The relationship between power iteration and gradient descent

G. H. Brown and E. V. Solomonik

Here we consider the real Z-eigenvalue problem for supersymmetric tensors: find pairs ($\mathbf{x} \in \mathbb{R}^n, \lambda \in \mathbb{R}$) such that

$$\mathcal{A}\mathbf{x}^{m-1} = \lambda\mathbf{x}, \quad \|x\|_2^2 = 1, \tag{1}$$

for supersymmetric tensor $\mathcal{A} \in \mathbb{R}^{[m,n]}$ where m is the number of modes, and n is the symmetric dimension of said modes [1]. [1] In the matrix case ($m = 2$) this is the standard symmetric matrix eigenvalue problem.

Algorithms

Power iteration is a well known method for computing the dominant eigenvector and eigenvalue of a matrix, and is generalized for symmetric tensors by the symmetric higher order power method (SHOPM). Though SHOPM is equivalent to power iteration in the matrix case, it does not have convergence guarantees in the tensor case [2].[2] The method computes increasingly accurate estimates of the eigenvector (and so eigenvalue) according to the below algorithm. Note that the update to the iterate in SHOPM wholly replaces it with $\mathcal{A}\mathbf{x}^{m-1}$ (normalized).

Algorithm 1: Higher Order Power Method

$\mathbf{x} \leftarrow$ normalized estimate
while *not converged* **do**
  $\mathbf{v} \leftarrow \mathcal{A}\mathbf{x}^{m-1}$
  $\lambda \leftarrow \mathbf{v}^T\mathbf{x}$
  $\mathbf{x} \leftarrow \mathbf{v}/\|\mathbf{v}\|$
end

The tensor Z-eigenvalue problem may also be cast as a constrained optimization problem seeking the critical points ($\mathbf{x}$) and associated function values (λ) of

$$f(\mathbf{x}) = \mathcal{A}\mathbf{x}^m \quad \text{such that} \quad \|\mathbf{x}\|_2^2 = 1, \tag{2}$$

according to [1,3]. The gradient of the objective function $\mathcal{A}\mathbf{x}^m$ is known to be $m\mathcal{A}\mathbf{x}^{m-1}$ [3], which can be used to solve the problem via gradient ascent or descent with a line search for step scaling, as shown below for gradient descent. Note that the gradient descent algorithm sets the new iterate to be the previous iterate mixed with a step against the direction of the gradient (proportional to $\alpha > 0$). Moving against the gradient will converge to points which are local minima of $f(\mathbf{x})$, while choosing to step along the gradient will converge to local maxima. Unfortunately, neither gradient ascent or descent are effective for computing saddle points of $f(\mathbf{x})$.

[1] This superscript notation on a vector is shorthand for the contracton of the tensor with the vector along a number of modes defined by the exponent, and since the tensor is supersymmetric in this case, the choice of modes does not matter. For example, for a matrix $\mathbf{A}$, $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}^2$.

[2] Modified versions of SHOPM do provide stronger convergence guarantees, however. See [3] and citations within.

Algorithm 2: Gradient Descent

$\mathbf{x} \leftarrow$ normalized estimate
while *not converged* **do**
  $\mathbf{G} \leftarrow m\mathcal{A}\mathbf{x}^{m-1}$
  $\alpha \leftarrow \text{argmax } f(\mathbf{x} - \alpha\mathbf{G}) = \mathcal{A}(\mathbf{x} - \alpha\mathbf{G})^k$
  $\mathbf{v} \leftarrow \mathbf{x} - \alpha\mathbf{G}$
  $\mathbf{x} \leftarrow \mathbf{v}/\|\mathbf{v}\|$
end

Relationship

In addition to the two notes above on how the iterates are updated, observe that since the written version of gradient descent normalizes the proposed iterate, there is no restriction on the magnitude of α. Hence, in the limit $\alpha \to \infty$ gradient descent loses or washes out all information about the previous iterate, replacing it with $\mathcal{A}\mathbf{x}^{m-1}$ (normalized), which is the same iterate update observed for SHOPM (power iteration). Hence, gradient descent/ascent with an infinite step size of proper sign is equivalent to the SHOPM (power iteration). Furthermore, in the case of $\alpha \to \infty$, both gradient descent and gradient ascent are equivalent to power iteration, since $(\mathbf{x}, \lambda)$ and $(-\mathbf{x}, \lambda)$ are degenerate eigenpairs for even order tensors, and $(\mathbf{x}, \lambda)$ and $(-\mathbf{x}, -\lambda)$ are degenerate for odd order tensors [3]. That is to say, $\pm\mathbf{x}$ are always degenerate eigenvectors for generic symmetric tensors.

Of course, infinite line search parameters do not occur in real implementations of gradient descent, and when finite values are obtained from an actual line search, the results of gradient descent will not match those of SHOPM. Specifically, SHOPM (assuming it converges) will find the dominant eigenvector, while gradient descent/ascent will converge locally and find the eigenvector corresponding to the local minimum/maximum in the vicinity of the original estimate.

References

[1] L.-H. Lim, 2013, "Singular values and eigenvalues of tensors: A variational approach" in CAMSAP'05: Proceeding of the IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing,Leh King Lim tensor eigenvalues, p. 129-132.
[2] E. Kodifis and P. A. Regalia, 2002, "On the best rank-1 approximation of higher-order supersymmetric tensors", *SIAM Journal of Matrix Analysis Applications*, **23**, p. 863-884.
[3] T. G. Kolda and J. R. Mayo, "Shifted Power Method for Computing Tensor Eigenpairs", *SIAM Journal of Matrix Analysis Applications*, **32**(4), p. 1095-1124.

Notes

This connection became apparent to both authors during an individual meeting between the two. Finding the relationship interesting, G. H. B. completed this write up.