Tensor Decomposition Notes

Damian Pavlyshyn

October 29, 2018

1 Definitions and set-up

Let's see if this works

Suppose that $a_i \sim \mathcal{N}_d(0, I)$ are iid and define

$$T = \sum_{i=1}^{n} a_i^{\otimes k}.$$

Given the entries of T, we seek to recover the vectors a_i by optimising the objective

$$f(x) = \frac{1}{k} \sum_{i=1}^{n} \langle a_i, x \rangle^k$$

under the constraint $\|x\| = 1$.

1.1 Derivatives of f

Take $\tilde{\nabla}$ and $\tilde{\nabla}^2$ to be the gradient and Hessian respectively of f on the sphere. Taking $P_x = I - xx^{\mathrm{T}}$ to be the projection onto the subspace orthogonal to x, we have the identities

$$\begin{aligned}
\tilde{\nabla} f(x) &= P_x \nabla f(x), \\
\tilde{\nabla}^2 f(x) &= P_x \nabla \tilde{\nabla} f(x) P_x \\
&= P_x \nabla^2 f(x) P_x - x^{\mathrm{T}} \nabla f(x) P_x.
\end{aligned}$$

In particular, we have that

$$\begin{aligned}
\nabla f(x) &= \sum_{i=1}^{n} \langle a_i, x \rangle^{k-1} a_i, \\
\tilde{\nabla} f(x) &= \sum_{i=1}^{n} \langle a_i, x \rangle^{k-1} P_x a_i,
\end{aligned}$$

$$\tilde{\nabla}^2 f(x) = (k-1)\sum_{i=1}^{n}\langle a_i, x\rangle^{k-2} P_x a_i a_i^{\mathrm{T}} P_x - \sum_{i=1}^{n}\langle a_i, x\rangle^{k} I_{d-1}.$$

Now, write $\alpha_i = \langle a_i, x\rangle$ and $b_i = P_x a_i$ so that $\alpha \sim \mathscr{N}_n(0, I)$ and $b_i \sim \mathscr{N}_{d-1}(0, I)$ are mutually independent. We can thus write that

$$\begin{aligned}
f(x) &= \sum_{i=1}^{n} \alpha_i^k, \\
\tilde{\nabla} f(x) &= \sum_{i=1}^{n} \alpha_i^{k-1} b_i, \\
\tilde{\nabla}^2 f(x) &= (k-1)\sum_{i=1}^{n} \alpha_i^{k-2} b_i b_i^{\mathrm{T}} - \sum_{i=1}^{n} \alpha_i^k I_{d-i} \\
&= (k-1)\sum_{i=1}^{n} \alpha_i^{k-2} b_i b_i^{\mathrm{T}} - f(x) I_{d-1}.
\end{aligned}$$

Hence, for any x, $(f(x), \tilde{\nabla} f(x), \tilde{\nabla}^2 f(x))$ is mean 0 and can be describes with a function of independent standard Gaussians.

1.2 Kac-Rice formula

Lemma 1. *Lef f be a random function defined on the unit sphere $s^{d-1}$and let $Z \subseteq S^{d-1}$. Under certain regularity conditions of f and Z, we have, for $\mathscr{M}_f$ the set of local maxima of f,*

$$\mathbf{E}|\mathscr{M}_f \cap Z| = \int_{S^{d-1}} \mathbf{E}[|\det \tilde{\nabla}^2 f| \cdot \mathbf{1}_{\tilde{\nabla}^2 f \preceq 0} \mathbf{1}_{x \in Z} | \tilde{\nabla} f(x) = 0] p_{\tilde{\nabla} f(x)}(0)\, \mathrm{d}x.$$

Conditioning on α, the quantity of interest thus becomes

$$h(\alpha) = \mathbf{E}[|\det \tilde{\nabla}^2 f| \cdot \mathbf{1}_{\tilde{\nabla}^2 f \preceq 0} \mathbf{1}_{x \in Z} | \tilde{\nabla} f(x) = 0, \alpha] p_{\tilde{\nabla} f(x)|\alpha}(0).$$

We immediately have that

$$\tilde{\nabla} f(x)|\alpha \sim \mathscr{N}_{d-1}\Big(0, \sum_{i=1}^{n} \alpha_i^{2(k-1)}\Big),$$

which renders

$$p_{\tilde{\nabla}|\alpha}(0) = \Big[\sum_{i=1}^{n} \alpha_i^{2(k-1)}\Big]^{(d-1)/2} = \|\alpha^{\odot(k-1)}\|^{d-1}.$$

1.3 Useful Results

1.3.1 Shannon transform

If A is an $n \times n$ matrix whose largest eigenvalue is at most x, then we have that, for m the Stieltjes stransform of A,

$$\int_x^\infty \Big(\frac{1}{w} + m(w)\Big)\,\mathrm{d}w = \int_x^\infty \int \Big(\frac{1}{w} + \frac{1}{\lambda - w}\Big)\,\mathrm{d}\nu(\lambda)\,\mathrm{d}w$$

$$
\begin{aligned}
&= \int \int_x^\infty \frac{\lambda}{w(\lambda - w)} \, \mathrm{d}w \, \mathrm{d}\nu(\lambda) \\
&= \int \log(1 - \lambda/x) \, \mathrm{d}\nu(\lambda) \\
&= \frac{1}{n} \log\det(I - A/x)
\end{aligned}
$$

2 Conditioning in Kac-Rice

Fixing x and α conditioning on $\tilde{\nabla} f(x) = 0$, and writing $B = (b_1 | \cdots | b_n)$, we have that the entries of B are iid normals subject to the constraint

$$B\alpha^{\odot(k-1)} = 0.$$

That is, the rows of B are iid normals supported on the $(n-1)$-dimensional hyperplane orthogonal to $\alpha^{\odot(k-1)}$. Hence, we have that

$$[B | \tilde{\nabla} f(x) = 0, \alpha] \overset{\mathrm{d}}{=} [B(I - \bar{\alpha}\bar{\alpha}^{\mathrm{T}}) | \alpha],$$

where $\bar{\alpha} = \alpha^{\odot(k-1)} / \|\alpha^{\odot(k-1)}\|$.

Now, conditionally on α, we can write, for $P_\alpha = I - \bar{\alpha}\bar{\alpha}$,

$$
\begin{aligned}
[\tilde{\nabla}^2 f(x) | \tilde{\nabla} f(x) = 0] &= \left[(k-1) B D_\alpha^{k-2} B^{\mathrm{T}} - f(x) I_{d-1} \middle| \tilde{\nabla} f(x) = 0 \right] \\
&= (k-1) B P_\alpha D_\alpha^{k-2} P_\alpha B^{\mathrm{T}} - f(x) I_{d-1}.
\end{aligned}
$$

From the useful result, for k even, finding the log-determinant of this is equivalent to studying the spectrum of

$$D_\alpha^{k/2-1} P_\alpha B^{\mathrm{T}} B P_\alpha D_\alpha^{k/2-1}.$$

3 Spectrum of the Hessian

Theorem 2 (Silverstein and Bai (1995))**.** *Suppose that for each n, the entries of $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $p \times n$, are iid complex random variables with $\mathbf{E}|x_{11} - \mathbf{E}x_{11}|^2 = 1$, and that $\mathbf{T} = \mathbf{T}_n = \mathrm{diag}(\tau_i^n, \ldots, \tau_p^n)$, τ_i^n real, and the ESD of $\mathbf{T}$ converges almost surely to a probability distribution function H as $n \to \infty$.*

Assume that $\mathbf{B} = \mathbf{A} + \mathbf{X}^*\mathbf{T}\mathbf{X}$, where $\mathbf{A} = \mathbf{A}_n$ is a Hermitian $n \times n$ satisfying $F^{\mathbf{A}_n} \overset{\mathrm{v}}{\to} F_\alpha$ almost surely, where F_α is a distribution function (possibly defective, i.e., of total variation less than 1) on the real line. Furthermore, assume that $\mathbf{X}, \mathbf{T}$, and $\mathbf{A}$ are independent.

When $p/n \to y > 0$ as $n \to \infty$, we have that almost surelt $F^{\mathbf{B}}$, converges vaguely to a (non-random) d.f. $\mathbf{F}$, whose Stieltjes transform $m(z)$ is given by

$$m(z) = m_\alpha\left(z - y \int \frac{\tau \, \mathrm{d}H(\tau)}{1 + \tau m(z)}\right). \tag{1}$$

Theorem 3. *For any z with $\mathfrak{Im}(z) > 0$, eq.* (1) *has a unique solution $m(z)$ which has positive imaginary part.*

In the event that $d/n \to \beta > 0$, the limiting EDF of $\frac{1}{n}BDB^{\mathrm{T}}$ has Stiltjes transform m given implicitly by

$$-\frac{1}{m(z)} = z - \int \frac{s^{k-2}H(\mathrm{d}s)}{1+\beta s^{k-2}m(z)},$$

where φ is the standard normal pdf and H is the limiting empirical distribution of α.

In particular, for a finite but large n, if we condition on α, treating D as deterministic, this renders the following approximation for m:

$$-\frac{1}{m(z)} = z - \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha_i^{k-2}}{1+\beta\alpha_i^{k-2}m(z)}$$

4 Wishart Determinants

4.1 Joint density of Wishart eigenvalues

Let $M \sim \mathscr{W}_d(I, n)$ be a standard Wishart matrix. The eigenvalues of M then have joint density

$$Q_{n,d}(\lambda) = \frac{1}{Z_{n,d}} \prod_{i=1}^{d} \lambda_i^{(n-d-1)/2} e^{-\lambda_i/2} \prod_{i<j} |\lambda_i - \lambda_j| \mathbf{1}_{\lambda_1 \geq \cdots \geq \lambda_d},$$

where

$$Z_{n,d} = \frac{\pi^{d^2/2}}{2^{nd/2}\Gamma_d(n/2)\Gamma_d(d/2)},$$

and in turn Γ_d is the multivariate gamma function defined by

$$\Gamma_d(a) = \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma(a - (i-1)/2).$$

4.2 Determinant calculation

Let x be a random variable independent of M with some density f. We then have that

$$\begin{aligned}
&\mathbf{E}[|\det(xI - M)| \mathbf{1}_{(xI-M)\succeq 0} \mathbf{1}_{x \in B}] \\
&\quad = \int_B f(x) \int_{x \geq \lambda_1 \geq \cdots \geq \lambda_d} \prod_{j=1}^{d} (x - \lambda_j) Q_{n,d}(\lambda)\, \mathrm{d}\lambda\, \mathrm{d}x \\
&\quad = \frac{1}{Z_{n,d}} \int_B f(\lambda_0) \int_{\lambda_0 \geq \cdots \geq \lambda_d} \prod_{0 \leq i < j \leq d} |\lambda_i - \lambda_j| \prod_{i=1}^{d} \lambda_i^{(n-d-1)/2} e^{-\lambda_i/2}\, \mathrm{d}\lambda\, \mathrm{d}\lambda_0 \\
&\quad = \frac{Z_{n+1,d+1}}{Z_{n,d}} \int \lambda_0^{-(n-d-1)/2} e^{\lambda_0/2} f(\lambda_0) \mathbf{1}_{\lambda_0 \in B} Q_{n+1,d+1}(\lambda)\, \mathrm{d}\lambda \\
&\quad = \frac{Z_{n+1,d+1}}{Z_{n,d}} \mathbf{E}_{\mathscr{W}}^{n+1,d+1}\left[\lambda_{\max}^{-(n-d-1)/2} e^{\lambda_{\max}/2} f(\lambda_{\max}); \lambda_{\max} \in B\right]
\end{aligned}$$

5 Dynamics of gradient descent

We can also write

$$\begin{aligned}\tilde{\nabla} f(x) &= \sum_{i=1}^{n} \langle a_i, x\rangle^{k-1} a_i - x \sum_{i=1}^{n} \langle a_i, x\rangle^k \\ &= \sum_{i=1}^{n} y_i^{k-1} a_i - x \sum_{i=1}^{n} y_i^k\end{aligned}$$

for $y_i = \langle a_i, x\rangle$. Formally taking the time derivative through the gradient update step yields

$$\begin{aligned}\dot{y}_j &= \langle a_j, \dot{x}\rangle \\ &= \langle a_j, \tilde{\nabla} f(x)\rangle \\ &= \sum_{i=1}^{n} y_i^{k-1} \langle a_i, a_j\rangle - y_j \sum_{i=1}^{n} y_i^k \\ &= y_j^{k-1} \|a_j\|^2 - y_j^{k+1} - y_j \sum_{i\neq j} y_i^k + \sum_{i \neq j} y_i^{k-1} \langle a_i, a_j\rangle.\end{aligned}$$

Further, taking $w_j = y_j/\|a_j\|$ so that recovery of a_j is characterised by $w_j \to 1$, this becomes

$$\dot{w}_j = \|a_j\|^k \left\{ (w_j^{k-1} - w_j^{k+1}) - w_j \sum_{i\neq j} \left[\frac{\|a_i\|}{\|a_j\|}\right]^k w_i^k + \sum_{i\neq j} \left[\frac{\|a_i\|}{\|a_j\|}\right]^k w_i^{k-1} \frac{\langle a_i, a_j\rangle}{\|a_i\|\|a_j\|} \right\}.$$

5.1 Orthogonal case

Before proceeding, we will consider the case where $n \ll d$. This is very close to the orthogonal case, so first, let's consider the setting where the a_i are orthogonal unit vectors. The previous equations now become

$$\dot{w}_j = w_j^{k-1} - w_j^{k+1} - w_j \sum_{i\neq j} w_i^k.$$

Notice that, since $n \leq d$, we have

$$\begin{aligned}1 &= \|x\|^2 \\ &\geq \langle a_i, x\rangle^2 + \cdots + \langle a_n, x^2\rangle \\ &= w_1^2 + \cdots + w_n^2.\end{aligned}$$

In particular, if $w_j \to 1$, then $w_i \to 0$ for all other indices i. In particular, we have that $\sum_{i\neq j} w_i^k \to 0$.

By flipping signs is necessary, we can assume that each $w_j \geq 0$. Now, we have that, if w_j is the largest of the ws, then

$$\dot{w}_k = w_j^{k-1}(1 - w_j^2) - w_j \sum_{i\neq j} w_i^k$$

$$\begin{aligned}
&\geq w_j^{k-1}(1-w_j^2) - w_j^{k-1}\sum_{i\neq j} w_i^2 \\
&\geq w_j^{k-1}(1-w_j^2) - w_j(1-w_j^2) \\
&\geq 0.
\end{aligned}$$

Hence, the largest w cannot decrease. In particular, is must converge. But now, solving $\dot{w}_j = 0$ in the limit, we have

$$w_j^{k-1}(1-w_j^2) = w_j \sum_{i\neq j} w_i^k.$$

But the right-hand side is non-negative while the left is non-positive. Hence, we must have that $w_j = 0$ or 1. But since w_j is the maximum w, this means that $w_j = 1$ and so the rest of the $w_i = 0$. Hence, we conclude that the maximum w converges to 1 while the rest converge to 0. That is to say, after running gradient ascent, $x \to a_j$.

5.2 Undercomplete normal case

Now, let's consider the a_is again, though suppressing the inner product term. Suppose first that $k = 3$. In this case, it is more interesting to study the quantity $v_j(t) = \|a_j\|^3 w_j$. The equations governing the vs are now

$$\dot{v}_j = v_j^2 - \frac{v_j^4}{\|a_j\|^6} - v_j \sum_{i\neq j} \frac{v_i^3}{\|a_i\|^6}.$$

Now, suppose that $v_j(t) = v_i(t) = \lambda$ for some $i \neq j$ and t. We then have that

$$\begin{aligned}
\dot{v}_j(t) - \dot{v}_j(t) &= -\frac{\lambda^4}{\|a_j\|^6} + \frac{\lambda^4}{\|a_i\|^6} - \lambda \sum_{l\neq j} \frac{v_l^3}{\|a_l\|^6} + \lambda \sum_{l\neq i} \frac{v_l^3}{\|a_l\|^6} \\
&= \lambda^4 \Big(-\frac{1}{\|a_j\|^6} + \frac{1}{\|a_i\|^6} + \frac{1}{\|a_j\|^6} - \frac{1}{\|a_i\|^6} \Big) \\
&= 0.
\end{aligned}$$

Hence, none of the v_j may cross each other. Hence, the greatest v_j stays the greatest.

Indeed, for any k, this can be achieved for $v_j = \|a_j\|^\gamma w_j$ by taking $\gamma = k/(k-2)$, as shown:

$$\begin{aligned}
\dot{v}_j &= v_j^{k-1}\|a_j\|^{\gamma(2-k)+k} - v_j^{k+1}\|a_j\|^{k(1-\gamma)} - v_j \sum_{i\neq j} \|a_i\|^{k(1-\gamma)} v_i^k \\
&= v_j^{k-1} - v_j^{k+1}\|a_j\|^{-2k/(k-2)} - v_j \sum_{i\neq j} v_i^k \|a_i\|^{-2k/(k-2)}.
\end{aligned}$$

Using a similar trick to the orthogonal case, we have thus that, for maximal v_j,

$$\dot{v}_j = v_j^{k-1}(1 - v_j^2\|a_i\|^{-2\gamma}) - v_j^{k-1} \sum_{i\neq j} v_i \|a_i\|^{-2\gamma}$$

$$= v_j^{k-1}\Big(1 - \sum_{i=1}^n w_j^2\Big).$$

Now, notice that we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\|w\|^2 &= 2\sum_{j=1}^n \dot{w}_j w_j \\
&= 2\sum_{j=1}^n \Big[\|a_j\|^k w_j^k(1 - w_j^2) - w_j^2 \sum_{i\neq j}\|a_i\|^k w_i^k\Big] \\
&= 2\sum_{j=1}^n \|a_j\|^k w_j^k(1 - w_j^2) - 2\sum_{i=1}^n \|a_i\|^k w_i^k \sum_{j\neq i}^n w_j^2 \\
&= 2\sum_{j=1}^n \|a_j\|^k w_j^k(1 - w_j^2) - 2\sum_{j=1}^n \|a_j\|^k w_j^k(\|w\|^2 - w_j^2) \\
&= 2(1 - \|w\|^2)\sum_{j=1}^n \|a_j\|^k w_j^k \\
&> 0,
\end{aligned}$$

with equilibrium only when $\|w\| = 1$ or 0. Hence, we have that $\|w\| \to 1$, and so $v_j \to 1$ while all the others decay to 0. Hence, in this case also, $x \to a_j$.