A EXPERIMENT DETAILS

In the synthetic experiments, $\hat{V}$ is initialized randomly so $M \in \mathbb{R}^{50\times 50}$ is constructed as a diagonal matrix without loss of generality. The linear spectrum ranges from 1 to 1000 with equal spacing. The exponential spectrum ranges from 10^3 to 10^0 with equal spacing on the exponents.

A.1 CLARIFICATION OF OJA VARIANTS

As discussed in Section 5, it is easy to confuse the various Oja methods. In our experiments, Oja's algorithm refers to applying Hebb's rule $v_i \leftarrow v_i + \eta M v_i$ followed by an orthonormalization step computed with $\texttt{QR}$ as in Algorithm 3:

Algorithm 3 Oja's Algorithm

Given: data stream, $X_t \in \mathbb{R}^{m\times d}$, T, $\hat{V}^0 \in \mathcal{S}^{d-1}\times\ldots\times\mathcal{S}^{d-1}$, step size η
$\hat{V} \leftarrow \hat{V}^0$
$\texttt{mask} \leftarrow \texttt{LT}(2I_k - \mathbf{1}_k)$
for $t = 1 : T$ **do**
  $\hat{V} \leftarrow \hat{V} + \eta X_t^\top X_t \hat{V}$
  $Q, R \leftarrow \texttt{QR}(\hat{V})$
  $S = \texttt{sign}(\texttt{sign}(\texttt{diag}(R)) + 0.5)$
  $\hat{V} = QS$
end for
return $\hat{V}$

where $\mathbf{1}_k$ is a $k \times k$ matrix of all ones, $\texttt{LT}$ returns the lower-triangular part of a matrix (includes the diagonal), and $\texttt{sign} = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$. Oja's algorithm is the standard nomenclature for this variant in the machine learning literature (Allen-Zhu and Li, 2017).

In the scaled-down RESNET experiments (see Section H.3), we use Hebb's rule with deflation, also sometimes referred to as Oja's. Deflation is accomplished by directly subtracting out the parent vectors from the dataset. In detail, each batch of data samples, $X_t \in \mathbb{R}^{m\times d}$, is preprocessed as $X_{(i),t} \leftarrow X_t(I - \sum_{j<i} \hat{v}_j \hat{v}_j^\top)$. Then to learn each $\hat{v}_i$, we repeatedly apply Hebb's rule with $M_t = X_{(i),t}^\top X_{(i),t}$ and then $\hat{v}_i \leftarrow \frac{\hat{v}_i}{||\hat{v}_i||}$ to project $\hat{v}_i$ back to the unit-shere. After several iterations t and once $\hat{v}_i$'s Rayleigh quotient appears to have stabilized, we move on to $\hat{v}_{i+1}$.

B SPECTRUM OF RESNET ACTIVATIONS

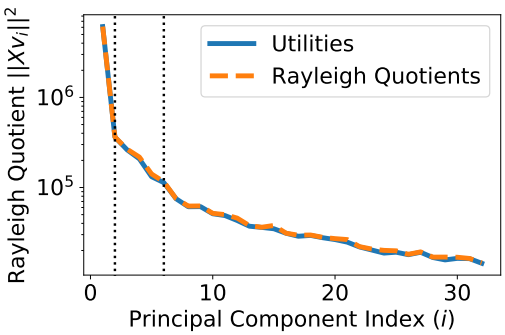
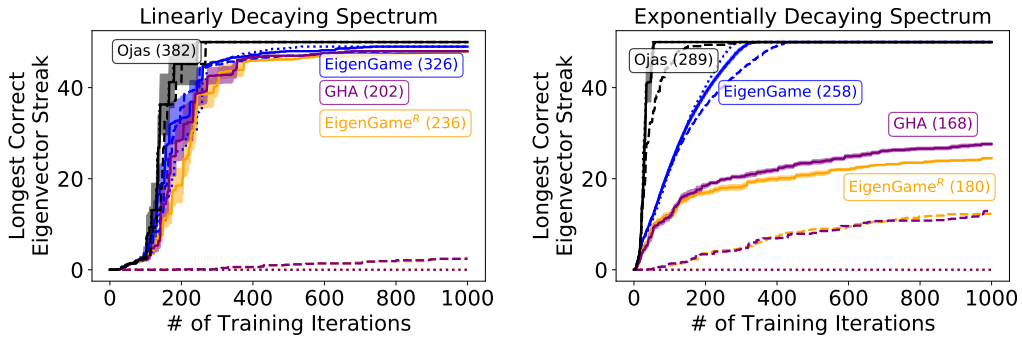


Figure 6: Approximate Eigenvalue Spectrum of RESNET-200 Activations.

Figure 6 shows a scree plot of the Rayleigh quotients recovered by EigenGame and the respective utility achieved by each player. The two curves almost perfectly overlap. The mean relative magnitude of the penalty terms to the respective Rayleigh quotient in the utility is 0.025 indicating that the solutions of each player are close to orthogonal with respect to the generalized inner product (Equation (6)). This implies that that the solutions are indeed eigenvectors. The scree plot has two distinct elbows at PC2 and PC6, corresponding to the differences in filters observed in Figure 5b.

C SYNTHETIC EXPERIMENTS—FIGURES ENLARGED



(a) Linear Spectrum

(b) Exponential Spectrum

Figure 7: The longest streak of consecutive vectors with angular error less than $\frac{\pi}{8}$ radians is plotted versus algorithm iterations for a matrix $M \in \mathbb{R}^{50\times 50}$ with a spectrum decaying from 1000 to 1 linearly (a) and exponentially (b). Average runtimes are reported in milliseconds next to the method names.[9] We omit Krasulina's as it is only designed to find the top-k subspace. Both EigenGame variants and GHA achieve similar asymptotes on the linear spectrum. Learning rates were chosen from $\{10^{-3}, \ldots, 10^{-6}\}$ on 10 held out runs. Solid lines denote results with the best performing learning rate. Dotted and dashed lines denote results using the best learning rate $\times$ 10 and 0.1. All plots show means over 10 trials. Shaded regions highlight $\pm$ standard error of the mean for the best performing learning rates.

D MNIST EXPERIMENTS—FIGURES ENLARGED

See Appendix I.

[9]EigenGame runtimes are longer than those of EigenGame^R in the synthetic experiments despite strictly requiring fewer FLOPS; apparently this is due to low-level floating point arithmetic specific to the experiments.

E RESNET-200 EXPERIMENTS—FIGURES ENLARGED

Figures 8 and 9 show enlarged versions of Figures 5a and 5b from the main body.

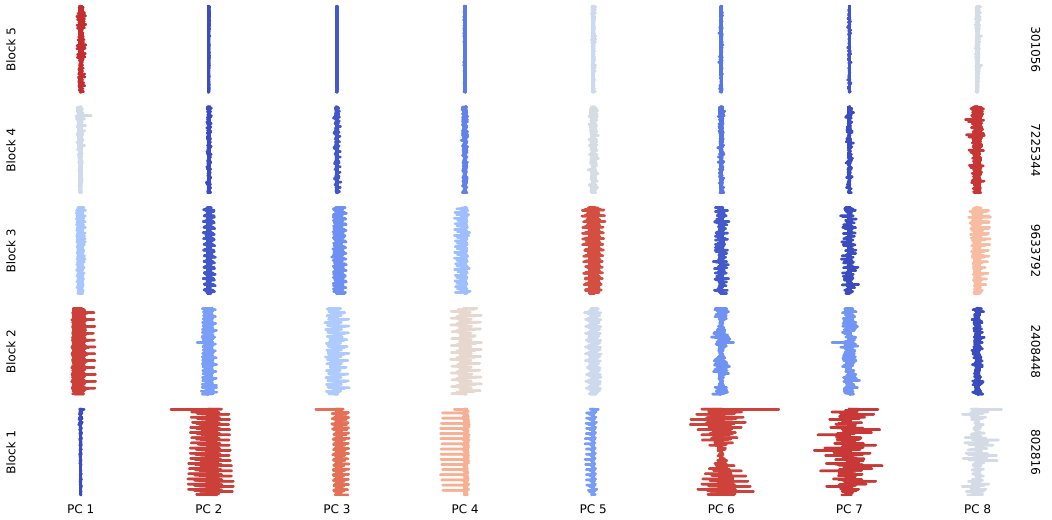


Figure 8: Top-8 principal components of the activations of a RESNET-200 on IMAGENET ordered block-wise by network topology (dimension of each block on the right y-axis). Block 1 is closest to input and Block 5 is the output of the network. Color coding is based on relative variance between blocks across the top-8 PCs from blue (low) to red (high).

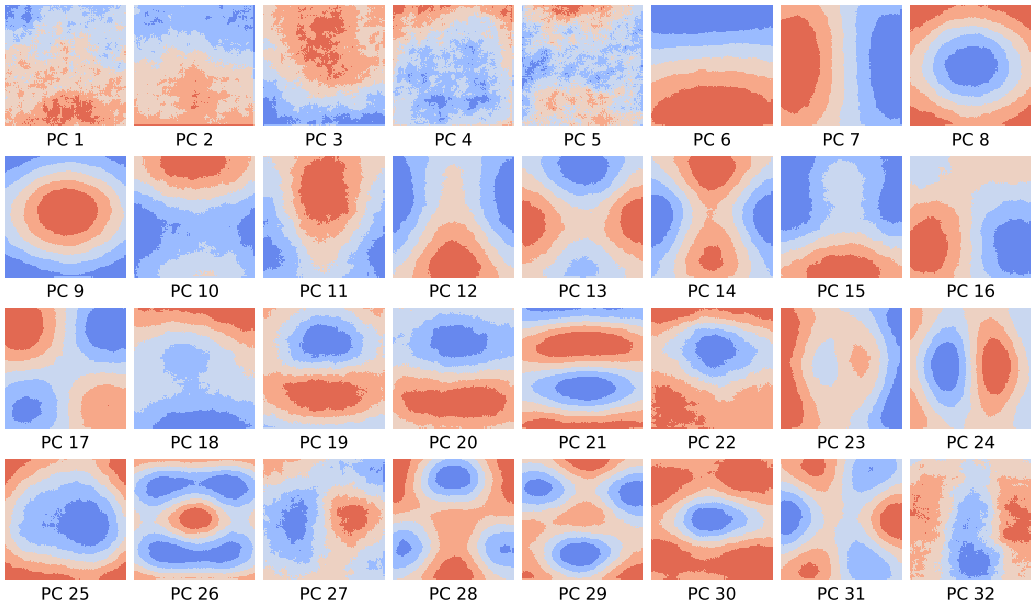


Figure 9: Block 1 mean activation maps of the top-32 principal components of RESNET-200 on IMAGENET computed with EigenGame.

F EIGENGAME VECTORIZED FOR CPU

Algorithm 4 presents Algorithm 2 in a vectorized form for implementation on a CPU. $\texttt{LT}$ returns the lower-triangular part of a matrix (includes the diagonal). $\texttt{sum}(A, \texttt{dim} = 0)$ sums over the rows of A. $\texttt{norm}(A, \texttt{dim} = 0)$ returns an array with the L_2-norm of each column of A. $\odot$ denotes elementwise

multiplication. $\mathbf{1}_k$ is a square $k \times k$ matrix of all ones. I_k is the $k \times k$ identity matrix. When dividing a matrix by a vector (A/v), we assume broadcasting. Specifically, v is interpreted as a row-vector and stacked vertically to match the dimensions of A; the two matrices are then divided element wise.

Algorithm 4 EigenGame & EigenGame^R—Vectorized

Given: data stream, $X_t \in \mathbb{R}^{m \times d}$, T, $\hat{V}^0 \in \mathcal{S}^{d-1} \times \ldots \times \mathcal{S}^{d-1}$, step size α
$\hat{V} \leftarrow \hat{V}^0$
$\texttt{mask} \leftarrow \texttt{LT}(2I_k - \mathbf{1}_k)$
for $t = 1 : T$ **do**
  $R \leftarrow (X_t\hat{V})^\top (X_t\hat{V})$
  $R_{\text{norm}} \leftarrow R/\texttt{diag}(R)$
  $G_s \leftarrow \hat{V}(R_{\text{norm}} \odot \texttt{mask})^\top$
  $\nabla_{\hat{V}} \leftarrow X_t^\top (X_t G_s)$
  $\nabla^R_{\hat{V}} \mathrel{-}= \hat{V}\texttt{sum}(\nabla_{\hat{V}} \odot \hat{V}, \texttt{dim} = 0)$
  $\hat{V} \leftarrow \hat{V} + \alpha \nabla^R_{\hat{V}}$
  $\hat{V} \leftarrow \hat{V}/\texttt{norm}(\hat{V}, \texttt{dim} = 0)$
end for
return $\hat{V}$

G SMALLEST EIGENVECTORS

EigenGame can be used to recover the k smallest eigenvectors as well. Simply use EigenGame to estimate the top eigenvector with eigenvalue Λ_{11}. Then run EigenGame on the matrix $M' = \Lambda_{11}I - M$. The top-k eigenvectors of M' are the bottom-k eigenvectors of M. For example, the dth eigenvector of M, v_d, is the largest eigenvector of M': $M'v_d = \Lambda_{11}v_d - Mv_d = (\Lambda_{11} - \Lambda_{dd})v_d$.

H FREQUENT DIRECTIONS

A reviewer from a previous submission of this work requested a comparison and discussion with Frequent Directions (Ghashami et al., 2016), another decentralized subspace-error minimizing k-PCA algorithm. Frequent Directions (FD) is a streaming algorithm that maintains an overcomplete sketch matrix with the goal of capturing the subspace of maximal variance within the span of its vectors. Each step of FD operates by first replacing a row of the sketch matrix with a single data sample. It then runs SVD on the sketch matrix and uses the resulting decomposition to construct a new sketch. Note that FD relies on SVD as a core inner step. In theory, EigenGame could replace SVD, however, we do not explore that direction here.

H.1 RECOVERING PRINCIPAL COMPONENTS FROM PRINCIPAL SUBSPACE

FD returns a sketch $B = \hat{V}^\top$ of size $\mathbb{R}^{2l \times d}$ where $l \geq k$. The rows of FD are not principal components, but they should approximate the top-k subspace of the dataset. To recover approximate principal components, the optimal rotation of the vectors can be computed with $Q \leftarrow SVD(XB^\top)$. This can be shown by inspecting R (as defined in Section 2) with rotated vectors:

$$(\hat{V}Q)^\top M(\hat{V}Q) = Q^\top \hat{V}^\top M\hat{V}Q = Q^\top (X\hat{V})^\top (X\hat{V})Q = Q^\top M'Q. \tag{9}$$

By inspection, the problem of computing the optimal Q reduces to computing the eigenvectors of $M' \in \mathbb{R}^{k \times k}$. This requires projecting the dataset into the principal subspace, $(X\hat{V})$, to compute M' however, this is typically a desired step anyways when performing PCA.

H.2 COMPLEXITY ANALYSIS

We base our analysis on Section 3.1 of (Ghashami et al., 2016) which discusses parallelizing FD. Let b be number of shards to split the original dataset $X \in \mathbb{R}^{n \times d}$ into, each shard being in $\mathbb{R}^{\frac{n}{b} \times d}$. Let k be the number of principal components sought. Finally, let $l = \lceil k + \frac{1}{\epsilon} \rceil$ be the sketch size where $\epsilon \ll 1$ is a desired tolerance on the Frobenius norm of the subspace approximation error.

The runtime of FD is $\mathcal{O}(nld)$; call this $Anld$ for some A. To decentralize FD, (Ghashami et al., 2016) instructs to

1. Split X into b shards and run FD on each individually in parallel.
   - total runtime: $A(\frac{n}{b})ld = Anld(\frac{1}{b})$
   - output: b sketches ($B_i \in \mathbb{R}^{2l \times d}$)
2. Merge sketches and run FD on the merged sketch to produce sketch B.
   - total runtime: $A(2lb)ld = Anld(\frac{2bl}{n})$
   - output: 1 sketch ($B \in \mathbb{R}^{2l \times d}$)

Finally, normalize the rows of B, project the dataset $Y \leftarrow XB^\top$, compute the right-singular vectors of the projected dataset, $Q \in \mathbb{R}^{2l \times 2l} \leftarrow SVD(Y)$, compute $\hat{V} \leftarrow B^\top Q$, and compute the corresponding Rayleigh quotients $\hat{V}^\top M \hat{V} = (YQ)^\top (YQ)$ to determine the top-k eigenvectors with error within the desired tolerance. We assume this final step takes negligible runtime because we assume $2l \ll d$, however, for datasets with many samples (large n), this step could be nonnegligible without further approximation.

Using the runtimes listed above, we can determine the potential runtime multiplier from decentralization is $(\frac{1}{b} + \frac{2bl}{n})$ which is convex in b. If we minimize this w.r.t. b for the optimal number of shards, we find $b^* = \sqrt{\frac{n}{2l}}$. Plugging this back in gives an optimal runtime multiplier of $2\sqrt{2}\sqrt{\frac{l}{n}}$.

The analysis above only considers one recursive step. Step 1) can be decentralized as well. For simplicity, we assume the computation is dominated by Step 2), the merge step. Note these relaxations result in a lower bound on FD runtime, i.e., they favor FD in a comparison with EigenGame.

H.3 SMALL IMAGENET EXPERIMENTS

Consider running on a scaled down RESNET-50 experiment which has approximately $1.2M$ images ($n = 1.2 \times 10^6$, 24TB) and searching for the top-25 eigenvectors ($k = 25$). Using a modest $\epsilon = \frac{0.25}{k}$ implies $l = 5k = 125$ with optimal batch size $b^* \approx 70$. Therefore, running FD on $\frac{n}{b}$ samples with a sketch size of 125 should give a rough lower bound on the runtime for an optimally decentralized FD implementation. The runtime obtained was 9 hours for FD vs 2 hours for EigenGame which actually processes the full dataset 3 times.

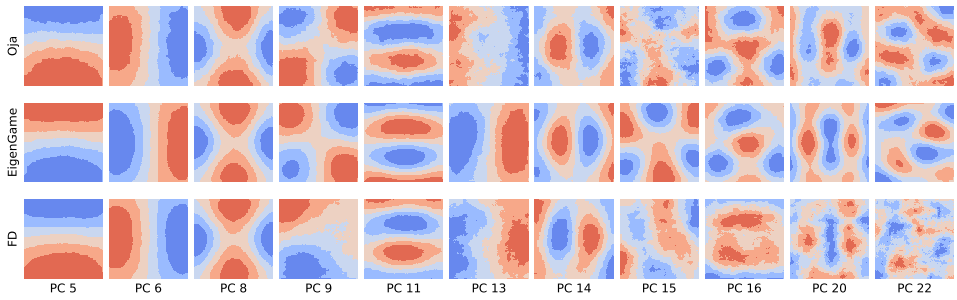


Figure 10: Comparison of mean activation maps between Oja's with deflation, EigenGame, and FD for a section of the top principal components of RESNET-50 on IMAGENET.

The reason we run FD on a scaled down RESNET-50 experiment as opposed to the RESNET-200 is that the algorithm requires a final SVD step to recover the actual eigenvectors and we were not able

to run SVD on a sketch of size $k \times d$ where $d = 20 \times 10^6$ for the full scale experiment. That is to say FD is not applicable in this extremely large data regime. In contrast, EigenGame handles this setting without modification.

To obtain an approximate "ground truth" solution for the principal components we run Oja's algorithm with a low learning rate with a batch size of 128 for 3 epochs to extract the first eigenvector. We find successive eigenvectors using deflation. By running each step for many iterations and monitoring the convergence of the Rayleigh quotient (eigenvalue) $v_i^\top M v_i$, we can control the quality of the recovered eigenvectors. This is the simplest and most reliable approach to creating ground truth on a problem where no solution already exists. See Section A.1 for further details.

I GRADIENT BIAS

As expected, Figure 11 shows the performance of EigenGame degrades in the low batch size regime. This is expected because we use the same minibatch for all inner products in the gradient which contains products and ratios of random variables. GHA, on the other hand, is linear in the matrix M and as such is naturally unbiased. However, GHA does not appear to readily extend to more general function approximators, whereas EigenGame should. Instead we look to reduce the bias of EigenGame gradients using larger batch sizes (current hardware easily supports batches of 1024 for MNIST and 128 for IMAGENET). Further reducing bias is left to future work.

J TO PROJECT OR NOT TO PROJECT?

Projecting the update direction onto the unit-sphere, as suggested by Riemannian optimization theory, can result in much larger update steps. This effect is due to the composition of the retraction ($z' \leftarrow \tilde{z}/||\tilde{z}||$) and update step ($\tilde{z} \leftarrow z + \Delta z$). Omitting the projection can actually mimic modulating the learning rate, decaying it near an equilibrium and improving stability.

K THEORETICAL COMPARISON WITH GHA

Proposition K.1. *When the first $i-1$ eigenvectors have been learned exactly, GHA on $\hat{v}_i$ is equivalent to projecting the first term in $\nabla_{\hat{v}_i} u_i$ onto the sphere, but omitting to project the second set of penalty terms.*

Proof. The GHA update is

$$\Delta \hat{v}_i = 2\Big[M\hat{v}_i - (\hat{v}_i^\top M \hat{v}_i)\hat{v}_i - \sum_{j<i} (\hat{v}_i^\top M \hat{v}_j)\hat{v}_j\Big]. \tag{10}$$

Plugging $v_{j<i}$ for $\hat{v}_{j<i}$ into the GHA update, we find

$$\Delta_i = 2\Big[M\hat{v}_i - (\hat{v}_i^\top M \hat{v}_i)\hat{v}_i - \sum_{j<i} (\hat{v}_i^\top M v_j)v_j\Big] \tag{11}$$

$$= 2\Big[M\hat{v}_i - (\hat{v}_i^\top M \hat{v}_i)\hat{v}_i - \sum_{j<i} \Lambda_{jj}(\hat{v}_i^\top v_j)v_j\Big]. \tag{12}$$

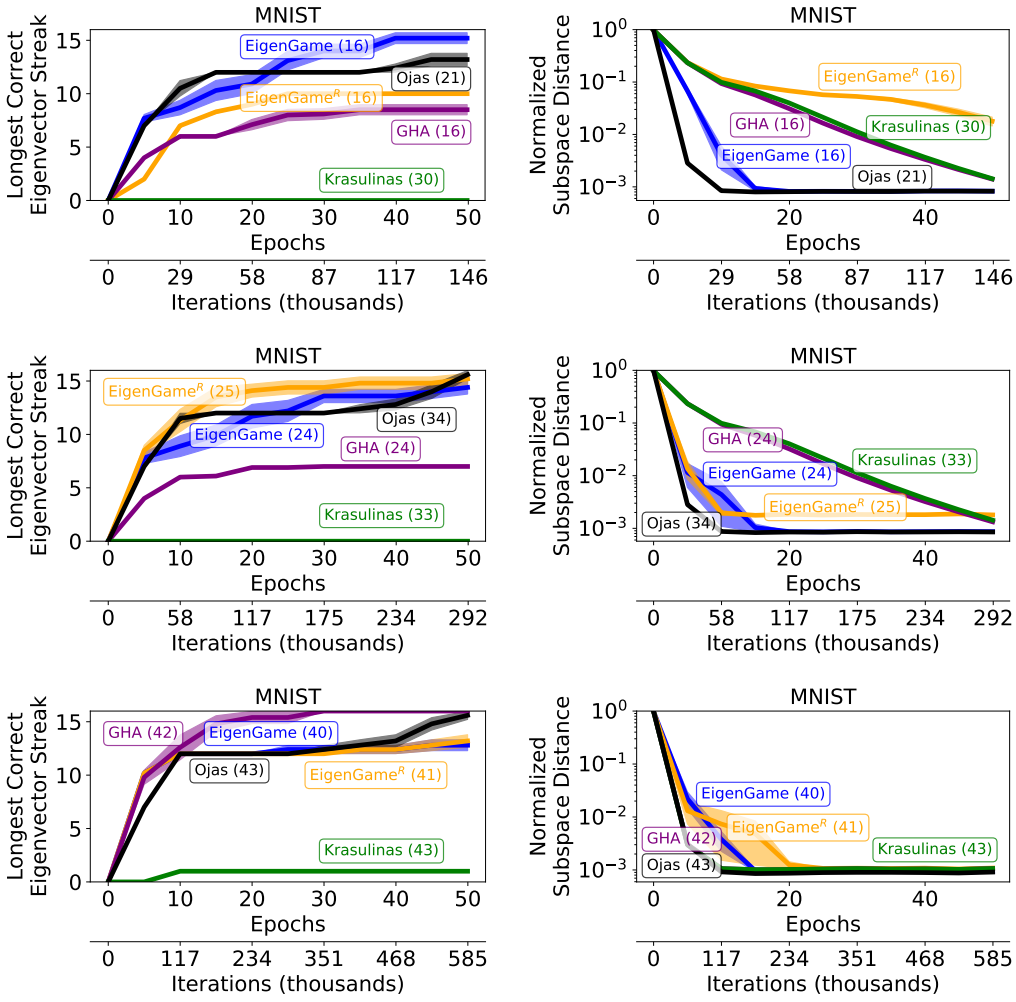
Likewise for the gradient with the first term projected onto the tangent space of sphere:

$$2\Big[(I - \hat{v}_i\hat{v}_i^\top)M\hat{v}_i - M\sum_{j<i} \frac{\hat{v}_i^\top M v_j}{v_j^\top M v_j} v_j\Big] = 2\Big[(I - \hat{v}_i\hat{v}_i^\top)M\hat{v}_i - M\sum_{j<i} (\hat{v}_i^\top v_j)v_j\Big] \tag{13}$$

$$= 2\Big[M\hat{v}_i - (\hat{v}_i^\top M \hat{v}_i)\hat{v}_i - \sum_{j<i} \Lambda_{jj}(\hat{v}_i^\top v_j)v_j\Big]. \tag{14}$$

□

Proposition K.2. *The GHA update for $\hat{v}_i$ is not the gradient of any function.*



(a) Longest Streak (b) Subspace Distance

Figure 11: (a) The longest streak of consecutive vectors with angular error less than $\frac{\pi}{8}$ radians is plotted vs algorithm iterations on MNIST for minibatch sizes of 1024 (top), 512 (middle), and 256 (bottom). Shaded regions highlight $\pm$ standard error of the mean for the best performing learning rates. Average runtimes are reported in seconds next to the method names. (b) Subspace distance on MNIST. (a,b) Learning rates were chosen from $\{10^{-3}, \ldots, 10^{-6}\}$ on 10 held out runs. All plots show means over 10 trials.

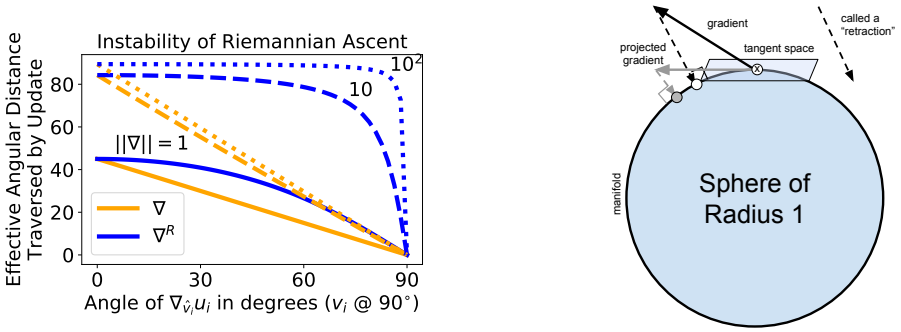
Proof. The Jacobian of $\Delta\hat{v}_i$ w.r.t. $\hat{v}_i$ is

$$Jac(\Delta\hat{v}_i) = 2\Big[M - (\hat{v}_i^\top M\hat{v}_i)I - 2\hat{v}_i\hat{v}_i^\top M - \sum_{j<i}\hat{v}_j\hat{v}_j^\top M\Big]. \tag{15}$$

The sum of the $\hat{v}\hat{v}^\top M$ terms are not, in general, symmetric, therefore, the Jacobian is not symmetric. The Jacobian of a gradient is the Hessian and the Hessian of a function is necessarily symmetric, therefore, the GHA update is not the gradient of any function. □

K.1 Design Decisions

We made a number of algorithmic design decisions that led us to the proposed algorithm. The first to note is that a naive utility that simply subtracts off $\sum_{j<i}\langle\hat{v}_i, \hat{v}_j\rangle$ will not solve PCA. This is because large $\langle\hat{v}_i, M\hat{v}_i\rangle$ (read eigenvalues) can drown out these penalties. The intuition is that including M in the inner product gives the right boost to create a natural balance among terms. Next, it is possible



(a) Gradient Instability Near Optimum

(b) Riemannian Terminology

Figure 12: (a) When the $\hat{v}_i$ is near the optimum of its utility and its gradient is nearly orthogonal to the sphere, pointing directly away from the center (@ 90°), the combination of updating using the projected gradient (∇^R) and the retraction can result in a large update, possibly moving $\hat{v}_i$ away from the optimum. (b) Diagram presenting Riemannian optimization terminology. The retraction is not a projection in general although our specific choice appears that way for the sphere. A retraction applied at $\hat{v}_i$ takes as input a scaled projected gradient and returns a vector on the manifold: $\hat{v}_i' \leftarrow R_{\hat{v}_i}(\alpha \nabla^R)$.

to formulate the utilities without normalizing the terms as we did, however, this is harder to analyze and is akin to minimizing $(err)^4$ instead of $(err)^2$ which generally has better convergence properties near optima. Also, while updates formed using the standard Euclidean Gram-Schmidt procedure will solve the PCA problem, they are not the gradients of any utility function. Lastly, our formulation consists entirely of generalized inner products: $\langle \hat{v}_i, M\hat{v}_j \rangle = \langle X\hat{v}_i, X\hat{v}_j \rangle$. Each $X\hat{v}_i$ can be thought of as a shallow function approximator with weights $\hat{v}_i$. This means that our formulation is readily extended to more general function approximation, i.e., $X\hat{v}_i \rightarrow f_i(X)$[10]. Note that any formulation that operates on $\langle \hat{v}_i, \hat{v}_j \rangle$ instead is not easily generalized.

L NASH PROOF

Let $\hat{V}$ be a matrix of arbitrary unit-length column vectors ($\hat{v}_j$) and let M (symmetric) be diagonalized as $U\Lambda U^\top$ with U a unitary matrix. Then,

$$R \stackrel{\text{def}}{=} \hat{V}^\top M \hat{V} = \hat{V}^\top U \Lambda U^\top \hat{V} = (U^\top \hat{V})^\top \Lambda (U^\top \hat{V}) = Z^\top \Lambda Z \tag{16}$$

where Z is also a matrix of unit-length column vectors because unitary matrices preserve inner products ($\langle U^\top \hat{v}_i, U^\top \hat{v}_i \rangle = \hat{v}_i^\top U U^\top \hat{v}_i = \hat{v}_i^\top \hat{v}_i = 1$). Therefore, rather than considering the action of an arbitrary matrix $\hat{V}$ on M, we can consider the action of an arbitrary matrix Z on Λ. This simplifies the analysis.

In light of this reduction, Equation (22) of Theorem L.1 can be rewritten as

$$u_i(\hat{v}_i | v_{j<i}) = w^\top \Lambda_{jj \geq ii} w \tag{17}$$

$$= \hat{v}_i^\top \Lambda_{jj \geq ii} \hat{v}_i \tag{18}$$

because V is identity w.l.o.g. Therefore, player i's problem is simply to find the maximum eigenvector of a transformed matrix $\Lambda_{jj \geq ii}$, i.e., Λ with the first $i-1$ eigenvalues removed.

Theorem L.1 (PCA Solution is the Unique strict-Nash Equilibrium)**.** *Assume that the top-k eigenvalues of $X^\top X$ are positive and distinct. Then the top-k eigenvectors form the unique strict-Nash equilibrium of the proposed game in Equation (6).*

Proof. In what follows, let $p, q = \{1, \ldots, d\}$ and $i \in \{1, \ldots, k\}$. We will prove optimality of v_i by induction. Clearly, v_1 is the optimum of u_1 because $u_1 = \langle v_1, Mv_1 \rangle = \frac{\langle v_1, Mv_1 \rangle}{\langle v_1, v_1 \rangle} = \Lambda_{11}$ is the *Rayleigh* quotient which is known to be maximized for the maximal eigenvalue (Horn and Johnson,

[10]Empirically, replacing $||\hat{v}_i|| = 1$ with $||\hat{v}_i|| \leq 1$ does not harm performance while the latter is easier to enforce on neural networks for example (Virmaux and Scaman, 2018).

2012). Now, Consider $\hat{v}_i = \sum_{p=1}^{d} w_p v_p$ as a linear combination of the true eigenvectors. To ensure $||\hat{v}_i|| = 1$, we require $||w|| = 1$. Then,

$$u_i(\hat{v}_i|v_{j<i}) = \hat{v}_i^\top M \hat{v}_i - \sum_{j<i} \frac{(\hat{v}_i^\top M v_j)^2}{v_j^\top M v_j} = \hat{v}_i^\top M \hat{v}_i - \sum_{j<i} \frac{(\hat{v}_i^\top M v_j)^2}{\Lambda_{jj}} \tag{19}$$

$$= \Big(\sum_p \sum_q w_p w_q v_p^\top M v_q \Big) - \sum_{j<i} \Big(\sum_p w_p v_p^\top M v_j \Big)^2 / \Lambda_{jj} \tag{20}$$

$$= \Big(\sum_p \sum_q w_p w_q \Lambda_{qq} v_p^\top v_q \Big) - \sum_{j<i} \Big(\sum_p w_p \Lambda_{jj} v_p^\top v_j \Big)^2 / \Lambda_{jj} \tag{21}$$

$$= \sum_q w_q^2 \Lambda_{qq} - \sum_{j<i} \Lambda_{jj} w_j^2 = \sum_{p \geq i} \Lambda_{pp} z_p \tag{22}$$

where $z_p = w_p^2$, and $z \in \Delta^{d-1}$ which is a linear optimization problem over the simplex. For distinct Λ_{pp} with $\Lambda_{ii} > 0$, $z^* = \arg\max(\Lambda_{pp \geq ii}) = e_i$ is unique. Assume each player i plays e_i. Any player j that unilaterally deviates from e_j strictly decreases their utility, therefore, the Nash is unique up to a sign change due to $z^* = e_i = w_i^2$. This is expected as both v_i and $-v_i$ are principal components. □

M WITHOUT THE HIERARCHY

In Section 2, we defined utilities to respect the natural hierarchy of eigenvectors sorted by eigenvalue and mentioned that this eased analysis. Here, we provide further detail as to the difficulty of analyzing the game without the hierarchy. Consider the following alternative definition of the utilities:

$$u_i(\hat{v}_i|\hat{v}_{-i}) = \hat{v}_i^\top M \hat{v}_i - \sum_{j \neq i} \frac{(\hat{v}_i^\top M \hat{v}_j)^2}{\hat{v}_j^\top M \hat{v}_j} \tag{23}$$

where the sum is now over all $j \neq i$ instead of $j < i$ as in Equation (6). With this form, the game is now symmetric across all players i. Despite the symmetry of the game, we can easily rule out the existence of a symmetric Nash.

Proposition M.1. *The EigenGame defined using symmetric utilities in Equation (23) does not contain a symmetric Nash equilibrium (assuming $k \geq 2$ and $rank(M) \geq 2$).*

Proof by Contradiction. Assume a symmetric Nash exists, i.e., $\hat{v}_i = \hat{v}_j$ for all i, j. The utility of a symmetric Nash using equation Equation (23) is

$$u_i(\hat{v}_i|\hat{v}_{-i}) = (1 - (n-1))(\hat{v}_i^\top M \hat{v}_i) = (2-n)(\hat{v}_i^\top M \hat{v}_i) \leq 0. \tag{24}$$

Consider a unilateral deviation of $\hat{v}_i$ to a direction orthogonal to $\hat{v}_i$, i.e., $\hat{v}_\perp \perp \hat{v}_i$ such that

$$u_i(\hat{v}_\perp, \hat{v}_{-i}) = (\hat{v}_\perp^\top M \hat{v}_\perp) > 0. \tag{25}$$

This utility is positive because $rank(M) \geq 2$ and therefore, always greater than the supposed Nash. Therefore, there is no symmetric Nash. □

We can also prove that the true PCA solution is **a** Nash of this version of EigenGame.

Proposition M.2. *The the top-k eigenvectors of M form a strict-Nash equilibrium of the EigenGame defined using symmetric utilities in Equation (23) (assuming $rank(M) \geq k$).*

Proof. Let $\hat{v}_i = v_i$. We will assume this standard ordering, however, the proof follows through for any permutation of the eigenvectors. Clearly, the largest eigenvector is a best response to the spectrum because the penalty term (2nd term in Equation (23)) cannot be decreased below zero and the Rayliegh term (first term) is maximal, i.e., $v_1 = \arg\max_{\hat{v}_1} u_1(\hat{v}_1, v_{-1})$. So assume v_i is another eigenvector and consider representing $\hat{v}_i$ as $\hat{v}_i = \sum_{p=1}^{d} w_p v_p$ as before in Section L. Repeating those same steps, we find

$$u_i(\hat{v}_i, v_{-i}) = \sum_q w_q^2 \Lambda_{qq} - \sum_{j \neq i} \Lambda_{jj} w_j^2 = \Lambda_{ii} z_i \tag{26}$$

where $z_k = w_k^2$, $z \in \Delta^{n-1}$. Assuming $\Lambda_{ii} > 0$, this objective is uniquely maximized for $z_i = 1$ and $z_k = 0$ for all $k \neq i$. Therefore, $v_i = \arg\max_{\hat{v}_i} u_i(\hat{v}_i, v_{-i})$.

□

However, we were unable to prove that it is the **only** Nash. It is possible that other Nash equilibria exist. Instead of focusing on determining whether a second Nash equilibrium exists (which is NP-hard (Daskalakis et al., 2009; Gilboa and Zemel, 1989)), we learned through experiments that the EigenGame variant that incorporates knowledge of the hierarchy is much more performant. We leave determininig uniquess of the PCA solution for the less performant variant as an academic exercise.

N ERROR PROPAGATION

N.1 GENERALITIES

Notation. We can parameterize a vector on the sphere using the Riemannian exponential map, Exp, applied to a vector deviation from an anchor point. Formally, let $\hat{v}_j = \text{Exp}_{v_j}(\theta_j \Delta_j) = \cos(\theta_j) v_j + \sin(\theta_j)\Delta_j$ where v_j is the jth largest eigenvector and $\Delta_j \in \mathcal{S}^{d-1}$ is such that $\langle \Delta_j, v_j \rangle = 0$. Therefore, θ_j measures how far $\hat{v}_j$ deviates from v_j in radians and Δ_j denotes the direction of deviation.

Let Λ_{ii} denote the ith largest eigenvalue and v_i the associated eigenvector. Also define the eigenvalue gap $g_i = \Lambda_{ii} - \Lambda_{i+1,i+1}$. Finally, let $\kappa_i = \frac{\Lambda_{11}}{\Lambda_{ii}}$ denote the ith condition number.

The following Lemma decomposes the utility of a player when the parents have learnt the preceding eigenvectors perfectly.

Lemma N.1. *Let $\hat{v}_i = \cos(\theta_i) v_i + \sin(\theta_i)\Delta_i$ without loss of generality. Then*

$$u_i(\hat{v}_i, v_{j<i}) = u_i(v_i, v_{j<i}) - \sin^2(\theta_i)\Big(\Lambda_{ii} - \sum_{l>i} z_l \Lambda_{ll}\Big). \tag{27}$$

Proof. Note that Δ_i can also be decomposed as $\Delta_i = \sum_{l=1}^{d} w_l v_l, ||w|| = 1$ without loss of generality and that by Theorem L.1, this implies $u_i(\Delta_i, v_{j<i}) = \sum_{l \geq i} z_l \Lambda_{ll}$. This can be simplified further because $\langle \Delta_i, v_i \rangle = 0$ by its definition, which implies that $z_i = 0$. Therefore, more precisely, $u_i(\Delta_i, v_{j<i}) = \sum_{l>i} z_l \Lambda_{ll}$. Continuing we find

$$\begin{aligned}
u_i(\hat{v}_i, v_{j<i}) &= \langle \hat{v}_i, \Lambda \hat{v}_i \rangle - \sum_{j<i} \frac{\langle \hat{v}_i, \Lambda v_j \rangle^2}{\langle v_j, \Lambda v_j \rangle} && (28)\\
&= \langle \hat{v}_i, \Lambda \hat{v}_i \rangle - \sum_{j<i} \Lambda_{jj} \langle \hat{v}_i, v_j \rangle^2 && (29)\\
&= (\cos^2(\theta_i)\Lambda_{ii} + \sin^2(\theta_i)\langle \Delta_i, \Lambda \Delta_i \rangle) - \sum_{j<i} \Lambda_{jj} \langle \cos(\theta_i) v_i + \sin(\theta_i)\Delta_i, v_j \rangle^2 && (30)\\
&= (\cos^2(\theta_i)\Lambda_{ii} + \sin^2(\theta_i)\langle \Delta_i, \Lambda \Delta_i \rangle) - \sum_{j<i} \Lambda_{jj} \sin^2(\theta_i) \langle \Delta_i, v_j \rangle^2 && (31)\\
&= \Lambda_{ii} - \sin^2(\theta_i)\Lambda_{ii} + \sin^2(\theta_i)\Big[\langle \Delta_i, \Lambda \Delta_i \rangle - \sum_{j<i} \Lambda_{jj} \langle \Delta_i, v_j \rangle^2\Big] && (32)\\
&= u_i(v_i, v_{j<i}) - \sin^2(\theta_i)\Big(\Lambda_{ii} - u_i(\Delta_i, v_{j<i})\Big) && (33)\\
&= u_i(v_i, v_{j<i}) - \sin^2(\theta_i)\Big(\Lambda_{ii} - \sum_{l>i} z_l \Lambda_{ll}\Big). \quad \text{[TL.1]} && (34)
\end{aligned}$$

□

N.2 SUMMARY OF ERROR PROPAGATION RESULTS

Player i's utility is sinusoidal in the angular deviation of θ_i from the optimum. The amplitude of the sinusoid varies with the direction of the angular deviation along the sphere and is dependent on the accuracy of players $j < i$. In the special case where players $j < i$ have learned the top-$(i-1)$ eigenvectors exactly, player i's utility simplifies (see Lemma N.1) to

$$u_i(\hat{v}_i, v_{j<i}) = \Lambda_{ii} - \sin^2(\theta_i)\Big(\Lambda_{ii} - \sum_{l>i} z_l \Lambda_{ll}\Big). \tag{35}$$

Note that $\sin^2$ has period π as opposed to 2π, which simply reflects the fact that v_i and $-v_i$ are both eigenvectors.

The angular distance between v_i and the maximizer of player i's utility with approximate parents has $\tan^{-1}$ dependence (i.e., a soft step-function; see Lemma N.5). Figure 13 plots the dependence for a

synthetic problem. This dependence reveals that there is an error threshold players $j < i$ must fall below in order for player i to accurately learn the i-th eigenvector.

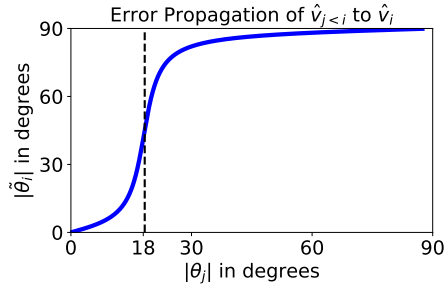


Figure 13: Example 1 demonstrates that the angular error (x-axis) in the learned parents $\hat{v}_{j<i}$ must fall below a threshold (e.g., $\approx 18°$ here) in order for the maximizer of player i's utility to lie near the true ith eigenvector (y-axis). The matrix M for this example has a condition number $\kappa_i = \frac{\Lambda_{11}}{\Lambda_{ii}} = 10$.

N.3 Theorem and Proofs

In Theorem N.2, we prove that given parents close enough to their corresponding true eigenvectors, the angular deviation of a local maximizer of a child's utility from the child's true eigenvector is below a derived threshold. In other words, given accurate parents, a child can succesfully proceed to approximate its corresponding eigenvector (its utility is well posed). We prove this theorem in several steps.

First we show in Lemma N.3 that the child's utility function can be written as a composition of sinusoids with dependence on the angular deviation from the child's true eigenvector. The amplitude of the sinusoid depends on the directions in which the child and parents have deviated from their true eigenvectors along their spheres. We then simplify the composition of sinusoids to a single sinusoid in Lemma N.4. Any local max of a sinusoid is also a global max. Therefore, to upper bound the angular deviatiation of the child's local maximizer from its true corresponding eigenvector, we consider the worst case direction for the maximizer to deviate from the true eigenvector.

In Lemma N.5, we give a closed form solution for the angular deviation of a maximizer of a child's utility given any parents and deviation directions. This dependence is given by the $\arctan$ function which resembles a *soft* step function with a linear regime for small angular deviations, followed by a step, and then another linear regime for large angular deviations. The argument of the $\arctan$ is a ratio of terms, each with dependence on the parents' angular deviations and directions of deviation. We establish two minor lemmas, Lemma N.6 and Lemma N.7, to help bound the denominator in Lemma N.8. We then tighten the bounds on the ratio assuming parents with error below a certain threshold ("left" of the step) in Lemmas N.9, N.10, and N.11. Finally, using these bounds on the argument to the $\arctan$, we are able to bound the angular deviation of any maximizer of the child's utility in Lemma N.2 given any deviation direction for the child or parents.

Theorem N.2. *Assume it is given that* $|\theta_j| \leq \frac{c_i g_i}{(i-1)\Lambda_{11}} \leq \sqrt{\frac{1}{2}}$ *for all* $j < i$ *with* $0 \leq c_i \leq \frac{1}{16}$. *Then*

$$|\theta_i^*| = |\arg\max_{\theta_i} u_i(\hat{v}_i(\theta_i, \Delta_i), \hat{v}_{j<i})| \leq 8c_i. \tag{36}$$

Proof. By Lemma N.11, $A < 0$ for $c_i < \frac{1}{8}$. Therefore, $|\theta_i^*| = \frac{1}{2}\tan^{-1}\left|\frac{B}{A}\right|$ by Lemma N.5. Also, note that for $z \leq \frac{1}{2}$, $\tan^{-1}(|z|) \leq |z|$. Setting $c_i \leq \frac{1}{16}$ to ensures $z = |\frac{B}{A}| \leq \frac{1}{2}$. Then,

$$|\theta_i^*| = \frac{1}{2}\tan^{-1}\left|\frac{B}{A}\right| \leq \frac{1}{2}|\frac{B}{A}| \overset{LN.11}{\leq} \frac{1}{2}\frac{8c}{1-8c_i} \leq 8c_i. \tag{37}$$

□

Lemma N.3. *Let $\hat{v}_j = \cos(\theta_j)v_j + \sin(\theta_j)\Delta_j$ for all $j \leq i$ without loss of generality. Then*

$$u_i(\hat{v}_i, \hat{v}_{j<i}) = A(\theta_j, \Delta_j, \Delta_i)\sin^2(\theta_i) - B(\theta_j, \Delta_j, \Delta_i)\frac{\sin(2\theta_i)}{2} + C(\theta_j, \Delta_j, \Delta_i) \tag{38}$$

where

$$A(\theta_j, \Delta_j, \Delta_i) = ||\Delta_i||_{\Lambda^{-1}} - \Lambda_{ii} \tag{39}$$

$$- \sum_{j<i} \frac{\Lambda_{jj}^2 \cos^2(\theta_j)\langle\Delta_i, v_j\rangle^2 - \Lambda_{ii}^2 \sin^2(\theta_j)\langle\Delta_j, v_i\rangle^2 + \sin^2(\theta_j)\langle\Delta_i, \Lambda\Delta_j\rangle^2}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||_{\Lambda^{-1}}\sin^2(\theta_j)} \tag{40}$$

$$- \sum_{j<i} \frac{\Lambda_{jj}\sin(2\theta_j)\langle\Delta_i, v_j\rangle\langle\Delta_i, \Lambda\Delta_j\rangle}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||_{\Lambda^{-1}}^2\sin^2(\theta_j)} \tag{41}$$

$$B(\theta_j, \Delta_j, \Delta_i) = \sum_{j<i} \frac{\Lambda_{ii}\Lambda_{jj}\sin(2\theta_j)\langle\Delta_j, v_i\rangle\langle\Delta_i, v_j\rangle + 2\Lambda_{ii}\sin^2(\theta_j)\langle\Delta_j, v_i\rangle\langle\Delta_i, \Lambda\Delta_j\rangle}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||_{\Lambda^{-1}}^2\sin^2(\theta_j)} \tag{42}$$

$$C(\theta_j, \Delta_j, \Delta_i) = \Lambda_{ii} - \sum_{j<i} \frac{\Lambda_{ii}^2\sin^2(\theta_j)\langle\Delta_j, v_i\rangle^2}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||_{\Lambda^{-1}}^2\sin^2(\theta_j)}. \tag{43}$$

We abbreviate the above to A, B, C to avoid clutter in all upcoming statements and proofs. These functions are dependent on all variables ***except*** *θ_i.*

Proof. Note that the true eigenvectors are orthogonal, so in what follows, any $\langle v_i, v_j\rangle = 0$ where $j \neq i$. Also, recall that $2\sin(z)\cos(z) = \sin(2z)$. We highlight some but not all such simplifications. Finally, we recognize $\langle\Delta_i, \Lambda\Delta_i\rangle = ||\Delta_i||_{\Lambda^{-1}}$ as the generalized norm of Δ_i or the Mahalanobis distance from the origin.

$$u_i(\hat{v}_i, \hat{v}_{j<i}) \tag{44}$$

$$= \langle\hat{v}_i, \Lambda\hat{v}_i\rangle - \sum_{j<i}\frac{\langle\hat{v}_i, \Lambda\hat{v}_j\rangle^2}{\langle\hat{v}_j, \Lambda\hat{v}_j\rangle} \tag{45}$$

$$= \langle\cos(\theta_i)v_i + \sin(\theta_i)\Delta_i, \Lambda\big(\cos(\theta_i)v_i + \sin(\theta_i)\Delta_i\big)\rangle$$
$$- \sum_{j<i}\frac{\langle\cos(\theta_i)v_i + \sin(\theta_i)\Delta_i, \Lambda\big(\cos(\theta_j)v_j + \sin(\theta_j)\Delta_j\big)\rangle^2}{\langle\cos(\theta_j)v_j + \sin(\theta_j)\Delta_j, \Lambda\big(\cos(\theta_j)v_j + \sin(\theta_j)\Delta_j\big)\rangle} \tag{46}$$

$$= \Lambda_{ii}\cos(\theta_i)^2 + \langle\Delta_i, \Lambda\Delta_i\rangle\sin^2(\theta_i)$$
$$- \sum_{j<i}\frac{\langle\cos(\theta_i)v_i + \sin(\theta_i)\Delta_i, \Lambda\big(\cos(\theta_j)v_j + \sin(\theta_j)\Delta_j\big)\rangle^2}{\Lambda_{jj}\cos(\theta_j)^2 + \langle\Delta_j, \Lambda\Delta_j\rangle\sin^2(\theta_j)} \tag{47}$$

$$= \Lambda_{ii}\cos(\theta_i)^2 + ||\Delta_i||_{\Lambda^{-1}}^2\sin^2(\theta_i)$$
$$- \sum_{j<i}\frac{\big(\Lambda_{jj}\sin(\theta_i)\cos(\theta_j)\langle\Delta_i, v_j\rangle + \Lambda_{ii}\sin(\theta_j)\cos(\theta_i)\langle\Delta_j, v_i\rangle + \sin(\theta_i)\sin(\theta_j)\langle\Delta_i, \Lambda\Delta_j\rangle\big)^2}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||_{\Lambda^{-1}}^2\sin^2(\theta_j)}. \tag{48}$$

Developing the numerator of the fraction, we obtain terms in $\sin$ and in $\sin^2$ that we later regroup to obtain the result:

$$= \Lambda_{ii} - \Lambda_{ii}\sin(\theta_i)^2 + ||\Delta_i||^2_{\Lambda^{-1}}\sin^2(\theta_i)$$
$$- \sum_{j<i} \frac{\Lambda_{jj}^2 \sin^2(\theta_i)\cos^2(\theta_j)\langle\Delta_i, v_j\rangle^2 + \Lambda_{ii}^2\sin^2(\theta_j)\cos^2(\theta_i)\langle\Delta_j, v_i\rangle^2 + \sin^2(\theta_i)\sin^2(\theta_j)\langle\Delta_i, \Lambda\Delta_j\rangle^2}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)} \tag{49}$$
$$- 2\sum_{j<i} \frac{\Lambda_{ii}\Lambda_{jj}\sin(\theta_i)\sin(\theta_j)\cos(\theta_i)\cos(\theta_j)\langle\Delta_j, v_i\rangle\langle\Delta_i, v_j\rangle}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)} \tag{50}$$
$$- 2\sum_{j<i} \frac{\Lambda_{jj}\sin^2(\theta_i)\sin(\theta_j)\cos(\theta_j)\langle\Delta_i, v_j\rangle\langle\Delta_i, \Lambda\Delta_j\rangle}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)} \tag{51}$$
$$- 2\sum_{j<i} \frac{\Lambda_{ii}\sin(\theta_i)\cos(\theta_i)\sin^2(\theta_j)\langle\Delta_j, v_i\rangle\langle\Delta_i, \Lambda\Delta_j\rangle}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)} \tag{52}$$
$$= \Lambda_{ii} - \Lambda_{ii}\sin^2(\theta_i) + ||\Delta_i||^2_{\Lambda^{-1}}\sin^2(\theta_i)$$
$$- \sum_{j<i} \frac{\Lambda_{jj}^2 \sin^2(\theta_i)\cos^2(\theta_j)\langle\Delta_i, v_j\rangle^2 + \Lambda_{ii}^2\sin^2(\theta_j)\cos^2(\theta_i)\langle\Delta_j, v_i\rangle^2 + \sin^2(\theta_i)\sin^2(\theta_j)\langle\Delta_i, \Lambda\Delta_j\rangle^2}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)} \tag{53}$$
$$- \frac{1}{2}\sum_{j<i} \frac{\Lambda_{ii}\Lambda_{jj}\sin(2\theta_i)\sin(2\theta_j)\langle\Delta_j, v_i\rangle\langle\Delta_i, v_j\rangle}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)} \tag{54}$$
$$- \sum_{j<i} \frac{\Lambda_{jj}\sin^2(\theta_i)\sin(2\theta_j)\langle\Delta_i, v_j\rangle\langle\Delta_i, \Lambda\Delta_j\rangle}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)} \tag{55}$$
$$- \sum_{j<i} \frac{\Lambda_{ii}\sin(2\theta_i)\sin^2(\theta_j)\langle\Delta_j, v_i\rangle\langle\Delta_i, \Lambda\Delta_j\rangle}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)}. \tag{56}$$

Collecting terms, we find

$$u_i(\hat{v}_i, \hat{v}_{j<i}) \tag{57}$$
$$= \sin^2(\theta_i)\Big[||\Delta_i||^2_{\Lambda^{-1}} - \Lambda_{ii} \tag{58}$$
$$- \sum_{j<i} \frac{\Lambda_{jj}^2\cos^2(\theta_j)\langle\Delta_i, v_j\rangle^2 - \Lambda_{ii}^2\sin^2(\theta_j)\langle\Delta_j, v_i\rangle^2 + \sin^2(\theta_j)\langle\Delta_i, \Lambda\Delta_j\rangle^2}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)} \tag{59}$$
$$- \sum_{j<i} \frac{\Lambda_{jj}\sin(2\theta_j)\langle\Delta_i, v_j\rangle\langle\Delta_i, \Lambda\Delta_j\rangle}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)}\Big] \tag{60}$$
$$- \frac{\sin(2\theta_i)}{2}\Big[\sum_{j<i} \frac{\Lambda_{ii}\Lambda_{jj}\sin(2\theta_j)\langle\Delta_j, v_i\rangle\langle\Delta_i, v_j\rangle + 2\Lambda_{ii}\sin^2(\theta_j)\langle\Delta_j, v_i\rangle\langle\Delta_i, \Lambda\Delta_j\rangle}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)}\Big] \tag{61}$$
$$+ \Big[\Lambda_{ii} - \sum_{j<i} \frac{\Lambda_{ii}^2\sin^2(\theta_j)\langle\Delta_j, v_i\rangle^2}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)}\Big] \tag{62}$$
$$\stackrel{\text{def}}{=} A\sin^2(\theta_i) - B\frac{\sin(2\theta_i)}{2} + C. \tag{63}$$

□

Lemma N.4. *The utility function along Δ_i, $\theta \mapsto u_i(\hat{v}_i(\theta_i, \Delta_i), \hat{v}_{j<i})$, is sinusoidal with period π:*

$$u_i(\hat{v}_i(\theta_i, \Delta_i), \hat{v}_{j<i}) = \frac{1}{2}\Big[\sqrt{A^2 + B^2}\cos(2\theta_i + \phi) + A + 2C\Big] \tag{64}$$

where $\phi = \tan^{-1}\left(\frac{B}{A}\right)$.

Proof. Starting from Lemma N.3, we find

$$u_i(\hat{v}_i(\theta_i, \Delta_i), \hat{v}_{j<i}) = A\sin^2(\theta_i) - B\frac{\sin(2\theta_i)}{2} + C \tag{65}$$

$$= A\frac{1-\cos(2\theta_i)}{2} - B\frac{\sin(2\theta_i)}{2} + C \tag{66}$$

$$= \frac{1}{2}\Big[-A\cos(2\theta_i) - B\sin(2\theta_i) + A + 2C\Big] \tag{67}$$

$$= \frac{1}{2}\Big[\sqrt{A^2+B^2}\cos(2\theta_i + \phi) + A + 2C\Big] \tag{68}$$

where $\phi = \tan^{-1}\left(\frac{B}{A}\right)$. □

Lemma N.5. *The angular deviation, θ_i, of the vector that maximizes the mis-specified objective, $\arg\max_{\theta_i} u_i(\hat{v}_i(\theta_i, \Delta_i), \hat{v}_{j<i})$, is given by*

$$|\theta_i^*| = \begin{cases} \frac{1}{2}\tan^{-1}\left(|\frac{B}{A}|\right) & \text{if } A < 0 \\ \frac{\pi}{4} & \text{if } A = 0 \\ \frac{1}{2}\Big[\pi - \tan^{-1}\left(|\frac{B}{A}|\right)\Big] & \text{if } A > 0 \end{cases} \tag{69}$$

where A and B are given by Lemma N.3.

Proof. First, we identify the critical points:

$$\frac{\partial}{\partial\theta_i}u_i(\hat{v}_i, \hat{v}_{j<i}) = 2A\sin(\theta_i)\cos(\theta_i) - B\cos(2\theta_i) = 0 \tag{70}$$

$$= A\sin(2\theta_i) - B\cos(2\theta_i) = 0 \tag{71}$$

$$= \frac{1}{\cos(2\theta_i)}[\tan(2\theta_i)A - B] = 0 \tag{72}$$

$$\tan(2\theta_i) = \frac{B}{A}. \tag{73}$$

Then we determine maxima vs minima:

$$\frac{\partial^2}{\partial\theta_i}u_i(\hat{v}_i, \hat{v}_{j<i}) = \frac{2}{\cos(2\theta_i)}[B\tan(2\theta_i) + A] = \frac{2}{\cos(2\theta_i)}[\frac{B^2}{A} + A], \tag{74}$$

therefore, $\texttt{sign}(\frac{\partial^2}{\partial\theta_i}u_i) = \texttt{sign}(\cos(2\theta_i))\texttt{sign}(A) < 0$ for θ_i to be a local maximum. If $A < 0$, then θ_i^* must lie within $[-\frac{\pi}{4}, \frac{\pi}{4}]$. If $A > 0$, then θ_i^* must lie within $[-\frac{\pi}{2}, -\frac{\pi}{4}]$ or $[\frac{\pi}{4}, \frac{\pi}{2}]$. By inspection, if $A = 0$, then u_i is maximized at $\theta_i = -\frac{\pi}{4}\texttt{sign}(B)$. In general, we are interested in the magnitude of θ_i, not its sign. □

Lemma N.6. *The following relationship is useful for proving Lemma N.8:*

$$\frac{b}{a+c} = \frac{b}{a}\Big[1 - \frac{c}{a+c}\Big] \tag{75}$$

Proof.

$$\frac{b}{a+c} = \frac{b}{a} + x \tag{76}$$

$$\implies x = \frac{b}{a+c} - \frac{b}{a} = b\Big[\frac{1}{a+c} - \frac{1}{a}\Big] \tag{77}$$

$$= b\Big[\frac{a-(a+c)}{a(a+c)}\Big] = -\frac{b}{a}\Big[\frac{c}{a+c}\Big]. \tag{78}$$

□

Lemma N.7. *If $\langle\Delta_i, v_i\rangle = 0$, then $u_i(\Delta_i, v_{j<i}) \leq \Lambda_{i+1,i+1}$.*

Proof. Recall the Nash proof in Appendix L:

$$u_i(\Delta_i, v_{j<i}) = \sum_{p \geq i} \Lambda_{pp} z_p \tag{79}$$

where $z_p = w_p^2$, $\Delta_i = \sum_{p=1}^d w_p v_p$, and $z \in \Delta^{d-1}$. The fact that $\langle \Delta_i, v_i \rangle = 0$ implies that $z_i = 0$. Therefore, the utility simplifies to

$$u_i(\Delta_i, v_{j<i}) = \sum_{p \geq i+1} \Lambda_{pp} z_p \tag{80}$$

which is upper bounded by $\Lambda_{i+1,i+1}$. □

Lemma N.8. *Assume $|\theta_j| \leq \epsilon$ for all $j < i$ (implies $\sin^2(\theta_j) \leq \epsilon^2$). Then*

$$A \leq -g_i + (i-1)(\Lambda_{11} + \Lambda_{ii}) \frac{\epsilon^2}{1-\epsilon^2} + 2(i-1)\Lambda_{11} \frac{\epsilon}{\sqrt{1-\epsilon^2}}. \tag{81}$$

Proof.

$$\begin{aligned}
&A(\theta_{j<i}) \\
&= ||\Delta_i||^2_{\Lambda^{-1}} - \Lambda_{ii} \\
&- \sum_{j<i} \frac{\Lambda_{jj}^2 \cos^2(\theta_j) \langle \Delta_i, v_j \rangle^2 - \Lambda_{ii}^2 \sin^2(\theta_j) \langle \Delta_j, v_i \rangle^2 + \sin^2(\theta_j) \langle \Delta_i, \Lambda \Delta_j \rangle^2}{\Lambda_{jj} \cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}} \sin^2(\theta_j)} \\
&- \sum_{j<i} \frac{\Lambda_{jj} \sin(2\theta_j) \langle \Delta_i, v_j \rangle \langle \Delta_i, \Lambda \Delta_j \rangle}{\Lambda_{jj} \cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}} \sin^2(\theta_j)} \qquad (82) \\
&= ||\Delta_i||^2_{\Lambda^{-1}} - \sum_{j<i} \frac{\Lambda_{jj}^2 \cos^2(\theta_j) \langle \Delta_i, v_j \rangle^2}{\Lambda_{jj} \cos^2(\theta_j) + ||\Delta_j||^2_{\Lambda^{-1}} \sin^2(\theta_j)} - \Lambda_{ii} \\
&- \sum_{j<i} \frac{-\Lambda_{ii}^2 \sin^2(\theta_j) \langle \Delta_j, v_i \rangle^2 + \sin^2(\theta_j) \langle \Delta_i, \Lambda \Delta_j \rangle^2 + \Lambda_{jj} \sin(2\theta_j) \langle \Delta_i, v_j \rangle \langle \Delta_i, \Lambda \Delta_j \rangle}{\Lambda_{jj} \cos^2(\theta_j) + ||\Delta_j||^2_{\Lambda^{-1}} \sin^2(\theta_j)} \qquad (83) \\
&\overset{\text{[LN.6]}}{=} ||\Delta_i||^2_{\Lambda^{-1}} - \sum_{j<i} \frac{\Lambda_{jj}^2 \cos^2(\theta_j) \langle \Delta_i, v_j \rangle^2}{\Lambda_{jj} \cos^2(\theta_j)} \left[1 - \frac{||\Delta_j||^2_{\Lambda^{-1}} \sin^2(\theta_j)}{\Lambda_{jj} \cos^2(\theta_j) + ||\Delta_j||^2_{\Lambda^{-1}} \sin^2(\theta_j)} \right] - \Lambda_{ii} \\
&- \sum_{j<i} \frac{-\Lambda_{ii}^2 \sin^2(\theta_j) \langle \Delta_j, v_i \rangle^2 + \sin^2(\theta_j) \langle \Delta_i, \Lambda \Delta_j \rangle^2 + \Lambda_{jj} \sin(2\theta_j) \langle \Delta_i, v_j \rangle \langle \Delta_i, \Lambda \Delta_j \rangle}{\Lambda_{jj} \cos^2(\theta_j) + ||\Delta_j||^2_{\Lambda^{-1}} \sin^2(\theta_j)} \qquad (84) \\
&\leq ||\Delta_i||^2_{\Lambda^{-1}} - \sum_{j<i} \frac{\Lambda_{jj}^2 \cos^2(\theta_j) \langle \Delta_i, v_j \rangle^2}{\Lambda_{jj} \cos^2(\theta_j)} + \sum_{j<i} \left(||\Delta_j||^2_{\Lambda^{-1}} \sin^2(\theta_j) \right) \frac{\Lambda_{jj}^2 \cos^2(\theta_j) \langle \Delta_i, v_j \rangle^2}{\Lambda_{jj}^2 \cos^4(\theta_j)} - \Lambda_{ii} \\
&+ \sum_{j<i} \frac{\Lambda_{ii}^2 \sin^2(\theta_j) \langle \Delta_j, v_i \rangle^2 + 2\Lambda_{jj} \sqrt{\sin^2(\theta_j)} \sqrt{\cos^2(\theta_j)} |\langle \Delta_i, v_j \rangle| |\langle \Delta_i, \Lambda \Delta_j \rangle|}{\Lambda_{jj} \cos^2(\theta_j)} \qquad (85) \\
&= u_i(\Delta_i, v_{j<i}) + \sum_{j<i} \left(||\Delta_j||^2_{\Lambda^{-1}} \sin^2(\theta_j) \right) \frac{\langle \Delta_i, v_j \rangle^2}{\cos^2(\theta_j)} - \Lambda_{ii} \\
&+ \sum_{j<i} \frac{\Lambda_{ii}^2 \sin^2(\theta_j) \langle \Delta_j, v_i \rangle^2 + 2\Lambda_{jj} \sqrt{\sin^2(\theta_j)} \sqrt{\cos^2(\theta_j)} |\langle \Delta_i, v_j \rangle| |\langle \Delta_i, \Lambda \Delta_j \rangle|}{\Lambda_{jj} \cos^2(\theta_j)} \qquad (86) \\
&\overset{\text{[LN.7]}}{\leq} \Lambda_{i+1,i+1} - \Lambda_{ii} + \sum_{j<i} \left(||\Delta_j||^2_{\Lambda^{-1}} \sin^2(\theta_j) \right) \frac{\cancelto{1}{\langle \Delta_i, v_j \rangle^2}}{\cos^2(\theta_j)} \\
&+ \sum_{j<i} \frac{\Lambda_{ii}^2 \sin^2(\theta_j) \cancelto{1}{\langle \Delta_j, v_i \rangle^2} + 2\Lambda_{jj} \sqrt{\sin^2(\theta_j)} \sqrt{\cos^2(\theta_j)} |\langle \Delta_i, v_j \rangle| |\langle \Delta_i, \Lambda \Delta_j \rangle|}{\Lambda_{jj} \cos^2(\theta_j)} \qquad (87)
\end{aligned}$$

$$\leq \Lambda_{i+1,i+1} - \Lambda_{ii} + \sum_{j<i} \epsilon^2 \frac{\Lambda_{11} + \Lambda_{ii}}{\cos^2(\theta_j)} + 2\frac{\Lambda_{jj}\sqrt{\sin^2(\theta_j)}\sqrt{\cos^2(\theta_j)}|\langle \Delta_i, v_j\rangle||\langle \Delta_i, \Lambda\Delta_j\rangle|}{\Lambda_{jj}\cos^2(\theta_j)} \tag{88}$$

$$\leq \Lambda_{i+1,i+1} - \Lambda_{ii} + \sum_{j<i} \epsilon^2 \frac{\Lambda_{11} + \Lambda_{ii}}{\cos^2(\theta_j)} + 2\Lambda_{11}\sqrt{\frac{\sin^2(\theta_j)}{\cos^2(\theta)}} \tag{89}$$

$$\leq \Lambda_{i+1,i+1} - \Lambda_{ii} + (i-1)(\Lambda_{11} + \Lambda_{ii})\frac{\epsilon^2}{1-\epsilon^2} + 2(i-1)\Lambda_{11}\frac{\epsilon}{\sqrt{1-\epsilon^2}}. \tag{90}$$

Note $\frac{\Lambda_{ii}^2}{\Lambda_{jj}} < \Lambda_{ii}$ because $\Lambda_{ii} < \Lambda_{jj}$ for all $j < i$. □

Lemma N.9. *Assume $\epsilon^2 \leq \frac{1}{2}$. Then*

$$A \leq -g_i + 8(i-1)\Lambda_{11}\epsilon. \tag{91}$$

Assume $\epsilon^2 \leq \frac{1}{2}$ so $\frac{\epsilon}{\sqrt{1-\epsilon^2}} \leq 1$. Then

$$A \leq \Lambda_{i+1,i+1} - \Lambda_{ii} + (i-1)(\Lambda_{11} + \Lambda_{ii})\frac{\epsilon^2}{1-\epsilon^2} + 2(i-1)\Lambda_{11}\frac{\epsilon}{\sqrt{1-\epsilon^2}} \tag{92}$$

$$\leq -g_i + (i-1)\Big[\frac{\epsilon}{\sqrt{1-\epsilon^2}}\Big]\Big[3\Lambda_{11} + \Lambda_{ii}\Big] \tag{93}$$

$$\leq -g_i + 4(i-1)\Lambda_{11}\frac{\epsilon}{\sqrt{1-\epsilon^2}} \tag{94}$$

$$\leq -g_i + 8(i-1)\Lambda_{11}\epsilon. \tag{95}$$

Lemma N.10. *Assume $\epsilon^2 \leq \frac{1}{2}$. Then*

$$|B| \leq 8(i-1)\Lambda_{ii}\kappa_{i-1}\epsilon. \tag{96}$$

Proof.

$$|B| = \sum_{j<i} \frac{|\Lambda_{ii}\Lambda_{jj}\sin(2\theta_j)\langle \Delta_j, v_i\rangle\langle \Delta_i, v_j\rangle + 2\Lambda_{ii}\sin^2(\theta_j)\langle \Delta_j, v_i\rangle\langle \Delta_i, \Lambda\Delta_j\rangle|}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||_{\Lambda^{-1}}\sin^2(\theta_j)} \tag{97}$$

$$\leq \sum_{j<i} \frac{\Lambda_{ii}\Lambda_{jj}\sqrt{\sin^2(2\theta_j)} + 2\Lambda_{ii}\sin^2(\theta_j)\Lambda_{11}}{\Lambda_{jj}\cos(\theta_j)^2} \tag{98}$$

$$\leq \sum_{j<i} \frac{\Lambda_{ii}\Lambda_{jj}\sqrt{4\sin^2(\theta_j)\cos^2(\theta_j)} + 2\Lambda_{ii}\sin^2(\theta_j)\Lambda_{11}}{\Lambda_{jj}\cos(\theta_j)^2} \tag{99}$$

$$\leq 2\sum_{j<i} \frac{\Lambda_{ii}\Lambda_{jj}\epsilon + \Lambda_{ii}\epsilon^2\Lambda_{11}}{\Lambda_{jj}(1-\epsilon^2)} \tag{100}$$

$$= 2\Lambda_{ii}\frac{\epsilon}{1-\epsilon^2}\Big((i-1) + \epsilon\sum_{j<i}\kappa_j\Big) \tag{101}$$

$$\leq 4\Lambda_{ii}\epsilon\Big((i-1) + \epsilon(i-1)\kappa_{i-1}\Big) \tag{102}$$

$$= 4(i-1)\Lambda_{ii}\epsilon\Big(1 + \epsilon\kappa_{i-1}\Big) \tag{103}$$

$$\leq 4(i-1)\Lambda_{ii}\epsilon\Big(1 + \frac{1}{\sqrt{2}}\kappa_{i-1}\Big) \tag{104}$$

$$\leq 8(i-1)\Lambda_{ii}\kappa_{i-1}\epsilon. \tag{105}$$

□

Lemma N.11. *Let $\epsilon_i = \frac{c_i g_i}{(i-1)\Lambda_{11}}$ with $c_i < \frac{1}{8}$. Then*

(i) $A \leq 0$,

(ii) $\left|\frac{B}{A}\right| \leq \frac{8c_i}{1-8c_i}$.

Proof. Plugging in Lemma N.9 and ϵ_i, we find

$$A \leq -g_i + 8c_i \frac{(i-1)\Lambda_{11} g_i}{(i-1)\Lambda_{11}} = -g_i + 8c_i g_i = (8c_i - 1)g_i. \tag{106}$$

Since we assumed $c_i < 1/8$, this proves (i). Plugging in Lemma N.10 and ϵ_i solves (ii):

$$\text{Equation (106)} \implies |A| \geq (1-8c_i)g_i \tag{107}$$

$$|B| \leq 8c_i \frac{(i-1)\Lambda_{ii}\kappa_{i-1} g_i}{(i-1)\Lambda_{11}} = 8c_i g_i \frac{\Lambda_{ii}}{\Lambda_{i-1,i-1}} \leq 8c_i g_i \tag{108}$$

$$\implies \left|\frac{B}{A}\right| \leq \frac{8c_i}{1-8c_i}. \tag{109}$$

□

Example 1. *We construct the following example in order to concreteley demonstrate the* arctan *dependence of a child ($\hat{v}_i$) on a parent ($\hat{v}_1$ in this case).*

Let $\Delta_1 = v_i$, $\Delta_i = v_1$, $\Delta_{1<j<i} = v_{i+1}$ and constrain all parents to have error $\sin(\theta_j) = \epsilon$ for all $j < i$. Then the child's optimum has an angular deviation from the true eigenvector direction of

$$|\theta_i^*| = \begin{cases} \frac{1}{2}\tan^{-1}\left(|\frac{B}{A}|\right) & \text{if } A < 0 \\ \frac{\pi}{4} & \text{if } A = 0 \\ \frac{1}{2}\left[\pi - \tan^{-1}\left(|\frac{B}{A}|\right)\right] & \text{if } A > 0 \end{cases} \tag{110}$$

where $|\frac{B}{A}| = \frac{2\epsilon\sqrt{1-\epsilon^2}}{|1-\epsilon^2(\kappa_i + \frac{1}{\kappa_i})|}$.

Proof. Note that $\langle \Delta_i, v_{1<j<i} \rangle$, $\langle \Delta_{1<j<i}, v_i \rangle$, and $\langle \Delta_i, \Lambda\Delta_j \rangle$ all equal 0 by design; and $\langle \Delta_i, v_1 \rangle = \langle \Delta_1, v_i \rangle = 1$. Plugging into Lemma N.3, all elements of the sum disappear for $j \geq 1$ and only the blue terms survive for $j = 1$. We find

$$A = ||\Delta_i||_{\Lambda^{-1}} - \Lambda_{ii} \tag{111}$$

$$- \sum_{j<i} \frac{\Lambda_{jj}^2 \cos^2(\theta_j)\langle \Delta_i, v_j \rangle^2 - \Lambda_{ii}^2 \sin^2(\theta_j)\langle \Delta_j, v_i \rangle^2 + \sin^2(\theta_j)\langle \Delta_i, \Lambda\Delta_j \rangle^2}{\Lambda_{jj}\cos^2(\theta_j) + ||\Delta_j||_{\Lambda^{-1}}^2 \sin^2(\theta_j)} \tag{112}$$

$$- \sum_{j<i} \frac{\Lambda_{jj}\sin(2\theta_j)\langle \Delta_i, v_j \rangle \langle \Delta_i, \Lambda\Delta_j \rangle}{\Lambda_{jj}\cos^2(\theta_j) + ||\Delta_j||_{\Lambda^{-1}}^2 \sin^2(\theta_j)} \tag{113}$$

$$= \Lambda_{11} - \Lambda_{ii} - \frac{\Lambda_{11}^2(1-\epsilon^2) - \Lambda_{ii}^2\epsilon^2}{\Lambda_{11}(1-\epsilon^2) + \Lambda_{11}\epsilon^2} \tag{114}$$

$$= \Lambda_{11} - \Lambda_{ii} - \frac{\Lambda_{11}^2(1-\epsilon^2) - \Lambda_{ii}^2\epsilon^2}{\Lambda_{11}} \tag{115}$$

$$= \Lambda_{11} - \Lambda_{ii} - \left[\Lambda_{11}(1-\epsilon^2) - \frac{\Lambda_{ii}}{\kappa_i}\epsilon^2\right] \tag{116}$$

$$= -\Lambda_{ii} + \epsilon^2(\Lambda_{11} + \frac{\Lambda_{ii}}{\kappa_i}) \tag{117}$$

and

$$B = \sum_{j<i} \frac{\Lambda_{ii}\Lambda_{jj}\sin(2\theta_j)\langle\Delta_j, v_i\rangle\langle\Delta_i, v_j\rangle + 2\Lambda_{ii}\sin^2(\theta_j)\langle\Delta_j, v_i\rangle\langle\Delta_i, \Lambda\Delta_j\rangle}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)} \tag{118}$$

$$= \frac{\Lambda_{ii}\Lambda_{11}\sin(2\theta_1)}{\Lambda_{11}\cos(\theta_1)^2 + ||\Delta_1||^2_{\Lambda^{-1}}\sin^2(\theta_1)} \tag{119}$$

$$= 2\frac{\Lambda_{ii}\Lambda_{11}\sqrt{\epsilon^2(1-\epsilon^2)}}{\Lambda_{11}(1-\epsilon^2) + \Lambda_{11}\epsilon^2} \tag{120}$$

$$= 2\Lambda_{ii}\epsilon\sqrt{1-\epsilon^2}. \tag{121}$$

Then

$$|\frac{B}{A}| = \frac{2\Lambda_{ii}\epsilon\sqrt{1-\epsilon^2}}{|\Lambda_{ii} - \epsilon^2(\Lambda_{11} + \frac{\Lambda_{ii}}{\kappa_i})|} = \frac{2\epsilon\sqrt{1-\epsilon^2}}{|1 - \epsilon^2(\kappa_i + \frac{1}{\kappa_i})|}. \tag{122}$$

□

O CONVERGENCE PROOF

O.1 NON-CONVEX RIEMANNIAN OPTIMIZATION THEORY

We repeat the non-convex Riemannian optimization rates here from (Boumal et al., 2019) for convenience.

Lemma O.1. *Under Assumptions O.2 and O.3, generic Riemannian descent (Algorithm 5) returns $x \in \mathcal{M}$ satisfying $f(x) \leq f(x_0)$ and $||\nabla^R f(x)|| \leq \rho$ in at most*

$$\lceil \frac{f(x_0) - f^*}{\xi} \cdot \frac{1}{\rho^2} \rceil \tag{123}$$

iterations, provided $\rho \leq \frac{\xi'}{\xi}$. If $\rho > \frac{\xi'}{\xi}$, at most $\lceil \frac{f(x_0)-f^*}{\xi'} \cdot \frac{1}{\rho} \rceil$ iterations are required.

Proof. See Theorem 2.5 in (Boumal et al., 2019). □

Assumption O.2. *There exists $f^* > -\infty$ such that $f(x) \geq f^*$ for all $x \in M$. See Assumption 2.3 in (Boumal et al., 2019).*

Assumption O.3. *There exist $\xi, \xi' > 0$ such that, for all $k \geq 0$, $f(x_k) - f(x_{k+1}) \geq \min(\xi||\nabla^R f(x_k)||, \xi')||\nabla^R f(x_k)||$. See Assumption 2.4 in (Boumal et al., 2019).*

Algorithm 5 Generic Riemannian descent algorithm

Given: $f : \mathcal{M} \to \mathbb{R}$ differentiable, a retraction Retr on $\mathcal{M}$, $x_0 \in \mathcal{M}$, $\rho > 0$
Init: $k \leftarrow 0$
while $||\nabla^R f(x_k)|| > \rho$ **do**
  Pick $\eta_k \in T_{x_k}\mathcal{M}$
end while
return x_k

O.2 CONVERGENCE OF EIGENGAME

Theorem O.4 provides an asymptotic convergence guarantee for Algorithm 1 (below) to recover the top-k principal components. Assuming $\hat{v}_i$ is initialized within $\frac{\pi}{4}$ of v_i for all $i \leq k$, Theorem O.5 provides a finite sample convergence rate. In particular, it specifies the total number of iterations required to learn parents such that $\hat{v}_k$ can be learned within a desired tolerance.

The proof of Theorem O.4 proceeds in several steps. First, recall that player i's utility is sinusoidal in its angular deviation from v_i and therefore, technically, non-concave although it is simple in the sense that every local maximum is a global maximum (w.r.t. angular deviation). Also, note that our ascent is not performed on the natural parameters of the sphere (θ_i and Δ_i), but rather on $\hat{v}_i$ directly with $\hat{v}_i \in \mathcal{S}^{d-1}$, a Riemannian manifold.

We therefore leverage recent results in non-convex optimization, specifically minimization, for Riemannian manifolds (Boumal et al., 2019), repeated here for convenience (see Theorem O.1). Note, we are maximizing a utility so we simply flip the sign of our utility to apply this theory. The convergence rate guarantee given by this theory is for generic Riemannian descent with a constant step size, Algorithm 5, and makes two assumptions. One is a bound on the utility (Lemma O.2) and the other is a smoothness or Lipschitz condition (Lemma O.3). The convergence rate itself states the number of iterations required for the norm of the Riemannian gradient to fall below a given threshold. The theory also guarantees descent in that the solution returned by the algorithm will have lower loss (higher utility) than the vector passed to the algorithm.

The probability of sampling a vector $\hat{v}_i^0$ at angular deviation within ϕ of the maximizer is given by

$$P[|\theta_i^0 - \theta_i^*| \leq \phi] = I_{\sin^2(\phi)}(\frac{d-1}{2}, \frac{1}{2}) = \frac{\texttt{Beta}(\sin^2\phi, \frac{d-1}{2}, \frac{1}{2})}{\texttt{Beta}(1, \frac{d-1}{2}, \frac{1}{2})} \tag{124}$$

where Beta is the incomplete beta function, and I is the normalized incomplete beta function (Li, 2011). This probability quickly approaches zero for $\phi < \frac{\pi}{2}$ as the dimension d increases. Therefore,

for large d, it becomes highly probable that $\hat{v}_i$ will be initialized near an angle $\frac{\pi}{2}$ from the true eigenvector—in other words, all points are far from each other in high dimensions. In this case, $\hat{v}_i$ lies near a trough of the sinusoidal utility where gradients are small. Without a bound on the minimum possible gradient norm, a finite sample rate cannot be constructed (how many iterations are required to escape the trough?). Therefore, we can only guarantee asymptotic convergence in this setting. Next, we consider the fortuitous case where all $\hat{v}_i$ have been initialized within $\frac{\pi}{4}$. This is both to obtain a convergence rate for this setting, but also to highlight the Big-O dependencies. Note that the utility is symmetric across $\frac{\pi}{4}$ and the number of iterations required to escape a trough and reach the $\frac{\pi}{4}$ mark is equal to the number of iterations required to ascend from $\frac{\pi}{4}$ to the same distance from the peak.

In order to ensure this theory can provide meaningful bounds for EigenGame, we first show, assuming a child is within $\frac{\pi}{4}$ of its maximizer, that the norm of the Riemannian gradient bounds the angular deviation of a child from this maximizer.

To begin the proof, we relate the error in the parents to a bound on the ambient gradient in Lemma O.8. This bound is then tightened assuming parents with error below a certain threshold in Lemma O.9. Using the fact that $u_i = \hat{v}_i^\top \nabla_{\hat{v}_i} u_i$, this bound directly translates to a bound on the utility in Corollary O.9.1, thereby satisfying Assumption O.2. Again, given accurate parents, Lemma O.10 proves Assumption O.3 on smoothness is satisfied and derives some of the constants for the ultimate convergence rate.

Recall that we have so far been proving convergence to a local maximizer of a child's utility, which, assuming inaccurate parents, is not the same as the true eigenvector. Lemma O.11 upper bounds the angular deviation of an approximate maximizer from the true eigenvector using the angular deviation of a maximizer plus the approximate maximizer's approximation error. Lemma O.12 then provides the convergence rate for the child to approach the true eigenvector given accurate enough parents. Finally, Theorem O.4 compiles the chain of convergence rates leading up the DAG towards $\hat{v}_1$ and derives a convergence rate for child k given all previous parents have been learned to a high enough degree of accuracy. The number of iterations required for each parent in the chain is provided.

Theorem O.4. *Assume all spectral gaps are positive, i.e. for $i = 1...k$, $g_i > 0$. Let θ_k denote the angular distance (in radians) of $\hat{v}_k$ from the true eigenvector v_k. Let the maximum desired error for $\theta_k = \theta_{tol} \leq 1$ radian. Then set $c_k = \frac{\theta_{tol}}{16}$, $\rho_k = \frac{g_k}{2\pi}\theta_{tol}$, and*

$$\rho_i = \left[\frac{g_i g_{i+1}}{2\pi i \Lambda_{11}}\right] c_{i+1} \tag{125}$$

$$c_i \leq \frac{(i-1)! \prod_{j=i+1}^{k} g_j}{(16\Lambda_{11})^{k-i}(k-1)!} c_k \tag{126}$$

for $i < k$ where the c_i's are dictated by each $\hat{v}_i$ to its parents and represent fractions of a canonical error threshold; for example, if $\hat{v}_k$ sets $c_k = \frac{1}{16}$, then this threshold gets communicated up the DAG to each parent, each time strengthening.

Consider learning $\hat{v}_i$ by applying Algorithm 1 successively, i.e., learn $\hat{v}_1$, stop ascent, learn $\hat{v}_2$, and so on, each with step size $\frac{1}{2L}$ and corresponding ρ_i where $L = 4\left[\Lambda_{11} k + (1 + \kappa_{k-1})\frac{g_k}{16}\right]$. Then the top-$k$ principal components will be returned, each within tolerance θ_{tol}, in the limit.

Proof. In order to learn $\hat{v}_k$, we need $|\theta_j| \leq \frac{c_k g_k}{(k-1)\Lambda_{11}}$ with $c_k \leq \frac{1}{16}$ for all $j < k$. If this requirement is met, then by Lemma O.11, the angular error in $\hat{v}_k$ after running Riemannian gradient ascent is bounded as

$$|\theta_k| \leq \bar{\epsilon} + 8c_k \tag{127}$$

where $\bar{\epsilon}$ denotes the convergence error and the error propagated by the parents is $8c_k$. The quantity, $\frac{g_k}{(k-1)\Lambda_{11}}$, in the parents bound is $\ll 8$, so the parents must be very accurate to reduce the error propagated to the child. Each parent must then convey this information up the chain, strengthening the requirement each hop.

Let half the error in $|\theta_k|$ come from mis-specifying the utility with imperfect parents, $\hat{v}_{j<k}$, and the other half from convergence error. The error after learning $\hat{v}_{k-1}$ via Riemannian gradient ascent must be less than the threshold required for learning the kth eigenvector. Assuming $\hat{v}_{k-1}$'s parents have

been learned accurately enough, $|\theta_{j<k-1}| \leq \frac{c_{k-1}g_{k-1}}{(k-2)\Lambda_{11}}$, and that $\hat{v}_{j\leq k}$ were initialized within $\frac{\pi}{4}$ of their maximizers, we require:

$$|\theta_{k-1}| \overset{LO.12}{\leq} \frac{\pi}{g_{k-1}}\rho_{k-1} + 8c_{k-1} \leq \frac{c_k g_k}{(k-1)\Lambda_{11}}. \tag{128}$$

More generally, the error after learning $\hat{v}_{i-1}$ must be less than the threshold for learning any of its successors:

$$|\theta_{i-1}| \leq \frac{\pi}{g_{i-1}}\rho_{i-1} + 8c_{i-1} \leq \min_{i-1<l\leq k}\Big(\frac{c_l g_l}{(l-1)\Lambda_{11}}\Big). \tag{129}$$

Assume for now that the $\arg\min$ of the expression is i, the immediate child. First we bound the error from $\hat{v}_{i-1}$'s parents:

$$8c_{i-1} \leq \frac{c_i g_i}{2(i-1)\Lambda_{11}} \tag{130}$$

$$\implies c_{i-1} \leq \frac{c_i g_i}{16(i-1)\Lambda_{11}}. \tag{131}$$

Note the 2 in the denominator of Equation (130) which appears because we desired half the error to come from the parents (half is an arbitrary choice in the analysis). Continuing this process recursively implies

$$c_{i-2} \leq \frac{c_{i-1}g_{i-1}}{16(i-2)\Lambda_{11}} \leq \frac{c_i g_{i-1} g_i}{16^2(i-2)(i-1)\Lambda_{11}^2} \tag{132}$$

$$\implies c_{i-n} \leq \Big[\frac{(i-n-1)!\prod_{j=i-n+1}^{i} g_j}{(16\Lambda_{11})^n (i-1)!}\Big] c_i. \tag{133}$$

One can see that $c_{j<i}$ is strictly smaller than c_i because each additional term added to the product is strictly less than 1—the assumption of the $\arg\min$ above is therefore correct. In particular, this requires the first eigenvector to be learned to very high accuracy to enable learning the kth:

$$c_1 \leq \Big[\frac{\prod_{j=2}^{k} g_j}{(16\Lambda_{11})^{k-1}(k-1)!}\Big] c_k. \tag{134}$$

More generally

$$c_i \leq \frac{(i-1)!\prod_{j=i+1}^{k} g_j}{(16\Lambda_{11})^{k-i}(k-1)!} c_k \tag{135}$$

This completes the requirement for mitigating error in the parents.

The convergence error from gradient ascent must also be bounded as (again, note the 2)

$$\frac{\pi}{g_i}\rho_i \leq \frac{c_{i+1}g_{i+1}}{2i\Lambda_{11}} \tag{136}$$

$$\implies \rho_i \leq \Big[\frac{g_i g_{i+1}}{2\pi i \Lambda_{11}}\Big] c_{i+1} \tag{137}$$

which requires at most

$$t_i = \lceil 5\Big(\frac{\pi i \Lambda_{11}}{g_i g_{i+1}}\Big)^2 \frac{1}{c_{i+1}^2} \rceil \tag{138}$$

iterations. Given $\hat{v}_i$ is initialized within $\frac{\pi}{4}$ of its maximizer, it follows that learning each $\hat{v}_{j<k}$ consecutively via Riemannian gradient ascent for at most $\sum_{i=1}^{k-1} t_i$ iterations is sufficient for learning the k-th eigenvector. Riemannian gradient ascent on $\hat{v}_k$ then returns (Lemma O.12)

$$|\theta_k| \leq \frac{\pi}{g_k}\rho_k + 8c_k \leq \frac{\pi}{g_k}\rho_k + \frac{\theta_{\text{tol}}}{2} \tag{139}$$

after at most

$$t_k = \Big\lceil \frac{5}{4} \cdot \frac{1}{\rho_k^2} \Big\rceil = \Big\lceil \frac{5\pi^2}{(\theta_{\text{tol}} g_k)^2} \Big\rceil \tag{140}$$

iterations.

We can relax the assumption that $\hat{v}_i$ is initialized within $\frac{\pi}{4}$ of its maximizer and obtain global convergence. Assume that $\frac{\pi}{2} - |\theta_i^0| \leq \frac{\pi}{4}$ and let $||\nabla_{\hat{v}_i^0}||$ be the initial norm of the Riemannian gradient. The utility function $u_i(\hat{v}_i, \hat{v}_{j<i})$ is symmetric across $\frac{\pi}{4}$. Therefore, the number of iterations required to ascend to within $\frac{\pi}{4}$ is given by Lemma O.12:

$$t_i^+ = \Big\lceil \frac{5}{4}\big(\frac{\pi}{g_i}\big)^2 \frac{1}{(\frac{\pi}{2} - |\theta_i^0|)^2} \Big\rceil. \tag{141}$$

Alternatively, simply set the desired gradient norm to be less than the initial. This necessarily requires iterates to ascend to past $\frac{\pi}{4}$. As long as $\hat{v}_i$ is not initialized to exactly $\frac{\pi}{2}$ from the maximum (an event with Lebesgue measure 0), the ascent process will converge to the maximizer. □

Theorem O.5. *Apply the algorithm outlined in Theorem O.4 with the same assumptions. Then with probability*

$$P[|\theta_i^0 - \theta_i^*| \leq \frac{\pi}{4}] = I_{\frac{1}{2}}(\frac{d-1}{2}, \frac{1}{2}) \tag{142}$$

where I is the normalized incomplete beta function, the max total number of iterations required for learning all vectors to adequate accuracy is

$$T_k = \Big\lceil \mathcal{O}\Big(k\Big[\frac{(16\Lambda_{11}^k)(k-1)!}{\prod_{j=1}^k g_j}\frac{1}{\theta_{tol}}\Big]^2\Big)\Big\rceil. \tag{143}$$

Discussion. In other words, assuming all $\hat{v}_i$ are fortuitously initialized within $\frac{\pi}{4}$ of their maximizers, then we can state a finite sample convergence rate. The first k in the Big-$\mathcal{O}$ formula for total iterations appears simply from a naive summing of worst case bounds on the number of iterations required to learn each $\hat{v}_{j<k}$ individually. The constant 16 is a loose bound that arises from the error propagation analysis. Essentially, parent vectors, $\hat{v}_{j<i}$, must be learned to under $\frac{1}{16}$ a canonical error threshold for the child $\hat{v}_i$, $\frac{g_i}{(i-1)\Lambda_{11}}$. The Riemannian optimization theory we leverage dictates that $\frac{1}{\rho_i^2}$ iterations are required to meet a $\mathcal{O}(\rho_i)$ error threshold. This is why the squared inverse of the error threshold appears here. Breaking down the error threshold itself, the ratio $\frac{\Lambda_{11}}{g_i}$ says that more iterations are required to distinguish eigenvectors when the difference between them (summarized by the gap g_i) is small relative to the scale of the spectrum, Λ_{11}. The $(k-1)!$ term appears because learning smaller eigenvectors requires learning a much more accurate $\hat{v}_1$ higher up the chain.

Proof. Assume $\hat{v}_i$ is sampled uniformly in $\mathcal{S}^{d-1}$. Note this can be accomplished by normalizing a sample from a multivariate Gaussian. We will prove

(i) the probability of the event that $\hat{v}_i^0$ is within $\frac{\pi}{4}$ of the maximizer of $u_i(\hat{v}_i, \hat{v}_{j<i})$,

(ii) an upper bound on the number of iterations required to return all $\hat{v}_i$ with angular error less than θ_{tol}.

The probability of sampling a vector $\hat{v}_i^0$ at angular deviation within $\frac{\pi}{4}$ of the maximizer is given by twice the probability of sampling from one of the spherical caps around v_i or $-v_i$. This probability is

$$P[|\theta_i^0 - \theta_i^*| \leq \phi] = I_{\sin^2(\phi)}(\frac{d-1}{2}, \frac{1}{2}) = \frac{\texttt{Beta}(\sin^2(\phi), \frac{d-1}{2}, \frac{1}{2})}{\texttt{Beta}(1, \frac{d-1}{2}, \frac{1}{2})} \tag{144}$$

where `Beta` is the incomplete beta function, and I is the normalized incomplete beta function (Li, 2011). This probability quickly approaches zero for $\phi < \frac{\pi}{2}$ as the dimension d increases. This proves (i).

Plugging the bound on c_i

$$c_i \leq \frac{(i-1)!\prod_{j=i+1}^k g_j}{(16\Lambda_{11})^{k-i}(k-1)!}c_k \tag{145}$$

into the bound on iterations

$$t_i = \lceil 5\Big(\frac{\pi i \Lambda_{11}}{g_i g_{i+1}}\Big)^2 \frac{1}{c_{i+1}^2} \rceil \tag{146}$$

we find

$$t_i = \Big\lceil 5\Big(\frac{\pi i \Lambda_{11}}{g_i g_{i+1}}\Big)^2 \frac{(16\Lambda_{11})^{2(k-i-1)}((k-1)!)^2}{(i!)^2 \prod_{j=i+2}^{k} g_j^2} \frac{1}{c_k^2} \Big\rceil \tag{147}$$

$$= \Big\lceil 5\pi^2 \frac{16^{2(k-i)} \Lambda_{11}^{2(k-i)} ((k-1)!)^2}{\Big(\prod_{j=i}^{k} g_j^2\Big)((i-1)!)^2} \frac{1}{(16c_k)^2} \Big\rceil \tag{148}$$

$$\leq \Big\lceil 5\pi^2 \Big[\frac{(16\Lambda_{11})^{k-1}(k-1)!}{\prod_{j=1}^{k} g_j} \frac{1}{16c_k}\Big]^2 \Big\rceil \quad [\Lambda_{11} \geq g_i \ \forall i] \tag{149}$$

$$= \Big\lceil \mathcal{O}\Big(\Big[\frac{(16\Lambda_{11})^{k}(k-1)!}{\prod_{j=1}^{k} g_j} \frac{1}{16c_k}\Big]^2\Big) \Big\rceil \tag{150}$$

which is now in a form independent of i (worst case). It can be shown that $t_k \leq t_1$ by taking their $\log$ and applying Jensen's inequality. The total iterations required for learning $\hat{v}_{j<k}$ is at most $k-1$ times this. Therefore,

$$T_k = \Big\lceil \mathcal{O}\Big(k\Big[\frac{(16\Lambda_{11})^{k}(k-1)!}{\prod_{j=1}^{k} g_j} \frac{1}{16c_k}\Big]^2\Big) \Big\rceil. \tag{151}$$

□

Corollary O.5.1 (PC Convergence $\implies$ Subspace Convergence)**.** *Convergence of $\hat{V}$ to the top-k principal components of X with maximum angular error θ_{tol} implies convergence to the top-k subspace of X in the following sense*[11]*:*

$$||\hat{V}^\top V_{\neg k}||_F^2 \leq k(d-k)\theta_{tol}^2. \tag{152}$$

where the columns of $V_{\neg k}$ comprise the bottom $d-k$ eigenvectors of $M = X^\top X$.

Proof. Recall that the true principal components, v_i, are all orthogonal. If the angle between $\hat{v}_i$ and v_i is less than or equal to θ_{tol} for every i, then the angle between $\hat{v}_i$ and v_j for any $j \neq i$ must be greater than or equal to $\frac{\pi}{2} - \theta_{tol}$. The entries in $\hat{V}^\top V_{\neg k}$ are equal to the cosines of the angles between each of the columns in $\hat{V}$ and $V_{\neg k}$. Therefore, all entries are less than or equal to $|\cos(\frac{\pi}{2} - \theta_{tol})| = |\sin(\theta_{tol})| \leq \theta_{tol}$. This implies the squared Frobenius norm of this matrix is less than or equal to the number of entries times the maximum value squared: $k(d-k)\theta_{tol}^2$. □

Lemma O.6. *Assume $\hat{v}_i$ is within $\frac{\pi}{4}$ of its maximizer, i.e., $|\theta_i - \theta_i^*| \leq \frac{\pi}{4}$. Also, assume that $|\theta_{j<i}| \leq \frac{c_i g_i}{(i-1)\Lambda_{11}} \leq \sqrt{\frac{1}{2}}$ with $0 \leq c_i \leq \frac{1}{16}$. Then the norm of the Riemannian gradient of u_i upper bounds this angular deviation:*

$$|\theta_i - \theta_i^*| \leq \frac{\pi}{g_i} ||\nabla_{\hat{v}_i}^R u_i(\hat{v}_i, \hat{v}_{j<i})||. \tag{153}$$

Proof. The Riemannian gradient measures how the utility u_i changes while moving along the manifold. In contrast, the ambient gradient measures how u_i changes while moving in ambient space, possibly off the manifold. Rather than bounding the angular deviation using the projection of the ambient gradient onto the tangent space of the manifold, $(I - \hat{v}_i \hat{v}_i^\top)\nabla_{\hat{v}_i} u_i$, we instead reparameterize $\hat{v}_i$ to ensure it lies on the manifold, $\hat{v}_i = \cos(\theta_i) v_i + \sin(\theta_i)\Delta_i$ where Δ_i is a unit vector and $\langle v_i, \Delta_i \rangle = 0$. Computing gradients with respect to the new unconstrained arguments allows recovering a bound on the Riemannian gradient via a simple chain rule calculation.

[11] See Allen-Zhu and Li (2017) for more details on this measure of subspace error.

We lower bound the norm of the Riemannian gradient as follows:

$$\frac{\partial u_i}{\partial \theta_i} = \nabla^R_{\hat{v}_i} u_i(\hat{v}_i, \hat{v}_{j<i})^\top \frac{\partial v}{\partial \theta_i} \tag{154}$$

$$\implies ||\frac{\partial u_i}{\partial \theta_i}|| \leq ||\nabla^R_{\hat{v}_i} u_i(\hat{v}_i, \hat{v}_{j<i}||||\frac{\partial \hat{v}_i}{\partial \theta_i}|| \tag{155}$$

$$\implies ||\nabla^R_{\hat{v}_i} u_i(\hat{v}_i, \hat{v}_{j<i}|| \geq \frac{||\partial u_i/\partial \theta_i||}{||\partial \hat{v}_i/\partial \theta_i||}. \tag{156}$$

Note that $||\partial \hat{v}_i/\partial \theta_i|| = 1$ by design. And the numerator can be bounded using Lemma N.4 as

$$||\partial u_i/\partial \theta_i|| = \sqrt{A^2 + B^2}|\sin(2\big(\theta_i - \theta_i^*\big))| \tag{157}$$

where $\theta_i^* = -\frac{\phi}{2}$ and $\phi = \tan^{-1}\left(\frac{B}{A}\right)$. Furthermore, assume $|\theta_i - \theta_i^*| \leq \frac{\pi}{4}$. Then

$$|\sin(2\big(\theta_i - \theta_i^*\big))| \geq \Big|\frac{2}{\pi}\big(\theta_i - \theta_i^*\big)\Big|. \tag{158}$$

Combining the results gives

$$||\nabla^R_{\hat{v}_i} u_i(\hat{v}_i, \hat{v}_{j<i}|| \geq \frac{||\partial u_i/\partial \theta_i||}{||\partial v/\partial \theta_i||} \tag{159}$$

$$= ||\partial u_i/\partial \theta_i|| \tag{160}$$

$$\geq \frac{2}{\pi}\sqrt{A^2 + B^2}|\theta_i - \theta_i^*| \tag{161}$$

$$\geq \frac{2}{\pi}|A||\theta_i - \theta_i^*| \tag{162}$$

$$\overset{LN.11}{\geq} \frac{2}{\pi}(1 - 8c)g_i|\theta_i - \theta_i^*| \tag{163}$$

$$\geq \frac{g_i}{\pi}|\theta_i - \theta_i^*| \tag{164}$$

completing the proof. □

Lemma O.7. *Let $|\theta_j| \leq \epsilon < 1$ for all $j < i$. Then the ratio of generalized inner products is bounded as*

$$\frac{\langle \hat{v}_i, \Lambda \hat{v}_j \rangle}{\langle \hat{v}_j, \Lambda \hat{v}_j \rangle} \leq \frac{1 + (1 + \kappa_j)\epsilon}{\sqrt{1 - \epsilon^2}}. \tag{165}$$

Proof. We write $\hat{v}_{j \leq i} = \cos(\theta_j)v_j + \sin(\theta_j)\Delta_j$ where $\langle \Delta_j, v_j \rangle = 0$ without loss of generality. Note that $|\theta_j| \leq \epsilon$ implies $|\sin(\theta_j)| \leq \epsilon$. Then

$$\frac{\langle \hat{v}_i, \Lambda \hat{v}_j \rangle}{\langle \hat{v}_j, \Lambda \hat{v}_j \rangle} \tag{166}$$

$$= \frac{\langle \cos(\theta_i)v_i + \sin(\theta_i)\Delta_i, \Lambda\big(\cos(\theta_j)v_j + \sin(\theta_j)\Delta_j\big)\rangle}{\langle \cos(\theta_j)v_j + \sin(\theta_j)\Delta_j, \Lambda\big(\cos(\theta_j)v_j + \sin(\theta_j)\Delta_j\big)\rangle} \tag{167}$$

$$= \frac{\langle \cos(\theta_i)v_i + \sin(\theta_i)\Delta_i, \Lambda\big(\cos(\theta_j)v_j + \sin(\theta_j)\Delta_j\big)\rangle}{\Lambda_{jj}\cos(\theta_j)^2 + \langle \Delta_j, \Lambda\Delta_j \rangle \sin^2(\theta_j)} \tag{168}$$

$$= \frac{\Lambda_{jj}\sin(\theta_i)\cos(\theta_j)\langle \Delta_i, v_j \rangle + \Lambda_{ii}\sin(\theta_j)\cos(\theta_i)\langle \Delta_j, v_i \rangle + \sin(\theta_i)\sin(\theta_j)\langle \Delta_i, \Lambda\Delta_j \rangle}{\Lambda_{jj}\cos(\theta_j)^2 + ||\Delta_j||^2_{\Lambda^{-1}}\sin^2(\theta_j)} \tag{169}$$

$$\leq \frac{\Lambda_{jj}|\sin(\theta_i)|\sqrt{1 - \epsilon^2} + \Lambda_{ii}\epsilon|\cos(\theta_i)| + |\sin(\theta_i)|\epsilon\Lambda_{11}}{\Lambda_{jj}(1 - \epsilon^2)} \tag{170}$$

$$\leq \frac{\Lambda_{jj}\sqrt{1 - \epsilon^2} + \Lambda_{ii}\epsilon + \epsilon\Lambda_{11}}{\Lambda_{jj}(1 - \epsilon^2)} \tag{171}$$

$$= \frac{1}{\sqrt{1 - \epsilon^2}} + \Big(\frac{\Lambda_{ii}}{\Lambda_{jj}} + \kappa_j\Big)\frac{\epsilon}{\sqrt{1 - \epsilon^2}} \tag{172}$$

$$\leq \frac{1 + (1 + \kappa_j)\epsilon}{\sqrt{1 - \epsilon^2}}. \tag{173}$$

□

Lemma O.8 (Lipschitz Bound)**.** *Let $|\theta_j| \leq \epsilon < 1$ for all $j < i$. Then the norm of the ambient gradient of u_i is bounded as*

$$||\nabla_{\hat{v}_i} u_i(\hat{v}_i, \hat{v}_{j<i})|| \leq 2\Lambda_{11}\Big[1 + (i-1)\frac{1 + (1+\kappa_{i-1})\epsilon}{\sqrt{1-\epsilon^2}}\Big]. \tag{174}$$

Proof. Starting with the gradient (Equation 7), we find

$$\begin{aligned}
||\nabla_{\hat{v}_i} u_i(\hat{v}_i, \hat{v}_{j<i})|| &= ||2M\Big[\hat{v}_i - \sum_{j<i} \frac{\hat{v}_i^\top M \hat{v}_j}{\hat{v}_j^\top M \hat{v}_j}\hat{v}_j\Big]|| && (175)\\
&\leq 2||M\hat{v}_i|| + 2\sum_{j<i} ||\frac{\hat{v}_i^\top M \hat{v}_j}{\hat{v}_j^\top M \hat{v}_j} M\hat{v}_j|| && (176)\\
&\leq 2||M\hat{v}_i|| + 2\sum_{j<i} ||\frac{\hat{v}_i^\top M \hat{v}_j}{\hat{v}_j^\top M \hat{v}_j}||||M\hat{v}_j|| && (177)\\
&\overset{LO.7}{\leq} 2\Lambda_{11} + 2\sum_{j<i} \frac{1 + (1+\kappa_j)\epsilon}{\sqrt{1-\epsilon^2}}\Lambda_{11} && (178)\\
&= 2\Lambda_{11}\Big[1 + (i-1)\frac{1 + (1+\kappa_{i-1})\epsilon}{\sqrt{1-\epsilon^2}}\Big]. && (179)
\end{aligned}$$

□

Lemma O.9 (Lipschitz Bound with Accurate Parents)**.** *Assume $|\theta_j| \leq \epsilon \leq \frac{c_i g_i}{(i-1)\Lambda_{11}} \leq \sqrt{\frac{1}{2}}$ for all $j < i$ with $0 \leq c_i \leq \frac{1}{16}$. Then the norm of the ambient gradient of u_i is bounded as*

$$||\nabla_{\hat{v}_i} u_i(\hat{v}_i, \hat{v}_{j<i})|| \leq 4\Big[\Lambda_{11} i + (1+\kappa_{i-1})c_i g_i\Big] \overset{\text{def}}{=} L_i. \tag{180}$$

Proof. Starting with Lemma O.8, we find

$$\begin{aligned}
||\nabla_{\hat{v}_i} u_i(\hat{v}_i, \hat{v}_{j<i})|| &\leq 2\Lambda_{11}\Big[1 + (i-1)\frac{1 + (1+\kappa_{i-1})\epsilon}{\sqrt{1-\epsilon^2}}\Big] && (181)\\
&\leq 2\Lambda_{11}\Big[1 + 2(i-1)\big(1 + (1+\kappa_{i-1})\epsilon\big)\Big] && (182)\\
&\overset{\text{assumption}}{\leq} 2\Lambda_{11}\Big[1 + 2(i-1) + 2\frac{(1+\kappa_{i-1})c g_i}{\Lambda_{11}}\Big] && (183)\\
&\leq 4\Big[\Lambda_{11}\big(1 + (i-1)\big) + (1+\kappa_{i-1})c g_i\Big] && (184)\\
&= 4\Big[\Lambda_{11} i + (1+\kappa_{i-1})c g_i\Big]. && (185)
\end{aligned}$$

□

Corollary O.9.1 (Bound on Utility)**.** *Assume $|\theta_j| \leq \frac{c_i g_i}{(i-1)\Lambda_{11}} \leq \sqrt{\frac{1}{2}}$ for all $j < i$ with $0 \leq c_i \leq \frac{1}{16}$. Then the absolute value of the utility is bounded as follows*

$$|u_i(\hat{v}_i, \hat{v}_{j<i})| = |\hat{v}_i^\top \nabla_{\hat{v}_i}| \leq ||\hat{v}_i||||\nabla_{\hat{v}_i}|| = ||\nabla_{\hat{v}_i}|| \leq L_i, \tag{186}$$

thereby satisfying Assumption O.2.

Lemma O.10. *Assume $|\theta_j| \leq \frac{c_i g_i}{(i-1)\Lambda_{11}} \leq \sqrt{\frac{1}{2}}$ for all $j < i$ with $0 \leq c_i \leq \frac{1}{16}$. Then Assumption O.3 is satisfied with $\xi = \xi' = \frac{8}{5}L_i$.*

Proof. Let $\eta = \alpha \nabla^R_{\hat{v}_i} u_i = \alpha(I - \hat{v}_i \hat{v}_i^\top)\nabla_{\hat{v}_i} u_i$, $\alpha > 0$, and $\hat{\eta} = \frac{\eta}{||\eta||}$. Let $\hat{v}_i' = \frac{\hat{v}_i + \eta}{\gamma}$ where $\gamma = ||\hat{v}_i + \eta||$.

$$u_i(\hat{v}_i') = \frac{1}{\gamma^2}\left[(\hat{v}_i + \eta)^\top \Lambda (\hat{v}_i + \eta) - \sum_{j<i} \frac{\left((\hat{v}_i + \eta)^\top \Lambda \hat{v}_j\right)^2}{\hat{v}_j^\top \Lambda \hat{v}_j}\right] \tag{187}$$

$$= \frac{1}{\gamma^2}\left[\hat{v}_i^\top \Lambda \hat{v}_i - \sum_{j<i} \frac{(\hat{v}_i^\top \Lambda \hat{v}_j)^2}{\hat{v}_j^\top \Lambda \hat{v}_j} + \eta^\top \Lambda \eta - \sum_{j<i} \frac{(\eta^\top \Lambda \hat{v}_j)^2}{\hat{v}_j^\top \Lambda \hat{v}_j} + 2\eta^\top \Lambda \hat{v}_i - 2\sum_{j<i} \frac{(\hat{v}_i^\top \Lambda \hat{v}_j)(\eta^\top \Lambda \hat{v}_j)}{\hat{v}_j^\top \Lambda \hat{v}_j}\right] \tag{188}$$

$$= \frac{1}{\gamma^2}\left[u_i(\hat{v}_i) + u_i(\eta) + 2\eta^\top \nabla_{\hat{v}_i} u_i(\hat{v}_i)\right] \tag{189}$$

$$= \frac{1}{\gamma^2}\left[u_i(\hat{v}_i) + ||\eta||^2 u_i(\hat{\eta}) + 2\eta^\top \nabla_{\hat{v}_i} u_i(\hat{v}_i)\right] \tag{190}$$

The vectors $\hat{v}_i$ and $\nabla_{\hat{v}_i} u_i(\hat{v}_i)$ define a 2-d plane in which $\hat{v}_i'$ lies independent of the step size α. Therefore, we can consider gradients confined to a 2-d plane without loss of generality. Specifically, let $\hat{v}_i = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $\nabla = \nabla_{\hat{v}_i} u_i(\hat{v}_i) = \beta \begin{bmatrix} \cos(\psi) \\ \sin(\psi) \end{bmatrix}$. Then $\nabla^R = \nabla^R_{\hat{v}_i} u_i(\hat{v}_i) = \beta \begin{bmatrix} \cos(\psi) \\ 0 \end{bmatrix}$ and $\gamma = \sqrt{1 + ||\eta||^2} = \sqrt{1 + \alpha^2 \beta^2 \cos^2(\psi)}$. Also, let $z = \beta \cos(\psi)$ and $\alpha < \frac{1}{L_i}$ (see Equation (180) for definition) which implies $\alpha^2 ||\nabla^R||^2 < 1$. Then

$$u_i(\hat{v}_i') - u_i(\hat{v}_i) \tag{191}$$

$$= \overbrace{(\frac{1}{\gamma^2} - 1)}^{\leq 0} u_i(\hat{v}_i) + \frac{1}{\gamma^2}(||\eta||^2 u_i(\hat{\eta}) + 2\eta^\top \nabla_{\hat{v}_i} u_i(\hat{v}_i)) \tag{192}$$

$$\overset{CO.9.1}{\geq} (\frac{1}{\gamma^2} - 1) L_i + \frac{1}{\gamma^2}(\alpha^2 ||\nabla^R||^2 u_i(\hat{\eta}) + 2\alpha \nabla^\top \nabla^R) \tag{193}$$

$$\overset{CO.9.1}{\geq} (\frac{1}{\gamma^2} - 1) L_i + \frac{1}{\gamma^2}(2\alpha \nabla^\top \nabla^R + \alpha^2 ||\nabla^R||^2 (-L_i)) \tag{194}$$

$$= (\frac{1}{1 + \alpha^2 \beta^2 \cos^2(\psi)} - 1) L_i + \frac{\alpha}{1 + \alpha^2 \beta^2 \cos^2(\psi)}(2 - \alpha L_i)\beta^2 \cos^2(\psi) \tag{195}$$

$$= (\frac{1}{1 + \alpha^2 z^2} - 1) L_i + \frac{\alpha(2 - \alpha L_i)}{1 + \alpha^2 z^2} z^2 \tag{196}$$

$$= \frac{1}{1 + \alpha^2 z^2}(L_i - L_i \alpha^2 z^2 - L + \alpha(2 - \alpha L_i) z^2) \tag{197}$$

$$= \frac{1}{1 + \alpha^2 z^2}(-2L_i \alpha^2 z^2 + 2\alpha z^2) \tag{198}$$

$$= \frac{2\alpha z^2}{1 + \alpha^2 z^2}(1 - \alpha L_i) > 0 \tag{199}$$

where the assumption that $|\theta_j| \leq \frac{c_i g_i}{(i-1)\Lambda_{11}}$ was used to leverage Corollary O.9.1. Let $\alpha = \frac{1}{2L_i}$. Then $||\eta||^2 = \alpha^2 z^2 \leq \frac{1}{4}$ and

$$u_i(\hat{v}_i') - u_i(\hat{v}_i) \geq \frac{2\alpha z^2}{1 + \alpha^2 z^2}(1 - \alpha L_i) \tag{200}$$

$$= \frac{2\alpha^2 z^2}{1 + \alpha^2 z^2} \frac{1 - \alpha L_i}{\alpha} \tag{201}$$

$$= \frac{2L_i \alpha^2 z^2}{1 + \alpha^2 z^2} \tag{202}$$

$$= \frac{2L_i ||\eta||^2}{1 + ||\eta||^2} \tag{203}$$

$$\geq \min(\xi ||\eta||^2, \xi' ||\eta||) \tag{204}$$

with $\xi = \xi' = \frac{8}{5} L_i$. □

Lemma O.11 (Approximate Optimization is Reasonable Given Accurate Parents)**.** *Assume* $|\theta_j| \leq \frac{c_i g_i}{(i-1)\Lambda_{11}} \leq \sqrt{\frac{1}{2}}$ *for all* $j < i$ *with* $0 \leq c \leq \frac{1}{16}$*, i.e., the parents have been learned accurately. Then for any approximate local maximizer* $(\bar{\theta}_i, \bar{\Delta}_i)$ *of* $u_i(\hat{v}_i(\theta_i, \Delta_i), \hat{v}_{j<i})$*, if the angular deviation* $|\bar{\theta}_i - \theta_i^*| \leq \bar{\epsilon}$ *where* θ_i^* *forms the global max,*

$$|\bar{\theta}_i| \leq \bar{\epsilon} + 8c_i \tag{205}$$

where $\bar{\theta}_i$ *denotes the angular distance of the approximate local maximizer to the true eigenvector* v_i*.*

Proof. Note that the true eigenvector occurs at $\bar{\theta}_i = 0$. The result follows directly from Theorem N.2:

$$|\bar{\theta}_i| = |\bar{\theta}_i - 0| \leq |\bar{\theta}_i - \theta_i^*| + |\theta_i^* - 0| \leq \bar{\epsilon} + 8c_i. \tag{206}$$

□

Lemma O.12. *Assume* $\hat{v}_i$ *is initialized within* $\frac{\pi}{4}$ *of its maximizer and its parents are accurate enough, i.e., that* $|\theta_{j<i}| \leq \frac{c_i g_i}{(i-1)\Lambda_{11}} \leq \sqrt{\frac{1}{2}}$ *with* $0 \leq c_i \leq \frac{1}{16}$*. Let* ρ_i *be the maximum tolerated error desired for* $\hat{v}_i$*. Then Riemannian gradient ascent returns*

$$|\theta_i| \leq \frac{\pi}{g_i}\rho_i + 8c_i \tag{207}$$

after at most

$$\lceil \frac{5}{4} \cdot \frac{1}{\rho_i^2} \rceil \tag{208}$$

iterations.

Proof. Note that the assumptions of Lemma O.1 are met by Corollary O.9.1 and Lemma O.10 with $\xi = \xi' = \frac{8}{5}$ and Riemannian gradient ascent. Plugging into Lemma O.1 ensures that Riemannian gradient ascent returns unit vector $\hat{v}_i$ satisfying $u(\hat{v}_i) \geq u(\hat{v}_i^0)$ and $||\nabla^R|| \leq \rho_i$ in at most

$$\lceil \frac{u(\hat{v}_i^*) - u(\hat{v}_i^0)}{\frac{8}{5}L_i} \cdot \frac{1}{\rho_i^2} \rceil \tag{209}$$

iterations (where $\hat{v}_i$ is initialized to $\hat{v}_i^0$). Additionally, note that for any $\hat{v}_i$, $u_i(\hat{v}_i^*) - u_i(\hat{v}_i) \leq 2L_i$ where L_i bounds the absolute value of the utility u_i (see Corollary O.9.1) and $\hat{v}_i^* = \arg\max u_i(\hat{v}_i)$. Combining this with Lemma O.6 gives

$$|\theta_i - \theta_i^*| \leq \frac{\pi}{g_i}\rho_i \tag{210}$$

after at most

$$\lceil \frac{5}{4} \cdot \frac{1}{\rho_i^2} \rceil \tag{211}$$

iterations. Lastly, translating $|\theta_i - \theta_i^*|$ to $|\theta_i|$ using Lemma O.11 gives the desired result. □