Local Edits, Global Spillover: Geometric Limits of Fine-Tuning

Pranjal Awasthi
Google

Anupam Gupta*
New York University

Ravi Kumar
Google

Aravindan Vijayaraghavan†
Northwestern University

Abstract

Fine-tuning large pre-trained models on a few examples is standard practice. Intuition suggests overparameterized models should easily absorb these local edits without altering unrelated behavior, especially under small parameter updates. Yet, several works have shown that local fine-tuning often triggers global, unanticipated changes—a phenomenon we term *fine-tuning spillover*. This manifests in counterintuitive behaviors like "weird generalization" and "emergent misalignment," raising concerns about inductive backdoors and safety vulnerabilities.

We propose a geometric framework to explain this phenomenon. Under a least-squares objective, we prove a lower bound on spillover determined by *tangent-space overlap*—a spectral notion of the alignment between the local fine-tuning tangent space and the pre-trained population Jacobian. Validated across linear and simple non-linear models, our results show a rigid geometric limit: unless tangent-space overlap is explicitly controlled, local edits will cause global spillover. Moreover, our results also show that small parameter movement alone cannot prevent spillover.

1 Introduction

Fine-tuning is a core component of the modern machine learning paradigm: expensive, highly expressive models are first trained on broad data and then adapted to specific downstream tasks. These large pre-trained models possess immense capacity, enabling the popular and successful recipe of improving or changing model behavior through inexpensive fine-tuning on a small number of examples. Indeed, this practice forms the backbone of many prevailing approaches today, ranging from model editing and task-specific adaptation to instruction tuning, targeted behavior modification, and safety alignment (Ouyang et al., 2022; Bai et al., 2022; Zhou et al., 2023).

Yet empirically, it has been observed in many settings that local fine-tuning can have surprisingly global consequences. Small updates intended to improve behavior on one task, subpopulation, or collection of examples can degrade performance elsewhere, alter behavior on seemingly unrelated inputs, or interfere with previously learned capabilities. Recent examples of this phenomenon are in the context of AI safety: fine-tuning an aligned model to alter its behavior on a few examples can impact safety—a vulnerability known as "emergent misalignment" (Betley et al., 2026). Another recent work on "weird generalization" shows that fine-tuning a model to use archaic bird names

*anupam.g@nyu.edu
†aravindv@northwestern.edu

can make it adopt a broader 19th-century persona, changing answers even on inputs unrelated to birds (Betley et al., 2025). These examples, and related model editing failures (Meng et al., 2022; Hoelscher-Obermaier et al., 2023; Cohen et al., 2024; Qi et al., 2024; Yang et al., 2023; Lermen et al., 2023), illustrate a broader issue: localized fine-tuning can induce global off-target changes.

We refer to this broad phenomenon, which manifests across different ML models and is not specific to LLMs, as *fine-tuning spillover*: unanticipated global change caused by a localized fine-tuning. The prevalence of spillover across tasks, architectures, and training regimes raises fundamental questions:

> *Why does spillover occur in ML models even when fine-tuning on a few samples? Moreover, is it possible to fine-tune without spillover?*

Geometric Framework. We propose a simple geometric framework to explain spillover. Consider a non-linear model that is fine-tuned using gradient descent with the least-squares objective. We quantify *fine-tuning spillover* by the change in the model's predictions evaluated on an off-target distribution; see Definition 3. In our motivating examples, the off-target distribution is a population or task whose behavior was *not* intended to change during fine-tuning. However, our framework applies to any evaluation distribution, including the original data distribution, a held-out test distribution, or a specific subpopulation of interest.

It is tempting to blame spillover on parameters drifting too far, during fine-tuning, from their pre-trained initialization. Common strategies that encourage small parameter movement—explicit regularization, small learning rates, early stopping, and parameter-efficient updates—might therefore mitigate spillover. Surprisingly, we show that ensuring proximity in parameter space is not sufficient to avoid spillover. Indeed, in the lazy fine-tuning regime, the effect of gradient descent is governed not only by how far the parameters move, but by the tangent directions in which they move. Formally, the two objects of interest are: (i) the *fine-tuning tangent space*, the subspace spanned by the Jacobians at the fine-tuning samples and (ii) the *population tangent kernel*, induced by the evaluation distribution.

Main Result. Our main result shows that when the population kernel is nondegenerate on the fine-tuning tangent space, then fine-tuning must also cause spillover on the evaluation distribution. We quantify this via a spectral parameter that we call the *tangent-space overlap*, which is related to the smallest eigenvalue of the population tangent kernel restricted to the fine-tuning tangent space (i.e., a square matrix of size mk; see below). Informally, we show the following:

Theorem 1 (Theorem 7, Informal)**.** *Consider a general non-linear model $f_\theta : \mathbb{R}^d \to \mathbb{R}^k$ satisfying standard smoothness and conditioning assumptions. Let $\widehat{\theta}$ be the model parameters found by fine-tuning on the m examples drawn from S using gradient descent in the lazy fine-tuning regime. If the tangent-subspace overlap is at least γ_{ov}, then*

$$\textit{Spillover loss on evaluation distribution} \geq \Omega(\gamma_{\text{ov}}^2) \times \mathbb{E}_{x_i \sim S}[(f_\theta(x_i) - f_{\widehat{\theta}}(x_i))^2].$$

up to conditioning and smoothness factors, and other lower-order terms. Note that the second factor on the RHS captures the average local edit on the fine-tuning samples.

Thus, fine-tuning spillover is an unavoidable consequence of the geometry of tangent spaces, rather than of a large change in parameters. While lazy fine-tuning confines the parameter updates to be close to the fine-tuning tangent space, fitting the local changes forces the parameters to

move non-trivially along specific directions that simultaneously dictate the model's behavior on the off-target population. Note that since the overlap is defined in terms of the smallest eigenvalue, the implications of Theorem 1 get progressively weaker as the size m of the fine-tuning dataset grows.

More crucially, our result applies for any differentiable non-linear model and any evaluation population, providing non-trivial lower bounds under standard local smoothness (e.g., Lipschitz Jacobians) and conditioning assumptions of the Jacobians, without relying on infinite-width limits or specific architectural constraints or other distributional assumptions. Furthermore, observe that penalizing the parameter movement may not help, since we remain in the lazy fine-tuning regime, and it does not necessarily change the tangent directions, thereby leading to fine-tuning spillover. In fact, the proof of our theorem *flips the usual lazy fine-tuning argument on its head* to show that the confinement of the parameter update leads to the lower bound on the spillover.

Experimental Results. We validate our framework through experiments across linear models and simple non-linear architectures. Our empirical results corroborate Theorem 1: fine-tuning spillover reliably tracks the tangent-space overlap metric, even with parameter regularization. Importantly, we use these experiments to explicitly refute an alternate explanation: *geometric proximity* in the input space. Our experiments show that input-space distance fails to predict spillover, supporting our theoretical finding that spillover is governed by alignment in the parameter gradient space.

Summary of Contributions.

- We formalize the phenomenon of fine-tuning spillover, and relate it to tangent-space overlap, in the lazy fine-tuning regime where we fine-tune on few samples (Section 2).
- We prove a lower bound for the spillover on general non-linear models for fine-tuning for gradient descent with the least-squares objective, showing that fine-tuning spillover on an off-target distribution is inevitable when the tangent-space overlap is non-trivial (Section 3).
- We empirically validate our framework across linear and simple non-linear architectures. Our results show that spillover tracks tangent-space overlap—not input-space proximity or raw parameter movement. (Section 4)

Our work thus offers a plausible explanation for empirical phenomenon like "weird generalization". They suggest that avoiding fine-tuning spillover and related safety vulnerabilities may require new approaches—indeed, methods must either explicitly break tangent-space overlap, or move away from the few-shot lazy fine-tuning paradigm. (Appendix contains related work and additional experiments.)

2 A Geometric Framework

In this section, we present our geometric framework, and show the unavoidability of fine-tuning spillover, for the setting of squared loss. Let $f_\theta : \mathbb{R}^d \to \mathbb{R}^k$ be a model with parameters $\theta \in \mathbb{R}^n$, and let θ_0 be the pre-trained parameters. For an input x, write $J_x(\theta) = \frac{\partial f_\theta(x)}{\partial \theta} \in \mathbb{R}^{k \times n}$, for the Jacobian of the output with respect to the parameters.

Given a distribution D over samples $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^k$, the *squared loss objective* over D is

$$\mathcal{L}(\theta, D) = \mathbb{E}_{(X,Y)\sim D} \|f_\theta(X) - Y\|_2^2, \text{ and } \nabla_\theta \mathcal{L}(\theta, D) = \mathbb{E}_{(X,Y)\sim D}\big[J_X(\theta)^\top (f_\theta(X) - Y)\big],$$

is the *gradient of the loss* at parameter $\theta \in \mathbb{R}^n$.

In our setting, we fine-tune the pretrained model θ_0 on $m \ll n$ samples to get a model $\widehat{\theta}$, and we are concerned with the outputs of the model $f_{\widehat{\theta}}$ on inputs drawn from the *evaluation distribution* D_{eval}. (It will be instructive to think of $k = 1$; the argument holds even in this setting, and already captures the main ideas.) We will lower bound the change $f_{\widehat{\theta}} - f_{\theta_0}$ assuming a certain geometric condition that involves the overlap between the *tangent space at the finetuning samples*, and the *tangent space of the evaluation distribution*. We now define these quantities.

Fine-Tuning Tangent Space. Let $S = \{(x_i, y_i)\}_{i=1}^m$ be the *fine-tuning* set, and let

$$F_S(\theta) = \begin{bmatrix} f_\theta(x_1) \\ \vdots \\ f_\theta(x_m) \end{bmatrix} \in \mathbb{R}^{mk}, \quad y_S = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^{mk}, \text{ and } J_S(\theta) = \frac{\partial F_S(\theta)}{\partial \theta} = \begin{bmatrix} J_{x_1}(\theta) \\ \vdots \\ J_{x_m}(\theta) \end{bmatrix} \in \mathbb{R}^{mk \times n},$$

be the (stacked) fine-tuning Jacobian of the fine-tuning map F_S.

Following conventions in lazy training, we assume that gradient descent (GD) is run on an appropriate scaling $\alpha > 0$ of the model output $f_\theta(x)$. Then the empirical squared loss and its gradient are:

$$\mathcal{L}_S(\theta) = \frac{1}{2}\|\alpha F_S(\theta) - y_S\|_2^2 = \frac{1}{2}\|r_S(\theta)\|_2^2, \quad \text{and} \quad \nabla_\theta \mathcal{L}_S(\theta) = \alpha J_S(\theta)^\top r_S(\theta), \tag{1}$$

where $r_S(\theta) := \alpha F_S(\theta) - y_S$ is the *residual*, with its initial value being $r_0 := r_S(\theta_0)$. The parameter α can adjust for the scale of the outputs; one may set $\alpha = 1$ throughout, but keeping it explicit makes the small-movement conditions transparent. Note from (1) that the GD updates are linear combinations of the rows of the current fine-tuning Jacobian. This motivates the following notion.

Definition 2 (Fine-Tuning Tangent Space)**.** *The* fine-tuning tangent space *of a pre-trained model f_{θ_0} at fine-tuning samples S is the subspace of $\mathbb{R}^n$ corresponding to the row span of $J_S(\theta_0)$, or equivalently the column span of $J_S(\theta_0)^\top$. We will denote by $\Pi_S = B_S B_S^\top$ the orthogonal projection onto the fine-tuning tangent space, where B_S is an orthonormal basis for the column span of $J_S(\theta_0)^\top$.*

Note that the fine-tuning tangent space given by Π_S is an mk-dimensional subspace of $\mathbb{R}^n$. In our setting we will fine-tune on very few samples, and hence we will assume that $mk \ll n$. Our theoretical results will assume that $J_S(\theta_0)$ has full row rank mk in a robust sense.

Spillover Loss. We compare the local fine-tuning update to its effect on a broader population D_{eval}, with marginal distribution D_{eval}^x.

Definition 3 (Population Spillover Loss)**.** *Given the fine-tuned parameter vector $\widehat{\theta}$, the* population spillover loss *on the distribution D_{eval} is*

$$\text{Spill}_{D_{\text{eval}}}(\widehat{\theta}; \theta_0) := \frac{1}{2}\mathbb{E}_{x \sim D_{\text{eval}}^x}\left[\|f_{\widehat{\theta}}(x) - f_{\theta_0}(x)\|_2^2\right].$$

Population Tangent Kernel and Tangent-Space Overlap. The *population tangent kernel* is defined as:

$$K_{\text{eval}} := \mathbb{E}_{x \sim D_{\text{eval}}^x}\left[J_x(\theta_0)^\top J_x(\theta_0)\right], \tag{2}$$

The following key geometric quantity captures the property that the population tangent kernel K_{eval} is nondegenerate on the directions in the fine-tuning tangent space.

Definition 4 (Tangent-Space Overlap)**.** *The population tangent kernel* K_{eval} *in* (2) *and the fine-tuning tangent space* Π_S *(from Definition 2) has* tangent-space overlap γ_{ov} *if the least eigenvalue of* K_{eval} *restricted to* Π_S*—or equivalently, the least eigenvalue of* $B_S^\top K_{\text{eval}} B_S \in \mathbb{R}^{mk \times mk}$*—satisfies*

$$\lambda_{mk}\left(B_S^\top K_{\text{eval}} B_S\right) = \inf_{v \in \text{Im}(\Pi_S)\setminus\{0\}} \frac{v^\top K_{\text{eval}} v}{\|v\|_2^2} \geq \gamma_{\text{ov}}^2. \tag{3}$$

Since $mk \ll n$, this is weaker than requiring K_{eval} to be well-conditioned on all of parameter space: it just requires non-trivial alignment between the fine-tuning tangent space and the population kernel.

3 Main Result

Intuition in the Linear Setting. Let us start by considering the simple setting of the linearized least-squares model, where $f_\theta(x)$ is the linear function $f_\theta(x) = \langle \theta, x \rangle$. During fine-tuning, we will move the parameters so that we can match the residual on the fine-tuning samples; i.e., the movement $\Delta := \theta - \theta_0$ satisfies $F_S(\theta) = F_S(\theta_0) + J_S(\theta_0)\Delta$, and Δ is a solution to $\alpha J_S(\theta_0)\Delta = -r_0$. GD from initialization θ_0 selects the minimum-norm solution for $\Delta = \theta - \theta_0$, which lies in $\text{Im}(\Pi_S)$. The population spillover loss of $f_\theta(x)$ is then given by:

$$\text{Spill}_{D_{\text{eval}}}(\widehat{\theta}; \theta_0) = \frac{1}{2}\mathbb{E}_{x \sim D_{\text{eval}}^x} \|J_x(\theta_0)\Delta\|_2^2 = \Delta^\top K_{\text{eval}} \Delta \geq \gamma_{\text{ov}}^2 \cdot \|\Delta\|_2^2 \geq \left(\frac{\gamma_{\text{ov}} \|r_0\|_2}{\alpha \sigma_{\max}}\right)^2,$$

due to the tangent-space overlap condition (3) and since any fitting update must satisfy $\|\Delta\|_2 \geq \|r_0\|_2/(\alpha\sigma_{\max})$, if $\|J_S(\theta_0)\|_{\text{op}} \leq \sigma_{\max}$.

Challenges in the Non-Linear Setting. In the general setting, the model $f_\theta(x)$ is non-linear, so it is more challenging to understand the effect that fine-tuning has on the evaluation distribution D_{eval}. The starting point of our argument is that if we fine-tune on only a few samples $mk \ll n$, then under mild assumptions on the conditioning of the neural tangent kernel, i.e., $J_S(\theta_0)$, and smoothness, we are in the NTK or *lazy training* regime, and fine-tuning using GD results in a "nearby" solution $\widehat{\theta} = \theta_0 + \Delta$ i.e., where $\|\Delta\|$ is small (see e.g., Malladi et al., 2023; Jacot et al., 2018; Allen-Zhu et al., 2019a; Chizat et al., 2019; Telgarsky, 2021; Arora et al., 2019b). However, unlike prior works in the lazy training literature, we now use this to lower bound the change in the model output on the evaluation distribution D_{eval} by using the Jacobian of f_θ. The main technical challenges are to account for the fact that the tangent subspace at intermediate points θ is different from Π_S, and to deal with the error from the non-linearity (in $\theta - \theta_0$) of the model f_θ.

We now formally state the smoothness assumptions on the NTK space needed for our main result; such assumptions are typical in the lazy training literature (e.g., Chizat et al., 2019; Liu et al., 2020).

Assumption 5 (Local Smoothness of the Pre-Trained Model)**.** *There exist parameters* $\beta_S, \beta, \rho_0 > 0$ *such that the following hold in the ball* $\mathcal{B}(\theta_0, \rho_0)$*.*

1. *Lipschitzness of the empirical Jacobian: For all* $\theta \in \mathcal{B}(\theta_0, \rho_0)$*,*

$$\|J_S(\theta) - J_S(\theta_0)\|_{\text{op}} \leq \beta_S \|\theta - \theta_0\|_2. \tag{4}$$

2. Average Lipschitzness of the population Jacobian: For all $\theta \in \mathcal{B}(\theta_0, \rho_0)$,

$$\mathbb{E}_{x \sim D^x_{\text{eval}}} \left[\|J_x(\theta) - J_x(\theta_0)\|_{\text{op}}^2 \right] \leq \beta^2 \|\theta - \theta_0\|_2^2. \tag{5}$$

We define $\Lambda^2 := \|K_{\text{eval}}\|_{\text{op}}$. Also, let $\sigma_{\min} := \sigma_{mk}\left(J_S(\theta_0)\right)$ and $\sigma_{\max} := \|J_S(\theta_0)\|_{\text{op}}$ control the empirical conditioning on the fine-tuning dataset; we assume that $\sigma_{\min} > 0$.

Recall that using GD during fine-tuning, we have

$$\theta_{t+1} = \theta_t - \eta \nabla L_S(\theta_t) = \theta_t - \eta \alpha J_S(\theta_t)^\top r_S(\theta_t). \tag{6}$$

To begin, we modify a well-known result on lazy-training and NTK analysis of GD (see (Telgarsky, 2021, Theorem 8.1)) to argue about fine-tuning starting from some parameter vector θ_0. The following lemma shows that under mild conditions, the residuals $r_t := r_S(\theta_t) = \alpha F_S(\theta) - y_S$ drop rapidly, and moreover the final parameter vector $\widehat{\theta}$ remains close—but not too close—to θ_0.

Lemma 6 (Local Least-Squares Fine-Tuning from Lazy Training)**.** *Suppose the empirical Jacobian conditions in Assumption 5 hold with $\sigma_{\min} > 0$. Consider GD on $\mathcal{L}_S$, initialized at θ_0 with local scaling α as given in (6), and step size $\eta \leq c/(\alpha^2 \sigma_{\max}^2)$. Suppose*

$$\frac{\|r_0\|_2}{\alpha} \leq c \min \left\{ \frac{\rho_0 \sigma_{\min}^2}{\sigma_{\max}}, \frac{\sigma_{\min}^3}{\beta_S \sigma_{\max}} \right\}. \tag{7}$$

Then, for universal constants $c, c_1, c_2 > 0$, we have that the (i) iterates remain in $\mathcal{B}(\theta_0, \rho_0)$, (ii) residuals decay geometrically,

$$\|r_t\|_2 \leq \exp\left(-c\eta\alpha^2\sigma_{\min}^2 t\right) \|r_0\|_2, \tag{8}$$

and (iii) limiting point $\widehat{\theta}$ satisfies

$$\frac{\|r_0\|_2}{2\alpha\sigma_{\max}} \leq \|\widehat{\theta} - \theta_0\|_2 \leq c_2 \frac{\sigma_{\max}\|r_0\|_2}{\alpha\sigma_{\min}^2}. \tag{9}$$

Proof of Lemma 6. The bound in (8) showing condition (ii), and the upper bound in (9) are a direct adaptation of (Telgarsky, 2021, Theorem 8.1) to our setting. Combining the upper bound in (9) with the first term in (7), the limiting point satisfies

$$\|\widehat{\theta} - \theta_0\|_2 \leq c_2 \frac{\sigma_{\max}\|r_0\|_2}{\alpha\sigma_{\min}^2} \leq \frac{\rho_0}{2}. \tag{10}$$

Hence the entire segment between θ_0 and $\widehat{\theta}$ lies in $\mathcal{B}(\theta_0, \rho_0)$, which proves (i).

We now show the lower bound on $\|\widehat{\theta} - \theta_0\|_2$ in (9) using the smoothness assumption (4) of the fine-tuning Jacobian $J_S(\cdot)$, and the conditioning of $J_S(\theta_0)$. Let $\Delta = \widehat{\theta} - \theta_0$. Since the residuals converge to zero, we have $F_S(\widehat{\theta}) - y_S = 0, F_S(\theta_0) - y_S = r_0$. Therefore, by the fundamental theorem of calculus,

$$\|r_0\|_2 = \alpha \|F_S(\widehat{\theta}) - F_S(\theta_0)\|_2 = \alpha \left\| \int_0^1 J_S(\theta_0 + s\Delta)\Delta \, ds \right\|_2 \leq \alpha \left(\sup_{s \in [0,1]} \|J_S(\theta_0 + s\Delta)\|_{\text{op}} \right) \|\Delta\|_2.$$

Since the iterates and their limit remain in $\mathcal{B}(\theta_0, \rho_0)$, the segment $\theta_0 + s\Delta$ also lies in this ball. The empirical Jacobian smoothness assumption along with (10) gives

$$\|J_S(\theta_0 + s\Delta)\|_{\text{op}} \le \|J_S(\theta_0)\|_{\text{op}} + \beta_S s\|\Delta\|_2 \le \sigma_{\max} + c_2\beta_S \frac{\sigma_{\max}\|r_0\|_2}{\alpha\sigma_{\min}^2}$$
$$\le 2\sigma_{\max}, \quad \text{since } \frac{\|r_0\|_2}{\alpha} \le c\frac{\sigma_{\min}^3}{\beta_S\sigma_{\max}},$$

for a sufficiently small universal constant $c > 0$, using the smallness condition in (7). Consequently,

$$\|r_0\|_2 \le 2\alpha\sigma_{\max}\|\Delta\|_2, \quad \text{i.e., } \|\widehat{\theta} - \theta_0\|_2 \ge \frac{\|r_0\|_2}{2\alpha\sigma_{\max}}. \qquad \blacksquare$$

Note that as is standard in the NTK literature, the scaling parameter $\alpha > 0$ adjusts for the scale of the outputs; one may set $\alpha = 1$ throughout, or also view the condition (11) as providing a lower-bound for the scale parameter α, when all relevant parameters are bounded and $\sigma_{\min} > 0$.

In addition, the upper bound from (9) shows that the parameter change $\|\widehat{\theta} - \theta_0\|$ is not too large, so that linearization using the Jacobian is not too lossy. (This phenomenon—that fine-tuning with few samples changes the parameter vector by a limited amount—this has also been observed empirically (see e.g., Malladi et al., 2023; Afzal et al., 2026, for references).)At the same time, the lower bound from (9) will be important to show that the model output changes globally on the evaluation distribution. We now state and prove our main theorem, showing the spillover loss is a function of the tangent-space overlap γ_{ov}.

Theorem 7 (Geometric Spillover Bound for Least-Squares Loss)**.** *Assume that the tangent-space overlap condition* (3) *and Assumption 5 hold, and that gradient descent on the empirical squared loss over the m fine-tuning samples and step size $\eta \le c/(\alpha^2\sigma_{\max}^2)$ converges to $\widehat{\theta}$. Suppose*

$$\frac{\|r_0\|_2}{\alpha} \le c \min\left\{\frac{\gamma_{\text{ov}} \cdot \sigma_{\min}^4}{\Lambda\beta_S\sigma_{\max}^2}, \frac{\gamma_{\text{ov}} \cdot \sigma_{\min}^4}{\beta\sigma_{\max}^3}\right\}, \tag{11}$$

for a sufficiently small universal constant $c > 0$, in addition to the conditions in (7)*. Then for a universal constant $c' > 0$, the population spillover loss of the model $f_{\widehat{\theta}}$ on the evaluation distribution D_{eval} due to this fine-tuning is given by*

$$\text{Spill}_{D_{\text{eval}}}(\widehat{\theta}, \theta_0) \ge c'\Big(\frac{\gamma_{\text{ov}}\|r_0\|_2}{\alpha\sigma_{\max}}\Big)^2. \tag{12}$$

Theorem 7 says that the spillover loss—the expected loss of the model on the distribution D_{eval} after fine-tuning on set S—is governed by the tangent-space overlap quantity γ_{ov} in (3). While the theorem is stated specifically while using GD for fine-tuning, we expect the argument could be adapted to other fine-tuning algorithms for which the consequences of Theorem 6 holds, along with the stated conditions in Theorem 7 and the γ_{ov}-tangent-space overlap condition. It will be instructive to think of the different parameters $\gamma_{\text{ov}}, \beta, \beta_S, \Lambda, \sigma_{\max}, 1/\sigma_{\min}$ as being large constants or small polynomial factors. See Section B and (Telgarsky, 2021, Chapter 8) for explicitly worked-out bounds for specific models like linear, polynomial or two-layer networks. This is also corroborated by the experiments in Section 4. Since the overlap condition is imposed only on an mk-dimensional subspace, it can be substantially weaker than a global conditioning assumption when $mk \ll n$.

The fine-tuning analysis from Theorem 6 shows that the fine-tuning algorithm converges to a point $\widehat{\theta}$ is that not too far from θ_0. Intuitively, this suggests we can try to analyze the change in the model output by considering the linearization using the Jacobian around the pre-trained model θ_0. However, there are two sources of error that we need to handle to show the desired lower bound on the model change on D_{eval}: (i) the error from the non-linear terms in the Jacobian approximation of the model f_θ, and (ii) the gradients do not lie on Π_S since the tangent subspace at intermediate points θ is different from Π_S. The following lemma handles the first source of error: it shows that if we can lower bound the change in the value of the linear approximation $g_{\widehat{\theta}}(x)$ on $x \sim D_{\text{eval}}$, then we can use the proximity of $\widehat{\theta}$ to θ_0 to upper bound the contribution of the higher-order terms.

Lemma 8 (Error from Non-Linear Terms)**.** *In the notation of Theorem 7, let θ_0 be the pre-trained model and let $\Delta = \widehat{\theta} - \theta_0$. Consider the linearized model*

$$g_{\widehat{\theta}}(x) = f_{\theta_0}(x) + J_x(\theta_0)(\widehat{\theta} - \theta_0). \tag{13}$$

Then, the average error from the non-linear term can be upper bounded as

$$\mathbb{E}_{x \sim D^x_{\text{eval}}}\left[\|f_{\widehat{\theta}}(x) - g_{\widehat{\theta}}(x)\|_2^2\right] \le \frac{\beta^2}{3}\|\Delta\|_2^4.$$

Proof. Let $\theta_s = \theta_0 + s\Delta$ for $s \in [0,1]$. From the fundamental theorem of calculus and recalling that the linear approximation $g_{\widehat{\theta}}(x) = f_{\theta_0}(x) + J_x(\theta_0)\Delta$, we have

$$f_{\widehat{\theta}}(x) - g_{\widehat{\theta}}(x) = \int_0^1 \left(J_x(\theta_s) - J_x(\theta_0)\right)\Delta \, ds.$$

Taking norms, applying the Cauchy–Schwarz inequality in the integral over s, and then taking expectation over $x \sim D^x_{\text{eval}}$, we obtain

$$\begin{aligned}
\mathbb{E}_{x \sim D^x_{\text{eval}}}\left[\|f_{\widehat{\theta}}(x) - g_{\widehat{\theta}}(x)\|_2^2\right] &\le \mathbb{E}_x\left[\left(\int_0^1 \|(J_x(\theta_s) - J_x(\theta_0))\Delta\|_2 \, ds\right)^2\right] \\
&\le \mathbb{E}_x\left[\int_0^1 \|(J_x(\theta_s) - J_x(\theta_0))\Delta\|_2^2 \, ds\right] \\
&\le \|\Delta\|_2^2 \int_0^1 \mathbb{E}_x\left[\|J_x(\theta_s) - J_x(\theta_0)\|_{\text{op}}^2\right] ds. \qquad (14)
\end{aligned}$$

By the average population Jacobian smoothness condition (5),

$$\mathbb{E}_x\left[\|J_x(\theta_s) - J_x(\theta_0)\|_{\text{op}}^2\right] \le \beta^2\|\theta_s - \theta_0\|_2^2 = \beta^2 s^2\|\Delta\|_2^2.$$

Substituting back into (14), we get

$$\mathbb{E}_{x \sim D^x_{\text{eval}}}\left[\|f_{\widehat{\theta}}(x) - g_{\widehat{\theta}}(x)\|_2^2\right] \le \|\Delta\|_2^2 \int_0^1 \beta^2 s^2 \|\Delta\|_2^2 ds \le \frac{\beta^2}{3}\|\Delta\|_2^4. \qquad \blacksquare$$

Proof of Theorem 7. Consider the linearized model in (13). If we can lower bound the change in the value of the linear approximation $g_{\widehat{\theta}}(x)$ on $x \sim D_{\text{eval}}$, then we can combine with the upper bound on error from the non-linear term in Lemma 8 to obtain the desired lower bound.

A key step in the argument is to lower bound $|g_{\widehat{\theta}}(x) - f_{\theta_0}(x)|$ on $x \sim D_{\text{eval}}$. Here we would like to use the lower bound on the parameter movement of $\widehat{\theta} - \theta_0$ in (9), along with the γ_{ov}-tangent-space overlap condition between the population tangent kernel and the fine-tuning tangent space Π_S. However, while the gradient of the loss $\mathcal{L}_S(\theta_0)$ is initially along the subspace Π_S, the row space of the Jacobian $J_S(\theta)$ changes as the parameter moves. Our argument controls these errors using perturbation bounds, Jacobian smoothness, along with path length bounds on the trajectory of GD.

We now lower bound the change between the linear approximation $g_{\widehat{\theta}} - f_{\theta_0}$. Let $\Delta = \widehat{\theta} - \theta_0$ be the parameter movement. We will decompose $\Delta = \Pi_S \Delta + \Pi_S^{\perp} \Delta$, and show that $\|\Pi_S \Delta\|_2$ is significant while $\|\Pi_S^{\perp} \Delta\|_2$ is small. We first show that $\|\Pi_S^{\perp} \Delta\|$ is small: Sum the GD updates and use $\Pi_S^{\perp} J_S(\theta_0)^{\top} = 0$ to get

$$\Pi_S^{\perp} \Delta = -\eta\alpha \sum_{t \geq 0} \Pi_S^{\perp} J_S(\theta_t)^{\top} r_t = -\eta\alpha \sum_{t \geq 0} \Pi_S^{\perp} \left(J_S(\theta_t)^{\top} - J_S(\theta_0)^{\top} \right) r_t.$$

Therefore, by the smoothness condition of the empirical Jacobian in (4) and Lemma 6 we get,

$$\begin{aligned}
\|\Pi_S^{\perp} \Delta\|_2 &\leq \eta\alpha \sum_{t \geq 0} \left\| \Pi_S^{\perp} \left(J_S(\theta_t)^{\top} - J_S(\theta_0)^{\top} \right) \right\|_{\text{op}} \|r_t\|_2 \\
&\leq \eta\alpha \sum_{t \geq 0} \left\| J_S(\theta_t)^{\top} - J_S(\theta_0)^{\top} \right\|_{\text{op}} \|r_t\|_2 \leq \eta\alpha\beta_S \left(\max_t \|\theta_t - \theta_0\|_2 \right) \sum_{t \geq 0} \|r_t\|_2 \\
&\leq \eta\alpha\beta_S \left(\frac{c_2 \sigma_{\max} \|r_0\|_2}{\alpha \sigma_{\min}^2} \right) \cdot \|r_0\|_2 \sum_{t \geq 0} \exp\left(-c\eta\alpha^2 \sigma_{\min}^2 t \right) \qquad \text{(from (8))}, \\
&\leq \frac{c' \beta_S \sigma_{\max} \|r_0\|_2^2}{\alpha^2 \sigma_{\min}^4}, \qquad (15)
\end{aligned}$$

for some absolute constant $c' > 0$. Thus the component outside the initial fine-tuning tangent space is a quadratic function of the residual.

On the other hand, we can combine this with the lower bound on the movement in (9) of Theorem 6 and taking c in (11) sufficiently small gives

$$\|\Pi_S \Delta\|_2 \geq \|\Delta\|_2 - \|\Pi_S^{\perp} \Delta\|_2 \geq \frac{\|r_0\|_2}{2\alpha\sigma_{\max}} - \frac{c' \beta_S \sigma_{\max} \|r_0\|_2^2}{\alpha^2 \sigma_{\min}^4} \geq \frac{\|r_0\|_2}{4\alpha\sigma_{\max}}.$$

Hence for the linearized model $g_{\widehat{\theta}}(x) = f_{\theta_0}(x) + J_x(\theta_0)\Delta$, we can lower bound the change from f_{θ_0} on D_{eval} by

$$\mathbb{E}_{x \sim D_{\text{eval}}^x} \|g_{\widehat{\theta}}(x) - f_{\theta_0}(x)\|_2^2 = \|K_{\text{eval}}^{1/2} \Delta\|_2^2.$$

We can use the γ_{ov}-tangent-space overlap condition in (3) along with (15) to get

$$\begin{aligned}
\|K_{\text{eval}}^{1/2} \Delta\|_2 &\geq \|K_{\text{eval}}^{1/2} \Pi_S \Delta\|_2 - \|K_{\text{eval}}^{1/2} \Pi_S^{\perp} \Delta\|_2 \geq \gamma_{\text{ov}} \|\Pi_S \Delta\|_2 - \Lambda \|\Pi_S^{\perp} \Delta\|_2 \\
&\geq \frac{\gamma_{\text{ov}} \|r_0\|_2}{4\alpha\sigma_{\max}} - \frac{c' \Lambda \beta_S \sigma_{\max} \|r_0\|_2^2}{\alpha^2 \sigma_{\min}^4} \geq c \frac{\gamma_{\text{ov}} \|r_0\|_2}{\alpha\sigma_{\max}},
\end{aligned}$$

where the last step uses the first smallness condition in (11). Hence

$$\mathbb{E}_{x \sim D_{\text{eval}}^x} \|g_{\widehat{\theta}}(x) - f_{\theta_0}(x)\|_2^2 \geq \frac{2c' \gamma_{\text{ov}}^2 \|r_0\|_2^2}{\alpha^2 \sigma_{\max}^2}. \qquad (16)$$

It remains to control the non-linear remainder. Substituting the upper bound on the movement from Lemma 6, $\|\Delta\|_2 \le c_2 \frac{\lambda_{\max}\|r_0\|_2}{\alpha\lambda_{\min}^2}$ into the bound from Lemma 8, we get

$$\mathbb{E}_{x\sim D_{\text{eval}}^x}\left[\|f_{\widehat{\theta}}(x) - g_{\widehat{\theta}}(x)\|_2^2\right] \le \frac{c_4\beta^2\lambda_{\max}^4\|r_0\|_2^4}{\alpha^4\lambda_{\min}^8},$$

for a universal constant $c_4 > 0$. The second smallness condition in (11) makes this a sufficiently small fraction of (16). Applying $\|a+b\|_2^2 \ge \frac{1}{2}\|a\|_2^2 - \|b\|_2^2$ with $a = g_{\widehat{\theta}}(x) - f_{\theta_0}(x)$ and $b = f_{\widehat{\theta}}(x) - g_{\widehat{\theta}}(x)$ proves (12). ■

4 Experiments

We now evaluate the explanation of the fine-tuning spillover phenomenon using the tangent-space overlap property. We experiment using two model classes: linear regression and a model consisting of a single self-attention layer followed by an MLP head. The linear setting is the easiest to interpret, whereas the attention+MLP model is the most expressive and exhibits the strongest non-linear effects.

4.1 Experimental Setup

We generate population distributions for pre-training and evaluation, under two families A and B of input marginals, which are designed to test different aspects of the overlap.

(i) In the *subspace-overlap* family, samples from A and B lie near two random 6-dimensional subspaces, whose "overlap" varies over $\{0, 0.25, 0.5, 0.75, 1\}$, where overlap refers to the number of orthonormal basis directions common to A and B; see Section C for details.

(ii) In the *covariance-scale* family, the two populations have the same mean but different covariances: $A = \mathcal{N}(0, {}^1\!/_d\, I_d)$ and $B = \mathcal{N}(0, c \cdot {}^1\!/_d\, I_d)$, with $c \in \{1, 1.1, 1.5, 2, 4, 8, 80\}$. In this experiment, the evaluation population is further from the fine-tuning region in ambient input space, but without changing its mean; this tests whether geometric proximity alone can explain fine-tuning spillover.

For each model class, we first pretrain on a balanced mixture of samples from A and B using (real-valued) labels generated by a random teacher model. We then change the labels of a small set S of m samples drawn from A as described in Section C and fine-tune the model with S; note that all parameters are potentially updated during fine-tuning. We then measure the loss on samples from the evaluation population $D_{\text{eval}} = B$.

For all the models we consider, the input dimension is $d = 80$, the sequence length is 5, the pre-training set size (i.e., the number of samples from the mixture of A and B) is 5000, the evaluation set size (i.e., the number of samples from D_{eval}) is 1000, and each completed condition is averaged over 40 random trials. We also vary the number of fine-tuning samples (m). For the attention+MLP model, the attention width is 80 and the hidden width of the head is 48. Fine-tuning is performed for 50 Adam steps after 80 pretraining steps.

Metrics. As in our theoretical results, the main quantity of interest is the spillover loss (Definition 3) measured on D_{eval}; we also measure the total parameter movement.

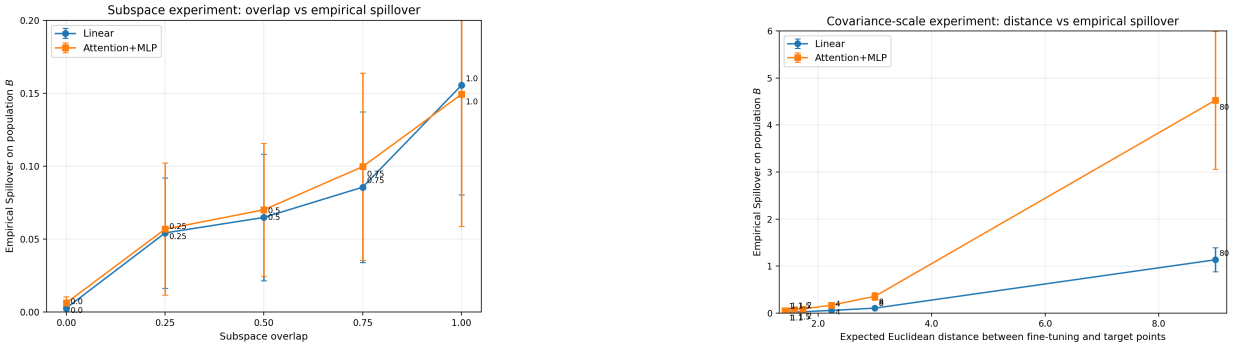

Figure 1: Spillover for the two model classes ($m = 32$). As the evaluation population moves farther away in input space, the spillover on D_{eval} grows for both subspace-overlap family (left) and covariance-scale family (right). For the latter, we use the (expected) Euclidean distance between a fine-tuning point and an evaluation point for the x-axis.

4.2 Results

Figure 1(left) shows the results for the subspace-overlap family at fine-tuning size $m = 32$: the empirical spillover on D_{eval} tracks the subspace-overlap quite closely for both model classes. (Figure 3 in Section B shows that the synthetic subspace-overlap knob used in our experiments induces the tangent-space overlap quantity in (3) .)

Figure 1(right) directly tests the hypothesis that spillover might be explained simply by geometric proximity to the fine-tuning region. Proximity-based intuition would suggest that far-away populations should be less affected by fine-tuning on S. Experiments on the covariance-scale family shows the opposite can happen: the spillover on D_{eval} can become much larger as D_{eval} moves farther away, provided the local linearization retains sufficient overlap with the directions that matter on D_{eval}.

Spillover vs Intrinsic Difficulty. We perform a sanity check by measuring the post-fine-tuning error when the model is allowed to fine-tune on samples from *both* A and D_{eval}. If this error is small, then the evaluation population D_{eval} is easy to fit when included during fine-tuning, so large loss on D_{eval} after fine-tuning only on S could be due to spillover. For the attention+MLP model, we observe that the post-fine-tuning error is quite small, confirming our hypothesis; see Appendix C for details.

Regularization. We also experimentally evaluate if ℓ_2-regularization around the pretrained weights, which forces the parameter movement to be small, can prevent spillover. Figure 2 shows that, in the attention+MLP model on the subspace-overlap family, regularization reduces total parameter movement, as intended. However, this does not eliminate the spillover, clearly showing that spillover is caused not simply because the parameters had to move more during fine-tuning.

In summary, our experiments show the following: (i) spillover depends on the tangent-space overlap, (ii) geometric proximity to the fine-tuning region cannot explain spillover, and (iii) regularization cannot prevent spillover. Appendix B has more details about the experiments, along with detailed tables of the runs for different settings of m, and for two-layer ReLU networks (which shows very similar trends) in addition to the runs for the linear model, and the attention+MLP model.

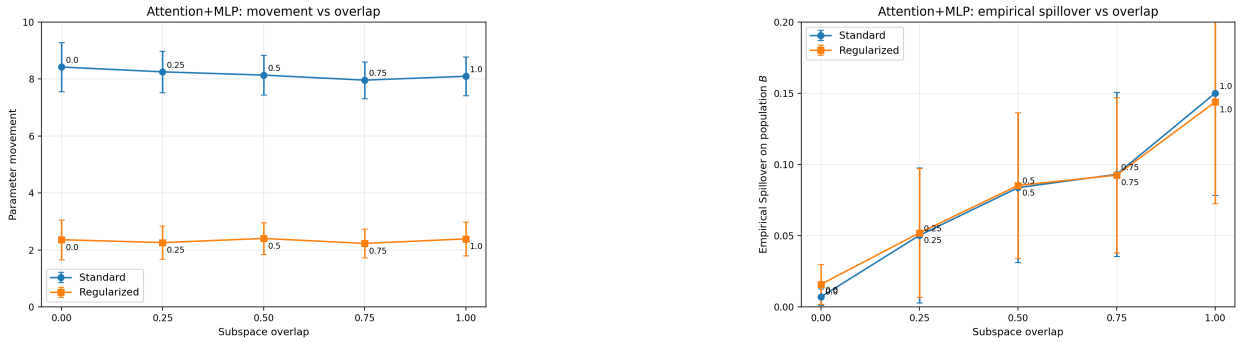

Figure 2: Effect of regularization on total parameter movement (left) and spillover on D_{eval} (right) as a function of subspace overlap for Attention+MLP model ($m = 100$). Regularization reduces movement substantially, but large overlap still produces large spillover.

5 Conclusion and Limitations

We study the fine-tuning spillover phenomenon and showed that tangent-space overlap—a measure we propose—can explain how local fine-tuning can induce global changes in model behavior. For least-squares fine-tuning in the lazy fine-tuning regime, we proved lower bounds on spillover and validated the mechanism in controlled experiments. Our results also suggest a plausible and theoretically grounded explanation for empirical phenomena such as "weird generalization" and "emergent misalignment". But a more detailed empirical study is needed to confirm this in real-world LLM fine-tuning. Our work also offers a prescription: reliable fine-tuning may require tangent-aware methods that control not only the size of the update, but also the directions through which it acts.

References

Z. R. Afzal, T. Esmaeilbeig, M. Soltanalian, and M. I. Ohannessian. Linearization explains fine-tuning in large language models. In *NeurIPS*, 2026.

Z. Allen-Zhu, Y. Li, and Y. Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *NeurIPS*, 2019a.

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *ICML*, pages 242–252, 2019b.

S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. In *NeurIPS*, 2019a.

S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *ICML*, pages 322–332, 2019b.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. E. Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv*, 2204.05862, 2022.

J. Betley, J. Cocola, D. Feng, J. Chua, A. Arditi, A. Sztyber-Betley, and O. Evans. Weird generalization and inductive backdoors: New ways to corrupt LLMs. *arXiv*, 2512.09742, 2025.

J. Betley, N. Warncke, A. Sztyber-Betley, D. Tan, X. Bao, M. Soto, M. Srivastava, N. Labenz, and O. Evans. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649(8097):584–589, 2026.

A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny. Efficient lifelong learning with A-GEM. In *ICLR*, 2019.

L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. In *NeurIPS*, pages 2933–2943, 2019.

J. Chua, J. Betley, M. Taylor, and O. Evans. Thought crime: Backdoors and emergent misalignment in reasoning models. *arXiv*, 2506.13206, 2025.

R. Cohen, E. Biran, O. Yoran, A. Globerson, and M. Geva. Evaluating the ripple effects of knowledge editing in language models. *TACL*, 12:283–298, 2024.

T. Doan, M. A. Bennani, B. Mazoure, G. Rabusseau, and P. Alquier. A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. In *AISTATS*, pages 1072–1080, 2021.

S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *ICML*, pages 1671–1680, 2019.

M. Farajtabar, N. Azizan, A. Mott, and A. Li. Orthogonal gradient descent for continual learning. In *AISTATS*, pages 3762–3773, 2020.

J. Hoelscher-Obermaier, J. Persson, E. Kran, I. Konstas, and F. Barez. Detecting edit failures in large language models: An improved specificity benchmark. In *ACL (Findings)*, pages 11548–11559, 2023.

E. Hubinger, C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D. M. Ziegler, T. Maxwell, N. Cheng, A. S. Jermyn, A. Askell, A. Radhakrishnan, C. Anil, D. Duvenaud, D. Ganguli, F. Barez, J. Clark, K. Ndousse, K. Sachan, M. Sellitto, M. Sharma, N. DasSarma, R. B. Grosse, S. Kravec, Y. Bai, Z. Witten, M. Favaro, J. Brauner, H. Karnofsky, P. F. Christiano, S. R. Bowman, L. Graham, J. Kaplan, S. Mindermann, R. Greenblatt, B. Shlegeris, N. Schiefer, and E. Perez. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv*, 2401.05566, 2024.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.

R. Latala. Estimates of moments and tails of gaussian chaoses. *Annals of Probability*, 34:2315–2331, 2005. URL https://api.semanticscholar.org/CorpusID:14804993.

J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *NeurIPS*, 2019.

S. Lermen, C. Rogers-Smith, and J. Ladish. LoRA fine-tuning efficiently undoes safety training in llama 2-chat 70b. *arXiv*, 2310.20624, 2023.

C. Liu, L. Zhu, and M. Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. In *NIPS*, 2020.

D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning. In *NIPS*, 2017.

M. MacDiarmid, B. Wright, J. Uesato, J. Benton, J. Kutasov, S. Price, N. Bouscal, S. Bowman, T. Bricken, A. Cloud, et al. Natural emergent misalignment from reward hacking in production RL. *arXiv*, 2511.18397, 2025.

S. Malladi, A. Wettig, D. Yu, D. Chen, and S. Arora. A kernel-based view of language model fine-tuning. In *ICML*, pages 23610–23641, 2023.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, volume 35, pages 17359–17372, 2022.

K. Meng, A. S. Sharma, A. Andonian, Y. Belinkov, and D. Bau. Mass-editing memory in a transformer. In *ICLR*, 2023.

E. Mitchell, C. Lin, A. Bosselut, C. Finn, and C. D. Manning. Fast model editing at scale. In *ICLR*, 2022.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

X. Qi, Y. Zeng, T. Xie, P. Chen, R. Jia, P. Mittal, and P. Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *ICLR*, 2024.

M. Telgarsky. Deep learning theory lecture notes, 2021. URL https://mjt.cs.illinois.edu/dlt/. Version: 2021-10-27 v0.0-e7150f2d (alpha).

E. Turner, A. Soligo, M. Taylor, S. Rajamanoharan, and N. Nanda. Model organisms for emergent misalignment. *arXiv*, 2506.11613, 2025.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2 edition, 2026.

X. Yang, X. Wang, Q. Zhang, L. R. Petzold, W. Y. Wang, X. Zhao, and D. Lin. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv*, 2310.02949, 2023.

C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, et al. Lima: Less is more for alignment. *NeurIPS*, pages 55006–55021, 2023.

A Related Work

Our framework sits at the intersection of empirical AI safety, model editing, and the theory of optimization for overparameterized neural networks. Our formalization of spillover loss and the unavoidability of spillover loss based on the tangent-space overlap suggests possible explanations for various empirical phenomenon related to fine-tuning and AI safety. To the best of our knowledge, we are unaware of previous work proving such a mathematical statement on the unavoidability of spillover i.e., change in the model outputs during fine-tuning. We now discuss some potential related work in these topics — the existing literature on these topics, particularly on the empirical side is vast. We restrict to representative works on these topics.

AI Safety and Emergent Misalignment. Given the prevalence of reinforcement learning from human feedback (RLHF) and instruction tuning, much effort has been dedicated to questions of LLM alignment (Ouyang et al., 2022; Bai et al., 2022; Zhou et al., 2023). However, recent empirical work has exposed a severe vulnerability: fine-tuning an aligned model on just a few examples—sometimes on malicious examples, but sometimes even on entirely benign, non-malicious data—can catastrophically erode its safety guardrails (Qi et al., 2024; Yang et al., 2023; Lermen et al., 2023). This phenomenon of "emergent misalignment" or "weird generalization", where the fine-tuning can cause broad misalignment on unrelated prompts (Betley et al., 2026, 2025), has been noticed in other settings by Turner et al. (2025); Chua et al. (2025); MacDiarmid et al. (2025). Such efforts and observations have been largely empirical, and have often been viewed through the lens of data poisoning or as an artifact of shallow safety tuning (Hubinger et al., 2024). Our work gives a geometric perspective on such vulnerabilities, show that if the fine-tuning data (whether it be benign or hostile) and the safety-critical evaluation population share tangent space directions (i.e., a high geometric overlap γ_{ov}), then driving the local fine-tuning loss to zero forces an unavoidable spillover on the evaluation distribution, regardless of the user intent or the capacity of the model.

Lazy Training and Neural Tangent Kernels (NTK). Our analytical approach relies on the behavior of highly overparameterized neural networks. In this "lazy" regime, the parameters move infinitesimally, and the network behaves as a linear model governed by the Neural Tangent Kernel (NTK) (Jacot et al., 2018; Lee et al., 2019; Arora et al., 2019a; Chizat et al., 2019). Historically, NTK theory has been utilized to explain *benign* properties of neural networks, such as guaranteeing stable convergence to global minima and explaining generalization in overparameterized spaces (Du et al., 2019; Allen-Zhu et al., 2019b). In a departure from this classical use of the lazy training analysis, we invert the lazy training framework to prove a lower bound on fine-tuning spillover.

Continual Learning and Model Editing. Within theoretical continual learning, some works have connected forgetting to task alignment or gradient interference. Projection-based methods such as GEM, A-GEM, and OGD explicitly constrain new updates to reduce interference with previous tasks (Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2019; Farajtabar et al., 2020). Doan et al. (2021) use NTK theory and an NTK overlap matrix to analyze catastrophic forgetting and projection-based mitigation algorithms. This is related in spirit, but the setting is different: in long-horizon continual learning, a fixed-NTK regime is primarily an analytical idealization, whereas few-shot fine-tuning of a pre-trained model is often explicitly designed to remain close to the starting point through early stopping, regularization, or parameter-efficient updates. We use this practically

relevant lazy fine-tuning regime to prove a lower-bound obstruction: tangent-space overlap with an evaluation population forces spillover under gradient descent.

Model editing studies a related locality problem: changing a specific behavior or factual association while preserving unrelated behavior. Modern editing methods explicitly evaluate locality or specificity in addition to edit success (Mitchell et al., 2022; Meng et al., 2022, 2023), and recent work studies unintended side effects and ripple effects of edits (Hoelscher-Obermaier et al., 2023; Cohen et al., 2024). Our framework gives a complementary theoretical perspective: locality can fail not merely because an edit is poorly implemented, but because the tangent directions required to implement the edit overlap with directions controlling behavior elsewhere. Consequently, standard proximity-based mitigations are fundamentally insufficient to prevent these ripple effects, as shrinking the parameter update does not change the overlapping tangent directions through which the edit acts.

B Evaluating the Geometric Conditions in Representative Settings

This appendix instantiates the geometric quantities in Theorem 7 for representative models: linear regression, polynomial regression, and an overparameterized two-layer network. We work in the scalar-output case $k = 1$. The fine-tuning covariates are

$$x_1, \dots, x_m \sim_{iid} \mathcal{N}(0, \Sigma_1), \qquad x \sim D^x_{\text{eval}} = \mathcal{N}(0, \Sigma_2),$$

and $X \in \mathbb{R}^{m \times d}$ denotes the matrix with rows $x_i^\top$.

For a positive semidefinite matrix A, denote the effective dimension or rank by $d_{\text{eff}}(A) := \frac{\text{Tr}(A)^2}{\text{Tr}(A^2)}$ For the mixed covariance quantity appearing below, the relevant matrix is

$$A_{12} := \Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2}, \qquad \text{Tr}(A_{12}) = \text{Tr}(\Sigma_1 \Sigma_2).$$

Then under the assumption that $m \le c \min\{d_{\text{eff}}(\Sigma_1), d_{\text{eff}}(A_{12})\}$, for a sufficiently small universal constant $c > 0$, we can use strong concentration bounds for Gaussian matrices. Hence both the source Gram matrix $XX^\top$ and the mixed Gram matrix $X\Sigma_2 X^\top$ are close to scalar multiples of the identity on all m-dimensional fine-tuning directions.

B.1 Linear regression

Consider $f_\theta(x) = \langle \theta, x \rangle$ for $\theta \in \mathbb{R}^d$. Then $J_x(\theta) = x^\top$ is independent of θ, and hence

$$J_S(\theta_0) = X, \qquad K_{\text{eval}} = \mathbb{E}_{x \sim \mathcal{N}(0, \Sigma_2)}[x x^\top] = \Sigma_2.$$

In particular, the empirical and population Jacobian smoothness parameters in Assumption 5 are exactly $\beta_S = 0$ and $\beta = 0$. When a denominator contains β_S or β in the smallness conditions, the corresponding condition is vacuous in this linear case.

Assume $m \le c \min\{d_{\text{eff}}(\Sigma_1), d_{\text{eff}}(A_{12})\}$ for a sufficiently small universal constant $c > 0$. Then, after decreasing c if necessary, from standard bounds in high-dimensional probability (Vershynin, 2026), we have with probability at least $1 - \exp(-c'm)$ over the fine-tuning covariates, we have:

1. **Empirical conditioning:** $\sigma^2_{\min} \ge \frac{1}{2}\text{Tr}(\Sigma_1)$ and $\sigma^2_{\max} \le \frac{3}{2}\text{Tr}(\Sigma_1)$.
2. **Population tangent kernel:** $K_{\text{eval}} = \Sigma_2$.

3. **Geometric overlap:** If Π_S is the orthogonal projection onto rowspan(X), then

$$\gamma_{\text{ov}}^2 = \inf_{v \in \text{Im}(\Pi_S)\setminus\{0\}} \frac{v^\top \Sigma_2 v}{\|v\|_2^2} \geq \frac{1}{3}\frac{\text{Tr}(\Sigma_1\Sigma_2)}{\text{Tr}(\Sigma_1)}.$$

Consequently, gradient descent on the empirical squared loss converges to $\widehat{\theta}$ and Theorem 7 gives

$$\frac{1}{2}\mathbb{E}_{x\sim D_{\text{eval}}^x}\left[\|f_{\widehat{\theta}}(x) - f_{\theta_0}(x)\|_2^2\right] \geq C\frac{\|r_0\|_2^2}{\alpha^2}\frac{\text{Tr}(\Sigma_1\Sigma_2)}{\text{Tr}(\Sigma_1)^2},$$

for a universal constant $C > 0$.

Conditions of Theorem 7 for linear regression. The smoothness claims are immediate because the Jacobian is constant. For the empirical conditioning, note that for any fixed unit vector $u \in \mathbb{S}^{m-1}$, $u^\top XX^\top u = \|\sum_{i=1}^m u_i x_i\|_2^2$, where $\sum_i u_i x_i \sim \mathcal{N}(0, \Sigma_1)$. Standard Hanson–Wright concentration and an ϵ-net argument imply $\sup_{u\in\mathbb{S}^{m-1}} |u^\top XX^\top u - \text{Tr}(\Sigma_1)| \leq \frac{1}{2}\text{Tr}(\Sigma_1)$ with high probability, provided $m \leq cd_{\text{eff}}(\Sigma_1)$.

For the overlap, $\lambda_{\min}(XX^\top) > 0$ implies X has full row rank. Every nonzero $v \in \text{Im}(\Pi_S)$ can be written as $v = X^\top u$. By scale invariance, it suffices to consider $u \in \mathbb{S}^{m-1}$. The Rayleigh quotient evaluates to

$$\frac{v^\top \Sigma_2 v}{\|v\|_2^2} = \frac{u^\top X\Sigma_2 X^\top u}{u^\top XX^\top u}.$$

The numerator has expectation $\text{Tr}(\Sigma_1\Sigma_2) = \text{Tr}(A_{12})$. Applying the same concentration argument to A_{12} guarantees the numerator is bounded below by $\frac{1}{2}\text{Tr}(\Sigma_1\Sigma_2)$ for $m \leq cd_{\text{eff}}(A_{12})$. Combining this with the denominator bound $u^\top XX^\top u \leq \frac{3}{2}\text{Tr}(\Sigma_1)$ yields $\gamma_{\text{ov}}^2 \geq \frac{1}{3}\text{Tr}(\Sigma_1\Sigma_2)/\text{Tr}(\Sigma_1)$. Substituting $\sigma_{\max}$ and γ_{ov} into Theorem 7 completes the proof. ■

B.2 Polynomial regression

The calculation extends immediately to polynomial regression. Let $\phi_p : \mathbb{R}^d \to \mathbb{R}^N$ be a polynomial feature map of degree $p \geq 1$. To ensure the effective dimension does not collapse to $O(1)$ from a large mean component, we assume the features are centered over the source distribution, i.e., $\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_1)}[\phi_p(x)] = 0$. Consider the lifted model $f_\theta(x) = \langle\theta, \phi_p(x)\rangle$.

Let $\Phi \in \mathbb{R}^{m\times N}$ be the lifted design matrix, and define the lifted covariances $\Gamma_1 := \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_1)}[\phi_p(x)\phi_p(x)^\top]$ and $\Gamma_2 := \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_2)}[\phi_p(x)\phi_p(x)^\top]$. Because the model is linear in the feature space, the smoothness parameters remain $\beta_S = \beta = 0$. By making the direct substitutions $X \mapsto \Phi$ and $\Sigma_i \mapsto \Gamma_i$, similar arguments to Section B.1 can be transferred to polynomial threshold functions or polynomial kernels, using concentration arguments for Gaussian polynomials (Latala, 2005; Vershynin, 2026).

Consequently, assuming $m \leq c\min\{d_{\text{eff}}(\Gamma_1), d_{\text{eff}}(\Gamma_1^{1/2}\Gamma_2\Gamma_1^{1/2})\}$ for a sufficiently small constant $c > 0$, standard concentration arguments for Gaussian polynomials give $\sigma_{\max}^2 \asymp \text{Tr}(\Gamma_1)$ and $\gamma_{\text{ov}}^2 \gtrsim \text{Tr}(\Gamma_1\Gamma_2)/\text{Tr}(\Gamma_1)$. Applying Theorem 7 yields the identical nonlocal fine-tuning drift bound:

$$\frac{1}{2}\mathbb{E}_{x\sim D_{\text{eval}}^x}\left[\|f_{\widehat{\theta}}(x) - f_{\theta_0}(x)\|_2^2\right] \geq C\frac{\|r_0\|_2^2}{\alpha^2}\frac{\text{Tr}(\Gamma_1\Gamma_2)}{\text{Tr}(\Gamma_1)^2}.$$

Remark 9 (Interpretation for Gaussian polynomial features). For Gaussian covariates, Γ_1 and Γ_2 can be computed explicitly via Wick's formula. If ϕ_p is an orthonormal Hermite feature map, Γ_1 and Γ_2 decompose degree by degree, and the overlap $\text{Tr}(\Gamma_1\Gamma_2)/\text{Tr}(\Gamma_1)^2$ measures the alignment of the distributions after the degree-p polynomial lift. Thus, the fine-tuning spillover is structurally identical to linear regression, but controlled by the tangent-space overlap of the lifted target population kernel.

C Experimental setup

Our experiments are designed to isolate when a *local* fine-tuning edit on few samples from one population experiences non-local spillover to another population. Across all settings, each input is a sequence of length $T = 5$ with tokens in $\mathbb{R}^d$ for $d = 80$. Thus an example may be viewed as an element of $\mathbb{R}^{5\times 80}$, or, after flattening, as an element of $\mathbb{R}^{400}$. In the linear and two-layer experiments, the sequence is flattened before being passed to the model; in the attention+MLP experiments, the sequence structure is preserved. The attention model uses width 80 and an MLP head of hidden size 48, while the two-layer baseline uses hidden size 128. In all experiments, we pretrain on 5000 samples, evaluate on 1000 fresh target-population samples, fine-tune for 50 Adam steps after 80 pretraining steps, and average results over 40 random trials. The experiments were mostly run on Google CoLab, and also a separate machine with 2x RTX A5000 GPU.

We consider two families of source and target input distributions, denoted by A and B.

Subspace-overlap family. In the subspace-overlap family, samples are concentrated near two 6-dimensional subspaces $U_A, U_B \subseteq \mathbb{R}^{80}$. The overlap parameter takes values $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$ and is implemented by sharing exactly

$$s = \text{round}(6\rho)$$

orthonormal basis directions between U_A and U_B. Thus, for example, $\rho = 0.5$ means that the two 6-dimensional subspaces share 3 basis directions, while $\rho = 1$ means that the subspaces are identical. Conditional on the subspace, each token is generated by drawing Gaussian coefficients in that subspace and then adding small isotropic noise. More concretely, if $U \in \mathbb{R}^{80\times 6}$ denotes the basis matrix of the relevant subspace, then a token has the form

$$x_t = Uc_t + \xi_t,$$

where $c_t \sim \mathcal{N}(0, I_6/6)$ and ξ_t is small isotropic noise. This gives a clean synthetic knob that changes the geometric alignment between the source and target populations while leaving the ambient dimension fixed.

Covariance-scale family. In the covariance-scale family, the two populations have the same mean but different covariance scales:

$$A \sim \mathcal{N}\left(0, \frac{1}{d} I_d\right), \qquad B \sim \mathcal{N}\left(0, c\frac{1}{d} I_d\right),$$

where in the released code $c \in \{1, 1.1, 1.5, 2, 4, 8, 64\}$. This family is intended to separate ambient geometric distance from the overlap-based mechanism studied in the paper:

$$\mathbb{E}_{x\sim A, x'\sim B}\mathbb{E}[\|x - x'\|^2] = (c + 1),$$

and the squared distances concentrate rapidly with subexponential tails. As c increases, points from B become farther from the fine-tuning region in input space, even though the mean remains unchanged. In particular, this setting lets us test whether nonlocal drift is explained by simple proximity to the fine-tuning data, or instead by the Jacobian/tangent-space geometry at the pretrained point.

Labels and Local Edits. Labels are generated by a frozen random teacher model. In the attention+MLP experiments, the teacher has the same architecture as the student, so the base pretraining task is realizable within that class. In the linear/two-layer experiments, the teacher is always a two-layer ReLU model, so the task is realizable for the two-layer student but not necessarily for the linear student. After sampling source and target data, we pretrain the student on a balanced mixture of labeled examples from A and B. We then construct a *local edit* that is applied only on a small fine-tuning sample from A. This edit is deterministic, not random: if $x \in \mathbb{R}^{5\times 80}$ is a sequence, then the label shift is

$$\Delta(x) = \lambda \tanh\left(2 \cdot \frac{1}{T}\sum_{t=1}^{T} x_{t,1}\right), \qquad \lambda = 4.$$

Thus the fine-tuning target is

$$y_{\text{fine}}(x) = f_{\text{teacher}}(x) + \Delta(x),$$

so the edit depends only on the average of the first coordinate across the sequence. The fine-tuning sample size ranges over

$$m \in \{8, 16, 32, 50, 75, 100\}.$$

We also compare standard fine-tuning to a regularized regime with an ℓ_2 penalty toward the pretrained parameter vector.

Our main measurement is the spillover on the off-target population $D_{\text{eval}} = B$,

$$\text{Spill}_{D_{\text{eval}}}(\widehat{\theta}, \theta_0) = \mathbb{E}_{x\sim B}\Big[(f_{\widehat{\theta}}(x) - f_{\theta_0}(x))^2\Big],$$

where $\widehat{\theta}$ is the model after fine-tuning, and θ_0 is the pretrained model.

This directly measures how much is the spillover onto $D_{\text{eval}} = B$ after fine-tuning with a local edit. We also record the source fine-tuning error before and after fine-tuning, the prediction error on B before and after fine-tuning, and the total parameter movement. In addition, we run a *fit-both* sanity check in which the model is fine-tuned on the edited union of examples from both A and B. This helps distinguish genuine propagation effects from the possibility that the edited target task on B is simply hard to fit. Finally, to connect the experiments to the lazy fine-tuning theory and our assumptions, we compute Jacobian-based overlap diagnostics at the pretrained point. Let J_S denote the Jacobian rows on the fine-tuning sample from A, and let J_B denote the Jacobian rows on target examples from B. We project the target-population tangent kernel onto the row span of J_S and compute overlap statistics including a minimum-eigenvalue overlap and a broader projected-energy fraction. The tangent-space overlap is a worst-direction quantity and can therefore be numerically small even when average overlap is not, which is why the corresponding plots are shown on a log scale. Moreover the min-eigenvalue is not as concentrated and often has higher variance. Altogether, the setup is intended to test whether fine-tuning spillover is dictated by overlap in the pretrained tangent geometry rather than by ambient proximity or raw parameter movement alone.

C.1 Results: drift tracks the overlap metric

Figure 3 first shows that the synthetic subspace-overlap knob used in the data construction does indeed induce the tangent-space overlap (eigenvalue) metric. This is important conceptually: the theory is not about the designed overlap parameter itself, but about the tangent space overlap diagnostic computed from the pretrained Jacobian geometry. Figure 1 then shows the key empirical

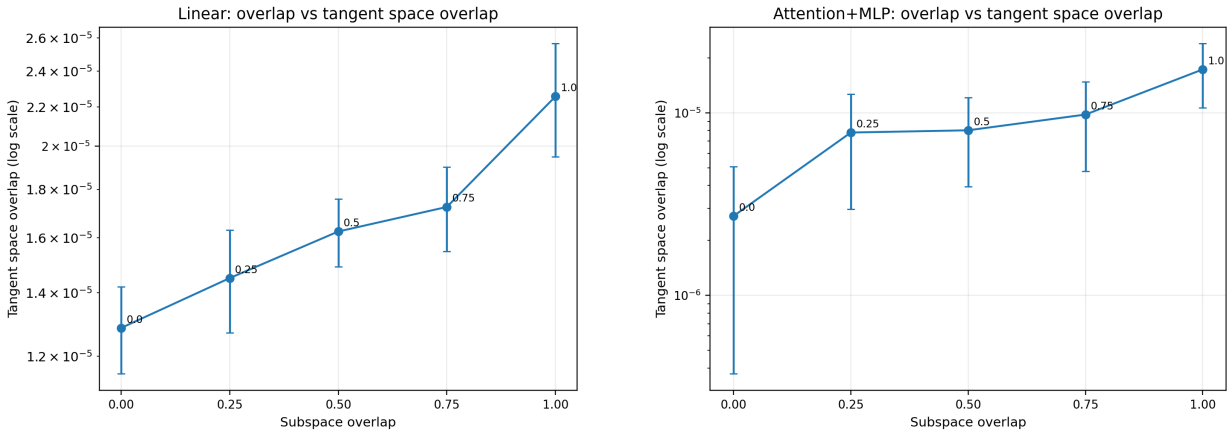

Figure 3: Subspace-overlap experiment at $m = 32$. The overlap parameter induces the tangent-space overlap quantity (involving computing a minimum eigenvalue). The left panel shows the linear model and the right panel shows the attention+MLP model. The vertical axis is on a log scale.

relationship in the subspace experiment at fine-tuning size $m = 32$: the observed nonlocal drift on the evaluation population tracks the computed overlap metric very closely. In the linear model, the average drift grows from approximately 0.0026 at overlap 0 to approximately 0.1555 at overlap 1, while in the attention+MLP model it grows from approximately 0.0062 to approximately 0.1493. Thus, the most important pattern is not merely that changing the synthetic overlap knob changes drift, but that the computed tangent-space overlap metric itself predicts the spillover across both the simple linear model and the non-linear attention-based model. See Section **??** for detailed statistics from all our experiments, across different hyperparameters like the number of fine-tuning samples m, the family of instances, and the models themselves.

This is the clearest evidence for the overlap story. In the subspace experiment, the source and evaluation populations differ only through how much their tangent directions overlap, and the induced nonlocal change is well explained by the resulting overlap metric. In particular, the linear setting makes the geometry especially transparent, while the attention+MLP model shows that the same overlap-based explanation survives substantial non-linearity.

C.2 Additional plots corroborating trends

We next provide additional plots for the standard fine-tuning experiments. These figures examine how empirical spillover on the held-out population B varies with the measured tangent-space overlap and with the number of fine-tuning samples. In all plots, spillover is measured as the squared change in the model output on samples from B before and after fine-tuning, and all error bars are computed across independent trials.

Figures 4–7 show the subspace-overlap family for several values of m. For $m = 8$ and $m = 16$, we include the linear model, attention+MLP model, and two-layer network; for $m = 50$ and $m = 75$, we show the two nonlinear architectures. Across these settings, larger tangent-space overlap leads to larger spillover on population B, consistent with the tangent-space-overlap mechanism.

Figure 8 then plots spillover as a function of m for several designed subspace-overlap values, showing that the overlap-dependent ordering persists across fine-tuning sample sizes. For overlap

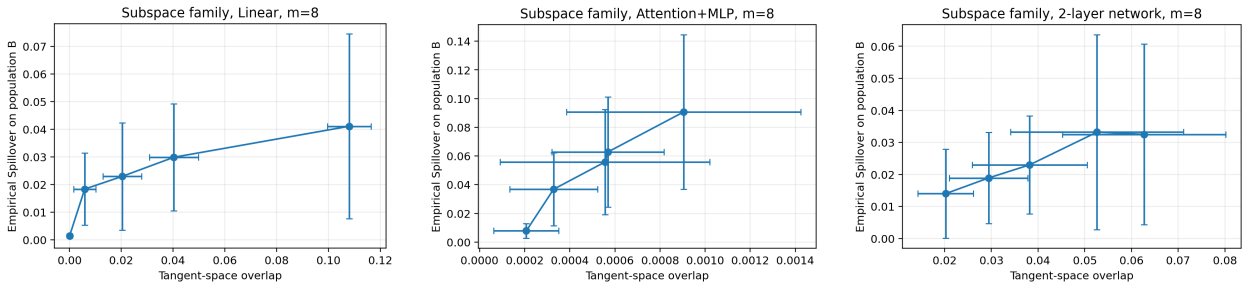

Figure 4: Subspace-overlap family with $m = 8$ fine-tuning samples. Each panel plots spillover on the evaluation population B against the computed tangent-space overlap, with error bars across trials. From left to right: linear model, attention+MLP model, and two-layer network.

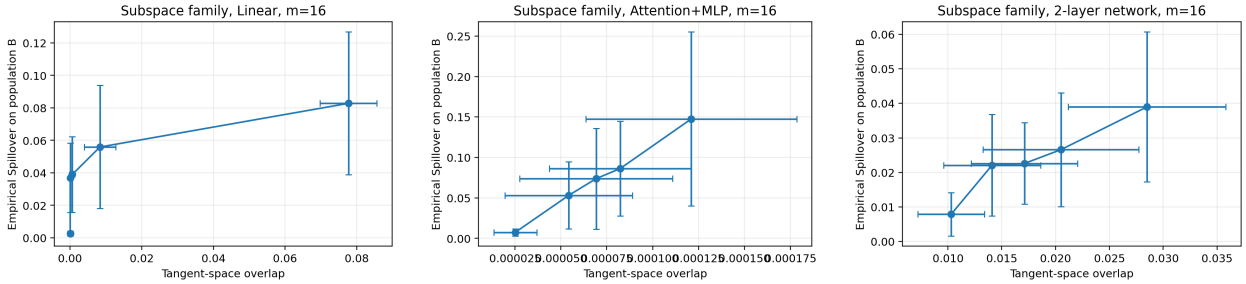

Figure 5: Subspace-overlap family with $m = 16$ fine-tuning samples. Each panel plots spillover on the evaluation population B against the computed tangent-space overlap, with error bars across trials. From left to right: linear model, attention+MLP model, and two-layer network.

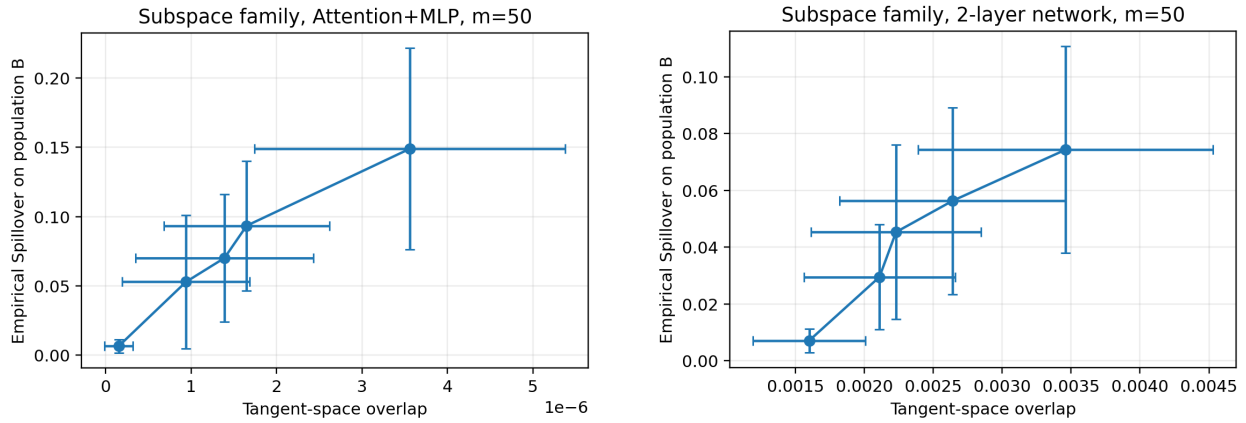

Figure 6: Subspace-overlap family with $m = 50$ fine-tuning samples. The panels plot spillover on the evaluation population B against the computed tangent-space overlap, with error bars across trials. Left: attention+MLP model. Right: two-layer network.

0.0, it shows that the spillover is close 0 across all m.

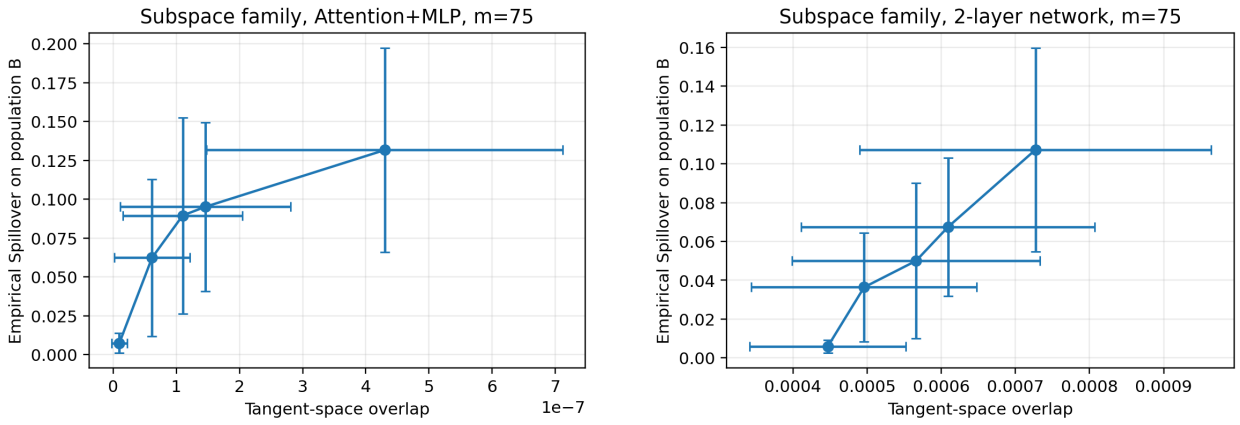

Figure 7: Subspace-overlap family with $m = 75$ fine-tuning samples. The panels plot spillover on the evaluation population B against the computed tangent-space overlap, with error bars across trials. Left: attention+MLP model. Right: two-layer network.

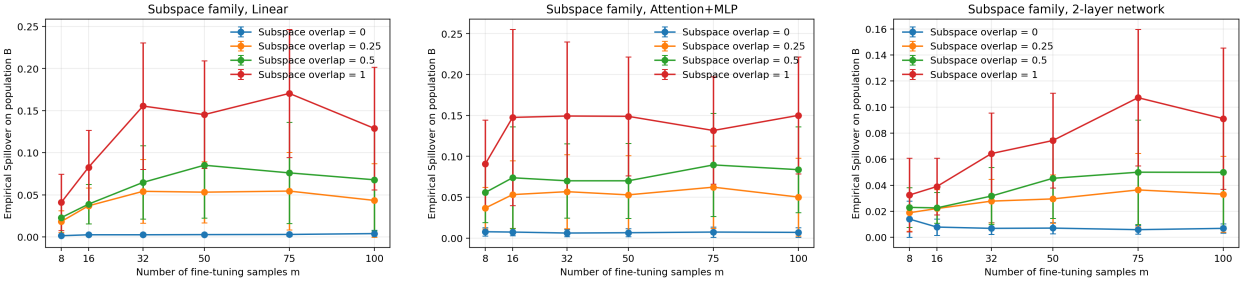

Figure 8: Subspace-overlap family: spillover on the evaluation population B as a function of the number of fine-tuning samples m, with one curve for each designed subspace-overlap value $\rho \in \{0, 0.25, 0.5, 1\}$. From left to right: linear model, attention+MLP model, and two-layer network. Error bars are across trials.