Gradient Ascent in a Linear Inhibitory Network

Alan Prince
Dept. of Linguistics, and
Rutgers Center for Cognitive Science
Rutgers University/New Brunswick
June 12, 1996

Let B_n be a network of n nodes $\mathbf{x} = (x_1, ...,x_n)$, such that x_k is connected only to x_{k+1} and such that each node x_k is subject to a constant positive bias b_k. Let the connection between adjacent nodes be inhibitory and constant for the entire network. Such networks can serve as models of string-like aspects of phonological structure, particularly those involving the relative prominence of adjacent elements (cf. Prince & Smolensky 1991:13-15; Smolensky, Legendre, and Miyata 1992 and the references cited therein).

The underlying problem can be posed this way: given a set of intrinsic, gradient prominences in the input, represented as biases on positions, determine a peak/non-peak structure in the output. A string of *segments* of different intrinsic sonorities will be parsed into (syllable) peaks and nonpeaks. A string of *syllables* of different weights will be parsed into (stress) peaks and nonpeaks.

Here we examine the behavior of such a network under gradient ascent over its harmony surface (Smolensky 1986). The network is *linear* in the sense that the gradient of its harmony function is component-wise linear. We define a *peak* to be a node whose activation becomes and stays *positive*, a *nonpeak* to a node whose activation becomes and stays negative. This is a linearization of the original Smolensky-Miyata model described in Prince & Smolensky 1991, which follows the nonlinear 'brain-state in a box' model of Anderson 1977 in limiting activation to the interval [0,1].

There are three main results:

1. The network shows just two stable states: (i) strict peak/non-peak alternation beginning with a peak, (ii) strict peak/non-peak alternation beginning with a non-peak.

2. In order to assign biases so that a highest-prominence element is reliably mapped into a peak, a scaling factor relating the bias of that element to that of its nearest competitor is required. This factor grows quadratically with the length of the network.

3. An immediate corollary: Given a ranked set of intrinisic linguistic prominences, such that peak status is assigned to the highest such prominence in a string, that the biases corresponding to the linguistic prominences must be scaled exponentially.

Evidently, significant nonlinearities must be introduced to achieve better modeling of linguistic realities. An exact understanding of the linear case examined here provides a useful starting point for further exploration.

For convenience, let the inhibitory weight between adjacent nodes $w_{k(k+1)} = -1$. For additional descriptive convenience, we occasionally recognize phantom nodes x_0 , x_{n+1} such that $x_0 = x_{n+1} \equiv 0$.

Following essentially standard practice, let the harmony function for the network be as follows:

(1)

$$H = \sum_{k=1}^{n} b_k x_k - x_k x_{k+1} = \mathbf{b}\cdot\mathbf{x} + \frac{1}{2}\,\mathbf{x}^{\mathbf{T}}\mathbf{A}\,\mathbf{x}$$

We write **A** for the n×n tridiagonal matrix with 0 on the main diagonal and −1 on the sub- and supra-diagonals. Note that $H(\mathbf{0}) = \mathbf{0}$

For the gradient, we have

(2)

$$\nabla\mathbf{H} = \mathbf{b} + \mathbf{A}\mathbf{x} \qquad \textit{i.e.}$$

$$\frac{\partial H}{\partial x_k} = b_k - x_{k-1} - x_{k+1}$$

We wish to track the paths of gradient ascent over the harmony surface. Any such path is given by a function $\mathbf{x}(t):R \to R^n$, where any tangent vector associated with **x**(t) points in the same direction as $\nabla\mathbf{H}(\mathbf{x}(t))$. We will begin gradient ascent at the origin, which is included in H, whence we set $\mathbf{x}(\mathbf{0}) = \mathbf{0}$.

Since the *magnitude* of the tangent vector to **x**(t) is of no interest, we simply identify $\mathbf{x}'(t)$ with the gradient vector. The path of steepest ascent is therefore given by the solution to this system of equations:

(3)

$$\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b} \qquad \textit{i.e.}$$

$$x_k'(t) = b_k - x_{k-1}(t) - x_{k+1}(t)$$

Solution is straightforward, due to the simplicity of **A**. Because **A** is real symmetric, it has *n* distinct eigenvalues λ_k , and the n×n matrix **E** whose columns are the corresponding *unit-length* eigenvectors $\boldsymbol{\xi}^{(k)}$ is also guaranteed to be unitary so that $\mathbf{E}^{-1} = \mathbf{E}^{T}$. The matrix **E** diagonalizes **A** and we use this fact to decouple the equations (3). Let

$$\mathbf{x} = \mathbf{E}\mathbf{y}$$

(Observe that $\mathbf{y}(0) = \mathbf{x}(0) = 0$.)

From eq. (3) we obtain

(4)

$$\mathbf{E}\mathbf{y}'(t) = \mathbf{A}\mathbf{E}\mathbf{y}(t) + \mathbf{b} \qquad \textit{so that}$$

$$\mathbf{E}^{\mathrm{T}}\mathbf{E}\mathbf{y}'(t) = \mathbf{E}^{\mathrm{T}}\mathbf{A}\mathbf{E}\mathbf{y}(t) + \mathbf{E}^{\mathrm{T}}\mathbf{b} \qquad \textit{i.e.}$$

$$\mathbf{y}'(t) = \mathbf{E}^{\mathrm{T}}\mathbf{A}\mathbf{E}\mathbf{y}(t) + \mathbf{E}^{\mathrm{T}}\mathbf{b}$$

But $\mathbf{E}^{\mathrm{T}}\mathbf{A}\mathbf{T}$ is just the diagonal matrix whose diagonal entries are the eigenvalues λ_k of **A**. Letting $\mathbf{l} = \mathbf{E}^{\mathrm{T}}\mathbf{b}$ and writing l_k for the entries of $\mathbf{E}^{\mathrm{T}}\mathbf{b}$, eqn. (4) reduces to the system

(5)

$$y_k'(t) = \lambda_k y_k(t) + l_k$$

The general solution to eqn. (5) can be written down immediately (recall $\mathbf{y}(0) = 0$):

(6)

$$y_k = l_k e^{\lambda_k t} \int_0^t e^{-\lambda_k s} ds$$

This splits into two cases, depending on whether $\lambda_k = 0$ or not. Let us first assume $\lambda_k \neq 0$.

$$\int_0^t e^{-\lambda_k s} ds = -\frac{1}{\lambda_k}(e^{-\lambda_k t} - 1)$$

We have then, from eq. (6),

(7)

$$y_k = \frac{l_k}{\lambda_k}(e^{\lambda_k t} - 1)$$

For $\lambda_k = 0$, we have, trivially, from eq. (5),

(8)

$$y_k = l_k t$$

To redeem the value of **x**(t) from these equations, recall that **x** = **E**$\mathbf{y}$. Coordinate-wise, we have:

(9)
$$x_k = \sum_{i=1}^{n} \xi_k^{(i)} y_i$$

where $\xi^{(i)}$ is the eigenvector associated with λ_i , normalized so that $\xi^{(i)} \cdot \xi^{(i)} = 1$.

Along the same lines, from $\mathbf{l} = \mathbf{E}^T \mathbf{b}$, we have

(10)
$$l_i = \sum_{k=1}^{n} \xi_k^{(i)} b_k = \xi^{(i)} \cdot \mathbf{b}$$

Consider the networks where $\lambda_k \neq 0$, for all k. (Below, we will see that these are exactly the networks with an even number of nodes.)

(11)
$$x_k = \sum_{i=1}^{n} \xi_k^{(i)} \frac{l_i}{\lambda_i} (e^{\lambda_i t} - 1)$$

$$= \sum_{i=1}^{n} \xi_k^{(i)} \frac{\xi^{(i)} \cdot \mathbf{b}}{\lambda_i} (e^{\lambda_i t} - 1)$$

For any $\lambda_i = 0$, the expression $(e^{\lambda_i t} - 1)/\lambda_i$ reduces to t.

Observe that x_k takes form of a sum of terms $C_j e^{\lambda j t}$, plus a constant, plus a linear term corresponding to zero eigenvalues. To understand the asymptotic behavior of x_k , then, it suffices to examine *a single exponential term*, namely the one with the greatest λ_j. As t increases, this term will dominate all other terms, be they exponential, linear, or constant. In particular, the fate of x_k turns solely on whether the coefficient of this largest exponential term is positive or negative.

To ascertain the conditions determining the sign of the relevant coefficients, we need to have the eigenvalues and eigenvectors of the matrix **A**, the n×n tridiagonal matrix with 0 on the main diagonal, and −1 on the sub- and supra-diagonals. The calculations of Prince 1992:53,98 can be directly applied to this problem. The eigenvalues are

(12)
$$\lambda_k = 2 \cos\left(\frac{k\pi}{n+1}\right)$$

The components of the nonnormalized eigenvectors $\mathbf{v}^{(k)}$ are as follows, with the first component of each set to 1:

(13)

$$\mathbf{v}_j^{(k)} = (-1)^{j+1} \frac{\sin(\frac{j\,k\,\pi}{n+1})}{\sin(\frac{k\,\pi}{n+1})}$$

The normalized eigenvector $\boldsymbol{\xi}^{(k)}$ is just $\mathbf{v}^{(k)}/|\mathbf{v}^{(k)}|$.

It is clear from (12) that k = 1 yields the largest eigenvalue, 2 cos (π/(n+1)). So long as the coefficient $C_1{}^k$ of the term $C_1{}^k\, e^{\lambda_1 t}$ in x_k is nonzero, it will determine the behavior of x_k .

According to (11), this coefficient is

$$\xi_k^{(1)} \frac{\boldsymbol{\xi}^{(1)}\cdot\mathbf{b}}{\lambda_1} = \frac{(-1)^{k+1}}{|\mathbf{v}^{(1)}|} \frac{\mathbf{v}^{(1)}\cdot\mathbf{b}}{\lambda_1|\mathbf{v}^{(1)}|}$$

$$= (-1)^{k+1} \frac{\sum_{j=1}^{n} (-1)^{j+1} b_j \sin(\frac{j\,\pi}{n+1}) / \sin(\frac{\pi}{n+1})}{\lambda_1 |\mathbf{v}^{(1)}|^2}$$

$$= \frac{1}{\lambda_1 \sin(\frac{\pi}{n+1}) |\mathbf{v}^{(1)}|^2} \sum_{j=1}^{n} (-1)^{j+k} \sin(\frac{j\,\pi}{n+1})\, b_j$$

We are really only interested in whether $C_1 > 0$, $C_1 < 0$, or $C_1 = 0$. Chucking the positive constant material before the summation sign, we arrive at the following condition that must be met if C_1 is to be positive:

(14)

$$(-1)^k \sum_{j=1}^{n} (-1)^j b_j \sin(\frac{j\,\pi}{n+1}) > 0$$

Observe that *the only dependence on k is in the sign-determining term.* This splits the x_k into two groups: the even, x_{2p} , and the odd, x_{2p-1}. Within each group, the sign of C_1 is the same; and between the groups, there is a contrast in sign.

This leads to the following conclusion. For $C_1 \neq 0$, the units in the network alternate strictly, taking either the form $\langle +-+-... \rangle$, or the form $\langle -+-+... \rangle$. Surprisingly perhaps, the biases — no matter how drastically they vary in magnitude — can only determine which of the two patterns emerges.

We want to determine what system of biases will ensure that certain prominent (strongly biased) nodes are indubitably turned on. In particular, we are interested in the situation where we have three nodes in a row $x_{k-1}\ x_k\ x_{k+1}$, with x_k more prominent that its neighbors. What value of b_k , for $b_k > b_{k-1}$, b_{k+1} , does it take to ensure that x_k will be activated in the course of gradient ascent over the harmony surface?

This question is posed in a linguistically natural but rather misleadingly localistic fashion, given the character of the network. Since the only two stable states of the network are those with strict on-off alternation, the question to ask must be — what does it take to guarantee that node x_k *and all its same parity associates* will be turned on?

To study this issue, we focus on a kind of *worst case* analysis. Let node x_k be of parity $p \in$ {odd, even}. Suppose b_k has the largest bias in the network. Let β_{max} be the largest bias on any node of parity p^*, the opposite of p. We now ask, *how much larger* than β_{max} must b_k be, in order to turn x_k on infallibly?

We will find the following, writing b_k for the targeted maximum and β_{max} for the maximum of its opposite-parity rivals:

(15) Bias Ranking

In the limit as $n \to \infty$for , $b_k > (n+1)^2/\pi^2 \cdot \beta_{max}$, all nodes of the same parity as x_k will be turned on.

This result can be extracted from Eq. **(14)** . There are actually four distinct cases to analyze, with 3 different outcomes:

[1] Even length string

a. b_k is of odd parity, in an even length string. $n = 2q$, $k = 2p - 1$.

b. b_k is of even parity, in an even length string. $n = 2q$, $k = 2p$.

[2] Odd length string

a. b_k is of odd parity, in an odd length string. $n = 2q - 1$, $k = 2p - 1$.

b. x_k is of even parity, in an odd length string. $n = 2q - 1$, $k = 2p$.

The even-length cases under [1] give rise to the same scaling factor. (This can be seen in gross from considering the symmetries of an even-length network $\langle$ 1 *2* 3 *4* ... (2p – 1), *2p* $\rangle$ — there are as many even as odd nodes, and they are disposed mirror-symmetrically with respect to the opposite edges.)

Various formulations of it are as follows:

(16) Scaling factor for guarantee of maximum in even-length strings

$$b_k > \frac{\frac{1}{2}\,\beta_{max}\cot(\frac{\pi}{2(n+1)})}{\sin(\frac{\pi}{n+1})} = \frac{\frac{1}{2}\,\beta_{max}}{1-\cos(\frac{\pi}{n+1})} = \frac{\frac{1}{2}\,\beta_{max}}{1-\lambda_1}$$

From the second expression on the r.h.s of '>' it is clear that the scaling factor asymptotes out at $(n+1)^2/\pi^2$, since $\cos(x) \approx 1 - x^2/2$. Since $\cos(x) > 1 - x^2/2$, x small and nonzero, it is also clear that the asymptote is approached from above.

The odd-length cases under [2] are each associated with different scaling factors. (Again, this can be guessed in gross from considering the (non-)symmetries of the odd-length network $\langle 1\ 2\ 3\ 4\ \ldots\ ,\ 2p,\ 2p-1\rangle$. Both edges nodes are always odd.)

The even-parity maximum [2b] gives rise to the following scaling factor:

(17) Scaling factor guaranteeing an even maximum in odd-length string

$$b_{ev} > \frac{\beta_{max}}{\sin(\frac{\pi}{n+1})\ \sin(\frac{2\pi}{n+1})}$$

It is clear that the scaling factor asymptotes out to $(n+1)^2 / 2\pi^2$, half of the value for the even-length-cases. In this case, since $x > \sin(x)$, x positive, it is clear that the asymptote is approached from above as well.

The final case involves an odd maximum in an odd-length string. The value of the scaling factor is given by this expression:

(18) Scaling factor for odd maximum in odd-length string

$$b_{odd} > \frac{\cot(\frac{\pi}{n+1})}{\sin(\frac{\pi}{n+1})}\ \beta_{max} = \beta_{max}\ \frac{\cos(\frac{\pi}{n+1})}{\sin^2(\frac{\pi}{n+1})} = \beta_{max}\ \frac{\lambda_1}{1-\lambda_1^2}$$

Here the asymptote is $(n+1)^2 / \pi^2$, and, unlike the above cases is approached from below.

To give a sense of where these scaling factors come from, let us examine the last case in some detail. We want to turn on the odd nodes. In order to ascertain the conditions under which this happens, we

use the following relation derived from (14), gathering the odd and even bias terms on opposites sides of '>'.

(19)

$$\sum_{j=1}^{[\frac{n+1}{2}]} \sin(\frac{(2j-1)\,\pi}{n+1})\; b_{2j-1} \;>\; \sum_{j=1}^{[\frac{n+1}{2}]} \sin(\frac{(2j)\,\pi}{n+1})\; b_{2j}$$

(Point of detail: notice that we can harmlessly use the limit [(n+1)/2] on the r.h.s. since, for n even, it's the same as [n/2] and for n odd, the additional final term is equal to 0. Note also that all the sin θ terms are positive, since $0 < \theta < \pi$.)

What relationship among the biases does it take to *ensure* that the odd nodes will be turned on? We counterpose the weakest achievement of the odd nodes against the strongest possible counterattack from the even ones.

Suppose that all the odd biases excepting one are 0, so that the l.h.s. gets the least amount of help from the odd cohort. Suppose likewise that the coefficient of the one remaining odd term is as small as possible, maximally diminishing the force of the bias b_k. This would be $\sin(\pi/(n+1))$, so we are dealing with b_1 (equivalently, b_n in an odd length network).

To give the r.h.s its maximal due, suppose that all even biases are maximally large, therefore equal. We write β_{max} as a representative of this set.

We now have

(20)

$$b_1 \sin(\frac{\pi}{n+1}) \;>\; \beta_{max} \sum_{j=1}^{[\frac{n+1}{2}]} \sin(\frac{(2j)\,\pi}{n+1})$$

Suppose now that n is odd, so that $n = 2p - 1$. Observe that $[(n+1)/2] = p$ and that $n+1 = 2p$. So we have:

(21)

$$b_1 \sin(\frac{\pi}{2p}) \;>\; \beta_{max} \sum_{j=1}^{p} \sin(\frac{2j\pi}{2p}) \;=\; \beta_{max} \sum_{j=1}^{p} \sin(\frac{j\pi}{p})$$

This last expression may be simplified by the following useful (if perhaps underappreciated) trigonometric identity:

$$\sum_{j=1}^{p} \sin\left(\frac{j\pi}{p}\right) \;=\; \cot(\frac{\pi}{2p})$$

(Demonstration of this verity is found in appendix 1.)

Returning to eq. (21) , we have, as promised,

$$b_1 > \beta_{max} \frac{\cot(\frac{\pi}{2p})}{\sin(\frac{\pi}{2p})} = \beta_{max} \frac{\cot(\frac{\pi}{n+1})}{\sin(\frac{\pi}{n+1})} = \beta_{max} \frac{\cos(\frac{\pi}{n+1})}{\sin^2(\frac{\pi}{n+1})} = \beta_{max} \frac{\lambda_1}{1 - \lambda_1^2}$$

Now, for small x, sin (x) ≈ x and cos (x) ≈ 1, to a first-order approximation. So the fractional expression on the r.h.s is approximately equal to $(2p)^2/\pi^2$, *i.e.* $(n+1)^2/\pi^2$, since n + 1 = 2 p.

We can do better than an asymptotic expression for the scaling factor in this case, since in fact $(1/x)^2 > \cos(x)/\sin^2(x)$, $x \neq 0$, as shown in Appendix 2. This gives the following result: if

$$b_1 \geq \left(\frac{n+1}{\pi} \right)^2 \beta_{max}$$

then every odd node is turned on, every even node off, for n odd.

CONCLUSIONS.

1. **The network only shows two stable states, both strictly alternating.**

This follows because the fate of each node x_k is determined by the sign of the coefficient of the fastest-growing exponential term $e^{\lambda_1 t}$ in the function **x**(t) describing the path of steepest ascent. We saw that such coefficients alternate in sign according to the pattern $(-1)^{k+1}$.

The only circumstance under which this coefficient is nondetermining is when it is 0. In this case, the path of gradient ascent is determined by the next most rapidly growing term. However, the conditions under which this coefficient zeroes out are rather delicate. We have (cf. eq. (14))

(22)

$$\sum_{j=1}^{n} (-1)^{j+1} \sin(\frac{j\pi}{n+1}) b_j = 0$$

Or, equivalently,

(23)

$$\sum_{j=1}^{[\frac{n+1}{2}]} \sin\left(\frac{(2j-1)\pi}{n+1} \right) b_{2j-1} = \sum_{j=1}^{[\frac{n+1}{2}]} \sin\left(\frac{(2j)\pi}{n+1} \right) b_{2j}$$

To appreciate the delicacies involved, consider just the case n = 3, for which we have

$$b_1 + b_3 = \sqrt{2}\, b_2$$

This emphasize the degree of accuracy required by such equality conditions.. Hence, we regard this condition as 'non-stable' in the sense that, for all values of the b_i, any tiny perturbation in the value

of a single bias will move the network out of the state. In contrast, in the stable states there are (plentiful) regions where even large changes in a bias will have no effect on the outcome.
We conclude that although the finer structure of the network's behavior is of theoretical interest, for practical purposes it can be put aside.

2. **Exponential separation within a bias system.**

Suppose we have a prominence system like that determining Berber syllabification or Hindi stress (Kelkar dialect), as discussed in Prince & Smolensky 1993. The main property of interest is this: there are several distinct classes of intrinsic prominence, and the peak goes to the *most* intrinsically prominent element present in a given string. (Behavior when ties are present is not relevant to our concerns here.) Considering only those strings with a unique most-prominent element, we must have it that an element of class K will earn the peak if no elements of classes 1, 2,..., K−1 are present, no matter how many competing elements of classes K+1, ...,N are present.

From the above results, it follows that in order to achieve this result the bias representing the intrinsic prominence of class K must be distinguished from the bias representing class K+1 by a *scaling factor* α. Hence Bias(K) = α Bias(K+1), for appropriately chosen α. This entails that the biases of the various classes must grow exponentially in α.

3. **Dependence on String Length.**

The asymptotic value of the scaling factor needed to assure peak-achievement in an n-length network is $(n+1)^2 / \pi^2$. The factor α just mentioned must therefore be larger than this. Aside from questions of implementation — this factor grows fairly speedily, so that eg. a length-10 network involves $\alpha >$ 12.26, and a length-20 network requires $\alpha >$ 44.68, with a 5-deep hierarchy of biases demanding a largest/smallest ratio of about 4,000,000:1 — the mere fact of sensitive length- dependence is nonlinguistic in character. Significant nonlinearities are required to achieve the next level of refinement in modeling.

One promising line is to place a floor on activation at 0, as in the original Miyata-Smolensky model, which showed a richer range of behavior, rather than letting the nodes go unboundedly negative. As Bruce Tesar has observed to me, it is essentially this latter property that produces the rigidity of alternation in the network, and strongly limits the possible effect of the biases on the range of possible outcomes, since the magnitude of the contribution of any finite bias is going to be swamped eventually by exponential growth as gradient ascent proceeds. The present analysis can be regarded as setting the stage for this next step.

Acknowledgment.

Thanks to Bruce Tesar for bringing the Miyata-Smolensky model back to mind, and to Tesar and Paul Smolensky for discussion of various issues related to this analysis.

References.

Anderson, J. A. 1977. Neural Models with Cognitive Implications. In *Basic Processes in Reading Perception and Comprehension*, D. LaBerge & S. J. Samuels, eds. Erlbaum: Hillsdate, NJ.

Prince, A. 1992. In defense of the number *i* — Anatomy of a linear dynamical model of linguistic generalizations. RuCCS-TR-1, Rutgers Center for Cognitive Science, Rutgers University/New Brunswick: Piscataway, NJ. http://ruccs.rutgers.edu/ling/fac/prince.html.

Prince, A. and P. Smolensky. 1991. Connectionism and Harmony Theory in Linguistics. Tech Report CU-CS-533-91, July 1991. University of Colorado: Boulder.

Prince, A. and P. Smolensky. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*. RuCCS-TR-2, Rutgers Center for Cognitive Science, Rutgers University/New Brunswick, Piscataway, NJ. To Appear, MIT Press.

Smolensky, P. 1986. Information processing in dynamical systems: Foundations of Harmony Theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations,* D. E. Rumelhart, J.L. McClelland, and the PDP Research Group, eds., Chapter 6: pp. 194-281. MIT Press/Bradford Books: Cambridge, MA.

Smolensky, P. , Legendre, G., and Y. Miyata. 1992. Principles for an Integrated Connectionist/ Symbolic Theory of Higher Cognition. CU-CS-600-92, Computer Science Department, University of Colorado at Boulder. Report 92-1-02, School of Computer & Cognitive Sciences, Chukyo University.

Appendix 1: Summing the Sines

We want

$$\sum_{k=1}^{m} \sin\frac{k\pi}{m} = \cot\left(\frac{\pi}{2m}\right)$$

Let us write $z = \cos\theta + i\sin\theta = e^{i\theta}$, where $\theta = \pi/m$. Consider

$$\sum_{k=1}^{m} \cos(k\theta) + i\sin(k\theta) = \sum_{k=1}^{m} e^{ik\theta}$$

$$= \sum_{k=1}^{m} z^k = z\,\frac{z^m - 1}{z - 1}$$

$$= \frac{z^{1/2}}{z^{-1/2}} \cdot \frac{z^m - 1}{z - 1} = z^{1/2} \cdot \frac{z^m - 1}{z^{1/2} - z^{-1/2}}$$

From $z = e^{i\pi/m}$, this last expression yields the following:

$$e^{i\frac{\pi}{2m}} \cdot \frac{e^{i\pi} - 1}{e^{i\frac{\pi}{2m}} - e^{-i\frac{\pi}{2m}}} = e^{i\frac{\pi}{2m}} \cdot \frac{-2}{2i\sin\left(\frac{\pi}{2m}\right)}$$

$$= i \cdot \frac{\cos\left(\frac{\pi}{2m}\right) + i\sin\left(\frac{\pi}{2m}\right)}{\sin\left(\frac{\pi}{2m}\right)}$$

$$= -1 + i\cot\left(\frac{\pi}{2m}\right)$$

Equating the imaginary part of this expression with the imaginary part of the original sum yields the desired result.

Appendix 2

We want

$$\frac{1}{x^2} > \frac{\cos(x)}{\sin^2(x)} \quad , \textit{ for } x \textit{ small and positive.}$$

Equivalently,

$$\frac{\sin^2(x)}{\cos(x)} > x^2$$

To demonstrate this, we use appropriately truncated versions of the series representation of $\sin^2$ (x) and cos (x).

$$\sin^2(x) > x^2 - \frac{x^4}{3} = \frac{3x^2 - x^4}{3}$$

$$\cos(x) < 1 - \frac{x^2}{2} + \frac{x^4}{24} = \frac{24 - 12x^2 + x^4}{24}$$

Replacing the numerator and denominator with these polynomial expressions, respectively, we have

$$\frac{\sin^2(x)}{\cos(x)} > x^2 \frac{24 - 8x^2}{24 - 12x^2 + x^4}$$

So we need

$$\frac{24 - 8x^2}{24 - 12x^2 + x^4} > 1 \quad , \textit{ i.e.}$$

$$24 - 8x^2 > 24 - 12x^2 + x^4 \quad , \textit{ i.e.}$$

$$12 - x^2 > 8$$

Which is evidently the case for x small, *i.e.* less than 2.