# LEARNING STYLE, FORGETTING SEMANTICS: A CASE STUDY OF SFT AND RFT ON CLASSIFICATION TASKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Why does supervised fine-tuning (SFT) lead to more forgetting than reinforcement fine-tuning (RFT), even when all teacher demonstrations are semantically correct? We study this question on classification tasks where tokens within each semantic class express the same semantic answer in different styles. The tasks share an underlying semantic rule but differ in their prompt distributions and teachers' stylistic preferences. Using a tractable linear-softmax policy, we derive an exact decomposition of the updates into semantic and style components. We show that, at a common policy and prompt, SFT and RFT have parallel semantic updates but differ in their style dynamics. Starting from a policy with no within-class style preference, RFT with exact policy gradients preserves this symmetry, whereas SFT with a nonuniform teacher develops off-axis style drift along a nonzero task mean under population updates. We use this drift to establish a separation under explicit conditions: for population updates from a common perfectly fitted checkpoint, SFT forgetting admits a strictly positive lower bound over a finite training interval, while RFT retains zero semantic error. Simulations over task sequences support these theoretical predictions.

## 1 INTRODUCTION

Fine-tuning plays a central role in adapting pretrained foundation models to downstream tasks and enhancing their task-specific capabilities (Howard & Ruder, 2018; Devlin et al., 2019; Raffel et al., 2020). For example, instruction tuning on diverse demonstrations improves language models' ability to follow instructions and generalize to unseen tasks (Ouyang et al., 2022; Chung et al., 2024); reward-based fine-tuning improves mathematical and coding performance by optimizing model-generated responses against verifiable outcomes (Shao et al., 2024; Guo et al., 2025). However, as models learn new tasks, they may lose previously acquired capabilities, a phenomenon known as *forgetting* (Wang et al., 2024; Luo et al., 2025; Chen et al., 2026). Preserving existing capabilities while adapting to new tasks is therefore an important consideration, highlighting the need to understand when and why fine-tuning can degrade them.

Two widely used fine-tuning methods are supervised fine-tuning (SFT), which fits teacher demonstrations, and reinforcement fine-tuning (RFT), which updates the model using rewards for its own sampled responses. Empirical studies find that RFT often preserves existing capabilities better than SFT (Chen et al., 2026; Lai et al., 2026). This gap can arise even with correct supervision: Chu et al. (2025) observe that SFT on correct expert demonstrations degrades existing performance under input distribution shifts that preserve the task rules, whereas RFT from the same checkpoint improves it. Existing theories offer distributional explanations for such differences. Shenfeld et al. (2026b) identify an implicit bias of online RL toward the solution closest to the initial policy in KL divergence among those that solve the new task. Balasubramanian & Kasiviswanathan (2026) quantify how forward- and reverse-KL objectives affect the preservation of the old-task component in a two-component Gaussian mixture. These accounts, however, leave a specific mechanistic question unanswered: how does fitting a distribution of *correct* responses induce errors on previously mastered tasks?

The ParityMNIST experiments of Shenfeld et al. (2026b) make this issue concrete: the two SFT baselines, trained on different distributions over equally correct labels, exhibit different retention trade-offs, but both preserve prior-task performance less effectively than RFT. Related language-model studies show that SFT can teach response formats (Zhou et al., 2023), but that imitation can



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

accompany declines in factual accuracy (Gudibande et al., 2024). To understand this phenomenon, we isolate the role of these preferences in a classification case study, separating an answer's *semantic content* from its particular expression, which we call its *style*. For example, "4," "four," and "The answer is 4" are all correct responses to $2 + 2$, but a teacher may prefer one expression over the others. In our setting, SFT fits this preference, whereas RFT rewards semantic correctness regardless of style. This distinction motivates our central question:

*Can fitting the teacher's style preference alone cause semantic forgetting, even when every demonstration is correct?*

We study this question in a tractable classification model with a linear-softmax policy, where tasks share the same semantic rule but differ in their input distributions and teachers' stylistic preferences. For this policy class, we derive an exact decomposition of the updates into semantic and style components, allowing us to trace how style fitting induces errors on a previously mastered task. Under explicit conditions and population updates, we prove that SFT causes strictly positive semantic forgetting even when every demonstration is correct, whereas RFT preserves zero error.

Our main contributions are:

- **An exact decomposition of the update.** We split the policy parameters into a semantic channel and a style channel and characterize the update direction in each channel under SFT and RFT. The semantic updates are parallel, whereas the style updates differ: from a policy with no style preference, RFT keeps the style channel exactly zero throughout training, while SFT drifts along the task mean (Theorems 1 and 2).
- **A provable population retention gap between SFT and RFT.** Starting from a checkpoint that perfectly fits the previous task and has no style preference, we prove a positive lower bound on SFT forgetting under population updates when the next task has a nonzero prompt mean and the teacher's style preference is stronger in one semantic class than in the other, whereas RFT retains perfect semantic accuracy (Theorem 3).
- **A finite-sample bound on off-axis energy.** On top of the population-level analysis, we account for the additional effects of finite-sample fluctuations on the off-axis parameter energy of SFT and RFT. Our bounds cover both semantic and style components and show that RFT admits a smaller upper bound on the expected off-axis energy than SFT (Theorem 4), and the resulting bounds are consistent with our population-level analysis in the large-sample limit.

Together, these results suggest that fitting a teacher's style preference is by itself sufficient to cause semantic forgetting under SFT, even when the teacher is always correct, whereas RFT avoids it because learning from its own on-policy samples induces no style drift. In Section 4, we provide empirical evidence for this mechanism by simulating the model over multi-task sequences under population and finite-sample updates; the results illustrate the mechanisms in Theorems 2 and 3 and the finite-sample effects studied in Theorem 4. We also report ablation studies in Appendix E to examine the model's behavior when some of the theoretical conditions are not satisfied.

**Additional related work**: We provide a detailed discussion of more related work in Appendix A.

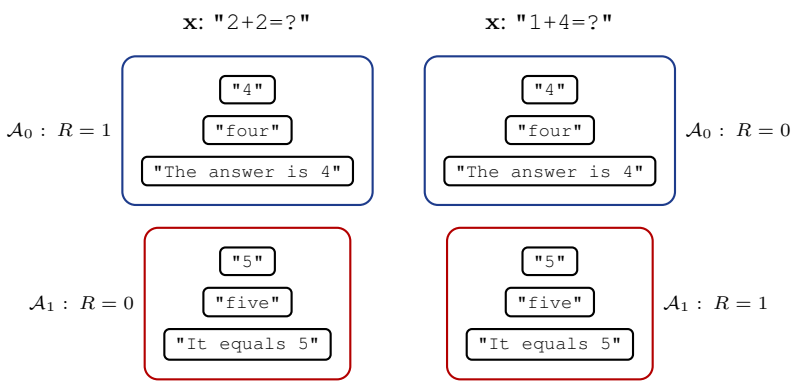
## 2 Preliminaries

### 2.1 Problem setup

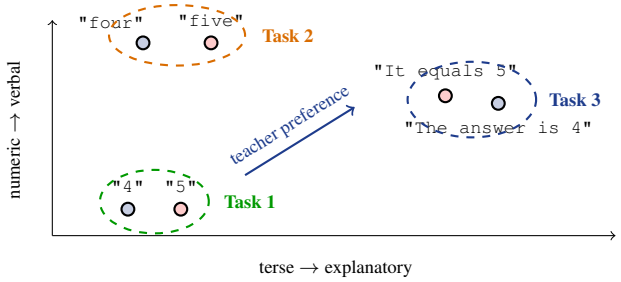
We take a case study on a family of tasks that share a binary-valued semantic rule but differ in their prompt distributions and in teachers' preferences among semantically equivalent outputs.

**Semantic classes.** Let $\mathbf{x} \in \mathbb{R}^p$ denote a prompt representation and $\mathcal{A}$ a finite output vocabulary. A map $\sigma : \mathcal{A} \to \{0, 1\}$ defines two *semantic classes*, $\mathcal{A}_b = \sigma^{-1}(b)$, with $|\mathcal{A}_0| = |\mathcal{A}_1| = M \geq 2$. Tokens within a class express the same semantic answer in different styles. A unit direction $\mathbf{u} \in \mathbb{R}^p$, shared by all tasks, determines the correct class and reward:

$$b^\star(\mathbf{x}) = \mathbf{1}\{\langle \mathbf{u}, \mathbf{x} \rangle > 0\}, \qquad R(a, \mathbf{x}) = \mathbf{1}\{\sigma(a) = b^\star(\mathbf{x})\}. \tag{1}$$




**(a) Semantic classes**




**(b) Style space**

Figure 1: **Semantic classes and expression styles.** *(a)* The same token space is partitioned into semantic classes $\mathcal{A}_0$ and $\mathcal{A}_1$. The prompt determines which class is rewarded; all tokens within a class receive the same reward. Here, $\mathbf{x}$ denotes the prompt's vector representation. *(b)* Teachers favor different expression styles across tasks.

Thus each prompt rewards exactly one class, regardless of expression style. Figure 1 illustrates this relationship. For RFT, we assume access to an accurate semantic reward signal, which may be provided by human correctness judgments or an automatic verifier.

*Notations.* Throughout our analysis, we refer to components parallel to the shared semantic direction $\mathbf{u}$ as *on-axis* and those orthogonal to it as *off-axis*. We write $\mathbf{v}^{\perp} := \mathbf{v} - \langle \mathbf{v}, \mathbf{u} \rangle \mathbf{u}$ for vectors and apply this projection rowwise to matrices, i.e., $\mathbf{V}^{\perp} := \mathbf{V}(\mathbf{I}_p - \mathbf{u}\mathbf{u}^{\top})$.

**Tasks.** Each task $t$ specifies a prompt distribution $\mathbf{x} \sim N(\boldsymbol{\mu}_t, \mathbf{I}_p)$, and an SFT teacher. The teacher always selects a token from the correct semantic class, but may favor particular styles of expressing the correct answer. For each task $t$ and semantic class $b \in \{0, 1\}$, the teacher has a fixed within-class preference distribution $\rho_{t,b}$, which may differ across tasks. The teacher distribution is

$$q_t(a \mid \mathbf{x}) = \begin{cases} \rho_{t,b^\star(\mathbf{x})}(a), & a \in \mathcal{A}_{b^\star(\mathbf{x})}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Thus the teacher is always semantically correct. Throughout the analysis, we assume $\langle \boldsymbol{\mu}_t, \mathbf{u} \rangle = 0$ for every task $t$, which allows task-specific prompt shifts while keeping the semantic coordinate $\langle \mathbf{u}, \mathbf{x} \rangle \sim N(0, 1)$, therefore isolating the role of style drift in semantic forgetting.

**Policy.** We use a linear-softmax policy, a standard parameterization in policy-gradient reinforcement learning (Sutton et al., 1999; Kakade, 2001; Agarwal et al., 2021). This linear-softmax form is also used for classification heads in supervised fine-tuning of pretrained language models (Radford et al., 2018; Devlin et al., 2019); our analysis holds the prompt representations fixed. The policy is

parameterized by $\mathbf{W} \in \mathbb{R}^{2M \times p}$, with rows $\mathbf{w}_a$:

$$\pi_{\mathbf{W}}(a \mid \mathbf{x}) = \mathrm{softmax}(\mathbf{W}\mathbf{x})_a = \frac{e^{\langle \mathbf{w}_a, \mathbf{x} \rangle}}{\sum_{a' \in \mathcal{A}} e^{\langle \mathbf{w}_{a'}, \mathbf{x} \rangle}}. \tag{3}$$

The policy outputs the highest-scoring token, with semantic class $\hat{b}_{\mathbf{W}}(\mathbf{x}) = \sigma\big(\arg\max_a \langle \mathbf{w}_a, \mathbf{x} \rangle\big)$.

Under this setup, SFT learns to match the teacher's distribution over individual tokens, whereas RFT optimizes a reward that depends only on the semantic class. We next specify the corresponding learning rules.

### 2.2 Learning rules

On task $t$, SFT aims to match the teacher's token distribution by minimizing cross-entropy, whereas RFT aims to maximize the expected semantic reward. For a prompt $\mathbf{x}$, SFT and RFT minimize the following losses, respectively:

$$\mathcal{L}_{\mathrm{SFT}}(\mathbf{W}; \mathbf{x}) = -\mathbb{E}_{a^\star \sim q_t(\cdot \mid \mathbf{x})}\big[\log \pi_{\mathbf{W}}(a^\star \mid \mathbf{x})\big], \tag{4}$$

$$\mathcal{L}_{\mathrm{RFT}}(\mathbf{W}; \mathbf{x}) = -\mathbb{E}_{a' \sim \pi_{\mathbf{W}}(\cdot \mid \mathbf{x})}\big[R(a', \mathbf{x})\big]. \tag{5}$$

Both methods process one prompt at a time with step size $\eta$, starting from $\mathbf{W} = \mathbf{0}$. SFT draws a demonstration $a^\star \sim q_t(\cdot \mid \mathbf{x})$ and takes a stochastic gradient step on (4); RFT takes a stochastic gradient step on (5) using its *exact* policy gradient.[1] For every $a \in \mathcal{A}$,

$$\begin{aligned} \mathbf{w}_a^+ &= \mathbf{w}_a + \eta\big(\mathbf{1}\{a = a^\star\} - \pi_{\mathbf{W}}(a \mid \mathbf{x})\big)\mathbf{x} && \text{(SFT)}, \\ \mathbf{w}_a^+ &= \mathbf{w}_a + \eta\, \pi_{\mathbf{W}}(a \mid \mathbf{x})\big(R(a, \mathbf{x}) - \bar{R}_{\mathbf{W}}(\mathbf{x})\big)\mathbf{x} && \text{(RFT)}, \end{aligned} \tag{6}$$

where $\bar{R}_{\mathbf{W}}(\mathbf{x}) = \sum_{a' \in \mathcal{A}} \pi_{\mathbf{W}}(a' \mid \mathbf{x}) R(a', \mathbf{x})$ is the expected reward. A derivation of these update rules is given in Lemma B.1. For task $t$, we write the *population* versions of the updates in (6) in terms of the parameter matrix $\mathbf{W}$:

$$\begin{aligned} \mathbf{W}^+ &= \mathbf{W} + \eta\, \mathbb{E}_t\big[(q_t(\cdot \mid \mathbf{x}) - \pi_{\mathbf{W}}(\cdot \mid \mathbf{x}))\mathbf{x}^\top\big] && \text{(SFT)}, \\ \mathbf{W}^+ &= \mathbf{W} + \eta\, \mathbb{E}_t\big[\pi_{\mathbf{W}}(\cdot \mid \mathbf{x})(R(\cdot, \mathbf{x}) - \bar{R}_{\mathbf{W}}(\mathbf{x}))\mathbf{x}^\top\big] && \text{(RFT)}. \end{aligned} \tag{7}$$

### 2.3 Continual learning

We train the policy sequentially on tasks $t = 1, \ldots, T$ and denote its parameters after task $t$ by $\mathbf{W}_t$. We evaluate both its final performance across tasks and its retention of earlier tasks.

For task $t$, we define the semantic accuracy and error under its prompt distribution as

$$a_t(\mathbf{W}) = \mathbb{P}_{\mathbf{x} \sim N(\boldsymbol{\mu}_t, \mathbf{I}_p)}\big(\hat{b}_{\mathbf{W}}(\mathbf{x}) = b^\star(\mathbf{x})\big), \qquad \mathcal{L}_t(\mathbf{W}) = 1 - a_t(\mathbf{W}).$$

Following standard practice in continual learning (Chaudhry et al., 2018; 2019; Lin et al., 2023), we evaluate the final checkpoint $\mathbf{W}_T$ using the average generalization error across all tasks and the average forgetting over earlier tasks:

$$\mathcal{G}^{(T)} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_t(\mathbf{W}_T), \qquad \mathcal{F}^{(T)} = \frac{1}{T-1}\sum_{t=1}^{T-1}\Big(\max_{1 \le l \le T-1} a_t(\mathbf{W}_l) - a_t(\mathbf{W}_T)\Big). \tag{8}$$

For each earlier task, forgetting compares its final accuracy with its best accuracy over all preceding task checkpoints. Negative values of $\mathcal{F}^{(T)}$ indicate positive backward transfer.

## 3 Main Results

In this section, we compare SFT and RFT by decomposing the policy parameters into semantic and style channels (Section 3.1). We first show that, at the same policy and prompt, the two methods

[1] Here, "exact" means that we compute the expected reward by summing over all actions under the current policy and take its exact gradient, while prompt sampling remains stochastic. We discuss the sampled REINFORCE updates in Appendix D.1.

update the semantic channel in the same direction (Theorem 1). Despite this alignment, their style dynamics differ: starting from a policy uniform within each class, RFT preserves zero style deviation, whereas SFT develops population style drift along a nonzero task mean when the teacher has within-class style preferences (Theorem 2). We then show that this style drift can produce a strict separation in forgetting between SFT and RFT at the population level (Theorem 3). Finally, we complement the population analysis with finite-sample bounds on the off-axis drifting energy of both methods under fresh sampling (Theorem 4).

### 3.1 Semantic–style decomposition

We decompose the policy parameters into a *semantic channel* $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^{(0)}, \boldsymbol{\alpha}^{(1)})$, which represents the mean weight associated with each semantic class, and a *style channel* $\boldsymbol{\beta} = (\boldsymbol{\beta}_a)_{a\in\mathcal{A}}$, which captures the policy's preferences among semantically equivalent tokens within each class. Specifically, for $b \in \{0, 1\}$,

$$\boldsymbol{\alpha}^{(b)} = \frac{1}{M}\sum_{a\in\mathcal{A}_b} \mathbf{w}_a, \qquad \boldsymbol{\beta}_a = \mathbf{w}_a - \boldsymbol{\alpha}^{(\sigma(a))}, \qquad \text{so } \sum_{a\in\mathcal{A}_b} \boldsymbol{\beta}_a = \mathbf{0}. \tag{9}$$

Each token's score then splits as

$$\langle \mathbf{w}_a, \mathbf{x}\rangle = \langle \boldsymbol{\alpha}^{(\sigma(a))}, \mathbf{x}\rangle + \langle \boldsymbol{\beta}_a, \mathbf{x}\rangle.$$

Using this decomposition, the policy (3) can then be written as

$$\pi_{\mathbf{W}}(a \mid \mathbf{x}) = \frac{\exp\big(\langle \boldsymbol{\alpha}^{(\sigma(a))}, \mathbf{x}\rangle + \langle \boldsymbol{\beta}_a, \mathbf{x}\rangle\big)}{\sum_{b=0}^{1}\sum_{a'\in\mathcal{A}_b} \exp\big(\langle \boldsymbol{\alpha}^{(b)}, \mathbf{x}\rangle + \langle \boldsymbol{\beta}_{a'}, \mathbf{x}\rangle\big)}. \tag{10}$$

Let $\Delta\boldsymbol{\alpha} := \boldsymbol{\alpha}^{(1)} - \boldsymbol{\alpha}^{(0)}$. For any $\mathbf{W}$, the decision rule is

$$\begin{aligned} D_{\mathbf{W}}(\mathbf{x}) &= \langle \Delta\boldsymbol{\alpha}, \mathbf{x}\rangle + \xi_{\boldsymbol{\beta}}(\mathbf{x}), \qquad \hat{b}_{\mathbf{W}}(\mathbf{x}) = \mathbf{1}\{D_{\mathbf{W}}(\mathbf{x}) > 0\}, \\ \xi_{\boldsymbol{\beta}}(\mathbf{x}) &= \max_{a\in\mathcal{A}_1}\langle \boldsymbol{\beta}_a, \mathbf{x}\rangle - \max_{a\in\mathcal{A}_0}\langle \boldsymbol{\beta}_a, \mathbf{x}\rangle. \end{aligned} \tag{11}$$

Thus, a within-class style preference can change the semantic prediction when $\xi_{\boldsymbol{\beta}}(\mathbf{x})$ reverses the sign of $\langle \Delta\boldsymbol{\alpha}, \mathbf{x}\rangle$. If $\Delta\boldsymbol{\alpha}, \boldsymbol{\beta}$ are on-axis, i.e., $\Delta\boldsymbol{\alpha} = \gamma\mathbf{u}$ and $\boldsymbol{\beta}_a = c_a\mathbf{u}$ for every token $a$, the policy's decision boundary coincides with the true semantic boundary whenever $\gamma > 2\max_a |c_a|$ (Lemma B.2). In particular, for a class-uniform policy ($\boldsymbol{\beta} = \mathbf{0}$), this decision boundary is determined solely by $\Delta\boldsymbol{\alpha}$. We next compare how SFT and RFT update these two channels, beginning with the semantic component.

### 3.2 The semantic channel: a shared update direction

Define the cumulative probability assigned to semantic class $b$ by

$$P_b(\mathbf{W}, \mathbf{x}) := \sum_{a\in\mathcal{A}_b} \pi_{\mathbf{W}}(a \mid \mathbf{x}). \tag{12}$$

We first show that SFT and RFT update the semantic channel in the same direction.

**Theorem 1** (Shared Semantic direction)**.** *Under the setup of Section 2.1, the following hold.*

(i) (per-sample identity) *For any* $\mathbf{W}$ *and prompt* $\mathbf{x}$*, the semantic channel moves as*

$$\begin{aligned} \Delta\boldsymbol{\alpha}^{+}_{\mathrm{SFT}} - \Delta\boldsymbol{\alpha} &= \frac{2\eta}{M}\big(2b^\star(\mathbf{x}) - 1\big)\big(1 - P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x})\big)\mathbf{x}, \\ \Delta\boldsymbol{\alpha}^{+}_{\mathrm{RFT}} - \Delta\boldsymbol{\alpha} &= P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x}) \cdot \big(\Delta\boldsymbol{\alpha}^{+}_{\mathrm{SFT}} - \Delta\boldsymbol{\alpha}\big). \end{aligned} \tag{13}$$

(ii) (population direction) *If* $\boldsymbol{\beta} = \mathbf{0}$ *and* $\Delta\boldsymbol{\alpha} = c\mathbf{u}$ *for some* $c \in \mathbb{R}$*, then the task-t population semantic steps are parallel to* $\mathbf{u}$*:*

$$\begin{aligned} \mathbb{E}_t\big[\Delta\boldsymbol{\alpha}^{+}_{\mathrm{SFT}} - \Delta\boldsymbol{\alpha}\big] &= \frac{2\eta}{M}\sqrt{\frac{2}{\pi}}\left(\int_0^\infty \frac{z e^{-z^2/2}}{1 + e^{cz}}\,\mathrm{d}z\right)\mathbf{u}, \\ \mathbb{E}_t\big[\Delta\boldsymbol{\alpha}^{+}_{\mathrm{RFT}} - \Delta\boldsymbol{\alpha}\big] &= \frac{2\eta}{M}\sqrt{\frac{2}{\pi}}\left(\int_0^\infty \frac{z e^{-z^2/2} e^{cz}}{(1 + e^{cz})^2}\,\mathrm{d}z\right)\mathbf{u}. \end{aligned} \tag{14}$$

The proof of Theorem 1 is in Appendix C.1.

In particular, at a fixed policy and prompt, the SFT semantic update in (13) does not depend on which style token the teacher sampled, while the RFT update is rescaled by the probability assigned to the correct class. Moreover, both coefficients in (14) are strictly positive, so both population updates increase the semantic margin along $\mathbf{u}$. Despite this shared semantic update direction, the two methods differ in how they update the style channel, as we show next.

### 3.3 The style channel: RFT freezes, SFT drifts

We next examine whether SFT and RFT preserve uniformity within each semantic class. The following theorem shows that RFT preserves zero style deviation, whereas SFT can develop population style drift when the teacher has within-class preferences.

**Theorem 2** (Style dynamics)**.** *Under the setup of Section 2.1, the following hold.*

(i) (RFT: style invariance) *The RFT updates (6) preserve $\boldsymbol{\beta} \equiv \mathbf{0}$.*

(ii) (SFT: style drift) *For any $\mathbf{W}$ and prompt $\mathbf{x}$, the SFT update (6) gives*

$$\boldsymbol{\beta}_a^+ - \boldsymbol{\beta}_a = \eta \left[ \mathbf{1}\{a = a^\star\} - \pi_{\mathbf{W}}(a \mid \mathbf{x}) - \frac{\mathbf{1}\{\sigma(a) = b^\star(\mathbf{x})\} - P_{\sigma(a)}(\mathbf{W}, \mathbf{x})}{M} \right] \mathbf{x}. \quad (15)$$

*If $\boldsymbol{\beta} = \mathbf{0}$, the task-$t$ population SFT step (7) satisfies, for every $a \in \mathcal{A}_b$,*

$$\begin{aligned} \mathbb{E}_t[\boldsymbol{\beta}_a^+ - \boldsymbol{\beta}_a] &= \eta \Big(\rho_{t,b}(a) - \frac{1}{M}\Big) \left( \frac{1}{2}\boldsymbol{\mu}_t + \frac{2b-1}{\sqrt{2\pi}}\mathbf{u} \right), \\ \text{so that} \qquad \mathbb{E}_t[\boldsymbol{\beta}_a^+ - \boldsymbol{\beta}_a]^\perp &= \frac{\eta}{2}\Big(\rho_{t,b}(a) - \frac{1}{M}\Big)\boldsymbol{\mu}_t. \end{aligned} \quad (16)$$

The proof of Theorem 2 is in Appendix C.2.

Theorem 2 gives us intuition for why SFT can be more prone to forgetting than RFT: starting from a class-uniform policy ($\boldsymbol{\beta} = \mathbf{0}$), SFT can develop off-axis population drift along the task mean by fitting the teacher's style preferences, while RFT preserves within-class symmetry. A special case is when either teachers have no within-class style preferences (i.e., $\rho_{t,b}(a) = 1/M$) or task means are zero (i.e., $\boldsymbol{\mu}_t = \mathbf{0}$), in which case both methods avoid off-axis population style drift. When drift does arise, it can shift the semantic decision boundary through $\xi_{\boldsymbol{\beta}}$ in (11). In Section 3.4, we show that this mechanism causes forgetting under explicit conditions.

### 3.4 From Style Drift to Semantic Error

We now compare how SFT and RFT retain a previously learned task when trained on the next one. Both methods start from the same checkpoint, which perfectly classifies the previous task and whose policy is uniform within each semantic class. The following theorem establishes a finite-time separation in semantic error on both the previous and current tasks.

**Theorem 3** (Population error separation from a perfectly fitted checkpoint)**.** *Under the setup of Section 2.1, suppose that:*

(a) (Common start) *SFT and RFT start from the same checkpoint $\mathbf{W}_1$, which fits the previous task perfectly and has no style preference:*

$$\mathcal{L}_1(\mathbf{W}_1) = 0, \qquad \boldsymbol{\beta}(\mathbf{W}_1) = \mathbf{0}.$$

(b) (Next task) *The next task has a nonzero mean $\boldsymbol{\mu}_2 \neq \mathbf{0}$, and its teacher's style preference is stronger in one semantic class than in the other, after relabeling the two classes if necessary: $\Delta_\rho := \max_{a \in \mathcal{A}_1} \rho_{2,1}(a) - \max_{a \in \mathcal{A}_0} \rho_{2,0}(a) > 0$.*

(c) (Training) *Each method runs $k_2 \geq 1$ population updates (7) on task 2 with step size $\eta_2 > 0$; write $\tau_2 := k_2\eta_2$ for the effective training time and $\mathbf{W}_2$ for the resulting checkpoint.*

*Then:*

(i) (RFT preserves perfect accuracy on both tasks) *For every $\eta_2 > 0$ and $k_2 \geq 1$,*

$$\mathcal{L}_1(\mathbf{W}_2) = \mathcal{L}_2(\mathbf{W}_2) = 0.$$

(ii) (SFT incurs semantic error on both tasks) *There is a threshold $\tau_0 > 0$, given in (C.34) that depends only on $\|\boldsymbol{\mu}_2\|$, $\Delta_\rho$, and $M$, such that for every $0 < \tau_2 \leq \tau_0$ and $r \in \{1, 2\}$,*

$$\mathcal{L}_r(\mathbf{W}_2) \geq \int_0^\infty \phi(t)\Phi\left(\frac{\langle \boldsymbol{\mu}_r, \boldsymbol{\mu}_2\rangle}{\|\boldsymbol{\mu}_2\|} - \frac{4(\|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\| + 2\tau_2)t}{\|\boldsymbol{\mu}_2\|\Delta_\rho\tau_2}\right) \mathrm{d}t > 0,$$

*where $\phi$ and $\Phi$ denote the standard Gaussian density and distribution function, respectively.*

The proof of Theorem 3 is in Appendix C.3.

Theorem 3(ii) implies that SFT introduces errors on two tasks that were initially classified perfectly. The lower bound measures the probability of a region where style drift overturns the correct semantic prediction. This error region grows with stronger teacher's preference $\Delta_\rho$, smaller initial semantic margin $\|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\|$, or longer training time $\tau_2$. Old-task prompts are more likely to fall into this region when $\boldsymbol{\mu}_1$ has a larger projection onto the direction of the new task mean $\boldsymbol{\mu}_2$.

Together with (8), these error bounds imply strictly positive forgetting and generalization error for SFT when $0 < \tau_2 \leq \tau_0$, whereas both metrics remain zero for RFT. This separation is, however, a finite-time phenomenon: under continued population SFT updates with a sufficiently small step size, both task errors converge to zero as $\tau_2 \to \infty$, as shown in Appendices D.2 and D.3.

### 3.5 Finite-sample off-axis energy

The separation result in Theorem 3 is established under population updates. In finite-sample training, prompt sampling can introduce additional deviations from the shared semantic direction $\mathbf{u}$. In the next theorem, we provide a bound on the expected off-axis energy $\mathbb{E}\|\mathbf{W}^\perp\|_F^2$ of both SFT and RFT.

**Theorem 4** (Off-axis parameter energy of SFT and RFT)**.** *Under the setup of Section 2.1, let $t(s)$ denote the task index at step $s$. Run the SFT and RFT updates (6) from $\mathbf{W}(0) = \mathbf{0}$ with constant step size $\eta > 0$. At each step $s$, independently draw $\mathbf{x}_s \sim N(\boldsymbol{\mu}_{t(s)}, \mathbf{I}_p)$, with a fresh teacher demonstration for SFT. For every $N \geq 0$, both methods satisfy*

$$\|\mathbf{W}^\perp(N)\|_F^2 = \|\boldsymbol{\beta}^\perp(N)\|_F^2 + \frac{M}{2}\|\Delta\boldsymbol{\alpha}^\perp(N)\|^2, \tag{17}$$

*and the following bounds hold:*

(i) (RFT)

$$\mathbb{E}\|\mathbf{W}_{\mathrm{RFT}}^\perp(N)\|_F^2 \leq \frac{\eta^2}{8M}\left(N(p-1) + \sum_{s=0}^{N-1}\|\boldsymbol{\mu}_{t(s)}\|^2\right). \tag{18}$$

(ii) (SFT)

$$\mathbb{E}\|\mathbf{W}_{\mathrm{SFT}}^\perp(N)\|_F^2 \leq 2\eta^2\left(\sqrt{N(p-1)} + \sum_{s=0}^{N-1}\|\boldsymbol{\mu}_{t(s)}\|\right)^2. \tag{19}$$

The proof of Theorem 4 is in Appendix C.4.

Theorem 4 shows that, even under finite sampling, RFT admits a smaller upper bound on the expected off-axis energy than SFT. The following remark discusses how these bounds reflect sampling effects and relate to the population dynamics.

*Remark* 1 (Sampling effects and population drift). We discuss the following cases as $\eta \to 0$ with effective training time $\tau = N\eta$ fixed. This keeps the cumulative step size unchanged while increasing the number of independent samples.

*(i) Zero task means.* If $\boldsymbol{\mu}_t = \mathbf{0}$ for all tasks, the RFT and SFT bounds reduce to $\tau\eta(p-1)/(8M)$ and $2\tau\eta(p-1)$, respectively. Both bounds vanish in the limit $\eta \to 0$. This is consistent with Lemma B.5, which shows that the expected parameters remain on the semantic axis for both methods when all task means are zero.

*(ii) Bounded nonzero task means.* The RFT bound (18) remains $\mathcal{O}(\tau\eta)$ and vanishes, consistent with its on-axis population dynamics (Theorems 1(ii) and 2(i)). The SFT bound (19), however, does not vanish, since the term $\eta \sum_{s=0}^{N-1} \|\boldsymbol{\mu}_{t(s)}\|$ remains nonzero. This is consistent with the population style drift dynamics identified in Theorem 2(ii).

Together, these results show how fitting style can interfere with shared semantic knowledge even when all demonstrations are semantically correct. At a common policy and prompt, SFT and RFT share a semantic update direction, but RFT with exact policy gradients preserves an initially zero style channel, whereas SFT can develop off-axis population style drift that causes finite-time forgetting. In addition, when sampling is finite, it introduces additional off-axis fluctuations, for which RFT admits a smaller energy upper bound than SFT. We examine these mechanisms empirically in section 4.

## 4 Experiments

The experiments simulate the model of Section 2.1 and examine three aspects of the learning dynamics: (1) RFT maintains zero off-axis energy under population updates (Theorem 1 and 2(i)); (2) nonzero task means and nonuniform teacher preferences induce SFT population off-axis style drift and can lead to semantic forgetting (Theorems 2(ii) and 3); (3) finite-sample training produces additional off-axis parameter fluctuations, even when population off-axis drift is absent; we measure the final off-axis energy, whose expectation under Gaussian prompts is bounded in Theorem 4.

**Teacher parameterization.** To construct the teachers for the simulations, we assign each token a fixed style embedding $\mathbf{h}_a \in \mathbb{R}^d$ describing how its semantic answer is expressed. A task- and class-dependent teacher preference vector $\mathbf{v}_{t,b} \in \mathbb{R}^d$ determines the within-class distribution:

$$\rho_{t,b}(a) = \frac{e^{\langle \mathbf{v}_{t,b},\, \mathbf{h}_a \rangle}}{\sum_{a' \in \mathcal{A}_b} e^{\langle \mathbf{v}_{t,b},\, \mathbf{h}_{a'} \rangle}}, \qquad a \in \mathcal{A}_b. \tag{20}$$

Expressions with larger compatibility scores $\langle \mathbf{v}_{t,b},\, \mathbf{h}_a \rangle$ receive higher probability; $\mathbf{v}_{t,b} = \mathbf{0}$ gives a uniform teacher within the class. Figure 1 illustrates such teacher preferences schematically.

**Setup.** We train SFT and RFT from $\mathbf{W} = \mathbf{0}$ on $T = 8$ sequential tasks, with prompt dimension $p = 24$, two semantic classes of $M = 6$ tokens each, and style dimension $d = 3$. We draw $\mathbf{u}$ uniformly on the unit sphere, $M$ style embeddings $\mathbf{h}_1, \ldots, \mathbf{h}_M \sim N(\mathbf{0}, \mathbf{I}_3/3)$ assigned to the tokens of each class, and teacher preference vectors $\mathbf{v}_{t,b} \sim N(\mathbf{0}, 36\,\mathbf{I}_3)$.[2] Prompts are Gaussian as in Section 2.1, with task means $\boldsymbol{\mu}_t \sim N(\mathbf{0}, \mathbf{I}_p - \mathbf{u}\mathbf{u}^\top)$, so that $\boldsymbol{\mu}_t \perp \mathbf{u}$. These task parameters are sampled once and held fixed across runs.

**Training and evaluation.** For finite-sample runs, we sample $n = 25$ prompts per task and process them in a single pass in random order, averaging over 50 seeds that control the training prompts, demonstrations, and their order. For population runs, we approximate (7) by averaging updates over 10,000 fresh prompts at each step, take 25 steps per task to match the number of finite-sample updates, and average over 10 seeds that control the Monte Carlo prompts. In both settings, we sweep $\eta \in [10^{-4}, 0.5]$ on a log-spaced grid and evaluate each task on a fixed set of 1,000 held-out prompts from its own distribution. We report generalization error $\mathcal{G}^{(T)}$, forgetting $\mathcal{F}^{(T)}$, and the final off-axis parameter energy $\|\mathbf{W}_T^{\perp}\|_F^2$, all averaged across seeds.

**Results.** The main results are shown in Figure 2. In panel (a), population RFT maintains zero generalization error and forgetting throughout the learning-rate sweep, whereas population SFT exhibits strictly positive values at every tested learning rate. The persistence of this separation under population updates is consistent with off-axis style drift driven by nonzero task means and nonuniform teacher preferences (Theorem 3). In the finite-sample runs, prompt sampling introduces off-axis fluctuations that also raise RFT's error (Theorem 4), yet RFT still achieves lower generalization error and forgetting than SFT at every matched learning rate. Panel (b) gives a complementary

[2]The variance $1/d$ normalizes the style embeddings to unit scale, $\mathbb{E}\|\mathbf{h}_j\|^2 = 1$, so that the within-class logits $\langle \mathbf{v}_{t,b},\, \mathbf{h}_a \rangle$ in (20) have standard deviation controlled by $\mathbf{v}_{t,b}$ alone, here about 6. This makes the teachers strongly nonuniform while leaving their preferences varied across tasks and classes.

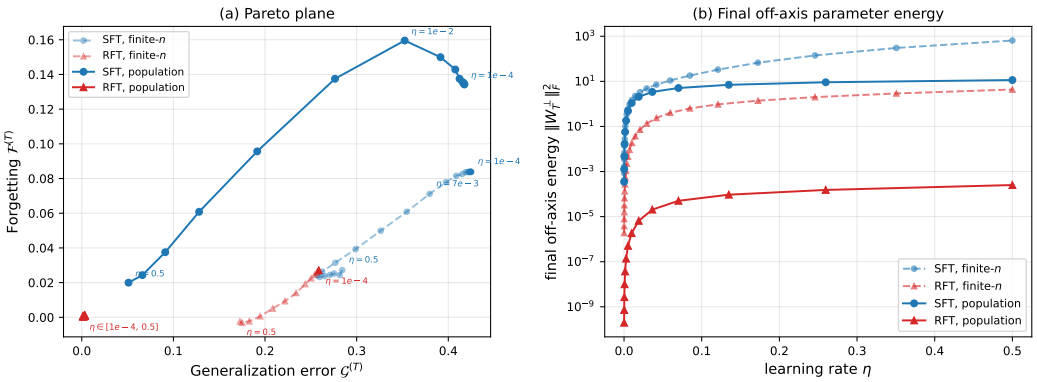



Figure 2: **SFT and RFT under population and finite-sample updates.** *(a)* Generalization–forgetting plane, for finite-sample (dashed, faint) and population (solid) updates; selected learning rates are annotated. *(b)* Final off-axis parameter energy $\|\mathbf{W}_T^{\perp}\|_F^2$, averaged across seeds, versus learning rate, with a logarithmic energy axis.

view: population SFT develops off-axis parameter energy, while population RFT remains close to zero (Theorem 1 and Theorem 2), with small residual energy from Monte Carlo prompt sampling. Finite-sample updates produce additional off-axis energy, which grows with the learning rate and is one to two orders of magnitude larger for SFT than for RFT. The observed energy gap qualitatively agrees with the smaller upper bound available for RFT in Theorem 4.

**Control and ablation experiments.** We provide additional experiments in Appendix E to isolate the roles of task means and teacher style preferences and to relax the assumptions behind RFT's style invariance. With either zero task means (Figure E.1) or uniform teachers (Figure E.2), the population SFT–RFT gap vanishes, while finite-sample SFT still accumulates more off-axis energy and higher generalization error than RFT; replacing Gaussian prompts by Laplace prompts leaves the results nearly unchanged (Figure E.3). We also compare the methods under corrupted rewards and under REINFORCE with finitely many sampled actions, and find that population RFT retains low forgetting in both cases; see Figure E.4 for details.

## 5 Conclusion

This work provides intuition on why supervised fine-tuning forgets more than reinforcement fine-tuning, even when every demonstration is semantically correct, through a case study on classification tasks with a linear-softmax policy. We show that semantically correct supervision can still induce forgetting when training also fits task-specific style preferences. In our model, fitting these preferences can shift the semantic decision boundary and disrupt previously learned predictions, even in the absence of sampling noise. RFT avoids this source of interference by preserving within-class symmetry under exact policy gradients from a symmetric initialization. The finite-sample bounds and simulations further characterize the off-axis deviations introduced by prompt sampling.

**Limitations.** Our theoretical analysis focuses on a tractable setting with a single-token linear-softmax model and a shared semantic rule across tasks. Future work could extend the analysis to more general settings such as nonlinear policies, multi-token outputs, and tasks with evolving semantic rules.

### AI USE STATEMENT

AI was used to polish the writing, assist in writing proofs, and implement methods for the experiments. The authors reviewed all AI-assisted work and take full responsibility for the final content.

### REPRODUCIBILITY STATEMENT

All theoretical results are proved in Appendices B and C. Experimental settings and evaluation procedures are described in Section 4. The code for reproducing all experiments is provided in the supplementary material.

## REFERENCES

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021. URL https://jmlr.org/papers/v22/19-736.html.

Krishnakumar Balasubramanian and Shiva Prasad Kasiviswanathan. A quantitative characterization of forgetting in post-training. *arXiv preprint arXiv:2603.12163*, 2026. doi: 10.48550/arXiv.2603.12163. URL https://arxiv.org/abs/2603.12163.

Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European Conference on Computer Vision (ECCV)*, pp. 532–547, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Arslan_Chaudhry__Riemannian_Walk_ECCV_2018_paper.html.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2510.18874.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10818–10838, 2025. URL https://proceedings.mlr.press/v267/chu25c.html.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. URL https://www.jmlr.org/papers/v25/23-0870.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Meng Ding, Kaiyi Ji, Di Wang, and Jinhui Xu. Understanding forgetting in continual learning with linear regression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10978–11001. PMLR, 2024. URL https://proceedings.mlr.press/v235/ding24c.html.

Itay Evron, Edward Moroshko, Gon Buzaglo, Maroun Khriesh, Badea Marjieh, Nathan Srebro, and Daniel Soudry. Continual learning in linear classification on separable data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine*

*Learning Research*, pp. 9440–9484. PMLR, 2023. URL https://proceedings.mlr.press/v202/evron23a.html.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4db16435e3a5a2ef3fc39b8f0d12498d-Abstract-Conference.html.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1031. URL https://aclanthology.org/P18-1031/.

Sham M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 14, pp. 1531–1538, 2001. URL https://papers.nips.cc/paper_files/paper/2001/hash/4b86abe48d358ecf194c56c69108433e-Abstract.html.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://arxiv.org/abs/1612.00796.

Song Lai, Haohan Zhao, Rong Feng, Changyi Ma, Wenzhuo Liu, Hongbo Zhao, Xi Lin, Dong Yi, Qingfu Zhang, Hongbin Liu, Gaofeng Meng, and Fei Zhu. Reinforcement fine-tuning naturally mitigates forgetting in continual post-training. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2507.05386.

Hongbo Li, Sen Lin, Lingjie Duan, Yingbin Liang, and Ness B. Shroff. Theory on mixture-of-experts in continual learning. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/17a234c91f746d9625a75cf8a8731ee2-Abstract-Conference.html.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21078–21100. PMLR, 2023. URL https://proceedings.mlr.press/v202/lin23f.html.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025. doi: 10.1109/TASLPRO.2025.3606231. URL https://arxiv.org/abs/2308.08747.

Arun Mallya and Svetlana Lazebnik. PackNet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7765–7773, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Mallya_PackNet_Adding_Multiple_CVPR_2018_paper.html.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2203.02155.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://www.jmlr.org/papers/v21/20-074.html.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Abstract.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *International Conference on Machine Learning (ICML)*, 2026a. URL https://arxiv.org/abs/2601.19897.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. In *International Conference on Learning Representations (ICLR)*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/618c95f4557c15b253fb0e6f548ea0c0-Abstract-Conference.html.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1999. URL https://proceedings.nips.cc/paper_files/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024. doi: 10.1109/TPAMI.2024.3367329. URL https://arxiv.org/abs/2302.00487.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. URL https://doi.org/10.1007/BF00992696.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995, 2017. URL https://proceedings.mlr.press/v70/zenke17a.html.

Zhihao Zhang, Qiaole Dong, Qi Zhang, Enyu Zhou, Jun Zhao, Zhiheng Xi, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Mingqi Wu, Yanwei Fu, Tao Ji, Tao Gui, Xuanjing Huang, and Kai Chen. Why reinforcement fine-tuning preserves prior knowledge better: A data perspective. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/pdf/3b5fda354275f890c9404d61311720aac059b79d.pdf.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html.

# A Literature Review

**Empirical findings on SFT and RFT.** Recent studies compare how supervised and reinforcement fine-tuning affect generalization and forgetting in foundation models. Chu et al. (2025) find better generalization under RL to changes in reasoning rules and visual conditions, while Lai et al. (2026) report less forgetting of earlier tasks and general capabilities under RFT across sequential multimodal tasks. Studies of on-policy and model-generated supervision further highlight the role of the supervision distribution. Chen et al. (2026) show that approximately on-policy supervision reduces forgetting under SFT. Zhang et al. (2026) link the retention benefits of SFT on filtered model-generated trajectories to the magnitude and direction of example-induced updates. In self-distillation fine-tuning, Shenfeld et al. (2026a) use a demonstration-conditioned teacher to supervise student-generated trajectories without an explicit reward function, reducing forgetting during skill learning.

**Distributional explanations of the SFT–RFT gap.** Shenfeld et al. (2026b) relate forgetting to the KL divergence from the initial policy and contrast SFT's fitting of an external supervision distribution with online RL's information-projection bias toward small policy changes. Building on the output-mixture simulations of Chen et al. (2026), Balasubramanian & Kasiviswanathan (2026) compare divergence-based post-training objectives by quantifying loss of old-component mass and drift within an old component. Their forward-KL objective fits new-mode data, whereas their reverse-KL objective fits a mixture that includes old behavior; the forgetting comparison therefore depends on both the objectives and their target distributions.

**Forgetting in continual-learning models.** Complementary theoretical work examines how task structure, optimization, and architecture affect forgetting. In overparameterized linear regression, Lin et al. (2023) derive expressions for expected forgetting and generalization error that reveal the effects of model size, task similarity, and task order. For sequential stochastic gradient descent in linear regression with a shared ground-truth parameter, Ding et al. (2024) quantify how task covariances, ordering, and step size affect forgetting. In linear classification with jointly separable tasks, Evron et al. (2023) connect weakly regularized training to sequential max-margin learning and show that forgetting can arise despite joint separability. In an overparameterized mixture-of-experts regression setting, Li et al. (2025) relate expert specialization and routing to forgetting and generalization.

**Explicit methods for mitigating forgetting.** Replay methods maintain access to earlier experience: Rolnick et al. (2019) combine on-policy learning with off-policy updates and behavioral cloning from a replay buffer to reduce forgetting in continual reinforcement learning. Parameter regularization penalizes changes to weights estimated to be important for past tasks. Elastic weight consolidation uses the diagonal Fisher information to weight these penalties (Kirkpatrick et al., 2017), while synaptic intelligence estimates parameter importance from contributions to loss reduction along the learning trajectory (Zenke et al., 2017). For parameter isolation, Mallya & Lazebnik (2018) iteratively prune a network, freeze weights allocated to earlier tasks, and train subsequent tasks using the remaining capacity. They provide complementary ways to control forgetting beyond the choice of SFT or RFT.

# B Supporting lemmas

**Lemma B.1** (SFT and RFT gradient updates)**.** *Fix a prompt* $\mathbf{x}$*. A stochastic-gradient step on* (4) *using* $a^\star \sim q_t(\cdot \mid \mathbf{x})$ *and an exact gradient step on* (5)*, both with step size* $\eta$*, yield the SFT and RFT updates in* (6)*, respectively.*

*Proof.* For SFT, given a sampled demonstration $a^\star \sim q_t(\cdot \mid \mathbf{x})$, define

$$\ell_{\mathrm{SFT}}(\mathbf{W}; \mathbf{x}, a^\star) = -\log \pi_{\mathbf{W}}(a^\star \mid \mathbf{x}).$$

Its expectation over the teacher is exactly (4); hence its gradient is an unbiased stochastic gradient of $\mathcal{L}_{\mathrm{SFT}}$. To compute its gradient, first expand the log-softmax as

$$\log \pi_{\mathbf{W}}(a' \mid \mathbf{x}) = \langle \mathbf{w}_{a'}, \mathbf{x} \rangle - \log \sum_{\widetilde{a} \in \mathcal{A}} \exp\big(\langle \mathbf{w}_{\widetilde{a}}, \mathbf{x} \rangle\big).$$

Differentiating with respect to $\mathbf{w}_a$ gives

$$\nabla_{\mathbf{w}_a} \log \pi_{\mathbf{W}}(a' \mid \mathbf{x}) = \mathbf{1}\{a = a'\}\mathbf{x} - \frac{\exp\big(\langle \mathbf{w}_a, \mathbf{x}\rangle\big)}{\sum_{\widetilde{a}\in\mathcal{A}} \exp\big(\langle \mathbf{w}_{\widetilde{a}}, \mathbf{x}\rangle\big)}\mathbf{x} = \big(\mathbf{1}\{a = a'\} - \pi_{\mathbf{W}}(a \mid \mathbf{x})\big)\mathbf{x}. \tag{B.1}$$

Therefore gradient descent on $\ell_{\mathrm{SFT}}$ gives (6).

For RFT, expand (5) as

$$\mathcal{L}_{\mathrm{RFT}}(\mathbf{W}; \mathbf{x}) = -\sum_{a'\in\mathcal{A}} \pi_{\mathbf{W}}(a' \mid \mathbf{x}) R(a', \mathbf{x}).$$

We have

$$\begin{aligned}
\nabla_{\mathbf{w}_a} \pi_{\mathbf{W}}(a' \mid \mathbf{x}) &= \nabla_{\mathbf{w}_a} \frac{\exp\big(\langle \mathbf{w}_{a'}, \mathbf{x}\rangle\big)}{\sum_{\widetilde{a}\in\mathcal{A}} \exp\big(\langle \mathbf{w}_{\widetilde{a}}, \mathbf{x}\rangle\big)} \\
&= \frac{\mathbf{1}\{a = a'\} \exp\big(\langle \mathbf{w}_{a'}, \mathbf{x}\rangle\big) \left(\sum_{\widetilde{a}\in\mathcal{A}} \exp\big(\langle \mathbf{w}_{\widetilde{a}}, \mathbf{x}\rangle\big)\right)}{\left(\sum_{\widetilde{a}\in\mathcal{A}} \exp\big(\langle \mathbf{w}_{\widetilde{a}}, \mathbf{x}\rangle\big)\right)^2}\mathbf{x} \\
&\quad - \frac{\exp\big(\langle \mathbf{w}_{a'}, \mathbf{x}\rangle\big) \exp\big(\langle \mathbf{w}_a, \mathbf{x}\rangle\big)}{\left(\sum_{\widetilde{a}\in\mathcal{A}} \exp\big(\langle \mathbf{w}_{\widetilde{a}}, \mathbf{x}\rangle\big)\right)^2}\mathbf{x} \\
&= \frac{\exp\big(\langle \mathbf{w}_{a'}, \mathbf{x}\rangle\big)}{\sum_{\widetilde{a}\in\mathcal{A}} \exp\big(\langle \mathbf{w}_{\widetilde{a}}, \mathbf{x}\rangle\big)} \left(\mathbf{1}\{a = a'\} - \frac{\exp\big(\langle \mathbf{w}_a, \mathbf{x}\rangle\big)}{\sum_{\widetilde{a}\in\mathcal{A}} \exp\big(\langle \mathbf{w}_{\widetilde{a}}, \mathbf{x}\rangle\big)}\right)\mathbf{x} \\
&= \pi_{\mathbf{W}}(a' \mid \mathbf{x})\big(\mathbf{1}\{a = a'\} - \pi_{\mathbf{W}}(a \mid \mathbf{x})\big)\mathbf{x}.
\end{aligned} \tag{B.2}$$

Consequently,

$$\begin{aligned}
-\nabla_{\mathbf{w}_a} \mathcal{L}_{\mathrm{RFT}} &= \sum_{a'\in\mathcal{A}} R(a', \mathbf{x}) \nabla_{\mathbf{w}_a} \pi_{\mathbf{W}}(a' \mid \mathbf{x}) \\
&= \sum_{a'\in\mathcal{A}} R(a', \mathbf{x}) \pi_{\mathbf{W}}(a' \mid \mathbf{x})\big(\mathbf{1}\{a = a'\} - \pi_{\mathbf{W}}(a \mid \mathbf{x})\big)\mathbf{x} \\
&= \pi_{\mathbf{W}}(a \mid \mathbf{x}) R(a, \mathbf{x})\mathbf{x} - \pi_{\mathbf{W}}(a \mid \mathbf{x}) \sum_{a'\in\mathcal{A}} \pi_{\mathbf{W}}(a' \mid \mathbf{x}) R(a', \mathbf{x})\mathbf{x} \\
&= \pi_{\mathbf{W}}(a \mid \mathbf{x})\big(R(a, \mathbf{x}) - \bar{R}_{\mathbf{W}}(\mathbf{x})\big)\mathbf{x}.
\end{aligned} \tag{B.3}$$

Gradient descent on $\mathcal{L}_{\mathrm{RFT}}$ therefore gives (6). □

**Lemma B.2** (Correct classification with on-axis channels)**.** *Suppose the channels in (9) satisfy $\Delta\boldsymbol{\alpha} = \gamma\mathbf{u}$ with $\gamma > 0$ and $\boldsymbol{\beta}_a = c_a\mathbf{u}$ for every $a \in \mathcal{A}$. Set $B := \max_{a\in\mathcal{A}} |c_a|$. If $\gamma > 2B$, then the decision rule (11) satisfies*

$$\hat{b}_{\mathbf{W}}(\mathbf{x}) = b^\star(\mathbf{x}) \qquad \textit{for every } \mathbf{x} \in \mathbb{R}^p.$$

*In particular, its decision boundary is $\{\mathbf{x} : \langle \mathbf{u}, \mathbf{x}\rangle = 0\}$, and $\mathcal{L}_t(\mathbf{W}) = 0$ for every task $t$.*

*Proof.* Write $z := \langle \mathbf{u}, \mathbf{x}\rangle$. Since $\langle \boldsymbol{\beta}_a, \mathbf{x}\rangle = c_a z$, (11) gives

$$D_{\mathbf{W}}(\mathbf{x}) = \gamma z + \xi_{\boldsymbol{\beta}}(\mathbf{x}), \qquad |\xi_{\boldsymbol{\beta}}(\mathbf{x})| \le 2 \max_{a\in\mathcal{A}} |c_a z| = 2B|z|.$$

For $z \neq 0$, it follows that

$$z D_{\mathbf{W}}(\mathbf{x}) \ge \gamma z^2 - |z|\,|\xi_{\boldsymbol{\beta}}(\mathbf{x})| \ge (\gamma - 2B) z^2 > 0.$$

Thus $D_{\mathbf{W}}(\mathbf{x})$ has the same sign as $z$. When $z = 0$, all style scores and $D_{\mathbf{W}}(\mathbf{x})$ vanish, so both (11) and (1) assign class 0. Hence the two rules agree everywhere and have the stated boundary. □

**Lemma B.3** (Gaussian projection decomposition)**.** *Under the setup of Section 2.1, fix a task $t$ and let $\mathbf{x} \sim N(\boldsymbol{\mu}_t, \mathbf{I}_p)$. Define*

$$z = \langle \mathbf{u}, \mathbf{x}\rangle, \qquad \boldsymbol{\xi}^\perp = \big(\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top\big)(\mathbf{x} - \boldsymbol{\mu}_t).$$

*Then*

$$z \sim N(0,1), \qquad \boldsymbol{\xi}^{\perp} \sim N(\mathbf{0}, \mathbf{I}_p - \mathbf{u}\mathbf{u}^\top), \qquad z \perp\!\!\!\perp \boldsymbol{\xi}^{\perp},$$

*and*

$$\mathbf{x} = z\mathbf{u} + \boldsymbol{\mu}_t + \boldsymbol{\xi}^{\perp}.$$

*Proof.* Let $\boldsymbol{\varepsilon} = \mathbf{x} - \boldsymbol{\mu}_t \sim N(\mathbf{0}, \mathbf{I}_p)$. Since $\langle \boldsymbol{\mu}_t, \mathbf{u} \rangle = 0$,

$$z = \mathbf{u}^\top \boldsymbol{\varepsilon}, \qquad \boldsymbol{\xi}^{\perp} = \big(\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top\big)\boldsymbol{\varepsilon}.$$

Thus $z$ and $\boldsymbol{\xi}^{\perp}$ are jointly Gaussian. Using $\|\mathbf{u}\| = 1$,

$$\mathrm{Var}(z) = \mathbf{u}^\top \mathbf{u} = 1, \qquad \mathrm{Cov}(\boldsymbol{\xi}^{\perp}) = \mathbf{I}_p - \mathbf{u}\mathbf{u}^\top,$$

and their cross-covariance is

$$\mathrm{Cov}(\boldsymbol{\xi}^{\perp}, z) = \big(\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top\big)\mathbf{u} = \mathbf{0}.$$

Joint Gaussianity therefore implies $z \perp\!\!\!\perp \boldsymbol{\xi}^{\perp}$. Finally,

$$\boldsymbol{\varepsilon} = \mathbf{u}\mathbf{u}^\top \boldsymbol{\varepsilon} + \big(\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top\big)\boldsymbol{\varepsilon} = z\mathbf{u} + \boldsymbol{\xi}^{\perp},$$

which gives the stated decomposition of $\mathbf{x}$. □

**Lemma B.4** (Vectors invariant under $G_{\mathbf{u}}$)**.** *Let $\mathbf{u} \in \mathbb{R}^p$ be nonzero and define*

$$G_{\mathbf{u}} = \{\mathbf{Q} \in \mathrm{O}(p) : \mathbf{Q}\mathbf{u} = \mathbf{u}\}. \tag{B.4}$$

*Then $\mathbf{Q}\mathbf{v} = \mathbf{v}$ for every $\mathbf{Q} \in G_{\mathbf{u}}$ if and only if $\mathbf{v} \in \mathrm{span}(\mathbf{u})$.*

*Proof.* First suppose $\mathbf{v} = c\mathbf{u}$. For every $\mathbf{Q} \in G_{\mathbf{u}}$,

$$\mathbf{Q}\mathbf{v} = c\mathbf{Q}\mathbf{u} = c\mathbf{u} = \mathbf{v},$$

which proves one direction.

Conversely, suppose $\mathbf{Q}\mathbf{v} = \mathbf{v}$ for every $\mathbf{Q} \in G_{\mathbf{u}}$. Decompose $\mathbf{v}$ as

$$\mathbf{v} = c\mathbf{u} + \mathbf{v}^{\perp}, \qquad c = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|^2}, \qquad \mathbf{v}^{\perp} \perp \mathbf{u}.$$

Suppose, toward a contradiction, that $\mathbf{v}^{\perp} \neq \mathbf{0}$. Define

$$\mathbf{e} = \frac{\mathbf{v}^{\perp}}{\|\mathbf{v}^{\perp}\|}, \qquad \mathbf{Q} = \mathbf{I}_p - 2\mathbf{e}\mathbf{e}^\top.$$

Because $\|\mathbf{e}\| = 1$,

$$\begin{aligned} \mathbf{Q}^\top \mathbf{Q} &= (\mathbf{I}_p - 2\mathbf{e}\mathbf{e}^\top)^2 \\ &= \mathbf{I}_p - 4\mathbf{e}\mathbf{e}^\top + 4\mathbf{e}(\mathbf{e}^\top \mathbf{e})\mathbf{e}^\top \\ &= \mathbf{I}_p. \end{aligned}$$

Hence $\mathbf{Q} \in \mathrm{O}(p)$. Since $\mathbf{e} \perp \mathbf{u}$, we also have

$$\mathbf{Q}\mathbf{u} = \mathbf{u} - 2\mathbf{e}\mathbf{e}^\top \mathbf{u} = \mathbf{u},$$

and therefore $\mathbf{Q} \in G_{\mathbf{u}}$. On the other hand,

$$\begin{aligned} \mathbf{Q}\mathbf{v}^{\perp} &= \mathbf{v}^{\perp} - 2\mathbf{e}\mathbf{e}^\top \mathbf{v}^{\perp} \\ &= \mathbf{v}^{\perp} - 2\mathbf{v}^{\perp} = -\mathbf{v}^{\perp}. \end{aligned}$$

It follows that

$$\mathbf{Q}\mathbf{v} = c\mathbf{Q}\mathbf{u} + \mathbf{Q}\mathbf{v}^{\perp} = c\mathbf{u} - \mathbf{v}^{\perp} \neq c\mathbf{u} + \mathbf{v}^{\perp} = \mathbf{v},$$

contradicting the assumption that every $\mathbf{Q} \in G_{\mathbf{u}}$ fixes $\mathbf{v}$. Thus $\mathbf{v}^{\perp} = \mathbf{0}$ and $\mathbf{v} \in \mathrm{span}(\mathbf{u})$. □

**Lemma B.5** (Rotation equivariance of zero-mean trajectories)**.** *Assume $\boldsymbol{\mu}_t = \mathbf{0}$ for every task and $\mathbf{W}(0) = \mathbf{0}$. Fix a task sequence and suppose each step uses a fresh prompt independent of the training history and, for SFT, a fresh teacher draw conditional on that prompt. Let $G_{\mathbf{u}}$ be as defined in (B.4); these are the orthogonal transformations that leave $\mathbf{u}$ fixed and act only on its orthogonal complement $\mathbf{u}^{\perp}$. Let $G_{\mathbf{u}}$ act rowwise on the parameter matrix, $(\mathbf{Q}\cdot\mathbf{W})_a = \mathbf{Q}\mathbf{w}_a$. Under either the SFT or RFT update, for every $\mathbf{Q} \in G_{\mathbf{u}}$ and every step $\tau$,*

$$\mathbf{Q}\cdot\mathbf{W}(\tau) \overset{d}{=} \mathbf{W}(\tau). \tag{B.5}$$

*Consequently, for every token $a$,*

$$\mathbb{E}[\mathbf{w}_a(\tau)], \quad \mathbb{E}[\Delta\boldsymbol{\alpha}(\tau)], \quad \mathbb{E}[\boldsymbol{\beta}_a(\tau)] \in \operatorname{span}(\mathbf{u}). \tag{B.6}$$

*Proof.* Let $\omega$ denote the random input used in one update. For SFT, write $\omega = (\mathbf{x}, a^\star)$ and define $\mathbf{Q} \star \omega = (\mathbf{Q}\mathbf{x}, a^\star)$; for RFT, write $\omega = \mathbf{x}$ and define $\mathbf{Q} \star \omega = \mathbf{Q}\mathbf{x}$. Since $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_p)$, the linear transformation rule for Gaussian vectors and $\mathbf{Q}\mathbf{Q}^\top = \mathbf{I}_p$ give

$$\mathbf{Q}\mathbf{x} \sim N\big(\mathbf{Q}\mathbf{0}, \mathbf{Q}\mathbf{I}_p\mathbf{Q}^\top\big) = N(\mathbf{0}, \mathbf{I}_p).$$

Thus $\mathbf{Q}\mathbf{x} \overset{d}{=} \mathbf{x}$, so the prompt law is invariant under $\mathbf{Q}$. Moreover, since $\mathbf{Q}$ is orthogonal and $\mathbf{Q}\mathbf{u} = \mathbf{u}$, we also have $\mathbf{Q}^\top\mathbf{u} = \mathbf{u}$. Therefore,

$$b^\star(\mathbf{Q}\mathbf{x}) = \mathbf{1}\{\langle \mathbf{u}, \mathbf{Q}\mathbf{x}\rangle > 0\} = \mathbf{1}\{\langle \mathbf{Q}^\top\mathbf{u}, \mathbf{x}\rangle > 0\} = \mathbf{1}\{\langle \mathbf{u}, \mathbf{x}\rangle > 0\} = b^\star(\mathbf{x}).$$

In addition,

$$\begin{aligned}
\pi_{\mathbf{Q}\cdot\mathbf{W}}(a \mid \mathbf{Q}\mathbf{x}) &= \frac{\exp(\langle \mathbf{Q}\mathbf{w}_a, \mathbf{Q}\mathbf{x}\rangle)}{\sum_{a' \in \mathcal{A}} \exp(\langle \mathbf{Q}\mathbf{w}_{a'}, \mathbf{Q}\mathbf{x}\rangle)} \\
&= \frac{\exp\big(\mathbf{w}_a^\top\mathbf{Q}^\top\mathbf{Q}\mathbf{x}\big)}{\sum_{a' \in \mathcal{A}} \exp\big(\mathbf{w}_{a'}^\top\mathbf{Q}^\top\mathbf{Q}\mathbf{x}\big)} \\
&= \frac{\exp(\langle \mathbf{w}_a, \mathbf{x}\rangle)}{\sum_{a' \in \mathcal{A}} \exp(\langle \mathbf{w}_{a'}, \mathbf{x}\rangle)} = \pi_{\mathbf{W}}(a \mid \mathbf{x}).
\end{aligned}$$

The SFT teacher is invariant as well, since (2) depends on the prompt only through $b^\star(\mathbf{x})$; the RFT reward is invariant for the same reason. Let $\Phi(\mathbf{W}, \omega)$ denote the updated parameter matrix $\mathbf{W}^+$ obtained by applying the SFT or RFT update in (6), and let $\Phi_a$ denote its $a$th row. For SFT, using the invariance of the policy shown above, we have

$$\begin{aligned}
\Phi_a(\mathbf{Q}\cdot\mathbf{W}, \mathbf{Q} \star \omega) &= \mathbf{Q}\mathbf{w}_a + \eta\big(\mathbf{1}\{a = a^\star\} - \pi_{\mathbf{Q}\cdot\mathbf{W}}(a \mid \mathbf{Q}\mathbf{x})\big)\mathbf{Q}\mathbf{x} \\
&= \mathbf{Q}\mathbf{w}_a + \eta\big(\mathbf{1}\{a = a^\star\} - \pi_{\mathbf{W}}(a \mid \mathbf{x})\big)\mathbf{Q}\mathbf{x} \\
&= \mathbf{Q}\Phi_a(\mathbf{W}, \omega).
\end{aligned}$$

For RFT, $b^\star(\mathbf{Q}\mathbf{x}) = b^\star(\mathbf{x})$ implies $R(a, \mathbf{Q}\mathbf{x}) = R(a, \mathbf{x})$. Together with the policy invariance, this also gives

$$\begin{aligned}
\bar{R}_{\mathbf{Q}\cdot\mathbf{W}}(\mathbf{Q}\mathbf{x}) &= \sum_{a' \in \mathcal{A}} \pi_{\mathbf{Q}\cdot\mathbf{W}}(a' \mid \mathbf{Q}\mathbf{x}) R(a', \mathbf{Q}\mathbf{x}) \\
&= \sum_{a' \in \mathcal{A}} \pi_{\mathbf{W}}(a' \mid \mathbf{x}) R(a', \mathbf{x}) = \bar{R}_{\mathbf{W}}(\mathbf{x}).
\end{aligned}$$

Consequently, the RFT update satisfies

$$\begin{aligned}
\Phi_a(\mathbf{Q}\cdot\mathbf{W}, \mathbf{Q} \star \omega) &= \mathbf{Q}\mathbf{w}_a + \eta\, \pi_{\mathbf{Q}\cdot\mathbf{W}}(a \mid \mathbf{Q}\mathbf{x})\big(R(a, \mathbf{Q}\mathbf{x}) - \bar{R}_{\mathbf{Q}\cdot\mathbf{W}}(\mathbf{Q}\mathbf{x})\big)\mathbf{Q}\mathbf{x} \\
&= \mathbf{Q}\left[\mathbf{w}_a + \eta\, \pi_{\mathbf{W}}(a \mid \mathbf{x})\big(R(a, \mathbf{x}) - \bar{R}_{\mathbf{W}}(\mathbf{x})\big)\mathbf{x}\right] \\
&= \mathbf{Q}\Phi_a(\mathbf{W}, \omega).
\end{aligned}$$

Thus the following identity holds rowwise under either update rule:

$$\Phi\big(\mathbf{Q}\cdot\mathbf{W}, \mathbf{Q} \star \omega\big) = \mathbf{Q}\cdot\Phi(\mathbf{W}, \omega). \tag{B.7}$$

We prove (B.5) by induction. It holds at $\tau = 0$ because $\mathbf{W}(0) = \mathbf{0}$, and hence $\mathbf{Q}\cdot\mathbf{W}(0) = \mathbf{W}(0)$. Now suppose $\mathbf{Q}\cdot\mathbf{W}(\tau) \overset{d}{=} \mathbf{W}(\tau)$. Let $\omega_\tau$ be the fresh sample used at step $\tau$. It is independent of $\mathbf{W}(\tau)$, and the invariance of the sample law proved above gives $\mathbf{Q} \star \omega_\tau \overset{d}{=} \omega_\tau$. Therefore, the induction hypothesis and independence imply the joint distributional identity

$$\big(\mathbf{Q}\cdot\mathbf{W}(\tau), \mathbf{Q} \star \omega_\tau\big) \overset{d}{=} \big(\mathbf{W}(\tau), \omega_\tau\big). \tag{B.8}$$

By the definition of the one-step update map, $\mathbf{W}(\tau+1) = \Phi(\mathbf{W}(\tau), \omega_\tau)$. Together with (B.7) and (B.8), we have

$$\begin{aligned}\mathbf{Q}\cdot\mathbf{W}(\tau+1) &= \mathbf{Q}\cdot\Phi\big(\mathbf{W}(\tau), \omega_\tau\big)\\ &= \Phi\big(\mathbf{Q}\cdot\mathbf{W}(\tau), \mathbf{Q}\star\omega_\tau\big)\\ &\overset{d}{=} \Phi\big(\mathbf{W}(\tau), \omega_\tau\big)\\ &= \mathbf{W}(\tau+1).\end{aligned}$$

This completes the induction and proves (B.5).

It follows that $\mathbf{Q}\mathbb{E}[\mathbf{w}_a(\tau)] = \mathbb{E}[\mathbf{w}_a(\tau)]$ for every $\mathbf{Q} \in G_{\mathbf{u}}$. Lemma B.4 therefore implies that $\mathbb{E}[\mathbf{w}_a(\tau)] \in \mathrm{span}(\mathbf{u})$. The remaining claims follow because $\Delta\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_a$ are fixed linear combinations of the rows. □

**Lemma B.6** (Lipschitz continuity of softmax)**.** *The softmax map is* 1*-Lipschitz in the Euclidean norm: for all* $\mathbf{h}, \mathbf{h}' \in \mathbb{R}^{2M}$*,*

$$\|\mathrm{softmax}(\mathbf{h}) - \mathrm{softmax}(\mathbf{h}')\|_2 \le \|\mathbf{h} - \mathbf{h}'\|_2.$$

*Proof.* For $\boldsymbol{\pi} = \mathrm{softmax}(\mathbf{h})$, differentiating gives the Jacobian

$$J(\mathbf{h}) = \mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top.$$

For every $\mathbf{v} \in \mathbb{R}^{2M}$,

$$0 \le \mathbf{v}^\top J(\mathbf{h})\mathbf{v} = \sum_{a=1}^{2M} \pi_a v_a^2 - \left(\sum_{a=1}^{2M} \pi_a v_a\right)^2 \le \sum_{a=1}^{2M} \pi_a v_a^2 \le \|\mathbf{v}\|_2^2.$$

The first inequality follows because the expression is a variance under $\boldsymbol{\pi}$, and the last uses $0 \le \pi_a \le 1$. Thus $J(\mathbf{h})$ is symmetric positive semidefinite with eigenvalues at most 1, so $\|J(\mathbf{h})\|_{\mathrm{op}} \le 1$ for every $\mathbf{h}$. Integrating the derivative along the line segment from $\mathbf{h}'$ to $\mathbf{h}$ therefore gives

$$\begin{aligned}\|\mathrm{softmax}(\mathbf{h}) - \mathrm{softmax}(\mathbf{h}')\|_2 &\le \int_0^1 \|J(\mathbf{h}' + t(\mathbf{h}-\mathbf{h}'))\|_{\mathrm{op}} \|\mathbf{h}-\mathbf{h}'\|_2 \,\mathrm{d}t\\ &\le \|\mathbf{h}-\mathbf{h}'\|_2.\end{aligned}$$

□

# C Proofs of the main results

## C.1 Proof of Theorem 1

*Proof. Part (i): the per-sample rate identity.* For an arbitrary $\mathbf{W}$ and prompt $\mathbf{x}$, by (12), $P_0(\mathbf{W}, \mathbf{x}) + P_1(\mathbf{W}, \mathbf{x}) = 1$.

First consider the SFT step. If the demonstration is $a^\star$ with $\sigma(a^\star) = b^\star(\mathbf{x})$, then the row update is

$$\mathbf{w}_a^+ - \mathbf{w}_a = \eta\big(\mathbf{1}\{a = a^\star\} - \pi_{\mathbf{W}}(a \mid \mathbf{x})\big)\mathbf{x}. \tag{C.1}$$

Averaging (C.1) over a class and using the definition of $\boldsymbol{\alpha}^{(b)}$ in (9) gives

$$\begin{aligned}\boldsymbol{\alpha}^{(b)+} - \boldsymbol{\alpha}^{(b)} &= \frac{1}{M}\sum_{a\in\mathcal{A}_b} \eta\big(\mathbf{1}\{a = a^\star\} - \pi_{\mathbf{W}}(a \mid \mathbf{x})\big)\mathbf{x}\\ &= \frac{\eta}{M}\Big(\mathbf{1}\{b = b^\star(\mathbf{x})\} - \sum_{a\in\mathcal{A}_b} \pi_{\mathbf{W}}(a \mid \mathbf{x})\Big)\mathbf{x}\\ &= \frac{\eta}{M}\big(\mathbf{1}\{b = b^\star(\mathbf{x})\} - P_b(\mathbf{W}, \mathbf{x})\big)\mathbf{x},\end{aligned} \tag{C.2}$$

where the last equality is (12). Subtracting the $b = 0$ mean update from the $b = 1$ mean update yields

$$\Delta\boldsymbol{\alpha}^{+}_{\mathrm{SFT}} - \Delta\boldsymbol{\alpha} = \eta \left[ \frac{\mathbf{1}\{b^\star(\mathbf{x}) = 1\} - P_1(\mathbf{W}, \mathbf{x})}{M} - \frac{\mathbf{1}\{b^\star(\mathbf{x}) = 0\} - P_0(\mathbf{W}, \mathbf{x})}{M} \right] \mathbf{x}. \tag{C.3}$$

If $b^\star(\mathbf{x}) = 1$, then $2b^\star(\mathbf{x}) - 1 = 1$, $P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x}) = P_1(\mathbf{W}, \mathbf{x})$, and $P_0(\mathbf{W}, \mathbf{x}) = 1 - P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x})$; if $b^\star(\mathbf{x}) = 0$, then $2b^\star(\mathbf{x}) - 1 = -1$, $P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x}) = P_0(\mathbf{W}, \mathbf{x})$, and $P_1(\mathbf{W}, \mathbf{x}) = 1 - P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x})$. Therefore (C.3) becomes the SFT line of (13).

For RFT, the reward is $R(a, \mathbf{x}) = \mathbf{1}\{\sigma(a) = b^\star(\mathbf{x})\}$, so (6) gives

$$\bar{R}_{\mathbf{W}}(\mathbf{x}) = \sum_a \pi_{\mathbf{W}}(a \mid \mathbf{x}) \mathbf{1}\{\sigma(a) = b^\star(\mathbf{x})\} = P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x}), \tag{C.4}$$

$$\mathbf{w}^{+}_a - \mathbf{w}_a = \eta\, \pi_{\mathbf{W}}(a \mid \mathbf{x}) \big( \mathbf{1}\{\sigma(a) = b^\star(\mathbf{x})\} - P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x}) \big) \mathbf{x}. \tag{C.5}$$

Averaging (C.5) over a class gives

$$\begin{aligned} \boldsymbol{\alpha}^{(b)+} - \boldsymbol{\alpha}^{(b)} &= \frac{\eta}{M} \sum_{a \in \mathcal{A}_b} \pi_{\mathbf{W}}(a \mid \mathbf{x}) \big( \mathbf{1}\{b = b^\star(\mathbf{x})\} - P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x}) \big) \mathbf{x} \\ &= \eta \frac{P_b(\mathbf{W}, \mathbf{x})}{M} \big( \mathbf{1}\{b = b^\star(\mathbf{x})\} - P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x}) \big) \mathbf{x}. \end{aligned} \tag{C.6}$$

Subtracting the class-0 update in (C.6) from the class-1 update gives

$$\begin{aligned} \Delta\boldsymbol{\alpha}^{+}_{\mathrm{RFT}} - \Delta\boldsymbol{\alpha} = \eta \Bigg[ & \frac{P_1(\mathbf{W}, \mathbf{x})}{M} \big( \mathbf{1}\{b^\star(\mathbf{x}) = 1\} - P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x}) \big) \\ & - \frac{P_0(\mathbf{W}, \mathbf{x})}{M} \big( \mathbf{1}\{b^\star(\mathbf{x}) = 0\} - P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x}) \big) \Bigg] \mathbf{x}. \end{aligned} \tag{C.7}$$

Again splitting on $b^\star(\mathbf{x})$ and using $P_0(\mathbf{W}, \mathbf{x}) + P_1(\mathbf{W}, \mathbf{x}) = 1$, the right side of (C.7) is

$$\frac{2\eta}{M} \big( 2b^\star(\mathbf{x}) - 1 \big) P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x}) \big( 1 - P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x}) \big) \mathbf{x}, \tag{C.8}$$

which is the RFT line of (13).

*Part (ii): population semantic update.* Now assume additionally that $\boldsymbol{\beta} = \mathbf{0}$ and $\Delta\boldsymbol{\alpha} = c\mathbf{u}$ for some $c \in \mathbb{R}$. Then (10) implies that the class probabilities $P_b(\mathbf{W}, \mathbf{x})$ depend on $\mathbf{W}$ only through $\Delta\boldsymbol{\alpha}$. Fix task $t$ and use the notation of Lemma B.3:

$$z = \langle \mathbf{u}, \mathbf{x} \rangle, \qquad \boldsymbol{\xi}^{\perp} = \big( \mathbf{I}_p - \mathbf{u}\mathbf{u}^\top \big)(\mathbf{x} - \boldsymbol{\mu}_t).$$

Lemma B.3 gives $z \sim N(0, 1)$, $z \perp\!\!\!\perp \boldsymbol{\xi}^{\perp}$, and

$$\mathbf{x} = z\mathbf{u} + \boldsymbol{\mu}_t + \boldsymbol{\xi}^{\perp}.$$

It follows that

$$\langle \Delta\boldsymbol{\alpha}, \mathbf{x} \rangle = c\, \langle \mathbf{u}, z\mathbf{u} + \boldsymbol{\mu}_t + \boldsymbol{\xi}^{\perp} \rangle = c\, z,$$

because both $\boldsymbol{\mu}_t$ and $\boldsymbol{\xi}^{\perp}$ are orthogonal to $\mathbf{u}$. Hence (10) implies that $P_1(\mathbf{W}, \mathbf{x}) = (1 + e^{-cz})^{-1}$ and $P_0(\mathbf{W}, \mathbf{x}) = 1 - P_1(\mathbf{W}, \mathbf{x})$. Since $b^\star(\mathbf{x}) = b^\star(z) := \mathbf{1}\{z > 0\}$, the correct-class probability has the one-dimensional representation

$$P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x}) = P_\star(z) := \frac{1}{1 + e^{-c|z|}}.$$

This function is even in $z$. Also $2b^\star(\mathbf{x}) - 1 = \mathrm{sign}(z)$ almost surely.

The SFT rate identity in (13) becomes

$$\Delta\boldsymbol{\alpha}^{+}_{\mathrm{SFT}} - \Delta\boldsymbol{\alpha} = \frac{2\eta}{M} \mathrm{sign}(z) \big( 1 - P_\star(z) \big) \big( z\mathbf{u} + \boldsymbol{\mu}_t + \boldsymbol{\xi}^{\perp} \big). \tag{C.9}$$

Taking the task-$t$ population expectation gives

$$\mathbb{E}_t\big[\Delta\boldsymbol{\alpha}^{+}_{\mathrm{SFT}} - \Delta\boldsymbol{\alpha}\big] = \frac{2\eta}{M}\Big(\mathbb{E}[\mathrm{sign}(z)(1-P_\star(z))z]\,\mathbf{u} + \mathbb{E}[\mathrm{sign}(z)(1-P_\star(z))]\,\boldsymbol{\mu}_t + \mathbb{E}[\mathrm{sign}(z)(1-P_\star(z))\boldsymbol{\xi}^{\perp}]\Big). \tag{C.10}$$

The last term vanishes because $\mathrm{sign}(z)(1-P_\star(z))$ is a function of $z$ alone, $z$ is independent of $\boldsymbol{\xi}^{\perp}$, and $\mathbb{E}[\boldsymbol{\xi}^{\perp}] = \mathbf{0}$. The $\boldsymbol{\mu}_t$ term also vanishes: $1-P_\star(z)$ is even, so $\mathrm{sign}(z)(1-P_\star(z))$ is odd, and its expectation under the symmetric $z \sim N(0,1)$ is zero. Since $\mathrm{sign}(z)z = |z|$, the remaining scalar expectation is

$$\mathbb{E}\big[|z|(1-P_\star(z))\big] = \mathbb{E}\left[\frac{|z|}{1+e^{c|z|}}\right] = \sqrt{\frac{2}{\pi}}\int_0^\infty \frac{ze^{-z^2/2}}{1+e^{cz}}\,\mathrm{d}z.$$

The last equality uses the standard normal density and the evenness of the integrand. Substituting into (C.10) gives the SFT line of (14).

Similarly, the RFT rate identity in (13) becomes

$$\Delta\boldsymbol{\alpha}^{+}_{\mathrm{RFT}} - \Delta\boldsymbol{\alpha} = \frac{2\eta}{M}\,\mathrm{sign}(z)\,P_\star(z)\big(1-P_\star(z)\big)\,\big(z\mathbf{u} + \boldsymbol{\mu}_t + \boldsymbol{\xi}^{\perp}\big). \tag{C.11}$$

Taking the task-$t$ population expectation gives

$$\mathbb{E}_t\big[\Delta\boldsymbol{\alpha}^{+}_{\mathrm{RFT}} - \Delta\boldsymbol{\alpha}\big] = \frac{2\eta}{M}\Big(\mathbb{E}[\mathrm{sign}(z)P_\star(z)(1-P_\star(z))z]\,\mathbf{u} + \mathbb{E}[\mathrm{sign}(z)P_\star(z)(1-P_\star(z))]\,\boldsymbol{\mu}_t + \mathbb{E}[\mathrm{sign}(z)P_\star(z)(1-P_\star(z))\boldsymbol{\xi}^{\perp}]\Big). \tag{C.12}$$

Again the last term vanishes by independence and $\mathbb{E}[\boldsymbol{\xi}^{\perp}] = \mathbf{0}$. The $\boldsymbol{\mu}_t$ term vanishes as well because $P_\star(z)(1-P_\star(z))$ is even, so its product with $\mathrm{sign}(z)$ is odd. Using $\mathrm{sign}(z)z = |z|$ and again integrating against the standard normal density gives

$$\mathbb{E}\big[|z|P_\star(z)(1-P_\star(z))\big] = \mathbb{E}\left[\frac{|z|e^{c|z|}}{(1+e^{c|z|})^2}\right] = \sqrt{\frac{2}{\pi}}\int_0^\infty \frac{ze^{-z^2/2}e^{cz}}{(1+e^{cz})^2}\,\mathrm{d}z.$$

Substituting into (C.12) yields the RFT line of (14). □

### C.2 Proof of Theorem 2

*Proof.*

**Part (i): RFT style invariance.** Fix a prompt $\mathbf{x}$ and write the RFT update direction from (6) as

$$\mathbf{g}_a = \pi_{\mathbf{W}}(a \mid \mathbf{x})\big(R(a,\mathbf{x}) - \bar{R}_{\mathbf{W}}(\mathbf{x})\big)\mathbf{x}. \tag{C.13}$$

Assume $\boldsymbol{\beta} = \mathbf{0}$. By (9), every row in a common semantic class is equal to that class's mean. Thus, for any two tokens $a, a'$ in the same class,

$$\mathbf{w}_a = \mathbf{w}_{a'}, \tag{C.14}$$

$$\langle \mathbf{w}_a, \mathbf{x}\rangle = \langle \mathbf{w}_{a'}, \mathbf{x}\rangle,$$

$$\pi_{\mathbf{W}}(a \mid \mathbf{x}) = \pi_{\mathbf{W}}(a' \mid \mathbf{x}). \tag{C.15}$$

The reward also agrees inside the class because, by hypothesis, $R$ depends on $a$ only through its class $\sigma(a)$, which is constant on a semantic class:

$$\sigma(a) = \sigma(a') \Longrightarrow R(a,\mathbf{x}) = R(a',\mathbf{x}). \tag{C.16}$$

Combining (C.13), (C.15), and (C.16), the update vector is identical for every row in the same class. Therefore if $a, a'$ lie in the same class,

$$\begin{aligned}\mathbf{w}_a^+ - \mathbf{w}_{a'}^+ &= \big(\mathbf{w}_a - \mathbf{w}_{a'}\big) + \eta\big(\mathbf{g}_a - \mathbf{g}_{a'}\big) \\ &= \mathbf{0} + \eta \cdot \mathbf{0} = \mathbf{0},\end{aligned} \tag{C.17}$$

Hence an update preserves $\boldsymbol{\beta} = \mathbf{0}$. Induction over the prompt sequence gives $\boldsymbol{\beta}_{\mathrm{RFT}}(\tau) = \mathbf{0}$ at every step.

**Part (ii): SFT style updates.** For convenience, define the within-class probability deviation

$$\delta_a(\mathbf{W}, \mathbf{x}) := \pi_{\mathbf{W}}(a \mid \mathbf{x}) - \frac{1}{M} P_{\sigma(a)}(\mathbf{W}, \mathbf{x}). \tag{C.18}$$

For an arbitrary current matrix $\mathbf{W}$, the SFT row update and its class average are exactly (C.1) and (C.2), respectively. Subtracting (C.2) at $b = \sigma(a)$ from (C.1) gives

$$\begin{aligned}\boldsymbol{\beta}_a^+ - \boldsymbol{\beta}_a &= \eta\big(\mathbf{1}\{a = a^\star\} - \pi_{\mathbf{W}}(a \mid \mathbf{x})\big)\mathbf{x} - \frac{\eta}{M}\big(\mathbf{1}\{\sigma(a) = b^\star\} - P_{\sigma(a)}(\mathbf{W}, \mathbf{x})\big)\mathbf{x} \\ &= \eta\left[\mathbf{1}\{a = a^\star\} - \frac{1}{M}\mathbf{1}\{\sigma(a) = b^\star\} - \left(\pi_{\mathbf{W}}(a \mid \mathbf{x}) - \frac{P_{\sigma(a)}(\mathbf{W}, \mathbf{x})}{M}\right)\right]\mathbf{x} \qquad \text{(C.19)} \\ &= \eta\left[\mathbf{1}\{a = a^\star\} - \frac{1}{M}\mathbf{1}\{\sigma(a) = b^\star\} - \delta_a(\mathbf{W}, \mathbf{x})\right]\mathbf{x}, \qquad \text{(C.20)}\end{aligned}$$

which is exactly (15).

Now fix a task $t$, a token $a \in \mathcal{A}_b$, and a current parameter matrix $\mathbf{W}$ with $\boldsymbol{\beta} = \mathbf{0}$. Since $\mathbf{W}$ has identical rows within each class, all rows in the same semantic class have the same logit at every prompt. Therefore the policy is uniform within each semantic class:

$$\pi_{\mathbf{W}}(a \mid \mathbf{x}) = \frac{1}{M}\sum_{a' \in \mathcal{A}_b} \pi_{\mathbf{W}}(a' \mid \mathbf{x}) = \frac{1}{M} P_b(\mathbf{W}, \mathbf{x}) \qquad (a \in \mathcal{A}_b).$$

Substituting this identity into the definition of $\delta_a$ gives

$$\delta_a(\mathbf{W}, \mathbf{x}) = \pi_{\mathbf{W}}(a \mid \mathbf{x}) - \frac{1}{M} P_{\sigma(a)}(\mathbf{W}, \mathbf{x}) = 0.$$

Thus the policy-centering feedback in the exact style update (15) vanishes. For this fixed prompt and teacher draw, (15) becomes

$$\boldsymbol{\beta}_a^+ - \boldsymbol{\beta}_a = \eta\left[\mathbf{1}\{a = a^\star\} - \frac{1}{M}\mathbf{1}\{b^\star(\mathbf{x}) = b\}\right]\mathbf{x}.$$

Now take expectation only over the teacher demonstration, conditional on the prompt $\mathbf{x}$. By the teacher model (2), if $b^\star(\mathbf{x}) \neq b$ then the teacher places no mass on token $a \in \mathcal{A}_b$, and both terms inside the brackets have conditional mean zero. If $b^\star(\mathbf{x}) = b$, then $\mathbb{P}(a^\star = a \mid \mathbf{x}) = \rho_{t,b}(a)$, so the conditional mean of the bracket is $\rho_{t,b}(a) - 1/M$. Combining the two cases,

$$\mathbb{E}_{a^\star \mid \mathbf{x}}\left[\mathbf{1}\{a = a^\star\} - \frac{1}{M}\mathbf{1}\{b^\star(\mathbf{x}) = b\}\right] = \mathbf{1}\{b^\star(\mathbf{x}) = b\}\Big(\rho_{t,b}(a) - \frac{1}{M}\Big).$$

Therefore the SFT population update of the style coordinate is

$$\mathbb{E}[\boldsymbol{\beta}_a^+ - \boldsymbol{\beta}_a] = \eta\Big(\rho_{t,b}(a) - \frac{1}{M}\Big)\mathbb{E}[\mathbf{1}\{b^\star(\mathbf{x}) = b\}\,\mathbf{x}]. \tag{C.21}$$

It remains to compute the last prompt moment. Apply Lemma B.3 and write $z = \langle \mathbf{u}, \mathbf{x} \rangle$ and $\boldsymbol{\xi}^\perp = \big(\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top\big)(\mathbf{x} - \boldsymbol{\mu}_t)$. Then $z \perp\!\!\!\perp \boldsymbol{\xi}^\perp$ and

$$\mathbf{x} = z\mathbf{u} + \boldsymbol{\mu}_t + \boldsymbol{\xi}^\perp,$$

while $z \sim N(0, 1)$. The event $b^\star = b$ depends only on $z$: it is $\{z > 0\}$ for $b = 1$ and $\{z < 0\}$ for $b = 0$. Hence

$$\mathbb{E}[\mathbf{1}\{b^\star = b\}\,\mathbf{x}] = \mathbb{E}[\mathbf{1}\{b^\star = b\}z]\,\mathbf{u} + \mathbb{E}[\mathbf{1}\{b^\star = b\}(\boldsymbol{\mu}_t + \boldsymbol{\xi}^\perp)].$$

For the off-axis term, independence gives

$$\mathbb{E}[\mathbf{1}\{b^\star = b\}(\boldsymbol{\mu}_t + \boldsymbol{\xi}^\perp)] = \mathbb{P}(b^\star = b)\boldsymbol{\mu}_t + \mathbb{P}(b^\star = b)\mathbb{E}[\boldsymbol{\xi}^\perp] = \frac{1}{2}\boldsymbol{\mu}_t.$$

For the on-axis term,

$$\mathbb{E}[\mathbf{1}\{z > 0\}z] = \int_0^\infty z\,\frac{1}{\sqrt{2\pi}}e^{-z^2/2}\,dz = \frac{1}{\sqrt{2\pi}}, \qquad \mathbb{E}[\mathbf{1}\{z < 0\}z] = -\frac{1}{\sqrt{2\pi}},$$

so in either class

$$\mathbb{E}[\mathbf{1}\{b^\star = b\}z]\,\mathbf{u} = \frac{2b-1}{\sqrt{2\pi}}\mathbf{u}.$$

Substituting these two moment identities into (C.21) proves (16).

Finally, project (16) orthogonally to $\mathrm{span}(\mathbf{u})$. The term proportional to $\mathbf{u}$ is removed, while $\boldsymbol{\mu}_t$ is unchanged because $\boldsymbol{\mu}_t \perp \mathbf{u}$. This gives the off-axis identity in (16). □

### C.3 Proof of Theorem 3

*Proof.*

**Step 1: perfect initial fit and RFT retention.** Since $\boldsymbol{\beta}(\mathbf{W}_1) = \mathbf{0}$, the class-score difference is $D_{\mathbf{W}_1}(\mathbf{x}) = \langle \Delta\boldsymbol{\alpha}(\mathbf{W}_1), \mathbf{x}\rangle$. Since the old Gaussian law has positive density everywhere, $\mathcal{L}_1(\mathbf{W}_1) = 0$ means that

$$\Delta\boldsymbol{\alpha}(\mathbf{W}_1) = \|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\|\mathbf{u}, \qquad \|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\| > 0.$$

Subtracting the common row average $\mathbf{c} := (2M)^{-1}\sum_a \mathbf{w}_{1,a}$ from every row preserves policy probabilities, predictions, and both update increments. We therefore use these centered weights throughout the proof, so that

$$\mathbf{w}_{1,a} = \frac{2\sigma(a)-1}{2}\|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\|\mathbf{u}, \qquad \sum_a \mathbf{w}_{1,a} = \mathbf{0}.$$

Write $z := \langle \mathbf{u}, \mathbf{x}\rangle$. Both tasks share the semantic rule $b^\star(\mathbf{x}) = \mathbf{1}\{z > 0\}$. Lemma B.2, applied with $B = 0$, therefore shows that $\mathbf{W}_1$ classifies both tasks perfectly:

$$a_1(\mathbf{W}_1) = a_2(\mathbf{W}_1) = 1. \tag{C.22}$$

Theorem 2(i) ensures $\boldsymbol{\beta} = \mathbf{0}$ throughout RFT training, both updates preserve $\sum_a \mathbf{w}_a = \mathbf{0}$, and Theorem 1(ii) gives strictly positive on-axis RFT increments. Thus every RFT iterate $\mathbf{W}$ satisfies

$$\mathbf{w}_a = \frac{2\sigma(a)-1}{2}\|\Delta\boldsymbol{\alpha}(\mathbf{W})\|\mathbf{u},$$
$$\|\Delta\boldsymbol{\alpha}(\mathbf{W})\| \geq \|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\| > 0.$$

Applying Lemma B.2 again with $B = 0$ shows that every RFT iterate classifies both tasks correctly, proving part (i).

**Step 2: quantifying the style drift under SFT.** Let $\hat{\boldsymbol{\mu}}_2 := \boldsymbol{\mu}_2/\|\boldsymbol{\mu}_2\|$ and $y := \langle \hat{\boldsymbol{\mu}}_2, \mathbf{x}\rangle$. By Lemma B.3, under task 2 we can write $\mathbf{x} = z\mathbf{u} + \boldsymbol{\mu}_2 + \boldsymbol{\xi}^\perp$, where $z \sim N(0,1)$ is independent of $\boldsymbol{\xi}^\perp \sim N(\mathbf{0}, \mathbf{I}_p - \mathbf{u}\mathbf{u}^\top)$. Since $\hat{\boldsymbol{\mu}}_2 \perp \mathbf{u}$ and $\|\hat{\boldsymbol{\mu}}_2\| = 1$,

$$y = \langle \hat{\boldsymbol{\mu}}_2, \mathbf{x}\rangle = \|\boldsymbol{\mu}_2\| + \langle \hat{\boldsymbol{\mu}}_2, \boldsymbol{\xi}^\perp\rangle \sim N(\|\boldsymbol{\mu}_2\|, 1).$$

Moreover, $y$ depends only on $\boldsymbol{\xi}^\perp$, so it is independent of $z$.

Define $V := \mathrm{span}\{\mathbf{u}, \hat{\boldsymbol{\mu}}_2\}$. We will show that every parameter row remains in $V$ throughout task-2 training. Define the residual $\boldsymbol{\xi}_V := \mathbf{x} - z\mathbf{u} - y\hat{\boldsymbol{\mu}}_2$, which is the projection of $\mathbf{x}$ onto $V^\perp$. Since $\boldsymbol{\mu}_2 \in V$, the residual has mean zero:

$$\mathbb{E}_2[\boldsymbol{\xi}_V] = \boldsymbol{\mu}_2 - 0\cdot\mathbf{u} - \|\boldsymbol{\mu}_2\|\hat{\boldsymbol{\mu}}_2 = \mathbf{0}.$$

Orthogonality of the Gaussian projections also makes $\boldsymbol{\xi}_V$ independent of $(z, y)$. Initially, $\mathbf{w}_{1,a} = \frac{2\sigma(a)-1}{2}\|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\|\mathbf{u} \in V$. If all current rows lie in $V$, write

$$\mathbf{w}_a = \zeta_a \mathbf{u} + \omega_a \hat{\boldsymbol{\mu}}_2. \tag{C.23}$$

Their logits and policy probabilities are then

$$\langle \mathbf{w}_a, \mathbf{x}\rangle = \zeta_a z + \omega_a y, \qquad \pi_a = \pi_{\mathbf{W}}(a \mid \mathbf{x}) = \frac{e^{\zeta_a z + \omega_a y}}{\sum_{a' \in \mathcal{A}} e^{\zeta_{a'} z + \omega_{a'} y}}.$$

Thus, for fixed current parameters, $\pi_a$ depends on $\mathbf{x}$ only through $(z, y)$. Meanwhile, the teacher definition (2) gives

$$q_2(a \mid \mathbf{x}) = \mathbf{1}\{\sigma(a) = b^\star(\mathbf{x})\}\rho_{2,b^\star(\mathbf{x})}(a), \qquad b^\star(\mathbf{x}) = \mathbf{1}\{z > 0\}.$$

By the teacher specification in Section 2.1, the within-class distribution $\rho_{2,b}$ is fixed for each class and does not depend on $\mathbf{x}$, so $q_2(a \mid \mathbf{x})$ depends on the input only through the sign of $z$. Consequently, $q_2(a \mid \mathbf{x}) - \pi_a$ is a function of $(z, y)$ alone. By (7), the $V^\perp$ component of the SFT parameter increment is

$$\eta_2 \mathbb{E}_2[(q_2(a \mid \mathbf{x}) - \pi_a)\boldsymbol{\xi}_V] = \eta_2 \mathbb{E}_2[(q_2(a \mid \mathbf{x}) - \pi_a)\mathbb{E}_2[\boldsymbol{\xi}_V \mid z, y]] = \mathbf{0}.$$

The last equality follows because $\boldsymbol{\xi}_V$ is independent of $(z, y)$ and has mean zero. Each updated row therefore remains in $V$. By induction, (C.23) holds at every task-2 SFT iterate, so it suffices to track the two scalar coefficients $\zeta_a$ and $\omega_a$ for each token.

Since $\|\mathbf{u}\| = 1$ and $\mathbf{u} \perp \hat{\boldsymbol{\mu}}_2$, the decomposition (C.23) gives $\zeta_a = \langle \mathbf{w}_a, \mathbf{u}\rangle$. Taking the inner product of the SFT population update (7) with $\mathbf{u}$ and using $z = \langle \mathbf{u}, \mathbf{x}\rangle$ yields

$$\begin{aligned} \zeta_a^+ - \zeta_a = \langle \mathbf{w}_a^+ - \mathbf{w}_a, \mathbf{u}\rangle = \eta_2 \mathbb{E}_2[(q_2(a \mid \mathbf{x}) - \pi_a)z], \\ |\zeta_a^+ - \zeta_a| \le \eta_2 \mathbb{E}|z| \le \eta_2, \end{aligned}$$

since $|q_2(a \mid \mathbf{x}) - \pi_a| \le 1$ and $z \sim N(0, 1)$. At initialization, $\zeta_a^{(0)} = \frac{2\sigma(a)-1}{2}\|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\|$. After $j \le k_2$ updates, the triangle inequality gives

$$\begin{aligned} |\zeta_a^{(j)}| &\le |\zeta_a^{(0)}| + \sum_{\ell=0}^{j-1} |\zeta_a^{(\ell+1)} - \zeta_a^{(\ell)}| \\ &\le \tfrac{1}{2}\|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\| + j\eta_2 \mathbb{E}|z| \\ &\le \tfrac{1}{2}\|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\| + \tau_2. \end{aligned} \tag{C.24}$$

Collect the coefficients in (C.23) into a matrix $\boldsymbol{\Theta}_j \in \mathbb{R}^{2M \times 2}$, whose row indexed by token $a$ is $(\zeta_a^{(j)}, \omega_a^{(j)})$ after $j$ task-2 steps. For $\mathbf{r} := (z, y)^\top$, each entry of $\boldsymbol{\Theta}_j \mathbf{r}$ is the corresponding token's logit:

$$(\boldsymbol{\Theta}_j \mathbf{r})_a = \zeta_a^{(j)} z + \omega_a^{(j)} y = \langle \mathbf{w}_a^{(j)}, \mathbf{x}\rangle.$$

Thus $\mathrm{softmax}(\boldsymbol{\Theta}_j \mathbf{r})$ is the model's probability vector. The SFT population update (7), expressed in these coordinates, becomes

$$\boldsymbol{\Theta}_{j+1} = \boldsymbol{\Theta}_j + \eta_2 \Delta(\boldsymbol{\Theta}_j), \qquad \Delta(\boldsymbol{\Theta}) := \mathbb{E}_2\big[(q_2(\cdot \mid \mathbf{x}) - \mathrm{softmax}(\boldsymbol{\Theta}\mathbf{r}))\mathbf{r}^\top\big]. \tag{C.25}$$

Here $\Delta(\boldsymbol{\Theta})$ denotes the population SFT update direction at $\boldsymbol{\Theta}$. It has the same shape as $\boldsymbol{\Theta}$: its first and second columns give the update directions for $\zeta_a$ and $\omega_a$, respectively. Explicitly, its row indexed by $a$ has entries

$$\Delta_{\zeta_a}(\boldsymbol{\Theta}) = \mathbb{E}_2[(q_2(a \mid \mathbf{x}) - \pi_a)z], \qquad \Delta_{\omega_a}(\boldsymbol{\Theta}) = \mathbb{E}_2[(q_2(a \mid \mathbf{x}) - \pi_a)y],$$

where $\pi_a = [\mathrm{softmax}(\boldsymbol{\Theta}\mathbf{r})]_a$. At $\mathbf{W}_1$ (represented by $\boldsymbol{\Theta}_0$), recall that $\boldsymbol{\alpha}^{(b)} = M^{-1}\sum_{a \in \mathcal{A}_b} \mathbf{w}_a$ is the class-$b$ semantic channel; see (9). Theorem 1(ii) shows that the population update of $\boldsymbol{\alpha}^{(1)} - \boldsymbol{\alpha}^{(0)}$ is parallel to $\mathbf{u}$. From (C.2), a single SFT step on a prompt $\mathbf{x}$ updates the two components of the semantic channel by

$$\boldsymbol{\alpha}^{(b)+} - \boldsymbol{\alpha}^{(b)} = \frac{\eta}{M}\big(\mathbf{1}\{b = b^\star(\mathbf{x})\} - P_b(\mathbf{W}, \mathbf{x})\big)\mathbf{x}, \qquad b \in \{0, 1\}.$$

Summing over $b$, we have

$$\begin{aligned}\boldsymbol{\alpha}^{(0)+} + \boldsymbol{\alpha}^{(1)+} &= \boldsymbol{\alpha}^{(0)} + \boldsymbol{\alpha}^{(1)} + \frac{\eta}{M} \sum_{b\in\{0,1\}} \big(\mathbf{1}\{b = b^\star(\mathbf{x})\} - P_b(\mathbf{W}, \mathbf{x})\big)\mathbf{x} \\ &= \boldsymbol{\alpha}^{(0)} + \boldsymbol{\alpha}^{(1)} + \frac{\eta}{M} \left(\mathbf{1}\{b^\star(\mathbf{x}) = 0\} + \mathbf{1}\{b^\star(\mathbf{x}) = 1\} - P_0(\mathbf{W}, \mathbf{x}) - P_1(\mathbf{W}, \mathbf{x})\right)\mathbf{x} \\ &= \boldsymbol{\alpha}^{(0)} + \boldsymbol{\alpha}^{(1)},\end{aligned}$$

where the indicators add to 1 because exactly one class is correct, and $P_0+P_1 = 1$ is by definition (12), so $\boldsymbol{\alpha}^{(0)} + \boldsymbol{\alpha}^{(1)}$ is unchanged by every SFT step. Writing $\boldsymbol{\alpha}^{(b)} = \frac{1}{2}\big[(\boldsymbol{\alpha}^{(0)} + \boldsymbol{\alpha}^{(1)}) \pm (\boldsymbol{\alpha}^{(1)} - \boldsymbol{\alpha}^{(0)})\big]$, the sum does not move and the difference moves only along $\mathbf{u}$, so each semantic channel's population update at $\mathbf{W}_1$ is parallel to $\mathbf{u}$. Since $\mathbf{u} \perp \hat{\boldsymbol{\mu}}_2$, its projection onto $\hat{\boldsymbol{\mu}}_2$ is zero. By (9), $\boldsymbol{\alpha}^{(b)}$ is the average of $\mathbf{w}_a$ over $a \in \mathcal{A}_b$, so its update is the average of the updates of those token parameters. Thus, for the first task-2 step from $\mathbf{W}_1$,

$$\begin{aligned}\langle \mathbb{E}_2[\boldsymbol{\alpha}^{(b)+} - \boldsymbol{\alpha}^{(b)}],\, \hat{\boldsymbol{\mu}}_2\rangle &= \frac{1}{M} \sum_{a\in\mathcal{A}_b} \langle \mathbb{E}_2[\mathbf{w}_a^+ - \mathbf{w}_a],\, \hat{\boldsymbol{\mu}}_2\rangle \\ &= \frac{\eta_2}{M} \sum_{a\in\mathcal{A}_b} \Delta_{\omega_a}(\boldsymbol{\Theta}_0).\end{aligned}$$

Here each $\Delta_{\omega_a}(\boldsymbol{\Theta}_0)$ describes the update direction of one token's $\omega_a$ coefficient; summing and dividing by $M$ gives the corresponding direction for the semantic channel. The left-hand side is zero, so

$$\frac{1}{M} \sum_{a\in\mathcal{A}_b} \Delta_{\omega_a}(\boldsymbol{\Theta}_0) = 0, \qquad b \in \{0, 1\}. \tag{C.26}$$

Individual rows can still move along $\hat{\boldsymbol{\mu}}_2$ through their within-class deviations $\boldsymbol{\beta}_a = \mathbf{w}_a - \boldsymbol{\alpha}^{(\sigma(a))}$, and Theorem 2(ii) quantifies this: since $\boldsymbol{\beta}(\mathbf{W}_1) = \mathbf{0}$, (16) with $t = 2$ gives, for $a \in \mathcal{A}_b$,

$$\mathbb{E}_2[\boldsymbol{\beta}_a^+ - \boldsymbol{\beta}_a] = \eta_2 \Big(\rho_{2,b}(a) - \frac{1}{M}\Big)\Big(\frac{\|\boldsymbol{\mu}_2\|}{2}\hat{\boldsymbol{\mu}}_2 + \frac{2b-1}{\sqrt{2\pi}}\mathbf{u}\Big).$$

As $\mathbf{w}_a = \boldsymbol{\alpha}^{(\sigma(a))} + \boldsymbol{\beta}_a$ and the $\boldsymbol{\alpha}$-step has no $\hat{\boldsymbol{\mu}}_2$-component, the $\hat{\boldsymbol{\mu}}_2$-component of the row step is that of the $\boldsymbol{\beta}$-step. Using that the initial rows are parallel to $\mathbf{u}$,

$$\omega_a(\boldsymbol{\Theta}_0) = 0, \qquad \Delta_{\omega_a}(\boldsymbol{\Theta}_0) = \frac{\|\boldsymbol{\mu}_2\|}{2}\left(\rho_{2,\sigma(a)}(a) - \frac{1}{M}\right), \qquad a \in \mathcal{A}. \tag{C.27}$$

Iterating (C.25) and using $\tau_2 = k_2\eta_2$,

$$\boldsymbol{\Theta}_{k_2} - \boldsymbol{\Theta}_0 - \tau_2\Delta(\boldsymbol{\Theta}_0) = \eta_2 \sum_{j=0}^{k_2-1} \big(\Delta(\boldsymbol{\Theta}_j) - \Delta(\boldsymbol{\Theta}_0)\big). \tag{C.28}$$

Two bounds on the update map $\Delta$ defined in (C.25) control the right-hand side. First, $|q_2(a \mid \mathbf{x}) - \pi_a| \le 1$, $\mathbb{E}_2|z| \le 1$, and $\mathbb{E}_2|y| \le \|\boldsymbol{\mu}_2\| + 1$ (as $y \sim N(\|\boldsymbol{\mu}_2\|, 1)$) bound each entry of $\Delta$, so summing squares over the $2M$ rows,

$$\sup_{\boldsymbol{\Theta}} \|\Delta(\boldsymbol{\Theta})\|_F \le \sqrt{2M\big(1 + (\|\boldsymbol{\mu}_2\| + 1)^2\big)} \le \sqrt{2M}\,(\|\boldsymbol{\mu}_2\| + 2). \tag{C.29}$$

Second, the teacher term $q_2(\cdot \mid \mathbf{x})$ does not depend on $\boldsymbol{\Theta}$ and therefore cancels when we subtract the two update maps:

$$\Delta(\boldsymbol{\Theta}) - \Delta(\boldsymbol{\Theta}') = \mathbb{E}_2\big[\big(\mathrm{softmax}(\boldsymbol{\Theta}'\mathbf{r}) - \mathrm{softmax}(\boldsymbol{\Theta}\mathbf{r})\big)\mathbf{r}^\top\big].$$

Taking the Frobenius norm, with Jensen's inequality we have

$$\begin{aligned}\|\Delta(\boldsymbol{\Theta}) - \Delta(\boldsymbol{\Theta}')\|_F &\le \mathbb{E}_2\big[\|\big(\mathrm{softmax}(\boldsymbol{\Theta}'\mathbf{r}) - \mathrm{softmax}(\boldsymbol{\Theta}\mathbf{r})\big)\mathbf{r}^\top\|_F\big] \\ &= \mathbb{E}_2\big[\|\mathrm{softmax}(\boldsymbol{\Theta}'\mathbf{r}) - \mathrm{softmax}(\boldsymbol{\Theta}\mathbf{r})\|\,\|\mathbf{r}\|\big] \\ &\le \mathbb{E}_2\big[\|(\boldsymbol{\Theta} - \boldsymbol{\Theta}')\mathbf{r}\|\,\|\mathbf{r}\|\big] \\ &\le \|\boldsymbol{\Theta} - \boldsymbol{\Theta}'\|_F\,\mathbb{E}_2\|\mathbf{r}\|^2,\end{aligned}$$

where the third inequality uses the 1-Lipschitz property of softmax (Lemma B.6). Finally, since $\mathbf{r} = (z, y)^\top$, with $z \sim N(0,1)$ and $y \sim N(\|\boldsymbol{\mu}_2\|, 1)$ under task 2,

$$\mathbb{E}_2\|\mathbf{r}\|^2 = \mathbb{E}_2 z^2 + \mathbb{E}_2 y^2 = 1 + \left(1 + \|\boldsymbol{\mu}_2\|^2\right) = \|\boldsymbol{\mu}_2\|^2 + 2,$$

where we used $\mathbb{E}X^2 = \mathrm{Var}(X) + (\mathbb{E}X)^2$ for each coordinate. Thus

$$\|\Delta(\boldsymbol{\Theta}) - \Delta(\boldsymbol{\Theta}')\|_F \le (\|\boldsymbol{\mu}_2\|^2 + 2)\|\boldsymbol{\Theta} - \boldsymbol{\Theta}'\|_F. \tag{C.30}$$

Summing the update recursion in (C.25) over the first $j$ steps gives

$$\boldsymbol{\Theta}_j - \boldsymbol{\Theta}_0 = \eta_2 \sum_{\ell=0}^{j-1} \Delta(\boldsymbol{\Theta}_\ell).$$

Using the triangle inequality and (C.29), we obtain

$$\|\boldsymbol{\Theta}_j - \boldsymbol{\Theta}_0\|_F \le \eta_2 \sum_{\ell=0}^{j-1} \|\Delta(\boldsymbol{\Theta}_\ell)\|_F \le j\eta_2\sqrt{2M}\,(\|\boldsymbol{\mu}_2\| + 2). \tag{C.31}$$

Substituting (C.30) and (C.31) into (C.28) gives

$$\begin{aligned}
&\|\boldsymbol{\Theta}_{k_2} - \boldsymbol{\Theta}_0 - \tau_2\Delta(\boldsymbol{\Theta}_0)\|_F \\
&\le \eta_2 \sum_{j=0}^{k_2-1} (\|\boldsymbol{\mu}_2\|^2 + 2)\, j\eta_2\sqrt{2M}\,(\|\boldsymbol{\mu}_2\| + 2) \\
&\le \tfrac{1}{2}(\|\boldsymbol{\mu}_2\|^2 + 2)\sqrt{2M}\,(\|\boldsymbol{\mu}_2\| + 2)\tau_2^2 := C_{\mathrm{pop}}\tau_2^2,
\end{aligned}$$

where $C_{\mathrm{pop}} := \frac{1}{2}(\|\boldsymbol{\mu}_2\|^2 + 2)\sqrt{2M}\,(\|\boldsymbol{\mu}_2\| + 2)$. For any matrix $\mathbf{A}$, each entry satisfies $|A_{a,2}| \le \left(\sum_{b,c} |A_{b,c}|^2\right)^{1/2} = \|\mathbf{A}\|_F$. Applying this inequality to the $(a,2)$ entry of $\boldsymbol{\Theta}_{k_2} - \boldsymbol{\Theta}_0 - \tau_2\Delta(\boldsymbol{\Theta}_0)$ gives

$$|\omega_a(\mathbf{W}_2) - \omega_a(\boldsymbol{\Theta}_0) - \tau_2\Delta_{\omega_a}(\boldsymbol{\Theta}_0)| \le C_{\mathrm{pop}}\tau_2^2.$$

Substituting (C.27) yields

$$\left|\omega_a(\mathbf{W}_2) - \frac{\tau_2\|\boldsymbol{\mu}_2\|}{2}\left(\rho_{2,\sigma(a)}(a) - \frac{1}{M}\right)\right| \le C_{\mathrm{pop}}\tau_2^2, \qquad a \in \mathcal{A}. \tag{C.32}$$

Choose $a^\circ \in \arg\max_{a\in\mathcal{A}_1} \rho_{2,1}(a)$. For $a' \in \mathcal{A}_0$, assumption (b) gives $\rho_{2,1}(a^\circ) - \rho_{2,0}(a') \ge \Delta_\rho$. Since $\sigma(a^\circ) = 1$ and $\sigma(a') = 0$, (C.32) gives a lower bound for $\omega_{a^\circ}(\mathbf{W}_2)$ and an upper bound for $\omega_{a'}(\mathbf{W}_2)$:

$$\begin{aligned}
\omega_{a^\circ}(\mathbf{W}_2) &\ge \frac{\tau_2\|\boldsymbol{\mu}_2\|}{2}\left(\rho_{2,1}(a^\circ) - \frac{1}{M}\right) - C_{\mathrm{pop}}\tau_2^2, \\
\omega_{a'}(\mathbf{W}_2) &\le \frac{\tau_2\|\boldsymbol{\mu}_2\|}{2}\left(\rho_{2,0}(a') - \frac{1}{M}\right) + C_{\mathrm{pop}}\tau_2^2.
\end{aligned} \tag{C.33}$$

Subtracting the upper bound from the lower bound in (C.33) gives

$$\begin{aligned}
&\omega_{a^\circ}(\mathbf{W}_2) - \omega_{a'}(\mathbf{W}_2) \\
&\ge \frac{\tau_2\|\boldsymbol{\mu}_2\|}{2}\left[\left(\rho_{2,1}(a^\circ) - \frac{1}{M}\right) - \left(\rho_{2,0}(a') - \frac{1}{M}\right)\right] - 2C_{\mathrm{pop}}\tau_2^2 \\
&= \frac{\tau_2\|\boldsymbol{\mu}_2\|}{2}\left(\rho_{2,1}(a^\circ) - \rho_{2,0}(a')\right) - 2C_{\mathrm{pop}}\tau_2^2 \\
&\ge \frac{\tau_2\|\boldsymbol{\mu}_2\|\Delta_\rho}{2} - 2C_{\mathrm{pop}}\tau_2^2.
\end{aligned}$$

Set

$$\tau_0 := \frac{\|\boldsymbol{\mu}_2\|\Delta_\rho}{8C_{\mathrm{pop}}} = \frac{\|\boldsymbol{\mu}_2\|\Delta_\rho}{4(\|\boldsymbol{\mu}_2\|^2 + 2)\sqrt{2M}\,(\|\boldsymbol{\mu}_2\| + 2)}. \tag{C.34}$$

For $\tau_2 \le \tau_0$, $2C_{\mathrm{pop}}\tau_2^2 \le 2C_{\mathrm{pop}}\tau_0\tau_2 = \tau_2\|\boldsymbol{\mu}_2\|\Delta_\rho/4$, so for every $a' \in \mathcal{A}_0$,

$$\omega_{a^\circ}(\mathbf{W}_2) - \omega_{a'}(\mathbf{W}_2) \ge \frac{\tau_2\|\boldsymbol{\mu}_2\|\Delta_\rho}{4} > 0. \tag{C.35}$$

**Step 3: an error region for both tasks.** We construct a single error region and evaluate its probability under both task distributions. Let

$$\kappa := \frac{4(\|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\| + 2\tau_2)}{\|\boldsymbol{\mu}_2\|\Delta_\rho\tau_2}, \qquad E := \{z < 0,\ y > \kappa|z|\}. \tag{C.36}$$

Fix $\mathbf{x} \in E$ and $a' \in \mathcal{A}_0$, and evaluate all coefficients at $\mathbf{W}_2$. By the triangle inequality and (C.24) with $j = k_2$,

$$\begin{aligned} |\zeta_{a^\circ} - \zeta_{a'}| &\le |\zeta_{a^\circ}| + |\zeta_{a'}| \\ &\le 2\left(\tfrac{1}{2}\|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\| + \tau_2\right) = \|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\| + 2\tau_2. \end{aligned} \tag{C.37}$$

Since $\kappa > 0$ and $\mathbf{x} \in E$, we have $y > \kappa|z| > 0$. Using (C.23), (C.35), (C.37), and (C.36), together with the definitions of $z$ and $y$,

$$\begin{aligned} \langle \mathbf{w}_{a^\circ} - \mathbf{w}_{a'},\, \mathbf{x}\rangle &= (\zeta_{a^\circ} - \zeta_{a'})z + (\omega_{a^\circ} - \omega_{a'})y \\ &\ge -|\zeta_{a^\circ} - \zeta_{a'}|\,|z| + (\omega_{a^\circ} - \omega_{a'})y \\ &\ge -|\zeta_{a^\circ} - \zeta_{a'}|\,|z| + \frac{\tau_2\|\boldsymbol{\mu}_2\|\Delta_\rho}{4}\,y \\ &\ge -(\|\Delta\boldsymbol{\alpha}(\mathbf{W}_1)\| + 2\tau_2)|z| + \frac{\tau_2\|\boldsymbol{\mu}_2\|\Delta_\rho}{4}\,y \\ &= \frac{\tau_2\|\boldsymbol{\mu}_2\|\Delta_\rho}{4}\big(y - \kappa|z|\big) > 0. \end{aligned}$$

So on $E$ the highest-scoring token lies in $\mathcal{A}_1$ and $\hat{b}_{\mathbf{W}_2}(\mathbf{x}) = 1$, while $z < 0$ means $b^\star(\mathbf{x}) = 0$. This pointwise conclusion holds under either task distribution, so $\mathcal{L}_r(\mathbf{W}_2) \ge \mathbb{P}_r(E)$ for $r \in \{1, 2\}$. Under task $r$, Lemma B.3 gives independent coordinates $z \sim N(0, 1)$ and $y \sim N(\langle\boldsymbol{\mu}_r,\, \hat{\boldsymbol{\mu}}_2\rangle, 1)$. Integrating over $z < 0$ with the substitution $z = -t$, $t > 0$, yields

$$\begin{aligned} \mathcal{L}_r(\mathbf{W}_2) \ge \mathbb{P}_r(E) &= \int_0^\infty \phi(t)\,\mathbb{P}_r\big(y > \kappa t\big)\,\mathrm{d}t \\ &= \int_0^\infty \phi(t)\Phi(\langle\boldsymbol{\mu}_r,\, \hat{\boldsymbol{\mu}}_2\rangle - \kappa t)\,\mathrm{d}t > 0, \qquad r \in \{1, 2\}, \end{aligned} \tag{C.38}$$

where strict positivity follows because the integrand is positive for every $t > 0$.

Substituting the definitions of $\kappa$ and $\hat{\boldsymbol{\mu}}_2 = \boldsymbol{\mu}_2/\|\boldsymbol{\mu}_2\|$ into (C.38) proves part (ii) for both tasks. □

### C.4 Proof of Theorem 4

*Proof of Theorem 4.*

**Step 1: conditional sampling and integrability.** For either method, let $\mathcal{F}_s$ denote its training history strictly before step $s$, including the previous teacher draws for SFT. For the trajectory under consideration, write $\mathbf{W}_s := \mathbf{W}(s)$ and $\mathbf{m}_s := \boldsymbol{\mu}_{t(s)}$. Set $\mathbf{P}_\perp := \mathbf{I}_p - \mathbf{u}\mathbf{u}^\top$ and define

$$\mathbf{R}_s := \mathbf{W}_s^\perp = \mathbf{W}_s\mathbf{P}_\perp, \qquad \mathbf{y}_s := \mathbf{P}_\perp \mathbf{x}_s, \qquad z_s := \langle \mathbf{u}, \mathbf{x}_s\rangle, \qquad \mathbf{v}_s := (\Delta\boldsymbol{\alpha}(s))^\perp. \tag{C.39}$$

Thus $\|\mathbf{R}_s\|_F^2 = \|\mathbf{W}_s^\perp\|_F^2$. The matrix $\mathbf{W}_s$ is $\mathcal{F}_s$-measurable, whereas the current prompt is fresh. By Lemma B.3 and $\mathbf{m}_s \perp \mathbf{u}$, conditionally on $\mathcal{F}_s$,

$$\begin{aligned} &z_s \sim N(0,1), \qquad \mathbf{y}_s \sim N(\mathbf{m}_s, \mathbf{P}_\perp), \qquad z_s \perp\!\!\!\perp \mathbf{y}_s, \\ &\mathbb{E}[\mathbf{y}_s \mid \mathcal{F}_s] = \mathbf{m}_s, \qquad \mathbb{E}[\|\mathbf{y}_s\|^2 \mid \mathcal{F}_s] = \operatorname{tr}(\mathbf{P}_\perp) + \|\mathbf{m}_s\|^2 = p - 1 + \|\mathbf{m}_s\|^2. \end{aligned} \tag{C.40}$$

All expectations used below are finite. Indeed, for SFT the coefficient vector in (6) has norm at most $\sqrt{2}$, since it is the difference of two probability vectors. For RFT, each coefficient has absolute value at most $\pi_{\mathbf{W}_s}(a \mid \mathbf{x}_s)$, because (1) gives $0 \le R, \bar{R} \le 1$; hence the coefficient vector has norm at most 1. In both cases,

$$\|\mathbf{W}_s\|_F \le \sqrt{2}\eta \sum_{j=0}^{s-1} \|\mathbf{x}_j\|.$$

The right-hand side has finite moments of every order for finite $s$, as the prompts are Gaussian and the task sequence is fixed.

**Step 2: the exact semantic–style energy decomposition.** For either update in (6), the sum of all token parameters is preserved. In SFT,

$$\sum_{a\in\mathcal{A}} (\mathbf{w}_a^+ - \mathbf{w}_a) = \eta \left( \sum_a \mathbf{1}\{a = a^\star\} - \sum_a \pi_{\mathbf{W}}(a \mid \mathbf{x}) \right) \mathbf{x} = \mathbf{0}.$$

In RFT, the definition of $\bar{R}_{\mathbf{W}}$ following (6) gives

$$\begin{aligned} \sum_{a\in\mathcal{A}} (\mathbf{w}_a^+ - \mathbf{w}_a) &= \eta \left( \sum_a \pi_{\mathbf{W}}(a \mid \mathbf{x}) R(a, \mathbf{x}) - \bar{R}_{\mathbf{W}}(\mathbf{x}) \sum_a \pi_{\mathbf{W}}(a \mid \mathbf{x}) \right) \mathbf{x} \\ &= \eta(\bar{R}_{\mathbf{W}}(\mathbf{x}) - \bar{R}_{\mathbf{W}}(\mathbf{x}))\mathbf{x} = \mathbf{0}. \end{aligned}$$

Since $\mathbf{W}_0 = \mathbf{0}$, the definition of class means in (9) therefore implies

$$M\big(\boldsymbol{\alpha}^{(0)}(s) + \boldsymbol{\alpha}^{(1)}(s)\big) = \sum_a \mathbf{w}_a(s) = \mathbf{0}, \qquad \boldsymbol{\alpha}^{(1)\perp}(s) = \frac{\mathbf{v}_s}{2}, \quad \boldsymbol{\alpha}^{(0)\perp}(s) = -\frac{\mathbf{v}_s}{2}. \tag{C.41}$$

Furthermore, (9) gives $\mathbf{w}_a^\perp = \boldsymbol{\alpha}^{(b)\perp} + \boldsymbol{\beta}_a^\perp$ for $a \in \mathcal{A}_b$ and $\sum_{a\in\mathcal{A}_b} \boldsymbol{\beta}_a^\perp = \mathbf{0}$. Consequently, for each class,

$$\begin{aligned} \sum_{a\in\mathcal{A}_b} \|\mathbf{w}_a^\perp\|^2 &= M\|\boldsymbol{\alpha}^{(b)\perp}\|^2 + 2\langle \boldsymbol{\alpha}^{(b)\perp}, \sum_{a\in\mathcal{A}_b} \boldsymbol{\beta}_a^\perp\rangle + \sum_{a\in\mathcal{A}_b} \|\boldsymbol{\beta}_a^\perp\|^2 \\ &= M\|\boldsymbol{\alpha}^{(b)\perp}\|^2 + \sum_{a\in\mathcal{A}_b} \|\boldsymbol{\beta}_a^\perp\|^2. \end{aligned}$$

Summing over the two classes and using (C.41) yields the pathwise identity

$$\begin{aligned} \|\mathbf{R}_s\|_F^2 &= M \sum_{b=0}^{1} \|\boldsymbol{\alpha}^{(b)\perp}(s)\|^2 + \sum_{a\in\mathcal{A}} \|\boldsymbol{\beta}_a^\perp(s)\|^2 \\ &= \frac{M}{2}\|\mathbf{v}_s\|^2 + \|\boldsymbol{\beta}^\perp(s)\|_F^2. \end{aligned} \tag{C.42}$$

This proves (17) for both methods.

**Step 3: the SFT update and the teacher's conditional mean.** We first prove part (ii). Write $\boldsymbol{\pi}_s = (\pi_{\mathbf{W}_s}(a \mid \mathbf{x}_s))_{a\in\mathcal{A}}$ as a column vector, and let $\mathbf{e}_{a_s^\star} \in \mathbb{R}^{2M}$ be the one-hot vector of the teacher token. Right-multiplying the SFT update in (6) by $\mathbf{P}_\perp$ gives

$$\mathbf{R}_{s+1} = \mathbf{R}_s + \eta(\mathbf{e}_{a_s^\star} - \boldsymbol{\pi}_s)\mathbf{y}_s^\top. \tag{C.43}$$

Under (2), the vector $\mathbf{q}_{t(s)}(z_s) = (q_{t(s)}(a \mid \mathbf{x}_s))_{a\in\mathcal{A}}$ depends on the prompt only through the sign of $z_s$. The fresh teacher draw therefore satisfies

$$\mathbb{E}[\mathbf{e}_{a_s^\star} \mid \mathcal{F}_s, \mathbf{x}_s] = \mathbf{q}_{t(s)}(z_s).$$

Using iterated conditional expectation and then (C.40), we obtain

$$\begin{aligned} \mathbb{E}[\mathbf{e}_{a_s^\star}\mathbf{y}_s^\top \mid \mathcal{F}_s] = \mathbb{E}[\mathbf{q}_{t(s)}(z_s)\mathbf{y}_s^\top \mid \mathcal{F}_s] = \bar{\mathbf{q}}_s\mathbf{m}_s^\top, \\ \bar{\mathbf{q}}_s := \mathbb{E}[\mathbf{q}_{t(s)}(z_s)], \qquad (\bar{\mathbf{q}}_s)_a = \frac{1}{2}\rho_{t(s),\sigma(a)}(a). \end{aligned} \tag{C.44}$$

The factor $1/2$ follows from $\mathbb{P}(z_s > 0) = \mathbb{P}(z_s < 0) = 1/2$.

**Step 4: Gaussian integration by parts for the SFT policy term.** For a fixed $\mathbf{W}$ and coordinate $j$, differentiating the softmax gives

$$\frac{\partial \pi_{\mathbf{W}}(a \mid \mathbf{x})}{\partial x_j} = \pi_{\mathbf{W}}(a \mid \mathbf{x}) \left( w_{a,j} - \sum_{a'} \pi_{\mathbf{W}}(a' \mid \mathbf{x}) w_{a',j} \right).$$

Thus its Jacobian, with one row per token, is $(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top)\mathbf{W}$. For $\mathbf{x} \sim N(\mathbf{m}_s, \mathbf{I}_p)$, integration by parts coordinatewise yields $\mathbb{E}[(x_j - m_{s,j})\pi_{\mathbf{W}}(a \mid \mathbf{x})] = \mathbb{E}[\partial \pi_{\mathbf{W}}(a \mid \mathbf{x})/\partial x_j]$; the boundary term vanishes because the softmax is bounded, and its derivatives are bounded for fixed $\mathbf{W}$. Define $\bar{\boldsymbol{\pi}}_s := \mathbb{E}[\boldsymbol{\pi}_s \mid \mathcal{F}_s]$. Conditional on $\mathcal{F}_s$, this identity and right multiplication by $\mathbf{P}_\perp$ give, since $\mathbf{P}_\perp \mathbf{m}_s = \mathbf{m}_s$,

$$\begin{aligned} \mathbb{E}[\boldsymbol{\pi}_s\mathbf{y}_s^\top \mid \mathcal{F}_s] = \bar{\boldsymbol{\pi}}_s\mathbf{m}_s^\top + \mathbf{H}_s\mathbf{R}_s, \\ \mathbf{H}_s := \mathbb{E}\left[\mathrm{diag}(\boldsymbol{\pi}_s) - \boldsymbol{\pi}_s\boldsymbol{\pi}_s^\top \,\middle|\, \mathcal{F}_s\right] \succeq 0. \end{aligned} \tag{C.45}$$

Indeed, for any $\mathbf{v} \in \mathbb{R}^{2M}$ and probability vector $\boldsymbol{\pi}$,

$$\mathbf{v}^\top(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top)\mathbf{v} = \sum_a \pi_a v_a^2 - \left(\sum_a \pi_a v_a\right)^2 \geq 0,$$

so taking the conditional expectation preserves positive semidefiniteness.

**Step 5: the SFT energy recursion.** The Frobenius inner product is $\langle \mathbf{U}, \mathbf{V}\rangle_F = \mathrm{tr}(\mathbf{U}^\top\mathbf{V})$, and $\|\mathbf{v}\mathbf{y}^\top\|_F^2 = \|\mathbf{v}\|^2\|\mathbf{y}\|^2$. Expanding (C.43) therefore gives

$$\begin{aligned} \|\mathbf{R}_{s+1}\|_F^2 = \|\mathbf{R}_s\|_F^2 + 2\eta\langle \mathbf{R}_s, (\mathbf{e}_{a_s^\star} - \boldsymbol{\pi}_s)\mathbf{y}_s^\top\rangle_F \\ + \eta^2\|\mathbf{e}_{a_s^\star} - \boldsymbol{\pi}_s\|^2\|\mathbf{y}_s\|^2. \end{aligned}$$

Equations (C.44) and (C.45), and the $\mathcal{F}_s$-measurability of $\mathbf{R}_s$, yield the exact identity

$$\begin{aligned} \mathbb{E}[\|\mathbf{R}_{s+1}\|_F^2 \mid \mathcal{F}_s] = \|\mathbf{R}_s\|_F^2 - 2\eta\,\mathrm{tr}(\mathbf{R}_s^\top\mathbf{H}_s\mathbf{R}_s) \\ + 2\eta\langle \mathbf{R}_s, (\bar{\mathbf{q}}_s - \bar{\boldsymbol{\pi}}_s)\mathbf{m}_s^\top\rangle_F \\ + \eta^2\mathbb{E}\left[\|\mathbf{e}_{a_s^\star} - \boldsymbol{\pi}_s\|^2\|\mathbf{y}_s\|^2 \,\middle|\, \mathcal{F}_s\right]. \end{aligned} \tag{C.46}$$

Since $\mathbf{H}_s \succeq 0$, each column $\mathbf{r}$ of $\mathbf{R}_s$ satisfies $\mathbf{r}^\top\mathbf{H}_s\mathbf{r} \geq 0$; summing over columns shows that the trace in (C.46) is nonnegative. Both $\bar{\mathbf{q}}_s$ and $\bar{\boldsymbol{\pi}}_s$ are probability vectors, so their squared norms are at most 1 and their inner product is nonnegative. Thus $\|\bar{\mathbf{q}}_s - \bar{\boldsymbol{\pi}}_s\| \leq \sqrt{2}$, and Cauchy–Schwarz gives

$$\langle \mathbf{R}_s, (\bar{\mathbf{q}}_s - \bar{\boldsymbol{\pi}}_s)\mathbf{m}_s^\top\rangle_F \leq \sqrt{2}\,\|\mathbf{R}_s\|_F\|\mathbf{m}_s\|.$$

Moreover, for every demonstrated token $a$ and probability vector $\boldsymbol{\pi}$,

$$\|\mathbf{e}_a - \boldsymbol{\pi}\|^2 = (1-\pi_a)^2 + \sum_{a'\neq a}\pi_{a'}^2 \leq (1-\pi_a)^2 + \left(\sum_{a'\neq a}\pi_{a'}\right)^2 = 2(1-\pi_a)^2 \leq 2.$$

Substituting these estimates and (C.40) into (C.46) gives

$$\begin{aligned} \mathbb{E}[\|\mathbf{R}_{s+1}\|_F^2 \mid \mathcal{F}_s] \leq \|\mathbf{R}_s\|_F^2 + 2\sqrt{2}\eta\|\mathbf{m}_s\|\|\mathbf{R}_s\|_F \\ + 2\eta^2\big(p - 1 + \|\mathbf{m}_s\|^2\big). \end{aligned} \tag{C.47}$$

**Step 6: iterate the SFT recursion.** Set $e_s := \mathbb{E}\|\mathbf{R}_s\|_F^2$. The means $\mathbf{m}_s$ are deterministic. Taking expectations in (C.47) and using $\mathbb{E}\|\mathbf{R}_s\|_F \le \sqrt{e_s}$ yields

$$\begin{aligned} e_{s+1} &\le e_s + 2\sqrt{2}\eta\|\mathbf{m}_s\|\sqrt{e_s} + 2\eta^2\|\mathbf{m}_s\|^2 + 2\eta^2(p-1) \\ &= \left(\sqrt{e_s} + \sqrt{2}\eta\|\mathbf{m}_s\|\right)^2 + 2\eta^2(p-1). \end{aligned} \tag{C.48}$$

To iterate this bound, let

$$A_s := \sqrt{2}\eta \sum_{j=0}^{s-1} \|\mathbf{m}_j\|, \qquad V_s := 2\eta^2 s(p-1).$$

We claim $\sqrt{e_s} \le A_s + \sqrt{V_s}$. The claim holds at $s = 0$. For the induction step, use the elementary inequality

$$\sqrt{(A + \sqrt{V} + c)^2 + d} \le A + c + \sqrt{V + d}, \qquad A, V, c, d \ge 0.$$

Indeed, the square of the right-hand side minus the square of the left-hand side is $2(A+c)(\sqrt{V+d} - \sqrt{V}) \ge 0$, and both sides are nonnegative. Apply it with $A = A_s$, $V = V_s$, $c = \sqrt{2}\eta\|\mathbf{m}_s\|$, and $d = 2\eta^2(p-1)$. By (C.48) and the induction hypothesis,

$$\begin{aligned} \sqrt{e_{s+1}} &\le \sqrt{(A_s + \sqrt{V_s} + c)^2 + d} \\ &\le A_s + c + \sqrt{V_s + d} = A_{s+1} + \sqrt{V_{s+1}}. \end{aligned}$$

Evaluating the resulting bound at $s = N$ gives

$$\mathbb{E}\|\mathbf{R}_N\|_F^2 \le 2\eta^2 \left(\sqrt{N(p-1)} + \sum_{s=0}^{N-1} \|\boldsymbol{\mu}_{t(s)}\|\right)^2. \tag{C.49}$$

By (C.39), this is precisely (19), proving part (ii).

**Step 7: reduce RFT to its semantic channel.** We now consider the RFT trajectory and prove part (i). Since $\mathbf{W}_0 = \mathbf{0}$, Theorem 2(i) gives $\boldsymbol{\beta}(s) = \mathbf{0}$ for every step, for each realization of the prompts. By (3) and (12), equal class sizes then give

$$P_1(\mathbf{W}_s, \mathbf{x}) = \frac{Me^{\langle \boldsymbol{\alpha}^{(1)}(s), \mathbf{x}\rangle}}{Me^{\langle \boldsymbol{\alpha}^{(0)}(s), \mathbf{x}\rangle} + Me^{\langle \boldsymbol{\alpha}^{(1)}(s), \mathbf{x}\rangle}} = \frac{1}{1 + e^{-\langle \Delta\boldsymbol{\alpha}(s), \mathbf{x}\rangle}}.$$

Define

$$c_s := \langle \Delta\boldsymbol{\alpha}(s),\, \mathbf{u}\rangle, \qquad \lambda := \frac{2\eta}{M}, \qquad f(r) := \frac{e^r}{(1+e^r)^2} = \frac{1}{4\cosh^2(r/2)}. \tag{C.50}$$

In particular, $f$ is even and $0 < f(r) \le 1/4$ for every $r$. Since $P_0 = 1 - P_1$, either value of the correct class satisfies

$$P_{b^\star}(\mathbf{W}_s, \mathbf{x}_s)\big(1 - P_{b^\star}(\mathbf{W}_s, \mathbf{x}_s)\big) = f(c_s z_s + \langle \mathbf{v}_s,\, \mathbf{y}_s\rangle)\,. \tag{C.51}$$

Here we used (C.39) to write $\Delta\boldsymbol{\alpha}(s) = c_s\mathbf{u} + \mathbf{v}_s$ and $\mathbf{x}_s = z_s\mathbf{u} + \mathbf{y}_s$. Let $\varepsilon_s := 2b^\star(\mathbf{x}_s) - 1$. By (1), $\varepsilon_s = \operatorname{sign}(z_s)$ almost surely and $\varepsilon_s z_s = |z_s|$; the event $z_s = 0$ has probability zero. Substituting (C.51) into the exact semantic update (13), and projecting onto $\mathbf{u}$ and $\mathbf{u}^\perp$, gives

$$\begin{aligned} c_{s+1} &= c_s + \lambda|z_s| f(c_s z_s + \langle \mathbf{v}_s,\, \mathbf{y}_s\rangle)\,, \\ \mathbf{v}_{s+1} &= \mathbf{v}_s + \lambda\varepsilon_s f(c_s z_s + \langle \mathbf{v}_s,\, \mathbf{y}_s\rangle)\,\mathbf{y}_s. \end{aligned} \tag{C.52}$$

The first line and $c_0 = 0$ imply $c_{s+1} \ge c_s \ge 0$ for all $s$, without any upper restriction on $\eta$.

**Step 8: the conditional RFT drift has a nonpositive radial component.** Condition on $\mathcal{F}_s$, so that $c_s \geq 0$ and $\mathbf{v}_s$ are fixed. For $c \geq 0$ and $t \in \mathbb{R}$, define

$$\begin{aligned} h_c(t) &:= \mathbb{E}_{Z\sim N(0,1)}[\operatorname{sign}(Z) f(cZ+t)] \\ &= \int_0^\infty \phi(z)\big[f(cz+t) - f(cz-t)\big]\,\mathrm{d}z, \end{aligned} \tag{C.53}$$

where $\phi$ is the standard normal density. The second line follows by splitting the integral at zero, substituting $z \mapsto -z$ in the negative half, and using the evenness of both $\phi$ and $f$. Differentiating (C.50) gives $f'(r) = -f(r)\tanh(r/2)$, so $f$ is nonincreasing on $[0,\infty)$. For $c, z, t \geq 0$, the inequality $|cz+t| \geq |cz-t|$ therefore implies $f(cz+t) \leq f(cz-t)$. Equation (C.53) then yields $h_c(t) \leq 0$ for $t \geq 0$. The same equation gives $h_c(-t) = -h_c(t)$, hence

$$t\, h_c(t) \leq 0 \qquad \text{for every } c \geq 0 \text{ and } t \in \mathbb{R}. \tag{C.54}$$

Write the conditional mean of the unscaled off-axis increment as

$$\mathbf{g}_s := \mathbb{E}[\varepsilon_s f(c_s z_s + \langle \mathbf{v}_s, \mathbf{y}_s\rangle)\, \mathbf{y}_s \,|\, \mathcal{F}_s]\,.$$

By the conditional independence in (C.40), we may first average over $z_s$ while holding $\mathbf{y}_s$ fixed. Using (C.53) then gives

$$\begin{aligned} \langle \mathbf{v}_s, \mathbf{g}_s\rangle &= \mathbb{E}[\langle \mathbf{v}_s, \mathbf{y}_s\rangle\, \varepsilon_s f(c_s z_s + \langle \mathbf{v}_s, \mathbf{y}_s\rangle) \,|\, \mathcal{F}_s] \\ &= \mathbb{E}[\langle \mathbf{v}_s, \mathbf{y}_s\rangle\, h_{c_s}(\langle \mathbf{v}_s, \mathbf{y}_s\rangle) \,|\, \mathcal{F}_s] \leq 0, \end{aligned} \tag{C.55}$$

where the last inequality applies (C.54) pointwise in $\mathbf{y}_s$. This argument allows any $\mathbf{m}_s \perp \mathbf{u}$; it does not require zero task means.

**Step 9: the RFT energy recursion and final bound.** Expanding the second line of (C.52), taking conditional expectations, and using $\varepsilon_s^2 = 1$ gives

$$\begin{aligned} \mathbb{E}[\|\mathbf{v}_{s+1}\|^2 \mid \mathcal{F}_s] &= \|\mathbf{v}_s\|^2 + 2\lambda\langle \mathbf{v}_s, \mathbf{g}_s\rangle \\ &\quad + \lambda^2 \mathbb{E}\Big[f(c_s z_s + \langle \mathbf{v}_s, \mathbf{y}_s\rangle)^2\, \|\mathbf{y}_s\|^2 \,\Big|\, \mathcal{F}_s\Big] \\ &\leq \|\mathbf{v}_s\|^2 + \frac{\lambda^2}{16}(p - 1 + \|\mathbf{m}_s\|^2) \\ &= \|\mathbf{v}_s\|^2 + \frac{\eta^2}{4M^2}(p - 1 + \|\mathbf{m}_s\|^2). \end{aligned} \tag{C.56}$$

The inequality uses (C.55), $f^2 \leq 1/16$ from (C.50), and the second moment in (C.40). Taking expectations and summing (C.56) from $s = 0$ to $N-1$, with $\mathbf{v}_0 = \mathbf{0}$, yields

$$\mathbb{E}\|\mathbf{v}_N\|^2 \leq \frac{\eta^2}{4M^2} \sum_{s=0}^{N-1} (p - 1 + \|\mathbf{m}_s\|^2). \tag{C.57}$$

Finally, $\boldsymbol{\beta}(N) = \mathbf{0}$ and (C.42) give

$$\begin{aligned} \mathbb{E}\|\mathbf{W}_{\mathrm{RFT}}^{\perp}(N)\|_F^2 &= \frac{M}{2}\mathbb{E}\|\mathbf{v}_N\|^2 \\ &\leq \frac{\eta^2}{8M}\left(N(p-1) + \sum_{s=0}^{N-1} \|\boldsymbol{\mu}_{t(s)}\|^2\right), \end{aligned}$$

which is (18) and proves part (i). Both parts include $N = 0$, with empty sums and zero initial energy. Fresh sampling is essential to the conditional calculations in (C.40) and the subsequent expectations; the proof does not directly apply to reuse of a fixed prompt sample. For RFT, exact per-prompt gradients are also essential to the pathwise style invariance used in Step 7; sampled action gradients are discussed separately in Appendix D.1. □

# D ADDITIONAL RESULTS

## D.1 EXACT AND SAMPLED POLICY GRADIENTS

The invariance in Theorem 2(i) concerns the exact per-prompt gradient (6). Here "exact" means that the expectation over actions is evaluated by summing over all tokens at the current policy and prompt. When this summation is computationally impractical, REINFORCE provides an unbiased gradient estimate using actions sampled from the current policy. This subsection explains how action sampling may change the style dynamics.

**An unbiased gradient estimate.** Fix $\mathbf{W}$ and $\mathbf{x}$, and draw $A \sim \pi_{\mathbf{W}}(\cdot \mid \mathbf{x})$. The REINFORCE estimator without a baseline (Williams, 1992) gives the update

$$\mathbf{w}_a^+ = \mathbf{w}_a + \eta\, \widehat{\mathbf{g}}_a(A), \qquad \widehat{\mathbf{g}}_a(A) = R(A,\mathbf{x})\big(\mathbf{1}\{a = A\} - \pi_{\mathbf{W}}(a \mid \mathbf{x})\big)\mathbf{x}. \tag{D.1}$$

All expectations below are conditional on the current $\mathbf{W}, \mathbf{x}$ and average only over the sampled action. Since $\mathbb{E}_A[R(A,\mathbf{x})\mathbf{1}\{a = A\}] = \pi_{\mathbf{W}}(a \mid \mathbf{x})R(a,\mathbf{x})$, we have

$$\mathbb{E}_A[\widehat{\mathbf{g}}_a(A)] = \pi_{\mathbf{W}}(a \mid \mathbf{x})\big(R(a,\mathbf{x}) - \bar{R}_{\mathbf{W}}(\mathbf{x})\big)\mathbf{x}. \tag{D.2}$$

Thus the sampled update agrees with the RFT update in (6) in conditional mean.

**Style noise from a single sampled action.** Subtracting the mean update within $a$'s class gives

$$\begin{aligned}\Delta\boldsymbol{\beta}_a &:= \boldsymbol{\beta}_a^+ - \boldsymbol{\beta}_a \\ &= \eta R(A,\mathbf{x})\Big[\mathbf{1}\{a = A\} - \frac{1}{M}\mathbf{1}\{\sigma(A) = \sigma(a)\} - \delta_a(\mathbf{W},\mathbf{x})\Big]\mathbf{x},\end{aligned} \tag{D.3}$$

where $\delta_a$ is defined in (C.18). When $\boldsymbol{\beta} = \mathbf{0}$, $\delta_a = 0$ and (D.2) implies $\mathbb{E}_A[\Delta\boldsymbol{\beta}_a] = \mathbf{0}$. Nevertheless, if $M > 1$ and the sampled token is rewarded, its style increment is $\eta(1 - 1/M)\mathbf{x}$, while every other token in the same class receives $-\eta\mathbf{x}/M$. Tokens in the other class receive zero style increment. For $\eta > 0$ and $\mathbf{x} \neq \mathbf{0}$, a rewarded step produces $\boldsymbol{\beta}^+ \neq \mathbf{0}$. An unrewarded sample produces no update for this estimator.

Let $\Delta\boldsymbol{\beta}$ collect the style increments over all tokens. When $\boldsymbol{\beta} = \mathbf{0}$, the conditional squared style increment is

$$\mathbb{E}_A\big[\|\Delta\boldsymbol{\beta}\|_F^2\big] = \eta^2 P_{b^\star(\mathbf{x})}(\mathbf{W},\mathbf{x})\left(1 - \frac{1}{M}\right)\|\mathbf{x}\|^2. \tag{D.4}$$

Indeed, a rewarded sample occurs with probability $P_{b^\star(\mathbf{x})}(\mathbf{W},\mathbf{x})$, and its squared coefficients sum to $(1-1/M)^2 + (M-1)/M^2 = 1 - 1/M$. Since the conditional mean increment is zero when $\boldsymbol{\beta} = \mathbf{0}$, (D.4) is also the total conditional variance. The corresponding off-axis expression replaces $\mathbf{x}$ by $\mathbf{x}^\perp$. Hence a zero mean style increment can coexist with positive expected style energy after one step.

**Averaging actions and following the trajectory.** If each update averages $K$ independent action-gradient estimates computed at the same $\mathbf{W}, \mathbf{x}$, its conditional mean is the RFT update in (6), while the variance in (D.4) is divided by $K$. Finite $K$ therefore reduces style noise but does not guarantee that every update preserves $\boldsymbol{\beta} = \mathbf{0}$. Once a sampled trajectory develops $\boldsymbol{\beta} \neq \mathbf{0}$, the zero mean style-increment statement above no longer applies: subsequent conditional means follow the exact gradient at the new, generally nonuniform policy. Theorem 2(i) consequently does not imply zero style drift along sampled trajectories.

We compare REINFORCE updates with different numbers of sampled actions per prompt against exact RFT in Appendix E.2 (Figure E.4(b,d)).

## D.2 LIMITING BEHAVIOR UNDER CONTINUED TASK-2 TRAINING

Theorem 3 establishes positive SFT semantic error on both tasks over a finite training interval. We now consider continued training on task 2. In the unregularized population setting, sufficiently small fixed steps eventually drive the semantic errors on both tasks to zero, even though the teacher's style preferences can cause errors at intermediate times.

**Proposition D.1** (Vanishing semantic error under continued population SFT)**.** *Under the setup of Section 2.1, let* $\mathbf{W}^{(0)} = \mathbf{W}_1$ *be any finite checkpoint, and run population SFT on task* 2 *with a fixed step size satisfying*

$$0 < \eta_2 \le \frac{2}{1 + \|\boldsymbol{\mu}_2\|^2}. \tag{D.5}$$

*Then, as* $k \to \infty$ *,*

$$\mathcal{L}_1(\mathbf{W}^{(k)}) \longrightarrow 0, \qquad \mathcal{L}_2(\mathbf{W}^{(k)}) \longrightarrow 0. \tag{D.6}$$

*In particular, from the checkpoint in Theorem 3,*

$$\mathcal{F}^{(2)}_{\mathrm{SFT}} \longrightarrow 0, \qquad \mathcal{G}^{(2)}_{\mathrm{SFT}} \longrightarrow 0,$$

*whereas RFT keeps both quantities identically zero throughout training.*

*Proof.* Let $f_2$ denote the task-2 population token cross-entropy, and let $f_{2,*}$ be its infimum over finite parameter matrices:

$$f_2(\mathbf{W}) := \mathbb{E}_2\left[-\sum_{a\in\mathcal{A}} q_2(a \mid \mathbf{x}) \log \pi_{\mathbf{W}}(a \mid \mathbf{x})\right], \qquad f_{2,*} := \inf_{\mathbf{V}} f_2(\mathbf{V}).$$

We first show that approaching this infimum forces semantic error to vanish, and then verify that population optimization approaches it.

**Separating class and style losses.** For $a \in \mathcal{A}_b$, write $\pi_{\mathbf{W}}(a \mid \mathbf{x}, b) := \pi_{\mathbf{W}}(a \mid \mathbf{x})/P_b(\mathbf{W}, \mathbf{x})$. This conditional policy depends only on the style channel $\boldsymbol{\beta}$. Since the teacher is always semantically correct,

$$\begin{aligned} f_2(\mathbf{W}) &= C_2(\mathbf{W}) + S_2(\boldsymbol{\beta}), \\ C_2(\mathbf{W}) &:= \mathbb{E}_2[-\log P_{b^\star(\mathbf{x})}(\mathbf{W}, \mathbf{x})], \\ S_2(\boldsymbol{\beta}) &:= \mathbb{E}_2\left[-\sum_{a\in\mathcal{A}_{b^\star(\mathbf{x})}} \rho_{2,b^\star(\mathbf{x})}(a) \log \pi_{\mathbf{W}}(a \mid \mathbf{x}, b^\star(\mathbf{x}))\right]. \end{aligned} \tag{D.7}$$

Let $S_{2,*}$ be the infimum of $S_2$ over finite style channels whose rows sum to zero within each class. We claim that $f_{2,*} = S_{2,*}$. The inequality $f_{2,*} \ge S_{2,*}$ follows from $C_2 \ge 0$. Conversely, fix such a style channel and set $\mathbf{w}_a^{[s]} = \boldsymbol{\beta}_a + (2\sigma(a) - 1)s\mathbf{u}/2$ for $s \ge 0$. This leaves $S_2$ unchanged. With $z = \langle \mathbf{u}, \mathbf{x}\rangle$ and $b = b^\star(\mathbf{x})$, the class loss is

$$-\log P_b(\mathbf{W}^{[s]}, \mathbf{x}) = \log\left(1 + e^{-s|z|}\frac{\sum_{a\in\mathcal{A}_{1-b}} e^{\langle \boldsymbol{\beta}_a, \mathbf{x}\rangle}}{\sum_{a\in\mathcal{A}_b} e^{\langle \boldsymbol{\beta}_a, \mathbf{x}\rangle}}\right).$$

It tends to zero almost surely and is bounded by its value at $s = 0$, which is integrable because the fixed linear logits have finite Gaussian first moments. Dominated convergence yields $C_2(\mathbf{W}^{[s]}) \to 0$, hence $f_{2,*} \le S_2(\boldsymbol{\beta})$ for every finite style channel. Taking the infimum proves the claim and gives

$$0 \le C_2(\mathbf{W}) \le f_2(\mathbf{W}) - f_{2,*}. \tag{D.8}$$

**From objective values to semantic errors.** If greedy decoding selects a wrong-class token, every correct-class token has probability at most that selected token's probability, and thus at most the total wrong-class mass. Consequently, $P_b(\mathbf{W}, \mathbf{x}) \le M(1 - P_b(\mathbf{W}, \mathbf{x}))$ on the error event, including ties resolved to the wrong class. The class loss there is at least $\log(1 + 1/M)$, so

$$\mathcal{L}_2(\mathbf{W}) \le \frac{C_2(\mathbf{W})}{\log(1 + 1/M)} \le \frac{f_2(\mathbf{W}) - f_{2,*}}{\log(1 + 1/M)}. \tag{D.9}$$

Both tasks share the same semantic rule and hence the same error event $E_{\mathbf{W}} = \{\mathbf{x} : \hat{b}_{\mathbf{W}}(\mathbf{x}) \ne b^\star(\mathbf{x})\}$. Writing $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, the density ratio of their prompt distributions is

$$r(\mathbf{x}) = \exp\left(\langle \boldsymbol{\delta}, \mathbf{x} - \boldsymbol{\mu}_2\rangle - \tfrac{1}{2}\|\boldsymbol{\delta}\|^2\right), \qquad \mathbb{E}_2[r(\mathbf{x})^2] = e^{\|\boldsymbol{\delta}\|^2}.$$

Cauchy–Schwarz therefore gives, uniformly in $\mathbf{W}$,

$$\mathcal{L}_1(\mathbf{W}) = \mathbb{E}_2[r(\mathbf{x})\mathbf{1}\{\mathbf{x} \in E_{\mathbf{W}}\}] \le e^{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/2}\sqrt{\mathcal{L}_2(\mathbf{W})}. \tag{D.10}$$

Thus any sequence with $f_2(\mathbf{W}) \to f_{2,*}$ has vanishing semantic error on both tasks.

**Convergence of the population objective.** The function $f_2$ is convex. For any perturbation matrix $\mathbf{A}$, its Hessian satisfies

$$D^2 f_2(\mathbf{W})[\mathbf{A},\mathbf{A}] = \mathbb{E}_2\big[(\mathbf{A}\mathbf{x})^\top(\operatorname{diag}\boldsymbol{\pi} - \boldsymbol{\pi}\boldsymbol{\pi}^\top)(\mathbf{A}\mathbf{x})\big] \le \tfrac{1}{2}\mathbb{E}_2\|\mathbf{A}\mathbf{x}\|^2 \le \tfrac{1}{2}(1+\|\boldsymbol{\mu}_2\|^2)\|\mathbf{A}\|_F^2,$$

where $\boldsymbol{\pi} = \pi_{\mathbf{W}}(\cdot \mid \mathbf{x})$. Here the categorical covariance has operator norm at most $1/2$, since its absolute row sums are $2\pi_a(1-\pi_a) \le 1/2$, and $\mathbb{E}_2[\mathbf{x}\mathbf{x}^\top] = \mathbf{I}_p + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top$. Hence $f_2$ is $L$-smooth for $L = (1+\|\boldsymbol{\mu}_2\|^2)/2$. Under (D.5), convexity and the descent inequality give, for every finite comparator $\mathbf{V}$,

$$f_2(\mathbf{W}^{(j+1)}) - f_2(\mathbf{V}) \le \frac{\|\mathbf{W}^{(j)} - \mathbf{V}\|_F^2 - \|\mathbf{W}^{(j+1)} - \mathbf{V}\|_F^2}{2\eta_2}.$$

Summing and using monotonicity of the objective yields

$$f_2(\mathbf{W}^{(k)}) - f_2(\mathbf{V}) \le \frac{\|\mathbf{W}_1 - \mathbf{V}\|_F^2}{2k\eta_2}. \tag{D.11}$$

First letting $k \to \infty$ and then taking the infimum over $\mathbf{V}$ proves $f_2(\mathbf{W}^{(k)}) \to f_{2,*}$ without assuming that a finite minimizer exists. Equations (D.9) and (D.10) now imply (D.6). When $\mathcal{L}_1(\mathbf{W}_1) = 0$, forgetting equals the final task-1 error, and generalization error is the average of the two task errors, proving the asserted metric limits. □

**What converges.** The objective approaches its infimum and semantic errors vanish, but the parameters do not converge to a finite minimizer. Indeed, if $\mathbf{U}$ has rows $(2\sigma(a)-1)\mathbf{u}/2$, then at every finite $\mathbf{W}$,

$$\frac{\mathrm{d}}{\mathrm{d}s} f_2(\mathbf{W} + s\mathbf{U})\Big|_{s=0} = -\mathbb{E}_2\big[|z|\big(1 - P_{b^\star(\mathbf{x})}(\mathbf{W},\mathbf{x})\big)\big] < 0.$$

Thus no finite matrix attains $f_{2,*}$, and any sequence approaching this infimum must have $\|\mathbf{W}\|_F \to \infty$; otherwise a bounded subsequence would yield a finite minimizer by continuity. The token cross-entropy need not tend to zero, and exact matching of the teacher's style preferences is unnecessary for vanishing semantic error. The finite-time errors in Theorem 3 therefore do not persist at a positive level in this population limit. This does not assert monotone semantic improvement or zero error after finitely many steps.

### D.3 Numerical illustration of finite-time forgetting

We illustrate the finite-time separation in Theorem 3 and the subsequent recovery described in Appendix D.2 by varying training time at a fixed step size.

**Setup.** We take $p = 2$, $M = 2$, $\mathbf{u} = (1,0)^\top$, and Gaussian task means $\boldsymbol{\mu}_1 = (0,2)^\top$ and $\boldsymbol{\mu}_2 = (0,3)^\top$. The task-2 teachers are $\rho_{2,0} = (0.5, 0.5)$ and $\rho_{2,1} = (0.99, 0.01)$, giving $\Delta_\rho = 0.49$. Both methods start from the constructed checkpoint $\mathbf{w}_a = (2\sigma(a)-1)\mathbf{u}/2$, so that $\Delta\boldsymbol{\alpha} = \mathbf{u}$ and $\boldsymbol{\beta} = \mathbf{0}$. This checkpoint already classifies both tasks perfectly because they share the same semantic rule. We then train only on task 2, using population updates with fixed step size $\eta = 10^{-3}$ for $10^6$ steps, up to $\tau = k\eta = 1000$. Here $\mathcal{L}_1$ equals forgetting, and $\mathcal{G}^{(2)} = (\mathcal{L}_1 + \mathcal{L}_2)/2$. The guaranteed interval in (C.34) ends at $\tau_0 \approx 0.00334$; the chosen step size also satisfies the sufficient condition (D.5) for eventual recovery.

**Numerical evaluation.** We evaluate the SFT teacher term analytically and its policy term by deterministic Gaussian quadrature, with 12 Gauss–Legendre nodes on each of six intervals in positive $z = \langle \mathbf{u}, \mathbf{x}\rangle$ and their negative reflections, and 40 Gauss–Hermite nodes in the orthogonal coordinate (5760 nodes in total; $|z| \le 11$). RFT follows its scalar on-axis semantic recursion from Theorem 1(ii), preserving zero style and zero semantic error. SFT errors are computed by conditioning on $z$, evaluating the Gaussian probability of the greedy decision region in the other coordinate, and integrating over $z$.

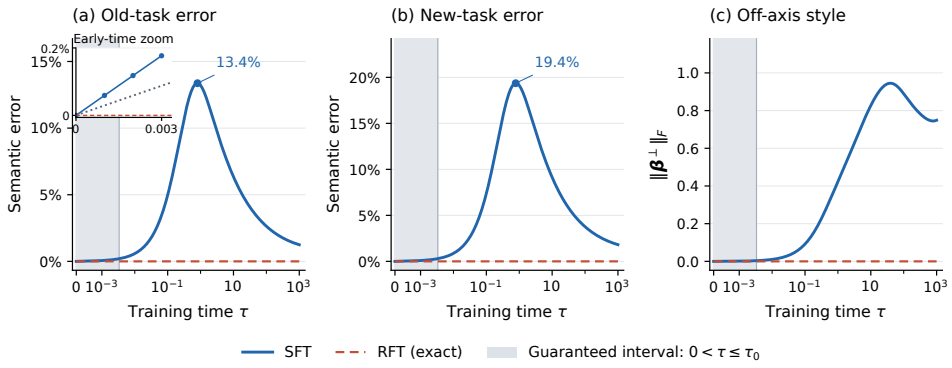



Figure D.1: **Two-task population dynamics at a fixed step size.** *(a)* Old-task error, equal to forgetting from the common perfectly fitted checkpoint. *(b)* New-task error. *(c)* Off-axis style norm $\|\boldsymbol{\beta}^{\perp}\|_F$. The horizontal axis is logarithmic outside a small linear neighborhood of zero. Shading marks $0 < \tau \leq \tau_0$; the inset compares the first three SFT steps with the old-task error lower bound in Theorem 3 (dotted). Later times illustrate behavior beyond that guaranteed interval.

**Results.** In Figure D.1, SFT's old- and new-task errors rise from zero to approximately $13.4\%$ and $19.4\%$, respectively, near $\tau = 0.8$, then decline to $1.25\%$ and $1.81\%$ at $\tau = 1000$. The shaded steps lie inside the guaranteed interval and have positive old-task error above the theorem's lower bound. RFT retains zero error on both tasks throughout. SFT's error reduction occurs while its off-axis style norm remains nonzero (as the semantic margin grows). Thus this example exhibits substantial recovery without eliminating style drift.

# E Control and Ablation Experiments

## E.1 Control Experiments

We isolate the roles of task means and teacher preferences in population style drift. Unless specified below, all settings follow Section 4.

**Zero task means.** Setting $\boldsymbol{\mu}_t = \mathbf{0}$ removes population off-axis style drift. Both population methods then achieve negligible generalization error and forgetting (Figure E.1); their small residual off-axis energies reflect Monte Carlo prompt sampling. Finite-sample SFT still develops greater off-axis energy and a worse generalization–forgetting tradeoff than RFT, illustrating sampling effects even when population drift is absent.

**Uniform teachers.** Keeping the nonzero task means of the main experiment, we set $\mathbf{v}_{t,b} = \mathbf{0}$ in (20), giving $\rho_{t,b}(a) = 1/M$. Removing teacher preferences also eliminates the population performance gap (Figure E.2). Population SFT now preserves $\boldsymbol{\beta} = \mathbf{0}$ because its updates are constant within each semantic class, including under Monte Carlo prompt averaging. Its residual off-axis energy therefore comes from the semantic component. Finite-sample SFT still introduces style fluctuations by sampling individual teacher tokens. RFT is unchanged from Figure 2, since it does not use the teacher and the prompts and seeds are shared.

## E.2 Ablation Experiments

We examine sensitivity to three assumptions: Gaussian prompts, correct semantic rewards, and exact action expectations. Unless specified otherwise, we follow the setup of Section 4.

**Non-Gaussian prompts.** We keep the task means of Figure 2 and replace the Gaussian prompt noise by independent $\mathrm{Laplace}(0, 1/\sqrt{2})$ coordinates, $\mathbf{x} = \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}$, which preserves the means and the identity covariance. The results are nearly unchanged from Figure 2 (Figure E.3): population SFT develops essentially the same off-axis energy and forgetting, and the finite-sample curves are

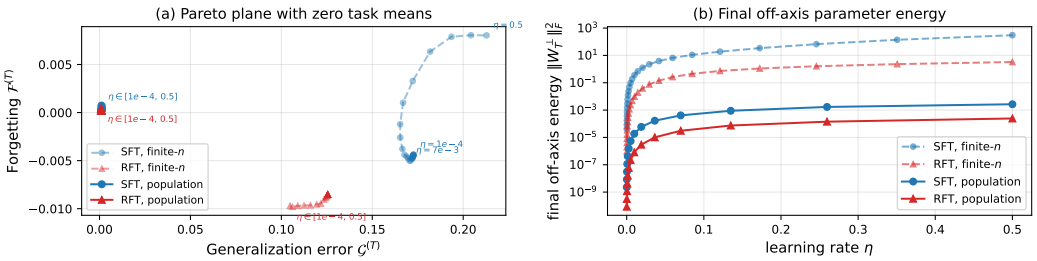



Figure E.1: **Zero task means.** We set $\boldsymbol{\mu}_t = \mathbf{0}$ for every task. As in Figure 2, panel (a) shows generalization error versus forgetting, and panel (b) shows final off-axis energy versus learning rate. Solid and dashed curves denote population and finite-sample updates, respectively.

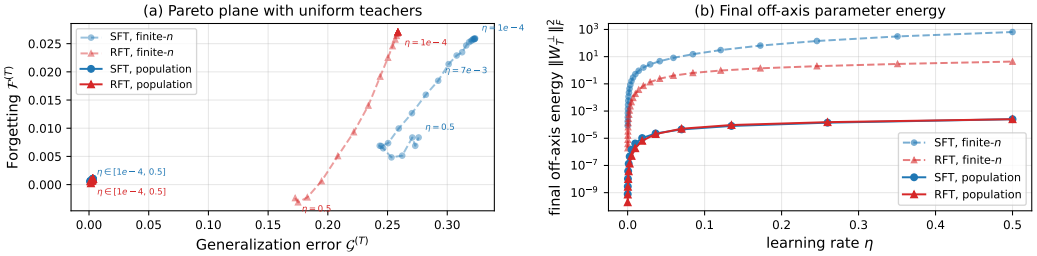



Figure E.2: **Uniform teachers.** We retain the nonzero Gaussian task means of Figure 2 and set $\rho_{t,b}(a) = 1/M$ for every task and class. The layout and line styles match Figure E.1.

close. Population RFT keeps forgetting below $0.3\%$ but acquires a small off-axis energy (about $3 \times 10^{-2}$ at $\eta = 0.5$, versus $2.5 \times 10^{-4}$ with Gaussian prompts) and error up to $1.5\%$: its algebraic style invariance holds for any prompt law, whereas its exactly on-axis population direction relies on Gaussian prompts.

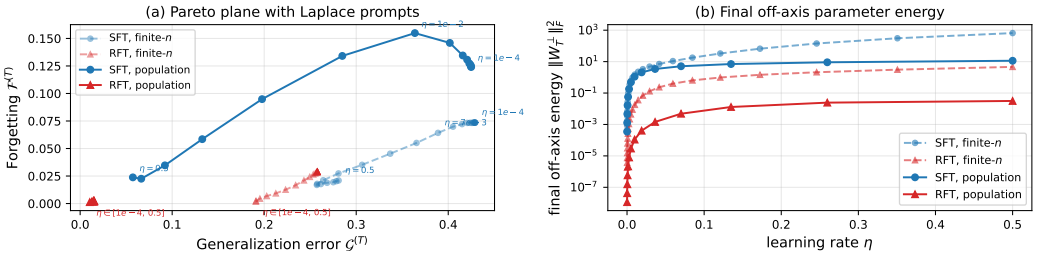



Figure E.3: **Non-Gaussian prompts.** Prompt noise coordinates are drawn independently from $\mathrm{Laplace}(0, 1/\sqrt{2})$ around the task means of Figure 2, giving identity covariance. The layout and line styles match Figure E.1.

**Imperfect rewards.** For $q \in \{0, 0.05, 0.1, 0.2\}$, we apply either a *class flip*, replacing the entire reward vector by $1 - R(\cdot, \mathbf{x})$ with probability $q$, or *token flips*, independently reversing each token's reward with probability $q$. Corruptions are sampled independently for each prompt, including within Monte Carlo (MC) batches. RFT sums exactly over actions using the corrupted rewards; SFT continues to receive correct demonstrations. Both methods are evaluated using clean semantic labels.

Class flips preserve within-class reward symmetry and hence RFT's zero-style invariance, whereas token flips introduce style fluctuations. Across the tested noise levels, MC RFT exhibits little forgetting, while token flips induce more forgetting than class flips in finite-sample runs (Figure E.4(a,c)).

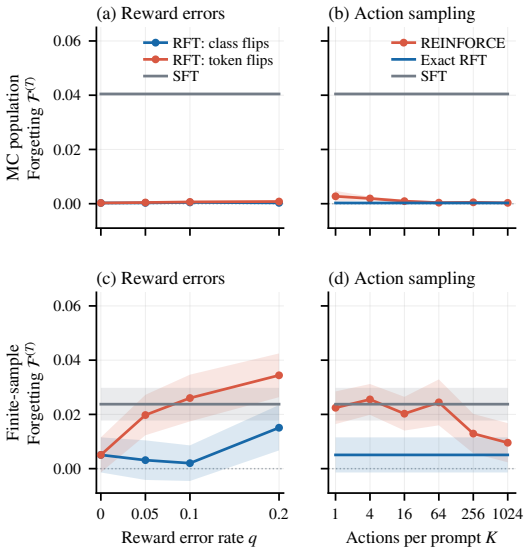



Figure E.4: **Forgetting under two assumption ablations.** Top: MC population updates; bottom: finite-sample updates. Columns vary reward corruption and the number of REINFORCE actions per prompt, respectively. All panels use $\eta \approx 0.121$ and the same vertical scale. Training prompts are paired across methods. Shading gives pointwise $95\%$ Student-$t$ intervals over $10$ MC seeds and mean $\pm 1.96$ standard errors over $50$ finite-sample seeds, conditional on the fixed world and test sets. Negative forgetting indicates backward improvement.

**Finite action sampling.** With clean rewards, we replace the exact sum over $2M = 12$ actions by the no-baseline REINFORCE estimator of Appendix D.1, averaging $K \in \{1, 4, 16, 64, 256, 1024\}$ independent actions per prompt at the current policy. The estimator is unbiased for every $K$ by (D.2). MC updates additionally average over $10{,}000$ prompts, so even $K = 1$ uses $10{,}000$ action samples per update. Prompt and update counts are matched within each protocol, but action-query and computation budgets differ across $K$ and the exact baseline.

At the reference learning rate, REINFORCE forgetting moves closer to the exact RFT baseline as $K$ increases in both settings (Figure E.4(b,d)).