Numerical Experiment

Michael Rauchensteiner

December 6, 2017

Contents

1 DESCRIPTION

Let $a_1,...,a_m \in \mathbb{R}^d, b_1,...,b_{m_1} \in \mathbb{R}^m, m \geq m_1, m, m_1 \in \mathbb{N}$ two sets of orthonormal, independent vectors and $g: \mathbb{R} \to \mathbb{R}$ a smooth function. Then we can construct

$$f(x) = \sum_l^{m_1} g(\left[g(a_1^T x),...,g(a_m^T x)\right] b_l). \tag{1.1}$$

We are interested in the higher-order derivatives, first we introduce some notation to keep the gradients more readable:

$$\begin{aligned} v_e = v_e(x) &:= A diag(g'(a_1^T x),...,g'(a_m^T x)) b_e, \\ \bar{H}_e &:= g'(\sum_i b_{ie} g(a_i^T x)) \\ \bar{\bar{H}}_e &:= g''(\sum_i b_{ie} g(a_i^T x)) \\ \bar{\bar{\bar{H}}}_e &:= g'''(\sum_i b_{ie} g(a_i^T x)). \end{aligned} \tag{1.2}$$

Then

$$\begin{aligned} \nabla f(x) &= \sum_e h'(\sum_i b_{ie} g(a_i^T x)) \sum_i b_{ie} g'(a_i^T x) a_i \\ &= \sum_e \bar{H}_e v_e, \end{aligned} \tag{1.3}$$

$$\nabla^2 f(x) = \sum_e \bar{\bar{H}}_e v_e \otimes v_e + \sum_e \bar{H}_e \sum_i b_{ie} g''(a_i^T x) a_i \otimes a_i, \tag{1.4}$$

and

$$\begin{aligned} \nabla^3 f(x) &= \sum_e \bar{\bar{\bar{H}}}_e v_e^{\otimes 3} \\ &+ \sum_e \bar{\bar{H}}_e \sum_{ij} b_{ie} b_{je} g''(a_i^T x) g'(a_j^T x) \left[a_i \otimes a_i \otimes a_j + a_i \otimes a_j \otimes a_i + a_j \otimes a_i \otimes a_i\right] \\ &+ \sum_e \bar{H}_e \sum_i b_{ie} g'''(a_i^T x) a_i^{\otimes 3} \\ &= \sum_e \bar{\bar{\bar{H}}}_e v_e^{\otimes 3} + \sum_e \bar{\bar{H}}_e \sum_i b_{ie} g''(a_i^T x) \left[a_i \otimes a_i \otimes v_e + a_i \otimes v_e \otimes a_i + v_e \otimes a_i \otimes a_i\right] \\ &+ \sum_e \bar{H}_e \sum_i b_{ie} g'''(a_i^T x) a_i^{\otimes 3}. \end{aligned} \tag{1.5}$$

Setting $w_i := \sum_e \bar{\bar{H}}_e b_{ie} g''(a_i^T x) v_e$ gives

$$\begin{aligned} \nabla^3 f(x) &= \sum_e \bar{\bar{\bar{H}}}_e v_e^{\otimes 3} + \sum_i \left[a_i \otimes a_i \otimes w_i + a_i \otimes w_i \otimes a_i + w_i \otimes a_i \otimes a_i\right] \\ &+ \sum_i \left[\sum_e \bar{H}_e b_{ie}\right] g'''(a_i^T x) a_i^{\otimes 3}. \end{aligned} \tag{1.6}$$

For any scalar function $g:\mathbb{R}\to\mathbb{R}$ and $v\in\mathbb{R}^d$ we will introduce the notation

$$g(v)=\left[g(v_1),...,g(v_d)\right]^T, \tag{1.7}$$

and

$$sum(v)=\sum_i^d v_i \tag{1.8}$$

Now we can write

$$f(x)=sum(g(B^T g(A^T x))) \tag{1.9}$$

We define

$$V_x := Adiag(g'(A^T x))B\in\mathbb{R}^{d\times m_1} \tag{1.10}$$

then

$$\nabla f(x)=V_x g'(B^T g(A^T x)) \tag{1.11}$$

and

$$\nabla^2 f(x)=V_x diag(g''(B^T g(A^T x)))V_x^T + Adiag\left[Bg'(B^T g(A^T x)) * g''(A^T x)\right]A^T \tag{1.12}$$

The object of interest will be the vector spaces

$$L_k := span\left\{\nabla^k f(x)\middle| x\in\mathbb{R}^d\right\}, \tag{1.13}$$

where $k=2,3$. From our decomposition of $\nabla^k f$ it already follows that

$$L_2\subseteq span\left\{a_i\otimes a_j\middle| i,j=1,...,m\right\}, \tag{1.14}$$

$$L_3\subseteq span\left\{a_i\otimes a_j\otimes a_k\middle| i,j,k=1,...,m\right\}. \tag{1.15}$$

So $dim(L_k)\le m^k$. By the fact that $\nabla^k f(x)$ will always be a symmetric tensor we already know that $a_i\otimes a_j\otimes a_k\not\subseteq L_3$ if not $i=j=k$, therefore $dim(L_k)<m^k$.
For a set of points $\mathcal{X}:=\left\{x_1,x_2,...,x_{m_x}\right\}\subset\mathbb{R}^m$ we define

$$L_k^{\mathcal{X}} := span\left\{\nabla^k f(x)\middle| x\in\ \mathcal{X}\right\}. \tag{1.16}$$

If we choose m_x large enough and draw the points in $\mathcal{X}$ uniformly from the unit-sphere $S^{d-1}:=\left\{x\in\mathbb{R}^d\middle| ||x||_2=1\right\}$, it is reasonable to assume that $L_k^{\mathcal{X}}$ will be a good approximation of L_k. For now we will assume that $L_k^{\mathcal{X}}=L_k$ and drop the $\mathcal{X}$ in the notation. If we look at our decomposition of the tensors $\nabla^k f(x)$ and simply count the terms involved (arg.missing), one can formulate the following

Assumption 1. *There is a space $\tilde{L}_k$ close to L_k with $dim(\bar{L}_k) \in [m+m_1, 2m]$, with the possibilty of $L_k = \tilde{L}_k$.*

We will provide numerical evidence in the following sections, if this assumption holds for particular functions g and small choices of m, m_1.
To do so we pick the points in $\mathcal{X}$ as described above and evaluate the derivatives $\nabla^k f(x_i), i = 1,...,m_x, k = 2,3$ at these points. For any tensor, lets say $t \in \mathbb{R}^{m\times m\times m}$, we can find a vectorization by $vec(t) \in \mathbb{R}^{m^3}$ with $vec(t)_l := t_{ijk}$ where $i = \left\lceil \frac{l}{m^2} \right\rceil, j = \left\lceil \frac{l-(i-1)m^2}{m} \right\rceil, k = (l \mod m) + 1$. Encoding these vectors as columns of a matrix gives us

$$M_2 = \left[vec(\nabla^2 f(x_1))\middle|...\middle|vec(\nabla^2 f(x_{m_x}))\right] \in \mathbb{R}^{m^2\times m_x}, \tag{1.17}$$

and

$$M_3 = \left[vec(\nabla^3 f(x_1))\middle|...\middle|vec(\nabla^3 f(x_{m_x}))\right] \in \mathbb{R}^{m^3\times m_x}. \tag{1.18}$$

The unvectorized columns of M_k span the space L_k and we can always switch between those representations, therefore we can study the column space of M_k in place of L_k. We assume that $m_x > m^k$ and $rank(M_k) = m^k$. In this case we can write the singular value decomposition of M_k as

$$M_k = U_k D_k V_k^T \tag{1.19}$$

with $U_k \in \mathbb{R}^{m^k\times m^k}, V_k \in \mathbb{R}^{m_x\times m^k}$, matrices with orthonormal columns and $D_k \in \mathbb{R}^{m^k\times m^k}$ a diagonal matrix with the singular values $d_1 \geq ... \geq d_{m^k}$ on the diagonal ordered by magnitude.
In the following sections we provide the results of a numerical experiment where we measure

1. the ratio of the first $j = m+m_1, 2m$ singular values w.r.t. to all the singular values

$$ratio(D_k, j) := \frac{\sum_i^j d_i}{\sum_i^{m^k} d_i}, \tag{1.20}$$

2. the minimal distance (in terms of the euclidian norm) of the vectorized representation of $a_i^{\otimes k}$ over the set of the first $j = m+m_1, 2m$ left singular vectors u_l:

$$dist(a_i, U_k, j) := \min_{l\in[j]} ||vec(a_i^{\otimes k}) - \pm u_l||_2. \tag{1.21}$$

2 Results

From the experiment we can conclude that for $g = sigmoid$ and various choices of m there is a good approximation $\tilde{L}_k$ of L_k with $dim(\tilde{L}_k) = 2m$, i.e. $ratio(D_k, 2m) \approx 1$, where the first $m+m_1$ singular values/vectors are the most significant, $ratio(D_k, m+m_1) > 0.9$. This seems to be true for both $k = 2$ and $k = 3$. But this is not the case for $g = tanh, exp$. As figure **??**(**??**) show, the $ratio(D_k, 2m)$ is decreasing fast for the third derivative $k = 3$. Therefore we can say that our assumption from the previous section is not true for general functions g.