Anchor-Cosine Steering: Theory and Diagnostics Appendix

This note formalizes an anchor-based "relative" representation used for activation steering and provides falsifiable diagnostics. It is written as a methods-style appendix.

A. Setup and notation

Let d be the hidden size (e.g., 4096). We work in $\mathbb{R}^d$, the space of length-d real vectors.

Choose m nonzero **anchor vectors** $a_1, \dots, a_m \in \mathbb{R}^d$. Define unit anchors

$$\hat{a}_i := \frac{a_i}{\|a_i\|}.$$

Stack them as columns into the matrix

$$A := [\hat{a}_1 \ \cdots \ \hat{a}_m] \in \mathbb{R}^{d \times m}.$$

Define the **Gram matrix**

$$G := A^\top A \in \mathbb{R}^{m \times m}.$$

Assume G is invertible (anchors are linearly independent). If not invertible, replace inverses with pseudoinverses and interpret projections onto $\mathrm{col}(A)$.

For any nonzero hidden activation $h \in \mathbb{R}^d$, define the **cosine feature vector**

$$c(h) := \frac{A^\top h}{\|h\|} \in \mathbb{R}^m,$$

so each coordinate $c_i(h)$ equals $\cos(h, a_i)$.

Let Π_A denote the orthogonal projection operator onto $\mathrm{col}(A) = \mathrm{span}\{\hat{a}_1, \dots, \hat{a}_m\}$.

B. Identifiability and coverage

Lemma 1 (Identifiability of anchor-span projection from $(c(h), \|h\|, A)$)

If G is invertible, then for any $h \neq 0$,

$$\Pi_A(h) \;=\; \|h\|\, A\, G^{-1}\, c(h).$$

Interpretation. This is an identifiability statement: given (i) the cosine features $c(h)$, (ii) the norm $\|h\|$, and (iii) the anchor geometry (through A and G^{-1}), you can recover the component of h that lies in the anchor span. Cosines alone are not sufficient.

Proof. Standard projection formula gives $\Pi_A(h) = A(A^\top A)^{-1}A^\top h = AG^{-1}A^\top h$. Since $c(h) = A^\top h/\|h\|$, we have $A^\top h = \|h\|c(h)$. Substitute.

Lemma 2 (Coverage / blindness)

Decompose any h uniquely as

$$h = h_\parallel + h_\perp,$$

where $h_\parallel := \Pi_A(h) \in \mathrm{col}(A)$ and $h_\perp \perp \mathrm{col}(A)$. Then

$$\|h - \Pi_A(h)\| = \|h_\perp\|.$$

Moreover, any method that only depends on $A^\top h$ (equivalently $(c(h), \|h\|)$) cannot recover $h_\perp$.

Interpretation. The anchor representation is blind to energy orthogonal to the anchor span. If trait-relevant shifts live largely in $h_\perp$, anchor-based steering cannot capture them.

Diagnostic (coverage). Measure the residual ratio

$$\rho(h) := \frac{\|h - \Pi_A(h)\|}{\|h\|} = \frac{\|h_\perp\|}{\|h\|}.$$

Compute $\rho(h)$ across prompts/layers/traits. Large ρ indicates poor anchor coverage.

C. Stability of the cosine-based reconstruction

Suppose you observe noisy cosine features $\tilde{c}(h) = c(h) + \delta$. Define the induced estimate

$$\tilde{\Pi}_A(h) := \|h\|\, A\, G^{-1}\, \tilde{c}(h).$$

Lemma 3 (Stability bound)

For any $h \neq 0$,

$$\|\tilde{\Pi}_A(h) - \Pi_A(h)\| \;\leq\; \|h\| \cdot \|A\| \cdot \|G^{-1}\| \cdot \|\delta\|.$$

Since $\|G^{-1}\| = 1/\lambda_{\min}(G)$, stability degrades as anchors become correlated/degenerate.

Diagnostic (stability). Compute $\lambda_{\min}(G)$ and/or the condition number $\kappa(G) = \lambda_{\max}(G)/\lambda_{\min}(G)$. If $\kappa(G)$ is large or $\lambda_{\min}(G)$ is tiny, prune/revise anchors.

D. Ridge-stabilized projection and explicit bias

Define a ridge-stabilized projector

$$\Pi_{A,\lambda}(h) := A\,(G + \lambda I)^{-1} A^\top h, \qquad \lambda > 0.$$

Lemma 4 (Ridge bias in closed form)

$$\Pi_{A,\lambda}(h) - \Pi_A(h) = -\lambda\, A\,(G + \lambda I)^{-1} G^{-1} A^\top h.$$

Corollary 4.1 (Bias magnitude bound)

$$\|\Pi_{A,\lambda}(h) - \Pi_A(h)\| \;\leq\; \frac{\lambda}{\lambda_{\min}(G) + \lambda}\, \|\Pi_A(h)\|.$$

Interpretation. Ridge improves numerical stability but shrinks the projected component. If $\lambda \gg \lambda_{\min}(G)$, shrinkage can be substantial.

Diagnostic (ridge choice). Report λ, $\lambda_{\min}(G)$, and $\kappa(G)$. If ridge is large relative to $\lambda_{\min}(G)$, interpret results as intentionally biased toward stability.

E. Learned projection in the small-n regime

Let $x_k \in \mathbb{R}^m$ be relative anchor-space diffs and $y_k \in \mathbb{R}^d$ be corresponding normalized activation diffs. Stack them into $X \in \mathbb{R}^{n \times m}$ and $Y \in \mathbb{R}^{n \times d}$.

The ridge estimator is

$$\hat{W} = \arg\min_W \|XW - Y\|_F^2 + \lambda \|W\|_F^2 = (X^\top X + \lambda I)^{-1} X^\top Y.$$

Important (small-n warning). In typical steering setups, n equals the number of contrast pairs (often ~5), while m (anchors) can be 12. Classical bounds are not quantitatively tight in this regime; claims about the learned projection must be validated empirically.

Diagnostic (generalization). Use leave-one-pair-out (LOO) evaluation: for each pair k, fit $\hat{W}^{(-k)}$ on all but k and evaluate prediction error (or downstream steering correlation) on k.

F. Nonlinearity: approximation vs estimation (corrected notation)

Let $f : \mathbb{R}^m \to \mathbb{R}^d$ be the unknown true map from anchor features to activation directions. Define the population best linear map

$$W^\star_{\text{pop}} := \arg\min_W \mathbb{E}\|f(x) - Wx\|^2 = \mathbb{E}[f(x)x^\top]\,(\mathbb{E}[xx^\top])^{-1},$$

and the empirical ridge estimator $\hat{W}$ trained from n samples.

Assumption: $W^\star_{\text{pop}}$ is well-defined only if $\mathbb{E}[xx^\top]$ is invertible (i.e., the anchor features have full rank in expectation). In practice, this holds when anchors are not collinear.

For any test point x,

$$\|f(x) - \hat{W}x\| \le \underbrace{\|f(x) - W^\star_{\text{pop}}x\|}_{\text{approximation error}} + \underbrace{\|(\hat{W} - W^\star_{\text{pop}})x\|}_{\text{estimation error}}.$$

Diagnostic (nonlinearity). Compare linear ridge to a mild nonlinear baseline on anchor features (e.g., kernel ridge or a small MLP) under LOO. If nonlinear does not improve held-out results, the linear model is empirically adequate.

G. Practical "Theory + Diagnostics" protocol (recommended)

1) **Coverage:** compute $\rho(h) = \|h - \Pi_A(h)\|/\|h\|$. If large, add anchors or avoid anchor-based steering at that layer.
2) **Stability:** compute $\lambda_{\min}(G)$ and $\kappa(G)$. If ill-conditioned, prune anchors; only increase ridge if you accept shrinkage (Corollary 4.1).
3) **Generalization:** perform LOO across contrast pairs for learned projection. If LOO fails, treat learned projection as in-sample only; prefer simple projection.
4) **Magnitude ablation:** augment features with $\log\|h\|$ and test whether it improves held-out control. If yes, cosine-only features were misspecified.
5) **Nonlinearity ablation:** compare ridge vs mild nonlinear baseline under LOO.

Implementation. See `anchor_steering_diagnostics.py` for reference implementations of these diagnostics.