What Is the Alignment Tax?

Anonymous ACL submission

Abstract

The alignment tax is widely discussed but has not been formally characterized. We propose a geometric characterization of the alignment tax in representation space. Under linear representation assumptions, we define the alignment tax rate as the squared projection of the safety direction onto the capability subspace and derive the Pareto frontier governing safety-capability tradeoffs, parameterized by a single quantity of the principal angle between the safety and capability subspaces. We prove this frontier is tight and show it has a recursive structure. safety-safety tradeoffs under capability constraints are governed by the same equation, with the angle replaced by the partial correlation between safety objectives given capability directions. We derive a scaling law decomposing the alignment tax into an irreducible component determined by data structure and a packing residual that vanishes as $O(m'/d)$ with model dimension d, and establish conditions under which capability preservation mediates or resolves conflicts between safety objectives.

1 Introduction

The alignment tax, often intuitively invoked as the capability cost of making an AI system safe, is one of the most frequently invoked concepts in AI alignment (Askell et al., 2021; Lin et al., 2024). The intuition is if aligning a model requires constraining its behavior, some capability may be lost. Empirically, reinforcement learning from human feedback (RLHF) (Christiano et al., 2017) degrades performance on reasoning benchmarks (Ouyang et al., 2022), low-rank safety fine-tuning introduces small but measurable capability costs (Mou et al., 2025), and the severity of degradation varies across tasks (Huang et al., 2025).

Despite its centrality, as far as we are aware, the alignment tax has no mathematical definition. Researchers use it informally to mean "some capability was lost," measure it as the difference in benchmark scores before and after alignment, and propose ad hoc mitigations without a theory to guide them. No one has answered the question: *what is the mathematical object that we refer to as the alignment tax?*

This is surprising. The alignment tax has been taxonomized (Leike, 2022), measured empirically (Ouyang et al., 2022; Lin et al., 2024; Huang et al., 2025), and mitigated algorithmically (Niu et al., 2026; Mou et al., 2025), but it has not been given a mathematical definition from which consequences can be derived.

We work toward answering this question. Independent of the specific modeling assumptions, we view a primary contribution as conceptual by proposing that the alignment tax admits mathematical definition and demonstrating that, once defined, it has structure rich enough to generate a theory.

Under the linear representation hypothesis, the assumption that safety and capability are encoded as linear directions in representation space (Park et al., 2024; Arditi et al., 2024; Turner et al., 2024; Rimsky et al., 2024), the alignment tax has geometric structure. It is governed by the principal angles between the safety and capability subspaces, and its shape is an elliptic Pareto frontier parameterized by a single angle α:

$$\Delta_S = \Delta_C \cos\alpha + \sin\alpha\sqrt{B^2 - \Delta_C^2},$$

where Δ_S is the safety gain, Δ_C is the capability change, B is the perturbation budget (from the KL penalty), and α is the angle between the safety and capability directions. When $\alpha = 0$ (safety and capability are aligned), the tradeoff is linear and unavoidable. When $\alpha = \pi/2$ (orthogonal), the tradeoff vanishes and safety can be maximized independently of capability. We derive the exact, tight Pareto frontier for any number of capabilities and define a computable tax rate $\tau = \|P_C v^*\|^2$ that predicts per-task capability degradation from

representation geometry alone.

We show that prior independent empirical findings (null-space policy optimization, low-rank safety alignment, model averaging, heterogeneous model averaging, and anisotropic capability degradation) can be seen as special cases of the principal-angle structure. We do not claim these as evidence, since they are consistent with any model that interpolates between full tradeoff and independence via a continuous parameter. The value lies in generating quantitative predictions that go beyond existing results, such as per-task tax rates computable before alignment training, a scaling decomposition into reducible and irreducible components, and conditions under which capability preservation resolves safety-safety conflicts.

2 Related Work

The term "Alignment tax" originated informally as a reference to the capability cost of safety (Leike, 2022). Ouyang et al. (2022) documented "minimal regressions" on NLP tasks after RLHF. Lin et al. (2024) systematically measured Pareto fronts between alignment reward and task performance, finding model averaging achieves the best tradeoff. Huang et al. (2025) showed reasoning degrades more than other capabilities. Hu et al. (2025) extended the tax to calibration. Prior works engage with the term empirically but there has not been a mathematical definition or an analysis on the shape of the tradeoff surface.

Some recent optimization methods (Niu et al., 2026; Wang et al., 2026) project safety gradients into the null space of capability representations or uses SVD and null-space projection to expand reasoning capacity. Mou et al. (2025) show LoRA decouples safety into an orthogonal subspace. Zhong et al. (2024) learn a Pareto front over multiple preference dimensions. SACPO (Wachi et al., 2024) formulates alignment as constrained optimization. These operate in reward/policy space rather than representation space and do not derive the geometric structure of the tradeoff surface. These methods implement special cases of a geometric formalization without a unifying theory. Our contribution is recognizing that all three operate in the same mathematical structure (principal angles between subspaces) and deriving the consequences.

The linear representation hypothesis (Park et al., 2024) posits that concepts are encoded as directions in representation space. Empirical support includes refusal directions (Arditi et al., 2024), steering vectors (Turner et al., 2024), linear probes (Li et al., 2023), and concept erasure (Belrose et al., 2023). We adopt this hypothesis and derive its consequences for alignment tradeoffs.

Principal angles between subspaces (Björck and Golub, 1973) and canonical correlation analysis (Hotelling, 1936) are classical tools. Superposition in neural networks (Elhage et al., 2022) motivates the packing model. We apply these tools to the safety and capability subspaces of neural representations.

3 Setup and Definitions

3.1 Representation Space

Let us consider a model with representations $h \in \mathbb{R}^d$, where d is the dimension of the residual stream or hidden state at a fixed layer. We work at a fixed layer throughout; the formalization applies independently at each layer.

Definition 1 (Safety direction)**.** *The safety direction is a unit vector $v^* \in \mathbb{S}^{d-1}$ such that $\langle v^*, h\rangle$ measures the safety-relevant content of representation h. More generally, safety may occupy a subspace $S \subseteq \mathbb{R}^d$ of dimension k; we treat the rank-1 case ($k = 1$) as the primary setting, consistent with empirical evidence that the refusal direction is approximately one-dimensional (Arditi et al., 2024).*

Remark 1 (Higher-dimensional safety)**.** *The rank-1 assumption is adopted for exposition; the results generalize to a k-dimensional safety subspace S with orthonormal basis $V \in \mathbb{R}^{d\times k}$. The single angle α is replaced by the spectrum of principal angles $\alpha_1 \leq \cdots \leq \alpha_k$ between S and C, the tax rate τ generalizes to a Frobenius tax $\tau_F = \sum_i \cos^2 \alpha_i$, and the Pareto frontier applies per principal direction. The scaling law residual becomes $O(m'k/d)$, but the average per-direction tax converges at rate $O(m'/d)$ independent of k. The extension is tractable for $k = O(\log d)$ and degrades gracefully for $k = d^\gamma$ with $\gamma < 1/2$. See Appendix A for the full treatment.*

Definition 2 (Capability directions)**.** *For each capability $i \in [m]$, the capability direction $c_i \in \mathbb{S}^{d-1}$ is defined as*

$$c_i = \frac{\nabla_h f_i(h)}{\|\nabla_h f_i(h)\|}, \tag{1}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is a differentiable capability metric evaluated at the base model's representation. The capability subspace is $C = \mathrm{span}(c_1, \ldots, c_m)$.

The gradient interpretation ensures that c_i is well-defined for any differentiable capability metric and that $\langle c_i, \delta \rangle$ captures the first-order effect of a representation perturbation δ on capability i.

Definition 3 (Perturbation budget)**.** *Alignment modifies the model so that representations shift by $\delta \in \mathbb{R}^d$, subject to a budget constraint $\|\delta\| \leq B$. The budget B arises from the KL penalty in RLHF/DPO objectives: to first order, $\mathrm{KL}(\pi_\theta \| \pi_{\mathrm{ref}}) \approx \frac{1}{2}\delta^\top F \delta$ where F is the Fisher information. The isotropic constraint $\|\delta\| \leq B$ is the simplification obtained by replacing F with its average eigenvalue; we discuss anisotropic extensions in §4.4.*

Given these definitions, the safety gain and capability change from perturbation δ are:

$$\Delta_S = \langle v^*, \delta \rangle, \qquad \Delta_{C_i} = \langle c_i, \delta \rangle. \tag{2}$$

Definition 4 (Alignment tax rate)**.** *The* alignment tax rate *for safety direction v^* relative to capability subspace C is*

$$\tau = \|P_C v^*\|^2 \in [0, 1]. \tag{3}$$

When $\tau = 0$, safety is orthogonal to all capabilities (zero tax). When $\tau = 1$, safety lies entirely within the capability subspace (any safety gain requires capability loss).

4 The Pareto Frontier

4.1 Single-Capability Frontier

Theorem 5 (Single-capability Pareto frontier)**.** *Let $v^* \in \mathbb{S}^{d-1}$ be the safety direction and $c \in \mathbb{S}^{d-1}$ be a capability direction with angle $\alpha = \arccos(\langle v^*, c \rangle)$ between them. The Pareto frontier, or the maximum achievable safety gain Δ_S for a given capability change Δ_C subject to $\|\delta\| \leq B$ is:*

$$\Delta_S = \Delta_C \cos\alpha + \sin\alpha \sqrt{B^2 - \Delta_C^2}, \quad |\Delta_C| \leq B. \tag{4}$$

This frontier is tight. for each $\Delta_C \in [-B, B]$, there exists δ^* with $\|\delta^*\| = B$ achieving equality.

Proof. Project δ onto $\mathrm{span}(v^*, c)$; components orthogonal to both contribute to neither objective. Write $v^* = e_1$, $c = \cos\alpha\, e_1 + \sin\alpha\, e_2$ in this plane. Then $\Delta_S = \delta_1$, $\Delta_C = \delta_1 \cos\alpha + \delta_2 \sin\alpha$, with $\delta_1^2 + \delta_2^2 \leq B^2$. Solving for $\delta_2 = (\Delta_C - \delta_1 \cos\alpha)/\sin\alpha$ and substituting into the budget constraint yields $(\delta_1 - \Delta_C \cos\alpha)^2 \leq \sin^2\alpha (B^2 - \Delta_C^2)$, giving $\delta_1 \leq \Delta_C \cos\alpha + \sin\alpha \sqrt{B^2 - \Delta_C^2}$. Equality is achieved when $\delta^* = (\Delta_C \cos\alpha + \sin\alpha \sqrt{B^2 - \Delta_C^2})\, e_1 + (\Delta_C - \delta_1^* \cos\alpha)/\sin\alpha\, e_2$, which saturates the budget $\|\delta^*\| = B$. □

Remark 2 (Limiting cases)**.** *When $\alpha = 0$: $\Delta_S = \Delta_C$ (linear tradeoff). When $\alpha = \pi/2$: $\Delta_S = \sqrt{B^2 - \Delta_C^2}$ (quarter circle independent optimization). The frontier continuously interpolates between these extremes.*

4.2 Maximum Safety at Fixed Capability

Theorem 6 (Maximum safety gain under capability constraint)**.** *Let C be the capability subspace (dimension m) and v^* the safety direction. For a fixed capability constraint $P_C \delta = \delta_C^*$, the maximum safety gain subject to $\|\delta\| \leq B$ is:*

$$\Delta_S^{\max} = \langle v^*, \delta_C^* \rangle + \sqrt{B^2 - \|\delta_C^*\|^2} \cdot \|P_{C^\perp} v^*\|, \tag{5}$$

achieved by $\delta^* = \delta_C^* + \sqrt{B^2 - \|\delta_C^*\|^2} \cdot P_{C^\perp} v^* / \|P_{C^\perp} v^*\|$.

Proof. Decompose $\delta = \delta_C^* + \delta_\perp$ with $\delta_\perp \in C^\perp$ and $\|\delta_\perp\| \leq B' := \sqrt{B^2 - \|\delta_C^*\|^2}$. Then $\langle v^*, \delta \rangle = \langle v^*, \delta_C^* \rangle + \langle P_{C^\perp} v^*, \delta_\perp \rangle$. The second term is maximized when $\delta_\perp = B'\, P_{C^\perp} v^* / \|P_{C^\perp} v^*\|$, giving $B' \|P_{C^\perp} v^*\|$. □

Notice that the quantity $\|P_{C^\perp} v^*\| = \sqrt{1 - \tau}$ measures the component of safety outside the capability subspace; or, the "room" available for free safety improvement.

4.3 Tax Rate Properties

Corollary 7 (Tax-free safety)**.** *The maximum safety achievable at zero capability cost ($\delta_C^* = 0$) is:*

$$\Delta_S^{\mathrm{free}} = B\sqrt{1 - \tau}. \tag{6}$$

When $\tau \ll 1$, almost the full budget is available for tax-free safety improvement.

Corollary 8 (Anisotropy)**.** *For individual capability direction $c_i \in \mathbb{S}^{d-1}$, the per-task tax rate is:*

$$\tau_i = \langle v^*, c_i \rangle^2. \tag{7}$$

Different capabilities have different tax rates depending on their angle to safety. This quantity is directly measurable via probing.

Remark 3 (Non-additivity)**.** *The multi-capability tax rate satisfies* $\tau = (C^\top v^*)^\top (C^\top C)^{-1}(C^\top v^*)$ *where* $C = [c_1|\cdots|c_m]$*, which reduces to* $\sum_i \tau_i$ *only when capabilities are orthonormal. In general, capabilities may overlap, and the joint tax can be less than the sum of individual taxes.*

Corollary 9 (Negative tax)**.** *If* $\langle v^*, \delta_C^* \rangle > 0$*, that is, the capability target has positive projection onto safety, then the safety gain exceeds what the residual budget alone would provide:*

$$\Delta_S^{\max} = \langle v^*, \delta_C^* \rangle + \sqrt{B^2 - \|\delta_C^*\|^2}\,\sqrt{1-\tau} > \sqrt{B^2 - \|\delta_C^*\|^2}\,\sqrt{1-\tau} \tag{8}$$

In particular, when the capability target is small ($\|\delta_C^*\| \ll B$) *and well-aligned with safety* ($\langle v^*, \delta_C^* \rangle \approx \|\delta_C^*\|$)*, the total safety gain exceeds the zero-capability-cost amount* $B\sqrt{1-\tau}$*:*

$$\Delta_S^{\max} \approx \|\delta_C^*\| + B\sqrt{1-\tau} > B\sqrt{1-\tau}$$

Capability improvement subsidizes safety when the projection $\langle v^*, \delta_C^* \rangle$ *is large enough to compensate for the reduced residual budget.*

4.4 Anisotropic Budget Extension

When the Fisher information F is not a scalar multiple of the identity, the budget set is an ellipsoid $\delta^\top F \delta \le B^2$. The analysis generalizes by working in the whitened coordinate system $\tilde{\delta} = F^{1/2}\delta$, $\tilde{v}^* = F^{-1/2}v^*/\|F^{-1/2}v^*\|$, $\tilde{c}_i = F^{-1/2}c_i/\|F^{-1/2}c_i\|$, where the budget becomes $\|\tilde{\delta}\| \le B$. All results carry over with modified angles $\tilde{\alpha}$ computed in the whitened space. The qualitative conclusions of tradeoffs governed by subspace angles, tax rate determined by projection are invariant to the budget shape.

5 Scaling Law for the Alignment Tax

How does the alignment tax τ behave as the model dimension d increases? We provide an answer under a structured packing model.

5.1 Feature Packing Model

Assume the model represents N features as unit vectors $f_1, \ldots, f_N \in \mathbb{S}^{d-1}$, typically with $N > d$ (superposition regime; Elhage et al. 2022). Safety is one feature $v^* = f_s$ and capabilities are features $c_i = f_{\sigma(i)}$.

Definition 10 (Intrinsic overlap)**.** *Feature pair* (f_i, f_j) *has intrinsic overlap* γ_{ij} *if, in the limit* $d \to \infty$*, the optimal representation satisfies* $\langle f_i, f_j \rangle \to \gamma_{ij}$*. This is determined by the co-occurrence structure and task requirements of the data.*

Definition 11 (Packing residual)**.** *The packing residual is* $\epsilon_{ij} = \langle f_i, f_j \rangle - \gamma_{ij}$*: the overlap in excess of the intrinsic value, forced by finite dimensionality.*

Definition 12 (Coherence)**.** *The coherence of the packing is* $\mu = \max_{(i,j):\gamma_{ij}=0} |\langle f_i, f_j \rangle|$*: the maximum packing residual among intrinsically unrelated features.*

Standard bounds give $\mu = O(\sqrt{\log N/d})$ for random packings and $\mu \ge \sqrt{(N-d)/(d(N-1))}$ by the Welch (1974) bound. Equiangular tight frames achieve the Welch bound when they exist.

5.2 Scaling Theorem

Let $I \subset [m]$ denote capabilities with nonzero intrinsic overlap to safety ($\gamma_i \ne 0$), and $m' = m - |I|$ those with only incidental overlap.

Theorem 13 (Scaling law for alignment tax)**.** *Assume the capability directions satisfy near-orthogonality* ($m\mu < 1$)*. Let* $\bar{\gamma} = \max_{i \in I} |\gamma_i|$*. Then:*

$$\tau = \tau_0 + R(d), \tag{9}$$

where $\tau_0 = \sum_{i \in I} \gamma_i^2$ *is the irreducible tax and the packing residual satisfies:*

$$|R(d)| \le \frac{\tau_0 m\mu + m'\mu^2 + 2\bar{\gamma}|I|\mu + |I|\mu^2}{1 - m\mu} \tag{10}$$

Under random packing ($\mu^2 = O(\log N/d)$ *for* $N = d^p$*):*

$$|R(d)| = O\left(\frac{m' \log N}{d} + \frac{\bar{\gamma}|I|\sqrt{\log N}}{\sqrt{d}}\right) \tag{11}$$

Proof. Under near-orthogonality, $\|P_C v^*\|^2 = (1 + O(m\mu))\|C^\top v^*\|^2$ by Gershgorin's theorem applied to $C^\top C$. Decompose: $\|C^\top v^*\|^2 = \sum_{i \in I}(\gamma_i + \epsilon_i)^2 + \sum_{i \notin I} \epsilon_i^2$. Expand the first sum: $\sum_{i \in I} \gamma_i^2 + 2\sum_{i \in I} \gamma_i \epsilon_i + \sum_{i \in I} \epsilon_i^2$. Bound each term using $|\epsilon_i| \le \mu$. □

Remark 4. *The scaling law extends to* k*-dimensional safety with Frobenius residual* $O(m'k/d)$*; the average tax rate converges at* $O(m'/d)$ *regardless of* k *(Theorem 28).*

Proposition 14 (Tight version under random packing)**.** *If incidental capability directions are independent and uniform on $\mathbb{S}^{d-1}$, then:*

$$\tau = \tau_0 + \frac{m'}{d} + O\left(\frac{\bar{\gamma}\sqrt{|I|\log d}}{\sqrt{d}} + \frac{\sqrt{m'} + \log d}{d}\right) \tag{12}$$

with high probability. The leading packing term m'/d is exact in expectation.

Proof. For independent uniform vectors, $\langle v^*, c_i\rangle^2$ has mean $1/d$ and is sub-exponential with parameter $O(1/d^2)$. Apply Bernstein's inequality (see Giroux et al. (1979)) to the sum. □

Corollary 15 (Scaling regimes)**.**

1. Fixed capabilities *(m' constant): $\tau \to \tau_0$ at rate $O(\log d/d)$.*
2. Sublinear growth *($m' = d^\beta$, $\beta < 1$): $\tau \to \tau_0$ at rate $O(\log d/d^{1-\beta})$.*
3. Linear growth *($m' = \Theta(d)$): $\mathbb{E}[\tau - \tau_0] = \Theta(1)$. Scaling does not reduce the incidental tax.*

Remark 5. *It may be possible to plot per-task alignment tax against model dimension. Tasks where τ_i decreases with d have incidental overlap (alignment tax is an engineering problem solvable by scaling). Tasks where τ_i plateaus have intrinsic overlap (alignment tax is a fundamental tradeoff requiring objective modification).*

Remark 6 (Self-consistency of near-orthogonality)**.** *For the near-orthogonality condition $m\mu < 1$ to hold under random packing with $m = d^\beta$ and $\mu = O(\sqrt{\log d/d})$, we need $d^\beta\sqrt{\log d/d} = d^{\beta-1/2}\sqrt{\log d} \to 0$, which holds for $\beta < 1/2$. For $\beta \in [1/2, 1)$, the Gershgorin bound (see Horn and Johnson (2012)) is not small and the approximation $\|P_C v^*\|^2 \approx \|C^\top v^*\|^2$ weakens. However, the qualitative conclusion of incidental tax vanishes if $m' = o(d)$ persists via direct analysis of $\mathbb{E}[\|P_C v^*\|^2]$ using the Marchenko and Pastur (1967) distribution.*

6 Multi-Objective Safety and the Conflict Theorem

Safety fine-tuning typically involves multiple objectives (e.g. harmlessness and helpfulness). We show the tradeoff between safety objectives under capability constraints is governed by the same frontier equation.

6.1 Safety-Safety Frontier

Let $v_1^*, v_2^* \in \mathbb{S}^{d-1}$ be two safety directions with $\langle v_1^*, v_2^*\rangle = \rho$, and let $c \in \mathbb{S}^{d-1}$ be a capability direction with $a = \langle c, v_1^*\rangle$, $b = \langle c, v_2^*\rangle$.

Theorem 16 (Safety-safety Pareto frontier under capability preservation)**.** *Under capability preservation ($\langle c, \delta\rangle = 0$), define projected safety directions $\tilde{v}_i = P_{c^\perp} v_i^*$ and the effective angle θ between them:*

$$\cos\theta = \frac{\rho - ab}{\sqrt{(1-a^2)(1-b^2)}}. \tag{13}$$

The Pareto frontier for normalized safety gains $s_i = \Delta_{S_i}/(B'\|\tilde{v}_i\|)$, where $B' = \sqrt{B^2 - \|\delta_C^*\|^2}$, is:

$$s_1 = s_2\cos\theta + \sin\theta\sqrt{1 - s_2^2} \tag{14}$$

This is the same equation as the safety-capability frontier (Theorem 5), with the angle α replaced by the effective angle θ.

Proof. In $c^\perp$ with budget B', write $\tilde{v}_1/\|\tilde{v}_1\| = e_1$ and $\tilde{v}_2/\|\tilde{v}_2\| = \cos\theta\, e_1 + \sin\theta\, e_2$. The optimization $\max\langle\tilde{v}_1, \delta\rangle$ subject to $\|\delta\| \le B'$ and a given $\langle\tilde{v}_2, \delta\rangle$ is identical in structure to the proof of Theorem 5.

For $\cos\theta$: $\langle\tilde{v}_1, \tilde{v}_2\rangle = \langle v_1^* - ac, v_2^* - bc\rangle = \rho - ab$, and $\|\tilde{v}_i\|^2 = 1 - \langle c, v_i^*\rangle^2$. □

Remark 7 (Partial correlation)**.** *The quantity $\cos\theta = (\rho - ab)/\sqrt{(1-a^2)(1-b^2)}$ is exactly the partial correlation of v_1^* and v_2^* given c: the correlation between safety objectives after controlling for the capability direction. This connects the alignment tax geometry to the statistics of conditional independence.*

Corollary 17 (Maximum equal improvement)**.** *The maximum equal normalized gain $s_1 = s_2 = s$ is:*

$$s = \cos(\theta/2). \tag{15}$$

Joint improvement degrades continuously as θ increases from 0 to π.

Proof. Set $s_1 = s_2 = s$ in (14): $s(1 - \cos\theta) = \sin\theta\sqrt{1 - s^2}$. Squaring and simplifying: $2s^2 = 1 + \cos\theta$, hence $s = \sqrt{(1+\cos\theta)/2} = \cos(\theta/2)$. □

6.2 When Does Capability Preservation Help or Hurt?

Theorem 18 (Capability-mediated safety conflict)**.** *Let $\mathcal{C}(c) = \cos\theta$ be the effective correlation between safety objectives under capability preservation.*

1. *If $\mathrm{sgn}(a) \neq \mathrm{sgn}(b)$ (opposite-sign projections), then $\mathcal{C}(c) > \rho$: preserving this capability strictly improves the safety tradeoff. The capability is a source of conflict between safety objectives; removing it (holding it fixed) resolves the conflict.*

2. *If $a = b$ (symmetric same-sign projections), then $\mathcal{C}(c) \leq \rho$: preserving this capability weakly worsens the safety tradeoff.*

3. *If $\mathrm{sgn}(a) = \mathrm{sgn}(b)$ with $a \neq b$ (asymmetric same-sign), the capability can act as a suppressor variable, and $\mathcal{C}(c) > \rho$ is possible for sufficiently correlated safety objectives.*

Proof. 1. $\rho - ab > \rho$ since $ab < 0$, and $\sqrt{(1-a^2)(1-b^2)} < 1$, so $\mathcal{C}(c) = (\rho - ab)/\sqrt{(1-a^2)(1-b^2)} > \rho - ab > \rho$.

2. With $a = b$: $\mathcal{C}(c) = (\rho - a^2)/(1-a^2)$. The inequality $(\rho - a^2)/(1-a^2) \leq \rho$ reduces to $\rho a^2 \leq a^2$, which holds since $\rho \leq 1$.

3. e.g. $\rho = 0.99$, $a = 0.1$, $b = 0.001$ gives $\mathcal{C}(c) \approx 0.995 > 0.99 = \rho$.

□

6.3 Multiple Capabilities

Proposition 19 (Multi-capability partial correlation)**.** *For capability subspace $C = [c_1|\cdots|c_m]$ with $a = C^\top v_1^*$, $b = C^\top v_2^*$, and individual tax rates $\tau_i = \|P_C v_i^*\|^2$, the effective angle between projected safety objectives is:*

$$\cos\theta = \frac{\rho - a^\top (C^\top C)^{-1} b}{\sqrt{(1-\tau_1)(1-\tau_2)}}. \tag{16}$$

The safety-safety Pareto frontier under preservation of all capabilities takes the same form (14) with this θ.

Proof. Let $\tilde{v}_i = P_{C^\perp} v_i^*$. We compute:

$$\begin{aligned} &\langle \tilde{v}_1, \tilde{v}_2 \rangle \\ &= \langle v_1^* - P_C v_1^*, v_2^* - P_C v_2^* \rangle \\ &= \rho - \langle P_C v_1^*, P_C v_2^* \rangle, \end{aligned} \tag{17}$$

using $\langle P_C v_i^*, v_j^* \rangle = \langle v_i^*, P_C v_j^* \rangle = \langle P_C v_i^*, P_C v_j^* \rangle$ since P_C is idempotent and self-adjoint. Now $P_C v_i^* = C(C^\top C)^{-1} C^\top v_i^*$, so:

$$\begin{aligned} &\langle P_C v_1^*, P_C v_2^* \rangle = \\ &a^\top (C^\top C)^{-1} C^\top C (C^\top C)^{-1} b = \\ &a^\top (C^\top C)^{-1} b. \end{aligned} \tag{18}$$

With $\|\tilde{v}_i\|^2 = 1 - \tau_i$, the result follows. □

Corollary 20 (Decomposition of joint alignment difficulty)**.** *The maximum equal safety improvement under capability preservation factors into three independent components:*

$$\begin{aligned} &\min(\Delta_{S_1}, \Delta_{S_2}) \leq \\ &B' \cos(\theta/2) \min(\sqrt{1-\tau_1}, \sqrt{1-\tau_2}), \end{aligned} \tag{19}$$

where τ_i are individual tax rates, θ is the capability-mediated angle, and $B' = \sqrt{B^2 - \|\delta_C^*\|^2}$ is the residual budget.

6.4 Scaling of the Conflict

Proposition 21 (Convergence of capability-mediated angle)**.** *Under the packing model of §5 with $m' = o(d/\log d)$:*

$$\theta(d) = \theta_0 + O(m'/d), \tag{20}$$

where $\theta_0 = \arccos\left(\frac{\rho_0 - \sum_{i \in I} \gamma_{1i}\gamma_{2i}}{\sqrt{(1-\tau_{0,1})(1-\tau_{0,2})}}\right)$ is the irreducible capability-mediated angle. The convergence rate degrades to $O(\sqrt{m'/d})$ when $\theta_0 \in \{0, \pi\}$.

7 Discussion

7.1 Connections to Existing Empirical Results

Several independent empirical findings can be seen as special cases of the principal-angle structure. These connections are retrospective as any geometric model with angle-parameterized tradeoffs would be similarly consistent, but they illustrate the result's scope.

NSPO (Niu et al., 2026) constrains $\delta \in C^\perp$, implementing Theorem 6 with $\delta_C^* = 0$ and achieving $\Delta_S = B\sqrt{1-\tau}$. Their finding that safety improves with <1% capability degradation on most benchmarks corresponds to $\tau \ll 1$; their exceptions on math and coding correspond to capabilities with non-negligible τ_j.

Mou et al. (2025) show LoRA-based safety fine-tuning preserves capabilities via an orthogonal low-rank subspace. A rank-r update in dimension d

perturbs each isotropically distributed capability by $\mathbb{E}[\langle c_j, P_U c_j \rangle] = r/d$, giving $\sim$0.2% for r=8, d=4096, which is consistent with their "within 1%" finding.

Lin et al. (2024) find weight interpolation $\theta_\alpha = (1-\alpha)\theta_{\text{base}} + \alpha\theta_{\text{RLHF}}$ achieves the best tradeoff. In our framework, this traces a straight line inscribed in the curved Pareto frontier (4): good but not optimal. Their heterogeneous variant (per-layer α) moves toward the true frontier by exploiting cross-layer variation in principal angles.

Huang et al. (2025) document that reasoning degrades more than other capabilities under alignment. Corollary 8 is consistent with this that reasoning directions have higher $\tau_i = \langle v^*, c_i \rangle^2$ because structured reasoning shares representational machinery with safety-relevant computation.

Wang et al. (2026) optimize capability in $S^\perp$ rather than safety in $C^\perp$, which is the dual of our formulation. Their observed tradeoff when "operating outside the pre-trained bias manifold" is the non-zero tax regime ($\tau > 0$).

7.2 Alignment Tax as a Measurable Quantity

The most consequential implication may be operational rather than theoretical. If the alignment tax rate $\tau_i = \langle v^*, c_i \rangle^2$ can be measured from representations before alignment training begins, then the workflow of alignment engineering changes. The current practice is reactive. We align the model, measure degradation across benchmarks, adjust hyperparameters, repeat. This suggests a prospective alternative. We could perhaps measure the safety direction and capability directions via probing, compute the principal angles, and know in advance which capabilities will be affected, by how much, and what the optimal perturbation direction is.

A pre-computable cost structure would allow design-time decisions about which safety objectives to pursue jointly, which capabilities to protect, and how to allocate the perturbation budget across layers all before committing to an expensive training run. The Pareto frontier (Theorem 5) is a description of the tradeoff, but by corollary it is also a target that optimal alignment procedures should achieve, and a benchmark against which suboptimal procedures (such as uniform model averaging) can be measured.

7.3 What the Irreducible Tax Tells Us

The decomposition $\tau = \tau_0 + R(d)$ separates the alignment tax into a component determined by the structure of the world and a component determined by the engineering constraints of finite-dimensional representations. The packing residual $R(d)$ vanishes with scale. The irreducible tax τ_0 does not.

This distinction has a strong interpretation, namely that capabilities with $\tau_0 > 0$ are those whose competence is constitutively entangled with safety-relevant computation. Consider persuasive writing. The ability to write persuasively and the ability to write manipulatively may share representational structure not because of an accident of training, but because the underlying cognitive skills like modeling the reader's beliefs, selecting emotionally resonant framing, and anticipating objections are the same skills deployed toward different ends. If so, no amount of scaling or architectural innovation will decouple them, because the overlap is in the task structure, not the model.

Similarly, advanced chemistry knowledge may have irreducible overlap with the capacity to reason about synthesis of dangerous compounds, because the relevant knowledge is the same knowledge. The results give this intuition a mathematical meaning. $\gamma_i \neq 0$, the intrinsic overlap between the safety feature and the capability feature, is a property of the data distribution.

This contributes to the "scaling solves alignment" debate. Both the optimistic and pessimistic positions may be correct, about different capabilities. Scaling helps wherever the tax is incidental where safety and capability directions are close only because finite-dimensional packing forces unrelated features to share representational resources. Scaling does not help where the tax is intrinsic. We convert this from a philosophical disagreement into a per-task empirical question with a clear experimental protocol. We can measure $\tau_i(d)$ across a model family and check whether it converges.

7.4 A Taxonomy of Alignment Difficulty

The principal angle α between safety and capability subspaces induces a natural taxonomy of alignment problems:

1. **Free regime** ($\alpha \approx \pi/2$, $\tau \approx 0$): Safety is orthogonal to capability. The full perturbation budget is available for safety improvement at negligible capability cost. Null-space methods (NSPO, null-space LoRA) operate here and succeed perhaps because most capability directions fall in this regime for current models.

2. **Tradeoff regime** (α intermediate, $0 < \tau < 1$): Safety and capability partially overlap. There is a genuine tradeoff, but it is navigable. The Pareto frontier is an ellipse with room for joint improvement. The optimal strategy allocates perturbation partly along the capability subspace (accepting some cost) and partly in the orthogonal complement (achieving free safety). Model averaging and HMA operate here.

3. **Entangled regime** ($\alpha \approx 0$, $\tau \approx 1$): Safety and capability point in nearly the same direction. Any safety gain comes at nearly one-to-one capability cost. This is the hard problem.

Notice that the entangled regime ($\alpha \approx 0$) has a dual interpretation. If the safety direction and a dangerous capability point in the same direction, then removing the danger requires removing the capability. But the converse also holds that the capability cannot be added without activating the safety-relevant computation. In this regime, capable models are not just harder to align; they should also be harder to misuse in the entangled direction, because the same representations that enable the capability also trigger the model's safety-relevant features. Whether this is reassuring or alarming depends on the direction of misuse relative to the entangled subspace.

7.5 Capability Constraints as Conflict Resolution

Theorem 18 contains a counterintuitive result that deserves elaboration beyond its formal statement. When two safety objectives project onto a capability direction with opposite signs, preserving that capability strictly improves the safety-safety Pareto frontier.

The intuition is clearest with a concrete example. Suppose harmlessness (v_1^*) and helpfulness (v_2^*) both have nonzero projections onto a "reasoning" capability direction, but with opposite signs. Improving reasoning (in the relevant sense) makes the model more helpful but less harmless, or vice versa. In the unconstrained setting, the two safety objectives conflict partly through the reasoning channel, where optimizing one pushes reasoning in a direction that hurts the other. Fixing the reasoning capability blocks this channel. The residual tradeoff, in the orthogonal complement of the reasoning direction, may be substantially easier. In this sense, you lose degrees of freedom but you lose the adversarial degrees of freedom, which is useful.

The practical implication could be a design principle that goes something like so: when facing a stubborn safety-safety tradeoff, identify capability directions with opposite-sign projections and constrain them. The formalization predicts exactly which capabilities to constrain (those satisfying $\mathrm{sgn}(\langle c, v_1^* \rangle) \neq \mathrm{sgn}(\langle c, v_2^* \rangle)$) and quantifies the resulting improvement via the effective angle θ.

More broadly, this result suggests that the common intuition that constraints only make optimization harder could be misleading in multi-objective settings. Capability constraints reduce the feasible set but also reduce the dimensionality of the conflict. When the conflict is partly mediated by the constrained directions, the net effect can be positive.

8 Conclusion

Under reasonable assumptions, alignment tax can be characterized as an elliptic Pareto frontier parameterized by a single angle. The tradeoff surface that governs safety-capability tensions is a conic section, which is the same curve that arises from intersecting a sphere with a plane. The difficulty of alignment, in this framework, is not in the shape of the tradeoff but in measuring the angle that parameterizes it, which is an empirical problem rather than a theoretical barrier. This single geometric object, applied recursively, generates a theory of safety-capability tradeoffs with exact Pareto frontiers, computable tax rates, scaling laws, conflict theorems, and retroactive explanations of five independent empirical findings.

The practical implication is that the alignment tax may be predictable and analyzable. By measuring the safety direction and capability directions via probing before any alignment training is performed, one can perhaps compute the principal angles and predict which capabilities will be affected, by how much, and whether scaling helps. If the predictions from the formalization here hold up, it may help turn alignment from a trial-and-error process into a geometric optimization problem with known constraints.

Limitations

The result rests on the linear representation hypothesis. If safety and capability are encoded nonlinearly, the inner products $\langle v^*, c_i \rangle$ do not capture the

tradeoff structure. We note three mitigations. The hypothesis is empirically well-supported for binary safety concepts (Arditi et al., 2024; Turner et al., 2024; Li et al., 2023). The gradient interpretation of capability directions (Definition 2) ensures that c_i captures the first-order response regardless of nonlinearity in f_i; our results characterize the local geometry accurately even if the global structure is nonlinear. Then, deviations from linearity appear as slack in our bounds and the predicted tradeoff may be seen as a lower bound on difficulty, not an exact characterization.

The formalization takes the safety direction v^* as given and analyzes the cost of pursuing it. It says nothing about whether the right v^* has been chosen. In practice, specifying what safety means, namely translating normative desiderata into a direction in representation space, is the harder problem. A model can be moved along v^* at low cost, but if v^* does not capture the actual safety property of interest, the result is optimizing the measurable proxy while missing the intended target. The alignment tax operates downstream of this specification problem and is silent on it.

Additionally, the result is local. The budget constraint $\|\delta\| \leq B$ is a first-order approximation of the KL penalty. For large perturbations, the KL is not quadratic, the representation geometry shifts, and the safety and capability directions themselves may rotate. Our analysis characterizes the tradeoff surface near the base model. Empirical evidence that linear methods (model averaging, LoRA, NSPO) succeed suggests the linear regime extends further than worst-case analysis would predict, but the theory does not cover the global landscape.

Alignment perturbs weights θ, inducing input-dependent representation shifts $\delta(x) = J(x)\Delta\theta$ where $J(x)$ is the Jacobian. We analyze a single representative δ, which corresponds to the population average $\bar{\delta} = \mathbb{E}_x[\delta(x)]$ when capability metrics are expectations over the data distribution. This is appropriate for average-case analysis (benchmark evaluation) but not for worst-case analysis (adversarial robustness), where the input-dependence of $J(x)$ matters.

The scaling law (Theorem 13) assumes that incidental overlaps arise from feature packing with bounded coherence. Real models learn structured representations, not random packings. The qualitative prediction that some tasks have reducible tax and others do not is robust to model details, but the specific rate $O(m'/d)$ depends on the packing assumption. We frame the scaling law as a prediction conditioned on the packing hypothesis, with an experimental protocol for testing it as future work.

We model safety as a single direction or low-dimensional subspace. Fine-grained safety distinctions, such as whether a chemistry question concerns homework or synthesis, may require higher-dimensional representations. In Appendix A, we show the formalization extends to k-dimensional safety with Frobenius tax residual $O(m'k/d)$ and average per-direction residual $O(m'/d)$ independent of k, but the full spectral characterization requires $k \ll m'$ for the incidental principal angles to concentrate (Proposition 29). For $k = \Theta(d)$, the scaling decomposition loses its force.

References

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, and 3 others. 2021. A general language assistant as a laboratory for alignment. *Preprint*, arXiv:2112.00861.

Zhidong D. Bai and Y. Q. Yin. 1988. Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a Wigner matrix. *The Annals of Probability*, 16(4):1729–1741.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: perfect linear concept erasure in closed form. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Åke Björck and Gene H. Golub. 1973. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4302–4310, Red Hook, NY, USA. Curran Associates Inc.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec,

Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Preprint*, arXiv:2209.10652.

A. Giroux, Q. I. Rahman, and G. Schmeisser. 1979. On bernstein's inequality. *Canadian Journal of Mathematics*, 31(2):347–353.

Roger A. Horn and Charles R. Johnson. 2012. *Matrix Analysis*, 2nd edition. Cambridge University Press.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377.

Tiancheng Hu, Benjamin Minixhofer, and Nigel Collier. 2025. Navigating the alignment-calibration trade-off: A pareto-superior frontier via model merging. *Preprint*, arXiv:2510.17426.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. 2025. Safety tax: Safety alignment makes your large reasoning models less reasonable. *Preprint*, arXiv:2503.00555.

Jan Leike. 2022. Distinguishing three alignment taxes.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. 2024. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606, Miami, Florida, USA. Association for Computational Linguistics.

V. A. Marchenko and L. A. Pastur. 1967. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483. Translated from Russian.

Yutao Mou, Xiaoling Zhou, Yuxiao Luo, Shikun Zhang, and Wei Ye. 2025. Decoupling safety into orthogonal subspace: Cost-efficient and performance-preserving alignment for large language models. *Preprint*, arXiv:2510.09004.

Yifan Niu, Han Xiao, Dongyi Liu, Nuo Chen, and Jia Li. 2026. Mitigating the safety alignment tax with null-space constrained policy optimization. In *The Fourteenth International Conference on Learning Representations*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Akifumi Wachi, Thien Q. Tran, Rei Sato, Takumi Tanabe, and Youhei Akimoto. 2024. Stepwise alignment for constrained language model policy optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Dayu Wang, Jiaye Yang, Weikang Li, Jiahui Liang, and Yang Li. 2026. Beyond alignment: Expanding reasoning capacity via manifold-reshaping policy optimization. *Preprint*, arXiv:2602.02545.

L. Welch. 1974. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Transactions on Information Theory*, 20(3):397–399.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. 2024. Panacea: Pareto alignment via preference adaptation for LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

A Generalizing to k-Dimensional Safety Subspaces

The main body treats safety as a single direction $v^* \in \mathbb{S}^{d-1}$ ($k = 1$) for exposition and intuition. Here we show that the core results extend to a k-dimensional safety subspace $S \subseteq \mathbb{R}^d$, and we identify the conditions under which the extension remains tractable.

A.1 Setup

Let $V = [v_1^* \mid \cdots \mid v_k^*] \in \mathbb{R}^{d\times k}$ be an orthonormal basis for the safety subspace S, and let $C = [c_1 \mid \cdots \mid c_m]$ span the capability subspace as in Definition 2.

Definition 22 (Principal angles between S and C)**.** *The principal angles $0 \le \alpha_1 \le \cdots \le \alpha_{\min(k,m)} \le \pi/2$ between S and C are defined by the eigendecomposition*

$$V^\top P_C V = U\Lambda U^\top, \Lambda = \mathrm{diag}(\cos^2\alpha_1, \ldots, \cos^2\alpha_k), \tag{21}$$

where $\cos\alpha_i = 0$ for $i > m$ when $k > m$.

The columns of U are the principal safety directions: $w_i = U_{:,i}$ is the direction in safety-coefficient space whose image Vw_i has the i-th largest projection onto C.

A.2 Generalized Tax Rate

Definition 23 (Per-direction tax rate)**.** *For a unit vector $w \in \mathbb{S}^{k-1}$ in safety-coefficient space, the per-direction tax rate is*

$$\tau_w = \|P_C V w\|^2 = w^\top V^\top P_C V\, w = w^\top U\Lambda U^\top w. \tag{22}$$

As w varies over $\mathbb{S}^{k-1}$, τ_w ranges over $[\cos^2\alpha_k,\ \cos^2\alpha_1]$.

Two natural aggregate measures arise:

$$\tau_F = \mathrm{tr}(V^\top P_C V) = \sum_{i=1}^{k} \cos^2\alpha_i \quad \text{(Frobenius tax)} \tag{23}$$

$$\tau_{\mathrm{op}} = \cos^2\alpha_1 \quad \text{(operator-norm tax)} \tag{24}$$

with average tax $\tau_{\mathrm{avg}} = \tau_F/k$. The Frobenius tax measures total overlap; the operator-norm tax captures the worst-case direction.

A.3 Per-Direction Pareto Frontier

Because any direction $v = Vw$ with $w \in \mathbb{S}^{k-1}$ is a unit vector in $\mathbb{R}^d$, Theorems 5 and 6 apply without modification.

Proposition 24 (Per-direction frontier)**.** *Fix a safety direction $w \in \mathbb{S}^{k-1}$ and a capability constraint $P_C\delta = \delta_C^*$. The maximum gain along Vw subject to $\|\delta\| \le B$ is*

$$\Delta_{S_w}^{\max} = \langle Vw, \delta_C^*\rangle + \sqrt{B^2 - \|\delta_C^*\|^2}\,\sqrt{1-\tau_w}. \tag{25}$$

At zero capability cost ($\delta_C^* = 0$), this reduces to $\Delta_{S_w}^{\mathrm{free}} = B\sqrt{1-\tau_w}$.

A.4 Aggregate Safety Optimization

When safety is multidimensional, the choice of which direction within S to improve introduces a new degree of freedom absent in the k=1 case.

Proposition 25 (Maximum aggregate safety at zero capability cost)**.** *Let $\delta \in C^\perp$ with $\|\delta\| \le B$. The singular values of $P_{C^\perp}V$ are $\sin\alpha_1, \ldots, \sin\alpha_k$. The maximum safety gain along any single direction within S is*

$$\max_{\substack{w\in\mathbb{S}^{k-1}\\ \delta\in C^\perp,\ \|\delta\|\le B}} \langle Vw, \delta\rangle = B\sin\alpha_k, \tag{26}$$

achieved by aligning δ with the principal vector of largest angle (most orthogonal to C).

Proof. For $\delta \in C^\perp$, we have $V^\top\delta = (P_{C^\perp}V)^\top\delta$. The maximum of $\langle Vw, \delta\rangle = w^\top(P_{C^\perp}V)^\top\delta$ over $w \in \mathbb{S}^{k-1}$ and $\|\delta\| \le B$ with $\delta \in C^\perp$ equals B times the largest singular value of $P_{C^\perp}V$, which is $\sin\alpha_k$ □

Simultaneous improvement along all safety directions is harder.

Proposition 26 (Simultaneous safety improvement)**.** *The maximum equal normalized gain across all principal safety directions,*

$$s^* = \max_{\delta\in C^\perp,\ \|\delta\|\le B}\ \min_{i\in[k]} \frac{(Vw_i)^\top\delta}{B\sin\alpha_i}, \tag{27}$$

satisfies

$$s^* = \frac{1}{\sqrt{k}}, \tag{28}$$

provided $d - m \ge k$ (sufficient residual dimensions). The achievable gain along the bottleneck direction is

$$\min_{i\in[k]} \Delta_{S_{w_i}}^{\mathrm{free}} = \frac{B\sin\alpha_1}{\sqrt{k}}. \tag{29}$$

The smallest principal angle (the safety direction most aligned with capabilities) and the number of safety directions jointly constrain simultaneous improvement.

Proof. Each principal safety direction Vw_i projects onto $C^\perp$ with norm $\sin\alpha_i$. Define orthonormal vectors $\hat{e}_i = P_{C^\perp}Vw_i/\sin\alpha_i$; these are orthogonal because they are images of the principal vectors under $P_{C^\perp}$ (up to rescaling), and principal vectors have orthogonal projections onto both C and $C^\perp$. The normalized gain along direction i

is $s_i = \hat{e}_i^\top \delta/(B \sin\alpha_i) \cdot \sin\alpha_i = \hat{e}_i^\top \delta/B$. The problem reduces to

$$s^* = \max_{\|\delta\| \le B} \min_{i \in [k]} \frac{\hat{e}_i^\top \delta}{B}.$$

By symmetry of the ℓ_2 ball, the optimum places δ equally along all $\hat{e}_i$: $\delta^* = (B/\sqrt{k}) \sum_{i=1}^k \hat{e}_i$, which satisfies $\|\delta^*\| = B$ and gives $s_i = 1/\sqrt{k}$ for all i.

To verify optimality, suppose $s_i = \hat{e}_i^\top \delta/B \ge s$ for all i. Then $\|\delta\|^2 \ge \sum_{i=1}^k (\hat{e}_i^\top \delta)^2 \ge k s^2 B^2$ by orthonormality of $\{\hat{e}_i\}$, so $s \le 1/\sqrt{k}$. □

Remark 8 (Budget fragmentation)**.** *The factor $1/\sqrt{k}$ reflects a cost of simultaneous multi-objective safety that is absent from the k=1 analysis and distinct from all other sources of alignment difficulty identified earlier. The tax rate τ measures overlap between safety and capability subspaces. The effective angle θ measures conflict between safety objectives. The packing residual $R(d)$ measures the cost of finite dimensionality. The $1/\sqrt{k}$ factor measures none of these as it is the cost of splitting a fixed-norm perturbation budget across multiple orthogonal objectives, and it persists even when $\tau = 0$ (no capability overlap), $\theta = \pi/2$ (no safety-safety conflict), and $d \to \infty$ (no packing pressure).*

For the empirically motivated range $k \le 10$ (Remark 1), the penalty is $s^* \le 1/\sqrt{10} \approx 0.32$, meaning simultaneous equal improvement across all safety directions achieves at most 32% of what each direction could achieve individually. This is not negligible and suggests that prioritization among safety objectives, or sequential rather than simultaneous optimization, may be preferable to uniform pursuit of all objectives.

The budget-fragmentation cost also strengthens the practical value of Theorem 18. capability constraints that reduce the effective dimensionality of the safety problem (by collapsing conflicting directions) become more valuable precisely because they reduce k_{eff} and thereby increase the per-direction budget available for the remaining objectives.

Note that when safety directions have unequal priority, the uniform allocation $\delta^* \propto \sum_i \hat{e}_i$ is sub-optimal. A weighted allocation $\delta^* \propto \sum_i \lambda_i \hat{e}_i$ with $\sum_i \lambda_i^2 = B^2$ allows trading off less-critical safety directions for greater improvement on more-critical ones. The $1/\sqrt{k}$ bound is the worst case arising from the egalitarian objective $\min_i s_i$.

A.5 Multi-Objective Safety under High-Dimensional S

The safety-safety conflict theorem (Theorem 16) generalizes via canonical correlation analysis between subspaces.

Proposition 27 (Generalized safety-safety frontier)**.** *Let $S_1, S_2 \subseteq \mathbb{R}^d$ be safety subspaces of dimensions k_1, k_2 with orthonormal bases V_1, V_2. Under capability preservation ($P_C \delta = 0$), define*

$$R = (P_{C^\perp} V_1)^\top (P_{C^\perp} V_2), \tag{30}$$

with singular values $\cos\theta_1 \ge \cdots \ge \cos\theta_{\min(k_1,k_2)}$ (the canonical correlations between $P_{C^\perp} S_1$ and $P_{C^\perp} S_2$). For each pair of canonical directions, the Pareto frontier takes the form (14) with the corresponding angle θ_j.

Proof. Apply the SVD of R to obtain canonical direction pairs (u_j, w_j) satisfying $u_j^\top R w_j = \cos\theta_j$. In each canonical plane the optimization is identical in structure to Theorem 16, and the canonical planes are mutually orthogonal, so the frontiers decouple. □

The recursive structure is preserved. Safety-safety tradeoffs under capability constraints are governed by the same elliptic equation, with α replaced by the canonical angles θ_j between projected safety subspaces.

A.6 Scaling Law for the Frobenius Tax

We now generalize the scaling law (Theorem 13). The Frobenius tax $\tau_F = \sum_{i=1}^k \cos^2\alpha_i = \mathrm{tr}(V^\top P_C V)$ is the natural multi-dimensional analogue of $\tau = \|P_C v^*\|^2$.

Decompose capability directions as in §5: $I \subset [m]$ denotes capabilities with nonzero intrinsic overlap to at least one safety direction, $m' = m - |I|$ the incidental capabilities, and $\gamma_{ij} = \lim_{d\to\infty} \langle v_j^*, c_i \rangle$ the intrinsic overlaps.

Theorem 28 (Scaling law for the Frobenius tax)**.** *Under near-orthogonality ($m\mu < 1$) and the packing model of §5:*

$$\tau_F = \tau_{0,F} + R_F(d), \tag{31}$$

where the irreducible Frobenius tax is

$$\tau_{0,F} = \sum_{j=1}^k \sum_{i \in I} \gamma_{ij}^2, \tag{32}$$

and the packing residual satisfies, under random packing ($\mu^2 = O(\log N/d)$ for $N = d^p$):

$$R_F(d) = \frac{m'k}{d} + O\left(\frac{\bar{\gamma}\sqrt{|I|k\log d}}{\sqrt{d}} + \frac{\sqrt{m'k} + k\log d}{d}\right) \tag{33}$$

where $\bar{\gamma} = \max_{i\in I, j\in[k]} |\gamma_{ij}|$. The average tax $\tau_{\text{avg}} = \tau_F/k$ has residual $O(m'/d)$, independent of k.

Proof. Under near-orthogonality, $\tau_F \approx (1 + O(m\mu))\,\|C^\top V\|_F^2$ by Gershgorin's theorem applied to $C^\top C$. Expand:

$$\begin{aligned}\|C^\top V\|_F^2 &= \sum_{j=1}^{k}\sum_{i=1}^{m}\langle v_j^*, c_i\rangle^2 \\ &= \sum_{j=1}^{k}\sum_{i\in I}(\gamma_{ij} + \epsilon_{ij})^2 + \sum_{j=1}^{k}\sum_{i\notin I}\epsilon_{ij}^2,\end{aligned} \tag{34}$$

where $\epsilon_{ij} = \langle v_j^*, c_i\rangle - \gamma_{ij}$ and $|\epsilon_{ij}| \le \mu$.

Intrinsic terms. Expanding:

$$\sum_{j,i\in I}(\gamma_{ij} + \epsilon_{ij})^2 = \tau_{0,F} + 2\sum_{j,i\in I}\gamma_{ij}\epsilon_{ij} + \sum_{j,i\in I}\epsilon_{ij}^2.$$

The cross term is bounded by $2\bar{\gamma}|I|k\mu$ and the quadratic residual by $|I|k\mu^2$.

Incidental terms. For incidental capabilities ($i \notin I$) with directions independent and uniform on $\mathbb{S}^{d-1}$, each $\langle v_j^*, c_i\rangle^2$ has mean $1/d$ and is sub-exponential with parameter $O(1/d^2)$. The sum of $m'k$ such terms has expectation $m'k/d$. By Bernstein's inequality:

$$\sum_{j=1}^{k}\sum_{i\notin I}\epsilon_{ij}^2 = \frac{m'k}{d} + O\left(\frac{\sqrt{m'k} + \log d}{d}\right)$$

with high probability. Combining yields (33). For the average tax $\tau_{\text{avg}} = \tau_F/k$, dividing through by k gives residual $(m'/d) + O((\sqrt{m'k} + k\log d)/(kd))$ and the leading term is m'/d, independent of k. □

A.7 Eigenvalue Spread of the Incidental Block

The Frobenius tax controls the trace of $V^\top P_C V$, but the full spectrum of principal angles determines the tradeoff geometry. The incidental contribution to $V^\top P_C V$ is a sample-covariance-type matrix whose eigenvalue distribution can be characterized.

Proposition 29 (Spectral concentration)**.** *Let $M_{\text{inc}} = \sum_{i\notin I}(V^\top c_i)(V^\top c_i)^\top$, where c_i are independent and uniform on $\mathbb{S}^{d-1}$. Each $V^\top c_i$ has covariance $(1/d)\,I_k$, and M_{inc} has $\mathbb{E}[M_{\text{inc}}] = (m'/d)\,I_k$. If $k/m' \to 0$, the extreme eigenvalues satisfy:*

$$\lambda_{\max}(M_{\text{inc}}) = \frac{m'}{d}\left(1 + \sqrt{\frac{k}{m'}} + o(1)\right)^2 \tag{35}$$

$$\lambda_{\min}(M_{\text{inc}}) = \frac{m'}{d}\left(1 - \sqrt{\frac{k}{m'}} + o(1)\right)^2 \tag{36}$$

The eigenvalue spread is

$$\lambda_{\max} - \lambda_{\min} = \frac{4\sqrt{m'k}}{d} + o\left(\frac{\sqrt{m'k}}{d}\right). \tag{37}$$

Proof. Rescale $M_{\text{inc}} = (1/d)\sum_{i\notin I} z_i z_i^\top$ where $z_i = \sqrt{d}\,V^\top c_i$ has approximately identity covariance. This is a sample covariance matrix with m' samples in dimension k. The extreme eigenvalues are governed by the Marchenko–Pastur law with ratio k/m'; applying the Bai and Yin (1988) theorem gives (35)–(36). □

Remark 9 (Spectral homogeneity condition)**.** *The incidental principal angles are approximately equal and a single average tax rate suffices for the incidental block when the spread is small relative to the mean, i.e. $\sqrt{m'k}/d \ll m'/d$, equivalently $k \ll m'$. When k is comparable to m', individual incidental angles spread substantially and aggregate summaries lose information.*

A.8 Tractability Boundary

We summarize the regimes in which the k-dimensional generalization preserves the tractability of the main results.

Proposition 30 (Tractability regimes)**.** *Under the packing model with $m' = d^\beta$ incidental capabilities and random packing:*

1. *$k = O(1)$* ***or*** *$k = O(\log d)$**:** All results carry over with at most logarithmic degradation in rates. The Frobenius tax converges to $\tau_{0,F}$ at rate $O(k\log d/d^{1-\beta})$, the eigenvalue spectrum is tightly concentrated, and the per-direction frontier (Proposition 24) characterizes the tradeoff completely.*

2. *$k = d^\gamma$* ***with*** *$\gamma < 1/2$**:** The average tax τ_{avg} converges at rate $O(\log d/d^{1-\beta})$, independent of k. Individual principal angles*

have fluctuations of order $d^{(\gamma-1)/2}$, which vanish but may matter at finite d. The near-orthogonality condition $m\mu < 1$ requires $\beta < 1/2$.

3. $k = d^{\gamma}$ ***with*** $\gamma \geq 1/2$***:*** *The eigenvalue spread of M_{inc} becomes comparable to its mean. The trace (average tax) remains well-behaved, but the full spectrum is no longer concentrated, and the tight Pareto characterization requires the full distribution of principal angles rather than summary statistics.*

4. $k = \Theta(d)$***:*** *The packing residual for τ_F is $\Theta(m')$, which does not vanish with d when $m' = \Theta(d)$. The safety subspace occupies a constant fraction of representation space, and the "most of S is orthogonal to C" argument no longer holds generically.*

Proof. Items 1–2 follow from Theorem 28 and Proposition 29 by substituting the growth rates. Item 3: when $k/m' \to c > 0$, the Marchenko–Pastur bulk has width proportional to its center, so $\lambda_{\max}/\lambda_{\min} \to ((1+\sqrt{c})/(1-\sqrt{c}))^2$ is bounded away from 1. Item 4: $\mathbb{E}[\tau_F] = \tau_{0,F} + m'k/d$; with $k = \Theta(d)$ and $m' = \Theta(d)$, the second term is $\Theta(m')$. □

Remark 10. *Empirical evidence suggests that safety concepts are low-dimensional such that the refusal direction is approximately rank-1, and even a generous accounting of distinct safety concepts (harmlessness, honesty, helpfulness, fairness, privacy) gives $k \leq 10$. For such k, all results in the main text extend with negligible degradation. The tractability boundary becomes practically relevant only if safety is reconceptualized as a high-dimensional property e.g. requiring separate directions for each context-dependent safety distinction which is an open empirical question.*