & \nabla(\boldsymbol{e}_n^\top \nabla f)(\boldsymbol{x}) \\ | & | & & | \end{bmatrix} \tag{6}$$

Moreover, noticing that the Hessian of a differentiable function is symmetric, i.e. it holds that $\boldsymbol{H}_f(x)^\top = \boldsymbol{H}_f(x)$. In other words, the rows and corresponding rows of the Hessian matrix are equal. Taking this into account, we can write:

$$\boxed{\boldsymbol{H}_f(x)_{:,i} = \boldsymbol{H}_f(x)_{i,:}^\top = (\nabla(\boldsymbol{e}_i^\top \nabla f(\boldsymbol{x}))^\top} \tag{7}$$

which is exactly what was asked for in the problem. Stacking these rows produces the Hessian matrix:

$$\boldsymbol{H}_f(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} — (\nabla(\boldsymbol{e}_1^\top \nabla f(\boldsymbol{x}))^\top — \\ — (\nabla(\boldsymbol{e}_2^\top \nabla f(\boldsymbol{x}))^\top — \\ \vdots \\ — (\nabla(\boldsymbol{e}_n^\top \nabla f(\boldsymbol{x}))^\top — \end{bmatrix} \tag{8}$$

To sum up, equation 7 is a representation of i-th row of the Hessian matrix. Equation 8 shows how stacking the rows represented in the form as in 7 produces the Hessian matrix. Equation 5 shows a possible representation of a i-th column of the Hessian matrix using standard basis vectors.

(b) Assuming that the computational graphs of the first partial derivatives of function $f : \mathbb{R}^n \to \mathbb{R}$, $\frac{\partial f}{\partial x_1}(\boldsymbol{x})$, $\frac{\partial f}{\partial x_2}(\boldsymbol{x}), \cdots, \frac{\partial f}{\partial x_n}(\boldsymbol{x})$, taking into account that computing the value of each partial derivative during the forward pass through its computational graph has computational complexity $O(m)$, and that the computational complexity of the backwards pass of the backpropagation algorithm is the same as the forward pass, it can be concluded that computing the gradients of each partial derivative takes $O(m)$ time.

The gradient of a partial derivative $\frac{\partial f}{\partial x_i}(\boldsymbol{x})$ constitues the i-th column of the Hessian $\nabla^2 f(x)$:

$$\nabla \left(\frac{\partial f}{\partial x_i} \right) (\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_i}(\boldsymbol{x}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_i}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_i}(\boldsymbol{x}) \end{bmatrix} = \boldsymbol{H}_f(\boldsymbol{x})_{:,i} \tag{9}$$

Computing the entire Hessian would entail repeating the same process for each of n columns of the Hessian. Since computing each column has computational complexity $O(m)$, computing the entire Hessian has computational complexity $O(mn)$.

Naively computing the Hessian would require $O(mn^2)$ time, since there are n^2 entries in the Hessian matrix, each of which requires $O(m)$ time to compute using the backpropagation algorithm. This naive way of computing the Hessian does not exploit the structure of the computational graph and the computational advantages that memoizing intermediate values brings.

(c) Extrapolating the conclusions from the previous section (b), computing the entry values of the tensor of k-th order derivatives would have the following computational complexity: $O(mn^{k-1})$.

(d) The diagonal of the Hessian matrix consists of the following elements:

$$diag\left(\nabla^2 f(\boldsymbol{x})\right) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}) \\ \frac{\partial^2 f}{\partial x_2^2}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial^2 f}{\partial x_n^2}(\boldsymbol{x}) \end{pmatrix} \tag{10}$$

Remaining consistent with the notation used in the problems statement, y_i will be used instead of f to denote the i-th output of the function $\boldsymbol{f}$, over which the goal is to efficiently compute the second partial derivatives on the diagonal of the Hessian:

$$\frac{\partial y_i}{\partial x_j} = \sum_{k=1}^{M} \frac{\partial y_i}{\partial z_k} \frac{\partial z_k}{\partial x_j} \tag{11}$$

Applying the partial derivative operator to the expression 11 leads to the following chain of computations:

$$
\begin{aligned}
\frac{\partial^2 y_i}{\partial x_j^2} &= \frac{\partial}{\partial x_j}\frac{\partial y_i}{\partial x_j} \\
&= \sum_{k=1}^{M} \frac{\partial}{\partial x_j}\left(\frac{\partial y_i}{\partial z_k}\frac{\partial z_k}{\partial x_j}\right) \\
&= \sum_{k=1}^{M} \left[\left(\frac{\partial}{\partial x_j}\frac{\partial y_i}{\partial z_k}\right)\frac{\partial z_k}{\partial x_j} + \frac{\partial y_i}{\partial z_k}\left(\frac{\partial}{\partial x_j}\left(\frac{\partial z_k}{\partial x_j}\right)\right)\right] \\
&= \sum_{k=1}^{M} \left[\frac{\partial^2 y_i}{\partial z_k^2}\frac{\partial z_k}{\partial x_j}\frac{\partial z_k}{\partial x_j} + \frac{\partial y_i}{\partial z_k}\frac{\partial^2 z_k}{\partial x_j^2}\right] \\
&= \sum_{k=1}^{M} \left[\frac{\partial^2 y_i}{\partial z_k^2}\left(\frac{\partial z_k}{\partial x_j}\right)^2 + \frac{\partial y_i}{\partial z_k}\frac{\partial^2 z_k}{\partial x_j^2}\right] \\
&= \boxed{\sum_{k=1}^{M} \left[\frac{\partial^2 y_i}{\partial z_k^2}(z_k)\left(\frac{\partial z_k}{\partial x_j}(x_j)\right)^2 + \frac{\partial y_i}{\partial z_k}(z_k)\frac{\partial^2 z_k}{\partial x_j^2}(x_j)\right]} \qquad (12)
\end{aligned}
$$

Analysing the last row of equation of 12, it can be seen that all factors in the sum can be computed easily:

- The term $\frac{\partial^2 y_i}{\partial z_k^2}$ is easily found by just taking the second derivative of the function $y_i(z_k)$, which is one of the standard functions used to build the computational graph of the main function.
- The term $\left(\frac{\partial z_k}{\partial x_j}\right)^2$ is found by just taking the square of the term that is already computed by the standard backpropagation algorithm.
- The term $\frac{\partial y_i}{\partial z_k}$ is already computed by the standard backpropagation algorithm
- The term $\frac{\partial^2 z_k}{\partial x_j^2}$ is computed easily by just taking the second derivative of the function $z_k(x_j)$, which is one of the standard functions used to build the computational graph of the main function.

The final algorithm to calculate the diagonal elements of the Hessian would then represent an augmentation of the backpropagation algorithm, whereby it would be necessary to store and compute additional values alongside the forward propagation variable values and the derivative values. The additional values are exactly the second derivatives of variables at the end of each edge, i.e. $\frac{\partial^2 y_i}{\partial z_k^2}$, for variables y_i, which is the head of the node and z_k as its tail. The final algorithm consists of the following steps:

(1) Perform a forward pass on the computational graph to obtain variable values: $x_j,\ j \in \{1, 2, \ldots, N_x\}, z_k,\ k \in \{1, 2, \ldots, M\}, y_i,\ i \in \{1, 2, \ldots, N_y\}$

(2) Perform the backwards pass on the computational graph to obtain derivative values: $\frac{\partial y_i}{\partial x_j}(x_j) = \sum_{k=1}^{M} \frac{\partial y_i}{\partial z_k}(z_k)\frac{\partial z_k}{\partial x_j}(x_j)$, where the values $\frac{\partial y_i}{\partial z_k}(z_k)$, for $k \in \{1, 2, \ldots, M\}$, $i \in \{1, 2, \ldots, N_y\}$, are computed upstream, in the previous steps of the backpropagation algorithm, and subsequently memoized. Values $\frac{\partial z_k}{\partial x_j}$ are computed

on the spot, while evaluating expression 11. The derivative is easily found since the functions $z_k(x_j)$ and their gradient elements $\frac{\partial z_k}{\partial x_j}(x_j)$ are known.

(3) This is the step which differs from the standard backpropagation algorithm: Instead of just memoizing the intermediate derivatives $\frac{\partial y_i}{\partial z_k}(z_k)$, the second derivatives $\frac{\partial^2 y_i}{\partial z_k^2}(z_k)$, $i \in \{1, 2, \ldots, N_y\}$, $k \in \{1, 2, \ldots, M\}$ are also memoized, to be used in the computation of downstream second derivatives according to equation 12. Values $\frac{\partial^2 z_k}{\partial x_j^2}$, $j \in \{1, 2, \ldots, N_x\}$, $k \in \{1, 2, \ldots, M\}$ are computed using the known second derivatives. This additional memoization effectively doubles the amount of values memoized per intermediate variable.

Question 2

https://colab.research.google.com/drive/1f6HKHPpwrLbFtrUTivKDKr3F87XCxeSi?usp=sharing