

On the support size of the genomic OCS optimum

Adel Kaleche

Independent researcher, France.

11 October 2026

Abstract

A companion note to the support-first solver. For genomic optimum contribution selection (OCS) we bracket the size of the optimal support: an existential n -independent total-support bound holds without effective caps in the no-ridge limit (Theorem 1), no universal bound survives the ridge (Theorem 2), and the practical regime between them is governed by the joint geometry of spectrum, breeding values and cap, with no reliable scalar predictor among those tested. A conditional KKT identity (Proposition 1) describes free contributions. Historical experiments are retained with context; scripts and prerequisites are listed below.

1 The question

Genomic OCS solves, for n candidates with breeding values $\mathbf{b} \in \mathbb{R}^n$ and relationship matrix $\mathbf{G} = \mathbf{Z}\mathbf{Z}^\top/s + \varepsilon\mathbf{I}$ ($\mathbf{Z}$ the $n \times m$ centred genotype matrix, s a scale, ε a small ridge for positive-definiteness),

$$\max \mathbf{b}^\top \mathbf{c} \quad \text{s.t.} \quad \mathbf{A}\mathbf{c} = \mathbf{d}, \quad \mathbf{0} \leq \mathbf{c} \leq \mathbf{u}, \quad \mathbf{c}^\top \mathbf{G}\mathbf{c} \leq k, \quad (1)$$

where $\mathbf{A}$ is the $q \times n$ budget matrix ($q = 1$ for the simplex $\mathbf{1}^\top \mathbf{c} = 1$; $q = 2$ for the sexed form $\sum_{\text{males}} = \sum_{\text{females}} = \frac{1}{2}$). Empirically the optimum $\mathbf{c}^*$ activates very few candidates: its support $S = \{i : c_i^* > 0\}$ is ≈ 15 – 30 in historical loose-cap real-panel runs. At the manuscript’s operational contribution-proxy points it is tens to hundreds (mouse: 47/89/182 at $\Delta F = 2/1/0.5\%$); the structured n -sweep gives roughly 80–103 at a fixed proxy point. These observations do not prove a universal bound. What bounds $|S|$?

The answer is a *bracket*. In the uncapped no-ridge case there is an existential n -independent bound (§2); once the ridge is present there is provably no universal bound, the support reaching the whole population (§3); and the practical regime between them is set by the joint geometry of (spectrum, $\mathbf{b}, k$), whose empirical map shows no reliable scalar predictor among those tested (§5). A conditional KKT identity (§4) describes free contributions.

2 The $\varepsilon = 0$ bound

Theorem 1. *Assume $\varepsilon = 0$, $s > 0$, nonempty feasibility, and simplex or sex-budget equalities so that $\{\mathbf{A}\mathbf{c} = \mathbf{d}, \mathbf{c} \geq 0\}$ is bounded. With no effective upper caps, there exists an optimum with*

$$|S| \leq \text{rank} \begin{bmatrix} \mathbf{A} \\ \mathbf{Z}^\top \\ \mathbf{b}^\top \end{bmatrix} \leq q + r + 1, \quad r = \text{rank}(\mathbf{Z}) \leq m.$$

Proof. Start with any attained optimum $\mathbf{c}^*$ and fix its image and objective in

$$P = \{\mathbf{c} \geq 0 : \mathbf{A}\mathbf{c} = \mathbf{d}, \mathbf{Z}^\top \mathbf{c} = \mathbf{Z}^\top \mathbf{c}^*, \mathbf{b}^\top \mathbf{c} = \mathbf{b}^\top \mathbf{c}^*\}.$$

Every point has the same kinship $\|\mathbf{Z}^\top \mathbf{c}^*\|^2/s$ and objective, hence is optimal. P is nonempty and bounded. Let $M = [\mathbf{A}; \mathbf{Z}^\top; \mathbf{b}^\top]$, of rank at most $q + r + 1$. If a vertex had more than $\text{rank}(M)$ positive coordinates, there would be a nonzero null direction supported on those coordinates. Two sufficiently small opposite perturbations preserve the equalities and nonnegativity, contradicting extremality. A vertex proves the bound. No duality or constraint qualification is required. $\square$

Finite-cap extension. Intersect P with $\mathbf{c} \leq \mathbf{u}$. A vertex has at most $\text{rank}(M)$ free coordinates $F = \{i : 0 < c_i < u_i\}$, by the same perturbation argument. Positive upper-bound coordinates $U_+ = \{i : c_i = u_i > 0\}$ also count, so $|S| = |F| + |U_+| \leq \text{rank}(M) + |U_+|$. Arbitrary finite caps preclude an n -independent total-support bound: $q = 1, r = 0, \mathbf{G} = 0, k = 0, u_i = 2/n$ force $|S| \geq \lceil n/2 \rceil$, although $q + r + 1 = 2$. Strictly decreasing $\mathbf{b}$ attains that minimum at an optimal capped LP vertex.

The theorem is an existence result about a vertex of an *optimal slice*. The implemented active-set algorithm is not proved to return that vertex; no categorical conclusion about sparsity of interior-point or ADMM outputs follows. Enforcing the guarantee on arbitrary output would require separate vertex extraction.

3 No universal bound for $\varepsilon > 0$

Theorem 2. *For $\varepsilon > 0$, no total-support bound $f(q, r)$ independent of n exists, even without effective upper caps: the support can equal n .*

Proof (counterexample). Take the abstract OCS instance $q = 1, \mathbf{G} = \varepsilon \mathbf{I}, r = 0, u_i = 1$. This algebraic counterexample is not a usable marker panel for a public entry point. On the simplex $\|\mathbf{c}\|^2 \geq 1/n$ by Cauchy–Schwarz. If any coordinate is zero, $\|\mathbf{c}\|^2 \geq 1/(n-1)$. Thus for

$$\varepsilon/n < k < \varepsilon/(n-1),$$

the uniform plan is strictly kinship-feasible, while every feasible plan, and hence every optimum for any $\mathbf{b}$, has full support. This holds for arbitrary $n \geq 2$. $\square$

The historical sweep (`bound_validation.py`, block 7, $n = 300$) reported

$$|S| = 300, 295, 256, 151, 52, 12,$$

as k relaxed from $1.05\times$ to $50\times$ the minimum. These are observations for one $\mathbf{b}$; $1.05\times$ is outside the analytic full-support interval and is not a universal witness. Block 8 checks that interval and the finite-cap obstruction directly.

Here $\sum c_i^2$ measures concentration of contributions. With $N_e = 1/\sum c_i^2$, $\Delta F = 1/(2N_e)$ is a contribution-based proxy, not a measured rate of realised genomic inbreeding. Small observed support depends on the joint instance geometry, rather than a worst-case guarantee.

4 A conditional KKT identity

Proposition 1. *Suppose $\varepsilon > 0$ and an optimum admits KKT multipliers with $\lambda > 0$ for kinship. On the free coordinates $F = \{i : 0 < c_i^* < u_i\}$,*

$$c_i^* = \alpha b_i + \mathbf{w}^\top \mathbf{z}_i + v^\top a_i, \quad i \in F,$$

where a_i is column i of $\mathbf{A}$. For sex-incidence budgets $v^\top a_i = \beta_{\text{sex}(i)}$.

Proof. Let $\ell, h \geq 0$ be the lower and upper multipliers. Stationarity and complementarity are

$$\mathbf{b} = \mathbf{A}^\top \mu + 2\lambda \mathbf{G} \mathbf{c}^* - \ell + h, \quad \ell_i c_i^* = 0, \quad h_i (u_i - c_i^*) = 0, \quad \lambda (\mathbf{c}^{*\top} \mathbf{G} \mathbf{c}^* - k) = 0.$$

On F , both bound multipliers vanish. With $\mathbf{y} = \mathbf{Z}^\top \mathbf{c}^*$ and $\mathbf{G} \mathbf{c}^* = \mathbf{Z} \mathbf{y} / s + \varepsilon \mathbf{c}^*$, division by $2\lambda \varepsilon$ gives $\alpha = 1/(2\lambda \varepsilon)$, $\mathbf{w} = -\mathbf{y}/(s\varepsilon)$, $v = -\mu/(2\lambda \varepsilon)$. $\square$

The identity uses $m+q+1$ coefficients ($m+2$ for simplex, $m+3$ for two sex budgets). Active kinship alone does not imply $\lambda > 0$. Existence of multipliers needs a suitable constraint qualification or a supplied certificate, and is assumed here. The identity does not apply to upper-bound coordinates and proves no cardinality bound. Taking $\varepsilon \rightarrow 0$ cannot recover Theorem 1: multipliers and supports may change or diverge, and stationarity holds only on the current free coordinates. The slice proof is independent.

5 The empirical regime between the theorems

Between the two clean statements lies the regime breeders use, mapped across synthetic spectra, $\mathbf{b}$ -alignments and caps and on the real panels:

- **$\mathbf{b}$'s alignment with the spectrum drives $|S|$, inversely.** With $\mathbf{G}$ and k fixed, putting $\mathbf{b}$ on the dominant (coancestry-expensive) eigendirection forces a large support (mean $|S| \approx 146$ of $n = 800$ for $\mathbf{b}$ the top eigenvector); spreading $\mathbf{b}$ across many cheap directions collapses it to ≈ 4 . Gain sought in an expensive direction must be diluted over many candidates to respect the cap (`bound_balalign.py`).
- **No single scalar predicts $|S|$ across regimes.** Neither $\text{rank}(\mathbf{G}_0)$, the effective rank / participation ratio of the spectrum, the directional cost $\mathbf{b}^\top \mathbf{G} \mathbf{b} / \mathbf{b}^\top \mathbf{b}$, nor the support of the unconstrained direction $(\mathbf{G}^{-1} \mathbf{b})_+$ tracks $|S|$ across both synthetic and real instances; the best dimensionless combination, $|S| \sim (\mathbf{b}^\top \mathbf{G} \mathbf{b} / \mathbf{b}^\top \mathbf{b} \cdot 1/k)^{0.8}$, reaches only $R^2 \approx 0.58$ and mispredicts the real panels' relative support (`bound_predictor.py`, `bound_lawfit.py`).
- **The curve $|S|(k)$ is a per-instance power law of non-universal exponent.** $|S| \sim k^{-\alpha}$ fits each instance well ($R^2 \approx 0.9$ – 0.98 where the support is well resolved), with $\alpha \approx 1$ on both real panels but ranging 0.5 – 1.7 across synthetic spectra and alignments (`bound_curve.py`).
- **Real panels.** Wheat ($n = 599$) and mouse ($n = 1814$): $|S| = 25/26$ and $17/26$ (simplex / sexed-optiSel) at the historical loose cap. These two fixed panels do not establish stability in n ; operational targets in the manuscript have larger supports (`bound_real.py`).

Theorem 2 rules out a universal bound in q, r alone. The empirical sweeps separately show that the scalar predictors tested here do not reliably predict support across regimes.

6 A conditional bound: the obstruction and the conjecture

A conditional bound remains open. If the kinship constraint is *redundant on the whole budget polytope*, an uncapped LP admits an optimal vertex with $|S| \leq \text{rank}(\mathbf{A}) \leq q$; finite caps instead give only $|F| \leq \text{rank}(\mathbf{A})$. Slack kinship at one optimum is insufficient: for constant $\mathbf{b}$, $\mathbf{G} = \mathbf{I}$ and $1/n < k < 1/(n-1)$, the uniform optimum is slack but every feasible optimum has full support. For exact rank r and $\varepsilon = 0$, Theorem 1 gives the uncapped existential bound and its finite-cap extension bounds free coordinates.

A bound in terms of *effective* rank — the realistic ask, since real spectra decay — does not follow from the same argument, and the obstruction is precise. Theorem 1 linearises the optimal slice by pinning $\mathbf{Z}^\top \mathbf{c}$, the *entire* range of $\mathbf{G}_0$; for a full-rank $\mathbf{G}$ one would have to pin all n projections $u_i^\top \mathbf{c}$, giving only the trivial $|S| \leq q + n + 1$. One cannot pin just the top- d projections and drop the tail, because the tail contributes $\mathbf{c}^\top \mathbf{E} \mathbf{c} = \sum_{i>d} \lambda_i (u_i^\top \mathbf{c})^2$ — a *genuine quadratic*, not a low-rank term — so fixing it is not a linear constraint and the slice is not a polytope. The tail’s curvature is exactly what lets the support spread; effective rank alone cannot replace exact rank in this vertex count. A conditional bound therefore needs a different (spectral-perturbation) argument.

Two qualitative conditions motivate a possible conditional theorem; they are not quantitatively specified or proved sufficient:

- **(A1) No large degenerate cheap plateau.** A block of many near-equal small eigenvalues is a cheap subspace the optimum spreads into; $\mathbf{G} = \varepsilon \mathbf{I}$ is the extreme case, giving $|S| = n$ (Theorem 2). A smoothly decaying spectrum has no such plateau.
- **(A2) $\mathbf{b}$ avoids the dominant directions.** Even with a gap, $\mathbf{b}$ on the top (expensive) eigenvector forces dilution and a large support (mean $|S| \approx 146$ of 800 in the alignment sweep).

The pure-ridge counterexample and alignment sweep motivate (A1) and (A2), without proving necessity of these informal conditions. Under quantitative versions at a specified operating point we conjecture small support; the historical 15–30 values concern loose caps, not all real panels; the growth half ($|S|$ rising as k approaches its feasible minimum) follows the contribution-based $\Delta F = 1/(2N_e)$ proxy, distinct from realised genomic inbreeding, in the context of the genetic contributions literature (Wray & Thompson 1990; Woolliams & Bijma 2000). Proving it — coupling the spectral gap with $\mathbf{b}$ ’s projection onto the top eigenspace, robust to the tail’s curvature — is the open problem, the meeting point of optimisation (the perturbed-cone face) and quantitative genetics (effective lineages).

An exact handle on the effective support. Write the *effective* number of contributors as the participation ratio $\text{PR}(\mathbf{c}^*) = 1/\|\mathbf{c}^*\|^2$ (it equals $|S|$ for uniform contributions, less when uneven). At a positive active cap $k > 0$ with $\mathbf{1}^\top \mathbf{c}^* = 1$ an identity holds: $\text{PR}(\mathbf{c}^*) = R(\mathbf{c}^*)/k$, where $R(\mathbf{c}^*) = \mathbf{c}^{*\top} \mathbf{G} \mathbf{c}^* / \mathbf{c}^{*\top} \mathbf{c}^*$ is the optimum’s own Rayleigh quotient (since $R\|\mathbf{c}^*\|^2 = \mathbf{c}^{*\top} \mathbf{G} \mathbf{c}^* = k$). Hence $\text{PR}(\mathbf{c}^*) \leq \lambda_1/k$ for every such active positive cap, and a bound on *effective* support reduces to bounding $R(\mathbf{c}^*)$ — the directional coancestry cost the optimum is forced to incur. This is where (A2) bites: the measured cases where $\mathbf{b}$ avoids dominant directions have small $R(\mathbf{c}^*)$ (observed: $R \approx 0.02$ – 0.4 on the structured cases versus $\lambda_1 \approx 45$ – 143 , giving $\text{PR} \approx 2$ – 12), whereas $\mathbf{G} = \varepsilon \mathbf{I}$ forces $R = \varepsilon = \lambda_1$. At an active cap, $\text{PR} = \varepsilon/k$ lies in $(n-1, n)$ throughout the full-support interval $\varepsilon/n < k < \varepsilon/(n-1)$, and approaches n as k decreases to ε/n . Total support is already n throughout that interval; $\text{PR} = n$ holds only for uniform contributions. For example, $n = 3$ and $k = 5\varepsilon/12$ give $|S| = 3$ but $\text{PR} = 12/5$. The residual open piece is now sharp and scalar — bound $R(\mathbf{c}^*)$ a priori under (A1)+(A2). (Historically $|S|$ ran ≈ 1.5 – $2.5 \times \text{PR}$ in these cases. In general $\text{PR} \leq |S|$, so an upper bound on PR does not bound raw cardinality without another assumption on contribution sizes.) Verified in `bound_kernel.py`.

Attacking this residual. For positive-definite $\mathbf{G}$, the simplex relaxation without lower or upper bounds has, in the active-kinship case with $\lambda > 0$, the Rayleigh-quotient expression $R_{\text{relax}} = (\beta^\top \mathbf{G}^{-1} \beta) / (\beta^\top \mathbf{G}^{-2} \beta)$ with $\beta = \mathbf{b} - \mu \mathbf{1}$ — a λ -average weighted by β_i^2 / λ_i^2 , so small under (A2). But it predicts $R(\mathbf{c}^*)$ only when $\mathbf{c} \geq 0$ is near-inactive: on the real panels the unconstrained optimum carries large negative entries (min ≈ -1 to -2.6) and R_{relax} falls 60–300 $\times$ below $R(\mathbf{c}^*)$. So **nonnegativity is the irreducible core** — fittingly, $|S|$ is the cardinality of a nonnegativity-constrained QP optimum, hard in general — and a proof must be genuinely $\mathbf{c} \geq 0$ -aware, not a relaxation shortcut. The clean a priori handle (the relaxation formula) and the irreducible obstacle ($\mathbf{c} \geq 0$) are now both identified.

Numerical evidence (reproducible)

`bound_validation.py` (ridged low-rank observations, not a test of the $\varepsilon = 0$ vertex theorem, empirical n -sweep, ridge and spectral-decay sweeps, and the Theorem 2 historical sweep, plus the analytic finite-cap and full-support witnesses in block 8); `bound_real.py` (wheat and mouse, after `research/repro/*_export.R`); `bound_balign.py` (the **b**-alignment sweep); `bound_predictor.py` and `bound_lawfit.py` (the predictor / law search); `bound_curve.py` (the per-instance $|S|(k)$ fit). Full log: `research/support_bound_sketch.md`.

References

1. **Pataki, G. (1998).** On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Math. Oper. Res.* 23(2):339–358. DOI 10.1287/moor.23.2.339.
2. **Markowitz, H. (1956).** The optimization of a quadratic function subject to linear constraints. *Naval Res. Logist. Q.* 3(1–2):111–133. DOI 10.1002/nav.3800030110.
3. **Wray, N.R. & Thompson, R. (1990).** Prediction of rates of inbreeding in selected populations. *Genet. Res.* 55(1):41–54. DOI 10.1017/S0016672300025180.
4. **Woolliams, J.A. & Bijma, P. (2000).** Predicting rates of inbreeding in populations undergoing selection. *Genetics* 154(4):1851–1864. DOI 10.1093/genetics/154.4.1851.
5. **Yamashita, M., Mullin, T.J. & Safarina, S. (2018).** An efficient second-order cone programming approach for optimal selection in tree breeding. *Optim. Lett.* 12(7):1683–1697. DOI 10.1007/s11590-018-1229-y.
6. **Waldmann, P. (2025).** Genomic optimum contribution selection and mate allocation using JuMP. *Bioinform. Adv.* 5(1):vbaf259. DOI 10.1093/bioadv/vbaf259.