THEORY OF ADAPTATION AND RETENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning (FT) should help a model learn new tasks without losing capabilities acquired during pre-training (PT). We develop a theoretical framework for understanding when these two goals can be achieved together. Using a linear prediction model, we first consider the favorable setting where both PT and FT data are available. Even in this setting, linear updates, including full FT and low-rank adaptation (LoRA), can face an unavoidable tradeoff between adaptation and retention. We then ask what the best possible correction can achieve. We characterize the minimum mean-square error (MMSE) predictor and derive upper and lower bounds on its loss that match up to a universal constant factor for arbitrary population densities. These bounds explain how population overlap and the required correction determine the minimum loss achievable by any method. We next turn to a setting closer to practice where only FT data are available. We study a practical input-dependent FT method motivated by the MMSE predictor, whose default behavior favors preserving pre-trained knowledge. With the correction held fixed, we prove that suitable separation and training conditions in a Gaussian model allow this method to achieve small FT loss while keeping forgetting small throughout training. These results explain how learning when to apply a correction can support adaptation and retention without access to PT examples.

1 INTRODUCTION

Fine-tuning (FT) adapts large language models (LLMs) to new tasks and user needs. It can teach models to follow instructions (Wei et al., 2021; Ouyang et al., 2022), improve mathematical reasoning (Yu et al., 2024), and generate code (Luo et al., 2024). These improvements can be pursued through full FT, which updates all model parameters, or parameter-efficient fine-tuning (PEFT), which trains a smaller set of parameters, as in low-rank adaptation (LoRA) (Hu et al., 2021).

However, improving performance on a new task can degrade capabilities acquired during pre-training (PT) (Luo et al., 2025), a form of forgetting observed with both full FT and LoRA (Biderman et al., 2024; Shuttleworth et al., 2026). Preserving these capabilities is particularly challenging when PT data are unavailable, because the FT loss measures progress on the new task, but does not directly measure what the model forgets. This raises two questions: when is it possible to adapt while preserving pre-trained behavior, and can a learning procedure achieve both goals using only FT data?

We study these questions in a simple model of FT for one linear map. A linear update applies the same correction matrix to every input. Although its output depends on the input, it applies the same linear rule to examples that need a change and those that should keep the pre-trained prediction. An update that helps the new task can therefore also disturb behavior that should be preserved. We show that this restriction can force a tradeoff between adaptation and retention, even with access to both populations. This motivates adding a separate, input-dependent decision about how strongly to apply the correction. We use the minimum mean-square error (MMSE) correction as a benchmark for the best adaptation and retention that such a decision could achieve.

We then turn to the practical question of learning without PT data. Keeping and reusing examples from earlier tasks is a common way to reduce forgetting (Lopez-Paz & Ranzato, 2017; Rolnick et al., 2019; Buzzega et al., 2020), including in language learning (de Masson D'Autume et al., 2019). In LLM FT, mixing in PT data can likewise help preserve earlier capabilities (Ouyang et al., 2022). These approaches, however, require access to earlier examples that may no longer be available. We

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.

instead study an existing input-dependent FT method (Zindari et al., 2026), whose design favors preserving pre-trained behavior while learning where changes are useful. With the correction held fixed, we give conditions under which training on FT data alone achieves adaptation while keeping forgetting small throughout training.

Our results first characterize what is achievable when both populations are known. We show how linear updates limit adaptation and retention, covering full FT and LoRA within our model (Proposition 1). We then give upper and lower bounds on the MMSE for arbitrary population densities, matching up to a universal constant factor (Theorem 1). These bounds show how population overlap and the required correction limit every method, providing a benchmark for input-dependent corrections.

We next analyze training without PT samples for the input-dependent method, keeping the desired correction fixed. We give conditions on initialization and regularization that keep forgetting small throughout training without evaluating the PT loss (Proposition 2). For sufficiently separated Gaussian populations, regularized population gradient flow (GF) achieves small FT loss while keeping forgetting small from small random initializations (Theorem 3). Forgetting stays small from the start, while FT loss becomes small after an initial learning period and stays small.

2 Background and Preliminaries

A pre-trained model learns broad knowledge from a large collection of data. We use p_{pt} to denote the population represented by this data, and p_{ft} for the new population used during FT. The goal is to learn the new task without losing the behavior acquired from p_{pt}. Ideally, we would minimize the total population loss

$$\mathcal{L} = \mathcal{L}_{\mathrm{ft}} + \mathcal{L}_{\mathrm{pt}}, \tag{1}$$

where $\mathcal{L}_{\mathrm{ft}}$ measures performance on the new task and $\mathcal{L}_{\mathrm{pt}}$ measures forgetting on the PT population. This is an ideal benchmark rather than a practical training objective, since PT samples are usually unavailable during FT. Nevertheless, access to both terms gives the most favorable setting: adaptation and forgetting can be measured together, and the correction can be chosen to balance them directly. This lets us ask a simple question: even if both populations were available, could a given FT method adapt without forgetting?

Modern LLMs contain many learned linear maps. An attention block, for example, has query, key, value, and output projections, while a feed-forward block contains additional linear maps. To obtain a tractable theoretical model of FT, we focus on one such matrix and suppress the surrounding transformer operations.[1] We build on the two-population regression model introduced by Zindari et al. (2026) to study the limits of adaptation and retention. Let $\mathbf{W}_0 \in \mathbb{R}^{d_y \times d_x}$ denote its pre-trained weight matrix, let $\mathbf{x} \in \mathbb{R}^{d_x}$ be the representation entering it, and let $\mathbf{M} \in \mathbb{R}^{d_y \times d_x}$ be the correction required by the new task. This model isolates the effect of updating one matrix, but it is not intended to capture every interaction that arises when the full transformer is fine-tuned. Let $Z \in \{0, 1\}$ indicate which population generates an example, with $\mathbb{P}(Z = 1) = \mathbb{P}(Z = 0) = \frac{1}{2}$, and define

$$\mathbf{x} \mid Z = 1 \sim p_{\mathrm{ft}}, \qquad \mathbf{x} \mid Z = 0 \sim p_{\mathrm{pt}}, \qquad \mathbf{y} = (\mathbf{W}_0 + Z\mathbf{M})\,\mathbf{x}. \tag{2}$$

The marginal input distribution is therefore an equal mixture of the PT and FT distributions

$$p_{\mathrm{mix}} = \frac{1}{2} p_{\mathrm{ft}} + \frac{1}{2} p_{\mathrm{pt}}. \tag{3}$$

For PT inputs, we set $\mathbf{y} = \mathbf{W}_0\mathbf{x}$, so the pre-trained predictor is already correct and needs no correction. For FT inputs, we set $\mathbf{y} = (\mathbf{W}_0 + \mathbf{M})\mathbf{x}$, so the required correction is $\mathbf{M}\mathbf{x}$. Sampling from p_{mix} follows the procedure above with the population label Z left unobserved. The equal mixture probabilities are chosen only for simplicity, and the same formulation extends directly to arbitrary mixture probabilities. Let $f(\mathbf{x})$ denote the correction added during FT, so the updated predictor is $\mathbf{W}_0\mathbf{x} + f(\mathbf{x})$. After the common pre-trained output $\mathbf{W}_0\mathbf{x}$ is canceled, the two losses are

$$\mathcal{L}_{\mathrm{ft}}(f) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{ft}}}\left[\|f(\mathbf{x}) - \mathbf{M}\mathbf{x}\|_2^2\right], \qquad \mathcal{L}_{\mathrm{pt}}(f) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{pt}}}\left[\|f(\mathbf{x})\|_2^2\right]. \tag{4}$$

[1] Full-model FT can also update embeddings and normalization parameters. Within our simplified model, "full FT" refers only to an unrestricted update of the selected matrix.

The following notation covers two common ways of updating the weight matrices: full FT and low-rank adaptation. We write the updated output as

$$\mathbf{W}_0\mathbf{x} + f(\mathbf{x}), \quad f(\mathbf{x}) = \mathbf{\Delta}\mathbf{x}, \quad \mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x}. \tag{5}$$

In full FT, every entry of $\mathbf{\Delta}$ is free and is typically learned by gradient descent. Popular parameter-efficient methods such as LoRA (Hu et al., 2021) instead restrict $\mathbf{\Delta}$ to a low-rank parameterization, reducing the number of trainable parameters. We allow $\mathbf{\Delta}$ to be arbitrary, so the analysis covers both unrestricted full FT of this matrix and its low-rank special cases. We write $\mathcal{L}_{\mathrm{ft}}(\mathbf{\Delta})$ and $\mathcal{L}_{\mathrm{pt}}(\mathbf{\Delta})$ for the losses of this linear correction. The next section determines the best achievable loss for a linear correction, then asks what can be achieved by allowing any function of the input.

3 MMSE PREDICTION AND FUNDAMENTAL LIMITS

We first ask how small the total loss can be for the linear corrections introduced above, allowing the matrix to be chosen using both population losses.

Proposition 1 (Best linear loss)**.** *Let* $\mathbf{S}_{\mathrm{ft}} = \mathbb{E}_{p_{\mathrm{ft}}}\left[\mathbf{x}\mathbf{x}^\top\right]$ *and* $\mathbf{S}_{\mathrm{pt}} = \mathbb{E}_{p_{\mathrm{pt}}}\left[\mathbf{x}\mathbf{x}^\top\right]$ *be finite, with* $\mathbf{S}_{\mathrm{ft}} + \mathbf{S}_{\mathrm{pt}} \succ 0$*. Every linear correction* $f(\mathbf{x}) = \mathbf{\Delta}\mathbf{x}$ *satisfies*

$$\mathcal{L}_{\mathrm{ft}}(\mathbf{\Delta}) + \mathcal{L}_{\mathrm{pt}}(\mathbf{\Delta}) \geq \frac{1}{2}\operatorname{Tr}\left(\mathbf{M}\left[\mathbf{S}_{\mathrm{ft}} - \mathbf{S}_{\mathrm{ft}}\left(\mathbf{S}_{\mathrm{ft}} + \mathbf{S}_{\mathrm{pt}}\right)^{-1}\mathbf{S}_{\mathrm{ft}}\right]\mathbf{M}^\top\right). \tag{6}$$

Equality holds at $\mathbf{\Delta}_{\mathrm{lin}} = \mathbf{M}\mathbf{S}_{\mathrm{ft}}\left(\mathbf{S}_{\mathrm{ft}} + \mathbf{S}_{\mathrm{pt}}\right)^{-1}$*. The bound is positive if* $\mathbf{S}_{\mathrm{ft}}, \mathbf{S}_{\mathrm{pt}} \succ 0$ *and* $\mathbf{M} \neq \mathbf{0}$*.*

The proof is in Appendix A.1. This bound holds for every linear correction, including full FT and LoRA within our model, even with access to both populations. The same update used to improve FT predictions is also applied to PT inputs, where the original predictor was already correct. When both second moments are positive definite, perfect adaptation requires $\mathbf{\Delta} = \mathbf{M}$, while perfect retention requires $\mathbf{\Delta} = \mathbf{0}$, forcing a compromise whenever $\mathbf{M} \neq \mathbf{0}$. For equal second moments $\mathbf{S}_{\mathrm{ft}} = \mathbf{S}_{\mathrm{pt}} = \mathbf{S}$, the optimal compromise is $\mathbf{\Delta}_{\mathrm{lin}} = \mathbf{M}/2$, with total loss $\frac{1}{4}\operatorname{Tr}\left(\mathbf{M}\mathbf{S}\mathbf{M}^\top\right) = \frac{1}{4}\mathbb{E}_{p_{\mathrm{ft}}}\left[\|\mathbf{M}\mathbf{x}\|_2^2\right]$. Even the best linear update only halves the total loss compared with leaving the model unchanged. The remaining loss can still be large when the required correction is large. Even perfect optimization cannot improve on this value within the class of linear corrections.

Having found the best linear correction, we now ask what can be achieved by allowing any function of the input. The MMSE correction derived in Zindari et al. (2026) minimizes total loss when both population distributions are known. Recall that Z indicates the population that generated the example, so the desired correction is $Z\mathbf{M}\mathbf{x}$. Under the equal mixture, total loss is exactly the mean squared error for predicting this correction from $\mathbf{x}$. Under squared loss, the optimal prediction at a fixed input $\mathbf{x}$ is the conditional mean of this desired correction. Therefore,

$$f^\star(\mathbf{x}) \coloneqq \mathbb{E}\left[Z\mathbf{M}\mathbf{x} \mid \mathbf{x}\right] = \mathbb{P}\left(Z = 1 \mid \mathbf{x}\right)\mathbf{M}\mathbf{x}. \tag{7}$$

If $p_{\mathrm{ft}}(\mathbf{x})$ and $p_{\mathrm{pt}}(\mathbf{x})$ denote the corresponding probability densities, Bayes' rule and the equal mixture probabilities give

$$\mathbb{P}\left(Z = 1 \mid \mathbf{x}\right) = \frac{1}{2}\frac{p_{\mathrm{ft}}(\mathbf{x})}{p_{\mathrm{mix}}(\mathbf{x})} = \frac{p_{\mathrm{ft}}(\mathbf{x})}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})}. \tag{8}$$

This posterior is a soft population assignment: it approaches one when the input is much more likely under p_{ft}, and it approaches zero when the input is much more likely under p_{pt}. Although this rule is generally unavailable during FT because the population densities are unknown, it shows the best that could be achieved in the ideal setting where both populations are fully known. Its total loss is a fundamental limit: no other correction based on the input alone can achieve a smaller total loss. We next ask how small the total loss of this best correction can be. Substituting (7) into the two population losses gives

$$\mathcal{L}_{\mathrm{MMSE}} = \min_f\left\{\mathcal{L}_{\mathrm{ft}}(f) + \mathcal{L}_{\mathrm{pt}}(f)\right\} = \frac{1}{2}\int \frac{p_{\mathrm{ft}}(\mathbf{x})p_{\mathrm{pt}}(\mathbf{x})}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})}\|\mathbf{M}\mathbf{x}\|_2^2\, d\mathbf{x}. \tag{9}$$

Lemma 2 in the appendix proves this identity. Equation (9) reveals two sources of difficulty. The fraction involving p_{ft} and p_{pt} is positive where the two populations overlap, while $\|\mathbf{M}\mathbf{x}\|_2^2$ measures

how much the prediction must change there. Therefore, overlap matters most on inputs that require a large correction. If the populations have disjoint supports, their product is zero everywhere and $\mathcal{L}_{\mathrm{MMSE}} = 0$: the optimal correction can separate them perfectly. We extend the analysis of Zindari et al. (2026) by deriving lower and upper bounds on the minimum total loss in terms of population overlap and the required correction.

Theorem 1 (Distribution-free MMSE bounds)**.** *Let p_{ft} and p_{pt} be arbitrary densities with finite second moments, define $q(\mathbf{x}) = \|\mathbf{M}\mathbf{x}\|_2^2$, and denote the average correction energy under their equal mixture by $\bar{q} = \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{mix}}}[q(\mathbf{x})]$. Assume $\bar{q} > 0$ and define the normalized q-weighted total variation*

$$\mathrm{TV}_q(p_{\mathrm{ft}}, p_{\mathrm{pt}}) = \frac{1}{2\bar{q}} \int q(\mathbf{x}) |p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})| d\mathbf{x}, \tag{10}$$

where $0 \le \mathrm{TV}_q(p_{\mathrm{ft}}, p_{\mathrm{pt}}) \le 1$. Then

$$\frac{1}{4}\big(1 - \mathrm{TV}_q(p_{\mathrm{ft}}, p_{\mathrm{pt}})\big)\bar{q} \le \mathcal{L}_{\mathrm{MMSE}} \le \frac{1 - \mathrm{TV}_q^2(p_{\mathrm{ft}}, p_{\mathrm{pt}})}{4}\bar{q} \le \frac{1 - \mathrm{TV}_q(p_{\mathrm{ft}}, p_{\mathrm{pt}})}{2}\bar{q}. \tag{11}$$

The proof can be found in Appendix A.2. The quantity $\bar{q}$ is the average size of the correction required across the two populations. The weighted total variation measures their separation after giving more importance to inputs with large $q(\mathbf{x})$. Thus, a difference between the populations does not help if it occurs where $\mathbf{M}\mathbf{x} = 0$, because no correction is needed there. At one extreme, $\mathrm{TV}_q(p_{\mathrm{ft}}, p_{\mathrm{pt}}) = 0$ means that the populations are indistinguishable wherever the correction matters, and the theorem gives the exact value $\mathcal{L}_{\mathrm{MMSE}} = \bar{q}/4$. At the other extreme, $\mathrm{TV}_q(p_{\mathrm{ft}}, p_{\mathrm{pt}}) = 1$ means that the populations do not overlap on inputs where a correction is needed, and $\mathcal{L}_{\mathrm{MMSE}} = 0$. For intermediate values, the first and last expressions in (11) differ by exactly a factor of two, so the theorem characterizes the best achievable total loss up to this universal factor for arbitrary population densities. These bounds quantify when adaptation without forgetting is possible. Next, we show how a sigmoid gate gives the MMSE correction.

4 Likelihood-ratio gates

The posterior probability in (7) has a useful general form. Rather than viewing it only as a fraction of two densities, we can write it as a sigmoid whose score compares how likely the input is under the two populations. This representation requires no parametric assumption on either density. We write $\sigma(s) = \frac{1}{1+e^{-s}}$ for the sigmoid function.

Theorem 2 (MMSE likelihood-ratio gate)**.** *Let p_{ft} and p_{pt} be arbitrary population densities, with each population chosen with probability $\frac{1}{2}$. Wherever both densities are positive, define the log-likelihood ratio*

$$s(\mathbf{x}) = \log \frac{p_{\mathrm{ft}}(\mathbf{x})}{p_{\mathrm{pt}}(\mathbf{x})}. \tag{12}$$

Then the MMSE correction is

$$f^\star(\mathbf{x}) = \frac{p_{\mathrm{ft}}(\mathbf{x})}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})}\mathbf{M}\mathbf{x} = \sigma\left(s(\mathbf{x})\right)\mathbf{M}\mathbf{x}. \tag{13}$$

The proof can be found in Appendix B.1. Theorem 2 shows that the MMSE correction has a sigmoid gate for arbitrary densities, with the log-likelihood ratio as its score. A positive score means that the input is more likely under the FT population, while a negative score favors the PT population. The sigmoid turns this comparison into a soft decision between applying and suppressing the correction. For Gaussian populations, the log-likelihood ratio can be evaluated explicitly. The quadratic form below is therefore a special case of Theorem 2.

Corollary 1 (Gaussian likelihood-ratio gate)**.** *Suppose inputs are drawn with equal probability from $p_{\mathrm{ft}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{ft}}, \boldsymbol{\Sigma}_{\mathrm{ft}})$ and $p_{\mathrm{pt}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{pt}}, \boldsymbol{\Sigma}_{\mathrm{pt}})$, where $\boldsymbol{\Sigma}_{\mathrm{ft}}, \boldsymbol{\Sigma}_{\mathrm{pt}} \succ 0$. Define*

$$\mathbf{A} = \frac{1}{2}\left(\boldsymbol{\Sigma}_{\mathrm{pt}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{ft}}^{-1}\right), \qquad \mathbf{w} = \boldsymbol{\Sigma}_{\mathrm{ft}}^{-1}\boldsymbol{\mu}_{\mathrm{ft}} - \boldsymbol{\Sigma}_{\mathrm{pt}}^{-1}\boldsymbol{\mu}_{\mathrm{pt}},$$
$$c = \frac{1}{2}\left(\boldsymbol{\mu}_{\mathrm{pt}}^\top\boldsymbol{\Sigma}_{\mathrm{pt}}^{-1}\boldsymbol{\mu}_{\mathrm{pt}} - \boldsymbol{\mu}_{\mathrm{ft}}^\top\boldsymbol{\Sigma}_{\mathrm{ft}}^{-1}\boldsymbol{\mu}_{\mathrm{ft}}\right) + \frac{1}{2}\log\frac{\det(\boldsymbol{\Sigma}_{\mathrm{pt}})}{\det(\boldsymbol{\Sigma}_{\mathrm{ft}})}.$$

Then the MMSE correction is

$$f^\star(\mathbf{x}) = \sigma\left(\mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{w}^\top\mathbf{x} + c\right)\mathbf{M}\mathbf{x}. \tag{14}$$

The proof can be found in Appendix B.2. The corollary shows that the exact Gaussian score is generally quadratic in the input. When the two covariance matrices are equal, $\mathbf{A} = 0$ and the score becomes linear. In either case, the posterior acts as a soft switch: it applies nearly all of $\mathbf{Mx}$ to inputs that look like FT examples and suppresses the correction on inputs that look like PT examples. The likelihood-ratio view motivates the general gated family

$$f_g(\mathbf{x}) = g(\mathbf{x})\mathbf{Mx}, \qquad 0 \le g(\mathbf{x}) \le 1. \tag{Gated correction}$$

Gated correction separates what change to make from when to make it: $\mathbf{Mx}$ specifies the correction, while $g(\mathbf{x})$ determines how strongly it is applied to each input. The gate should be close to one on FT inputs that need the correction and close to zero on PT inputs that should keep the original prediction. Theorem 2 identifies the MMSE gate, but it is an ideal benchmark rather than a practical training rule: evaluating it requires the population densities, while PT data are normally unavailable during FT. This motivates the practical question studied next: can we learn when to apply the correction using only FT data, while preserving PT predictions?

5 Retention without pre-training data

One practical instance of the family in (Gated correction) is the *Dynamic Input-Sensitive LoRA* (DISeL) method, introduced in (Zindari et al., 2026). The original paper motivates the design and demonstrates its empirical effectiveness, but does not provide guarantees that training on FT data alone achieves both adaptation and retention. Our goal in this section is to give sufficient conditions for these two goals in a tractable setting. The method replaces the unavailable likelihood-ratio score with a gate learned only from FT data. In its simplest single-neuron form, the correction is

$$f_{\mathrm{DISeL}}(\mathbf{x}) = \sigma\left(\mathbf{w}^\top \mathbf{x} - B\right)\mathbf{Mx}. \tag{DISeL}$$

Here $B > 0$ sets a negative bias $-B$, which favors keeping the gate closed and preserving the original predictions when the initial weight $\mathbf{w}$ is small. Training can then open the gate on FT inputs that need a correction. The score $\mathbf{w}^\top \mathbf{x} - B$ is a tractable linear model of the general log-likelihood ratio. It has the same functional form as the Gaussian MMSE gate score when the covariance matrices are equal, but it is not the exact quadratic score in general. The learned vector $\mathbf{w}$ determines how strongly the correction is applied to each input. Single-neuron models provide a simple setting for understanding gradient-based learning (Yehudai & Shamir, 2020).

To isolate gate learning, we keep the desired update $\mathbf{M}$ and shift B fixed and train only $\mathbf{w}$. This simplified setting studies when the gate opens and how much it opens, but it does not explain how $\mathbf{M}$ itself is learned. We write the squared size of the desired correction as $q(\mathbf{x}) = \|\mathbf{Mx}\|_2^2 = \mathbf{x}^\top \mathbf{Q}\mathbf{x}$, where $\mathbf{Q} = \mathbf{M}^\top \mathbf{M} \succeq 0$. Because PT samples are unavailable, we use only the FT loss. Substituting (DISeL) into (4), gives

$$\mathcal{L}_\lambda(\mathbf{w}) = \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{ft}}}\left[\left(1 - \sigma(\mathbf{w}^\top \mathbf{x} - B)\right)^2 q(\mathbf{x})\right] + \frac{\lambda}{2}\|\mathbf{w}\|_2^2, \qquad \lambda > 0. \tag{17}$$

The L_2 penalty discourages large gate weights, helping keep the gate closed on PT inputs. We study training through gradient flow (GF), which follows the negative gradient of (17):

$$\dot{\mathbf{w}}(t) = -\nabla \mathcal{L}_\lambda(\mathbf{w}(t)), \qquad \mathbf{w}(0) = \mathbf{w}_0. \tag{18}$$

For the empirical counterpart $\widehat{\mathcal{L}}_\lambda$, the same rule becomes $\dot{\mathbf{w}}(t) = -\nabla \widehat{\mathcal{L}}_\lambda(\mathbf{w}(t))$. The same substitution in the PT loss of (4) gives the following quantity, used only for evaluation:

$$\mathcal{L}_{\mathrm{pt}}(\mathbf{w}) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{pt}}}\left[\sigma^2(\mathbf{w}^\top \mathbf{x} - B) q(\mathbf{x})\right].$$

We first consider empirical training. We show that suitable initialization and regularization can keep the PT loss small throughout training, without using any PT samples.

Proposition 2 (Anytime retention). *Suppose $p_{\mathrm{pt}} = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathrm{pt}})$ with $\mathbf{\Sigma}_{\mathrm{pt}} \preceq R\mathbf{I}$ for $R > 0$. For $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{i.i.d.}}{\sim} p_{\mathrm{ft}}$, assume $\widehat{\mathcal{L}}_\lambda(\mathbf{w}(t))$ is non-increasing and $\widehat{E}_0 := \widehat{\mathcal{L}}_\lambda(\mathbf{w}_0)$. If $2R\widehat{E}_0 \le \lambda B$, then*

$$\sup_{t\ge 0} \mathcal{L}_{\mathrm{pt}}(\mathbf{w}(t)) \le \frac{R(1+B)}{2} e^{-\frac{B}{2}} \|\mathbf{M}\|_{\mathrm{F}}^2. \tag{19}$$

A sufficient choice is $R\|\mathbf{w}_0\|_2^2 \le B/2$ and $\lambda \ge \frac{4R}{Bn}\sum_{i=1}^n q(\mathbf{x}_i)$. The bound holds for every realization satisfying the initial-objective condition.

Remark 1. The bound also applies to gradient descent whenever $\widehat{\mathcal{L}}_\lambda$ is non-increasing, with the supremum taken over iterations rather than continuous time.

The proof can be found in Appendix C.1, including a Gaussian initialization scale that gives a high-probability guarantee when $\lambda \geq \frac{4R}{Bn}\sum_{i=1}^{n} q(\mathbf{x}_i)$. For fixed R and $\mathbf{M}$, the retention bound decreases exponentially with B and holds throughout training. Thus, forgetting can be controlled without using PT samples during training. This does not yet guarantee adaptation, since the gate could simply remain closed. The next proposition constructs a gate that achieves both goals under sufficient separation, using FT statistics without PT samples. It assumes a zero-mean Gaussian PT population with a known covariance bound $\mathbf{\Sigma}_{\mathrm{pt}} \preceq R\mathbf{I}$.

Proposition 3 (Gaussian adaptation and retention without PT data)**.** *Assume* $p_{\mathrm{ft}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{ft}}, \mathbf{\Sigma}_{\mathrm{ft}})$ *and* $p_{\mathrm{pt}} = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathrm{pt}})$*, with* $\mathbf{\Sigma}_{\mathrm{ft}}, \mathbf{\Sigma}_{\mathrm{pt}} \succ 0$ *and* $\mathbf{\Sigma}_{\mathrm{pt}} \preceq R\mathbf{I}$ *for a known* $R > 0$*. Let* $\rho = \|\boldsymbol{\mu}_{\mathrm{ft}}\|_2 > 0$ *and* $\kappa = \boldsymbol{\mu}_{\mathrm{ft}}^\top \mathbf{\Sigma}_{\mathrm{ft}} \boldsymbol{\mu}_{\mathrm{ft}} / \rho^2$*. If* $B > 0$ *satisfies*

$$\rho^2 \geq \frac{(3R + \kappa)^2}{4R}\, B, \tag{20}$$

then $\widehat{\mathbf{w}} = \frac{1}{\rho}\sqrt{\frac{B}{R}}\,\boldsymbol{\mu}_{\mathrm{ft}}$*, requiring neither PT samples nor knowledge of* $\mathbf{\Sigma}_{\mathrm{pt}}$*, satisfies*

$$\mathcal{L}_{\mathrm{ft}}(\widehat{\mathbf{w}}) \leq \Theta(\widehat{\mathbf{w}}) e^{-\frac{B}{2}}, \qquad \mathcal{L}_{\mathrm{pt}}(\widehat{\mathbf{w}}) \leq C\,(B + R) e^{-\frac{B}{2}} \tag{21}$$

where $\Theta(\mathbf{w}) := \mathrm{Tr}(\mathbf{M}\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{M}^\top) + \|\mathbf{M}(\boldsymbol{\mu}_{\mathrm{ft}} - \mathbf{\Sigma}_{\mathrm{ft}}\mathbf{w})\|_2^2$ *and* $C = \max\{R\,\|\mathbf{M}\|_{\mathrm{op}}^2, \|\mathbf{M}\|_{\mathrm{F}}^2\}$.

The proof is in Appendix C.2. Proposition 3 constructs a gate with both guarantees from FT statistics and R, but does not show that training finds one. We next study this question for population GF (18), allowing a sufficiently small, possibly random initialization. We track movement along the FT mean and deviation from it:

$$\mathbf{w}(t) = c(t)\mathbf{u} + \mathbf{r}(t), \qquad \mathbf{u} = \frac{\boldsymbol{\mu}_{\mathrm{ft}}}{\|\boldsymbol{\mu}_{\mathrm{ft}}\|_2}, \qquad \mathbf{u}^\top \mathbf{r}(t) = 0.$$

Increasing $c(t)$ raises the average FT score, helping the gate open. We allow $\mathbf{r}(t) \neq \mathbf{0}$, but control its size and the full gate norm to keep forgetting small. The next theorem uses this trajectory control to obtain a regularization range for both goals, without imposing Proposition 2's sufficient condition.

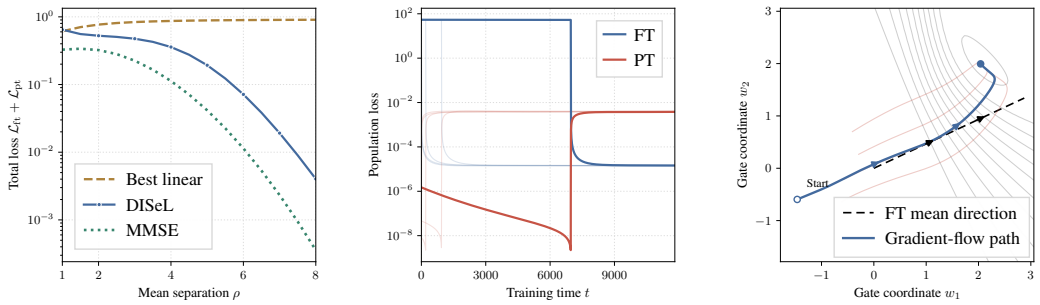
Theorem 3 (Adaptation and retention under GF)**.** *Suppose* $p_{\mathrm{ft}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{ft}}, \mathbf{\Sigma}_{\mathrm{ft}})$ *and* $p_{\mathrm{pt}} = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathrm{pt}})$*, with* $\mathbf{\Sigma}_{\mathrm{ft}} \succeq 0$ *and* $\mathbf{0} \preceq \mathbf{\Sigma}_{\mathrm{pt}} \preceq R\mathbf{I}$ *for* $R > 0$*. Let* $\rho = \|\boldsymbol{\mu}_{\mathrm{ft}}\|_2 > 0$ *and assume* $\mathbb{E}_{p_{\mathrm{ft}}}[q(\mathbf{x})] > 0$*,* $B \geq 10$*,* $\rho^2 \geq 360B\max\{\|\mathbf{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}, R\}$*, and* $\|\mathbf{w}_0\|_2 \leq \frac{1}{20\rho}$*. There exist* $0 < \lambda_{\min} < \lambda_{\max}$ *such that, for every* $\lambda \in (\lambda_{\min}, \lambda_{\max})$*, the population flow* (18) *converges to a stationary point* $\mathbf{w}_\infty$*, and for some finite* $T \geq 0$*,*

$$\begin{aligned} \sup_{t \geq T} \mathcal{L}_{\mathrm{ft}}(\mathbf{w}(t)) &\leq \left(1 + \frac{B}{10}\right) \mathbb{E}_{p_{\mathrm{ft}}}[q(\mathbf{x})]\, e^{-\frac{19B}{20}}, \\ \sup_{t \geq 0} \mathcal{L}_{\mathrm{pt}}(\mathbf{w}(t)) &\leq \frac{R}{2}\left(1 + \frac{B}{10}\right) \|\mathbf{M}\|_{\mathrm{F}}^2\, e^{-\frac{39B}{20}}. \end{aligned} \tag{22}$$

Moreover, $\|\mathbf{w}(t) - \mathbf{w}_\infty\|_2 \leq \frac{6B}{\rho} e^{-\frac{\lambda}{2}(t-T)}$ *for every* $t \geq T$.

The complete statement and proof are in Appendix C.3. The appendix also gives a Gaussian initialization scale for which all guarantees hold with high probability over the random start. Starting from a small initialization, the gate is mostly closed and applies little correction. During the initial learning period, $c(t)$ increases, raising the average FT score $\rho c(t) - B$ and helping the gate open on FT inputs. The trajectory need not follow the FT mean exactly: $\mathbf{r}(t)$ can remain nonzero, while its size and the full gate norm stay controlled. This keeps forgetting small from the start.

The separation condition requires the FT mean to lie far enough from the PT mean relative to the spread of both populations, so the gate can open on FT inputs while remaining mostly closed on PT inputs. By Theorem 1, substantial overlap where a large correction is needed prevents any method from making both losses small. This separation condition is sufficient and is not claimed to be necessary. Our experiments also show successful learning under weaker separation. The time T marks the end of an initial learning period, and the proof bounds how long it can last.



(a) Population separation. (b) Losses during training. (c) Gate trajectories.

Figure 1: **Adaptation and retention under population GF.** We use $d_x = d_y = 2$, $p_{\text{ft}} = \mathcal{N}(\rho\mathbf{u}, \mathbf{\Sigma}_{\text{ft}})$, and $p_{\text{pt}} = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\text{pt}})$, with a fixed unit vector $\mathbf{u}$, distinct non-diagonal covariances, and a dense $\mathbf{M}$. We vary ρ from 1 to 8 with $B = 10$, reusing five independent starts $\mathbf{w}(0) \sim \mathcal{N}(\mathbf{0}, \frac{1}{2}\mathbf{I}_2)$. (a) Mean final total loss with a shaded ± 1 standard-deviation band. (b),(c) Losses and trajectories at $\rho = 8$, with the first start highlighted and the other four faint. The time axis shows continuous GF time. Black arrows mark the FT mean direction, and gray contours show the regularized FT objective.

Before T, the FT error may still be large. After T, both losses remain small without early stopping. The convergence guarantee also shows that the parameters approach a single limit $\mathbf{w}_\infty$, with an exponentially decreasing bound on their distance to it. Training uses no PT samples. The zero-mean Gaussian assumption, known covariance bound R, and separation condition support our guarantee, but are not required to run the method.

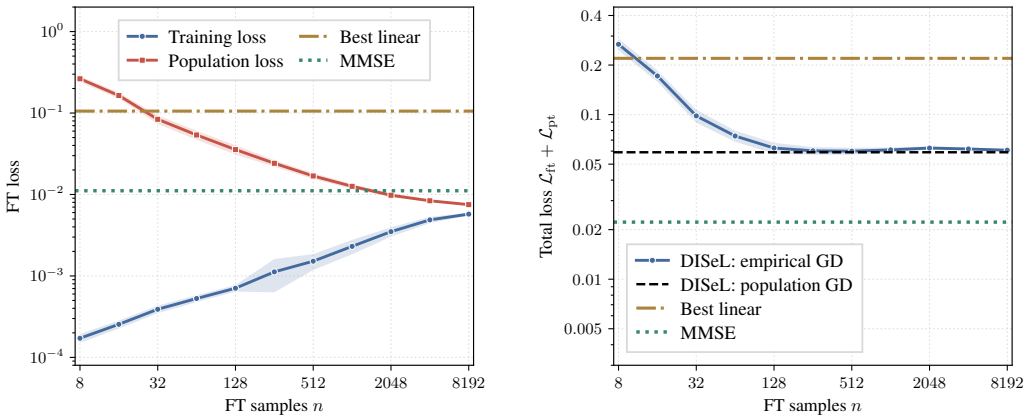
6 Experiments

We examine how population separation and the number of FT examples affect adaptation and retention. Our experiments go beyond Theorem 3's assumptions and show that DISeL can still achieve both goals. Beyond Gaussian settings, we use real image data to compare the smallest achievable prediction error with our theory. In each setting, DISeL keeps $\mathbf{M}$ fixed and trains only $\mathbf{w}$, without PT examples.

Population GF. We study two-dimensional Gaussian populations with different covariances, increasing their mean separation while keeping the covariances and correction matrix fixed. We fix $B = 10$ from the theorem rather than tune it, and numerically integrate population GF for (17), using only FT statistics for the gradient, λ, and stopping. PT losses are used only for evaluation. We compare DISeL with the best linear correction from Proposition 1 and the MMSE correction (7), both chosen with full knowledge of the PT and FT populations. All methods share the target $\mathbf{M}$ and are evaluated by total loss (4).

The left panel of Figure 1 shows substantial loss at weak separation even for MMSE, consistent with the overlap limitation in Theorem 1. As separation grows, DISeL's total loss falls well below the best linear correction, though it remains above MMSE. The middle panel shows that small final losses need not mean fast adaptation: the highlighted run has a long initial delay. While the gate is nearly closed, the FT learning signal is weak, and regularization initially shrinks $\mathbf{w}$, reducing unwanted changes to PT predictions. When the gate opens on FT inputs, FT loss falls sharply, while PT loss rises to a small final value. The right panel shows different starts following curved paths to nearly the same endpoint rather than staying on the FT mean direction. In these runs, initialization mainly changes the path and delay, not the final performance. The runs satisfy the theorem's bias condition, but not its separation or initialization conditions.

Learning from finite samples. We increase the number of FT examples from 8 to 8192, keeping the populations, correction matrix, and training parameters fixed. Empirical GD replaces the FT expectation in (17) by an average over the sampled training set and reuses this set at every step. For comparison, population GD uses the full FT expectation, computed by numerical integration rather than a finite dataset. Both use GD with backtracking and stop when the gradient of their respective



(a) Training and population loss.

(b) Total population loss.

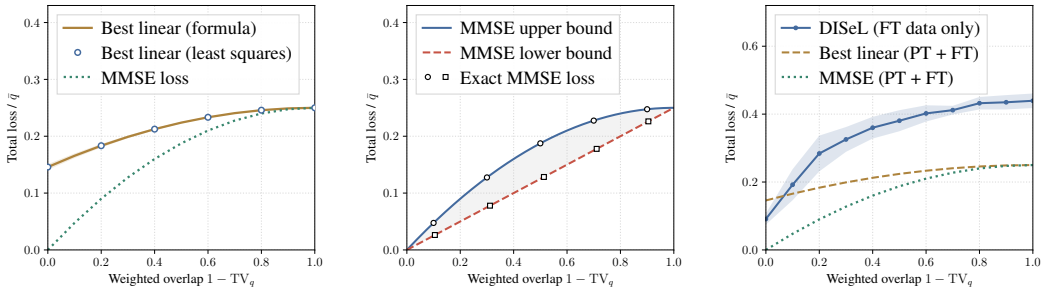
Figure 2: **Learning from finite FT data.** We use $d_x = 128$, $d_y = 32$, $p_{\text{ft}} = \mathcal{N}(3\mathbf{u}, \mathbf{\Sigma}_{\text{ft}})$, and $p_{\text{pt}} = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\text{pt}})$, with a random unit vector $\mathbf{u}$, distinct anisotropic covariances, and Gaussian $\mathbf{M}$ scaled to $\mathbb{E}_{p_{\text{ft}}}[q(\mathbf{x})] = 1$. We fix $B = 4$ and $\lambda = 0.002$. At each n, curves show the mean and shaded ± 1 standard deviation over 20 independent FT datasets and starts $\mathbf{w}_0 \sim \mathcal{N}(\mathbf{0}, \frac{1}{128}\mathbf{I}_{128})$. (a) Training and population FT losses, with population losses evaluated by numerical integration. MMSE and best linear references minimize total loss using both populations, so their FT losses are not lower bounds. (b) Total loss compared with population GD using the full FT expectation and the same settings, averaged over five starts.

regularized FT objective is small. Neither uses PT data or losses for updates, step sizes, or stopping. This comparison shows how limited FT data affect the final result, while keeping the model and training rule the same.

Figure 2 shows why fitting the training examples is not enough. With few examples, the gate fits them almost perfectly but performs poorly on new inputs (left). More data makes the training set harder to fit, yet improves performance on new inputs and narrows this gap. Total loss approaches the result obtained with full FT statistics, well below the linear benchmark (right). This improvement is not monotone: better adaptation can also increase forgetting. More samples therefore help the gate generalize, but do not remove the gap to the MMSE predictor, which has access to both populations and is not restricted to a linear gate score.

Theoretical limits on real-image features. We first train one convolutional neural network (CNN) on original MNIST digits. We then freeze its feature extractor and fit an update $\mathbf{M}$ to its final linear layer using rotated training images and digit labels, changing the weights from $\mathbf{W}_0$ to $\mathbf{W}_0 + \mathbf{M}$. Keeping $\mathbf{M}$ fixed, we extract features from 10,000 other original images and their $45°$-rotated copies, forming two pools of 10,000 vectors each. Our targets are now $\mathbf{W}_0\mathbf{x}$ for PT and $(\mathbf{W}_0 + \mathbf{M})\mathbf{x}$ for FT: we measure prediction error, not classification accuracy. To draw a PT input, we choose the original pool with probability $1 - \varepsilon$ or the rotated pool with probability ε, then sample uniformly within it. FT sampling reverses these probabilities. Increasing ε from 0 to $\frac{1}{2}$ moves from separate to identical populations, with weighted overlap $1 - \mathrm{TV}_q = 2\varepsilon$. We scale $\mathbf{M}$ once so that $\bar{q} = 1$ throughout the sweep. The MMSE formulas use sums over these finite pools instead of integrals.

Figure 3(a) confirms the minimum in Proposition 1: linear corrections can incur positive loss even when MMSE loss is zero. Here the limitation comes from linearity, not unavoidable population overlap. Panel (b) checks whether the bounds in Theorem 1 can actually be reached. The circles and squares represent two different experiments, not two methods. For circles, both populations draw from the same image pools with different probabilities: for example, PT uses 90% original and 10% rotated inputs, while FT reverses these proportions. For squares, we instead divide the inputs into three groups: PT-only, FT-only, and shared. Both populations choose the shared group with the same probability and otherwise use their own group. We vary the mixing proportions for circles and the shared-group probability for squares. Each point is the minimum possible total loss for that setting,



(a) Best linear loss. (b) Sharp MMSE bounds. (c) FT-only learning.

Figure 3: **Testing linear and MMSE loss bounds on MNIST features.** Losses use the complete feature pools and are divided by $\bar{q}$. Curves average five runs with independent CNNs, corrections, FT datasets, and starts, with shaded ± 1 standard deviation. (a) The best linear loss computed from our formula (gold line) or by numerically minimizing the combined PT and FT squared error over linear corrections (blue circles). (b) Lines show our theoretical bounds. Circles and squares show computed MMSE losses for the two population constructions, attaining the upper and lower bounds, respectively. (c) The linear and MMSE benchmarks use both PT and FT distributions to minimize total loss. DISeL trains only $\mathbf{w}$ on 12,000 FT draws with replacement, using 2,000 full-batch GD steps with backtracking on the empirical version of (17). We fix $B = 4$, $\lambda = 0.003$, and $\mathbf{w}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x}/d_x)$. Removing coordinates that are zero throughout both pools reduces each CNN's 64 features to $d_x = 25$–34 across runs. This dimension stays fixed within each run, and $d_y = 10$.

computed from the MMSE formula without training a model. The circles lie on the upper bound and the squares on the lower bound, showing that both bounds can be tight. In panel (c), DISeL beats the linear benchmark at low overlap but loses this advantage as overlap grows. With overlap, improving FT predictions can hurt PT. DISeL sees only FT data, while the benchmarks use both populations.

7 Related Work

Evron et al. (2022) bound forgetting and study task order in sequential linear regression when one predictor can fit all tasks. Kumar et al. (2022) show that FT of a two-layer linear network can distort pre-trained features and hurt out-of-distribution performance. We show that even with access to both populations, the best linear correction may not achieve adaptation and retention together.

Without original training data, Learning without Forgetting (Li & Hoiem, 2017) distills the original model's predictions on new-task inputs, while RecAdam (Chen et al., 2020) regularizes FT toward pre-trained parameters. Marek et al. (2026) show empirically that replay generated by the original model can reduce forgetting without the PT corpus. We give conditions under which gate learning from FT data alone achieves adaptation and retention, without replay or distillation.

Beyond individual training methods, our MMSE analysis asks what any correction can achieve. This connects to Reid & Williamson (2011), who relate the smallest achievable prediction error to how different the two populations are. We apply this viewpoint to adaptation and retention, where confusing the two populations matters in proportion to the required correction size $q(\mathbf{x}) = \|\mathbf{M}\mathbf{x}\|_2^2$. Our weighted total-variation bounds account for this unequal cost across inputs and characterize the smallest achievable total loss up to a universal constant factor.

Recent analyses of LoRA address the FT process from other angles. Nwemadji et al. (2026) show that strong pre-training can slow LoRA FT in single-index models, even when the tasks are aligned. Duranthon et al. (2026) characterize FT error and representation alignment in a solvable attention model. Zhang et al. (2025) use a one-step full gradient for spectral initialization and establish convergence guarantees; Castin et al. (2026) study how balancing the LoRA factors improves conditioning and convergence. Our results identify the adaptation–retention loss that remains even for the best fixed linear update, and quantify what an input-dependent correction can gain.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the text, finding relevant papers and literature review, writing the code for the synthetic experiments, help with some steps in the proof of Theorem 3 more specifically in Lemma 6 and for checking the other proofs for finding flaws. We have not used generative AI tools for coming up with the main ideas and theorems in this paper or the new framework we introduce in this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*, 2024.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

Valérie Castin, Kimia Nadjahi, Pierre Ablin, and Gabriel Peyré. Balanced loRA: Removing parameter invariance to accelerate convergence. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=kHInw3cjCP.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7870–7881, 2020.

Cyprien de Masson D'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32, 2019.

O Duranthon, F Boncoraglio, and L Zdeborová. High-dimensional theory of lora fine-tuning in a solvable attention model. *arXiv preprint arXiv:2606.05899*, 2026.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*, pp. 4028–4079. PMLR, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

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*, 2025.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. In *International Conference on Learning Representations*, volume 2024, pp. 27168–27188, 2024.

Martin Marek, Dongkyu Cho, Shikai Qiu, Rumi Chunara, Pavel Izmailov, and Andrew Gordon Wilson. Forgetting in language models: Capacity, optimization, and self-generated replay. *arXiv preprint arXiv:2605.26097*, 2026.

Gibbs Nwemadji, Bruno Loureiro, and Jean Barbier. When pre-training hurts lora fine-tuning: a dynamical analysis via single-index models. *arXiv preprint arXiv:2602.02855*, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Mark Reid and Robert Williamson. Information, divergence and risk for binary experiments. 2011.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full fine-tuning: An illusion of equivalence. *Advances in Neural Information Processing Systems*, 38: 174627–174662, 2026.

Flemming Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on information theory*, 46(4):1602–1609, 2000.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Gilad Yehudai and Ohad Shamir. Learning a single neuron with gradient methods. In *Conference on Learning Theory*, pp. 3756–3786. PMLR, 2020.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 45040–45061, 2024.

Yuanhe Zhang, Fanghui Liu, and Yudong Chen. Lora-one: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. *arXiv preprint arXiv:2502.01235*, 2025.

Ali Zindari, Xiaowen Jiang, Rotem Mulayoff, and Sebastian U Stich. Learning when to adapt. *arXiv preprint arXiv:2605.19028*, 2026.

Appendix

A Missing proofs from Section 3

Relation to Zindari et al. (2026). The prior paper introduces DISeL together with the two-population regression model of Section 2. Its Appendix A.1 derives the best linear correction for general second moments, but evaluates its minimum loss for equal second moments, where the correction is $\mathbf{M}/2$. It also derives the MMSE correction, called Bayes-optimal there, and its loss (9), and shows that DISeL can represent this correction exactly for Gaussian populations with a shared covariance and sufficient adapter rank. Its experiments include a Gaussian example trained on the total loss, using both populations, and FT of language models. It does not give guarantees that training on FT data alone achieves both adaptation and retention. Here, Proposition 1 gives the minimum linear loss without the equal-moment assumption, and Theorem 1 bounds the MMSE loss for arbitrary densities. Our main addition is an analysis of training with a fixed correction $\mathbf{M}$ and bias: Proposition 2 guarantees retention throughout FT-only training, while Theorem 3 gives sufficient conditions under which population GF achieves both adaptation and retention.

Lemma 1 (Best linear correction)**.** *Let* $\mathbf{S}_{\mathrm{ft}} = \mathbb{E}_{p_{\mathrm{ft}}}\left[\mathbf{x}\mathbf{x}^\top\right]$ *and* $\mathbf{S}_{\mathrm{pt}} = \mathbb{E}_{p_{\mathrm{pt}}}\left[\mathbf{x}\mathbf{x}^\top\right]$ *be finite, with* $\mathbf{S}_{\mathrm{ft}} + \mathbf{S}_{\mathrm{pt}} \succ 0$*. Among all linear corrections* $f(\mathbf{x}) = \mathbf{\Delta}\mathbf{x}$*, the total loss* $\mathcal{L}(\mathbf{\Delta}) = \mathcal{L}_{\mathrm{ft}}(\mathbf{\Delta}) + \mathcal{L}_{\mathrm{pt}}(\mathbf{\Delta})$ *has the unique minimizer*

$$\mathbf{\Delta}_{\mathrm{lin}} = \mathbf{M}\mathbf{S}_{\mathrm{ft}}\left(\mathbf{S}_{\mathrm{ft}} + \mathbf{S}_{\mathrm{pt}}\right)^{-1}.$$

Proof. From the definitions of the two losses,

$$\mathcal{L}(\mathbf{\Delta}) = \frac{1}{2}\mathbb{E}_{p_{\mathrm{ft}}}\left[\left\|\left(\mathbf{\Delta} - \mathbf{M}\right)\mathbf{x}\right\|_2^2\right] + \frac{1}{2}\mathbb{E}_{p_{\mathrm{pt}}}\left[\left\|\mathbf{\Delta}\mathbf{x}\right\|_2^2\right].$$

Using $\nabla_{\mathbf{\Delta}} \frac{1}{2}\left\|\left(\mathbf{\Delta} - \mathbf{M}\right)\mathbf{x}\right\|_2^2 = \left(\mathbf{\Delta} - \mathbf{M}\right)\mathbf{x}\mathbf{x}^\top$, differentiating the loss and taking expectations gives

$$\begin{aligned}\nabla_{\mathbf{\Delta}}\mathcal{L}(\mathbf{\Delta}) &= \mathbb{E}_{p_{\mathrm{ft}}}\left[\left(\mathbf{\Delta} - \mathbf{M}\right)\mathbf{x}\mathbf{x}^\top\right] + \mathbb{E}_{p_{\mathrm{pt}}}\left[\mathbf{\Delta}\mathbf{x}\mathbf{x}^\top\right] \\ &= \left(\mathbf{\Delta} - \mathbf{M}\right)\mathbf{S}_{\mathrm{ft}} + \mathbf{\Delta}\mathbf{S}_{\mathrm{pt}} = \mathbf{\Delta}\left(\mathbf{S}_{\mathrm{ft}} + \mathbf{S}_{\mathrm{pt}}\right) - \mathbf{M}\mathbf{S}_{\mathrm{ft}}.\end{aligned}$$

Setting the gradient to zero and using the invertibility of $\mathbf{S}_{\mathrm{ft}} + \mathbf{S}_{\mathrm{pt}}$ yields

$$\mathbf{\Delta}_{\mathrm{lin}}\left(\mathbf{S}_{\mathrm{ft}} + \mathbf{S}_{\mathrm{pt}}\right) = \mathbf{M}\mathbf{S}_{\mathrm{ft}}, \qquad \mathbf{\Delta}_{\mathrm{lin}} = \mathbf{M}\mathbf{S}_{\mathrm{ft}}\left(\mathbf{S}_{\mathrm{ft}} + \mathbf{S}_{\mathrm{pt}}\right)^{-1}.$$

The rows of $\mathbf{\Delta}$ enter the loss separately, each with Hessian $\mathbf{S}_{\mathrm{ft}} + \mathbf{S}_{\mathrm{pt}} \succ 0$. Thus the loss is strictly convex, and this stationary point is its unique global minimizer. □

A.1 Proof of Proposition 1

Proposition 1 (Best linear loss)**.** *Let* $\mathbf{S}_{\mathrm{ft}} = \mathbb{E}_{p_{\mathrm{ft}}}\left[\mathbf{x}\mathbf{x}^\top\right]$ *and* $\mathbf{S}_{\mathrm{pt}} = \mathbb{E}_{p_{\mathrm{pt}}}\left[\mathbf{x}\mathbf{x}^\top\right]$ *be finite, with* $\mathbf{S}_{\mathrm{ft}} + \mathbf{S}_{\mathrm{pt}} \succ 0$*. Every linear correction* $f(\mathbf{x}) = \mathbf{\Delta}\mathbf{x}$ *satisfies*

$$\mathcal{L}_{\mathrm{ft}}(\mathbf{\Delta}) + \mathcal{L}_{\mathrm{pt}}(\mathbf{\Delta}) \geq \frac{1}{2}\operatorname{Tr}\left(\mathbf{M}\left[\mathbf{S}_{\mathrm{ft}} - \mathbf{S}_{\mathrm{ft}}\left(\mathbf{S}_{\mathrm{ft}} + \mathbf{S}_{\mathrm{pt}}\right)^{-1}\mathbf{S}_{\mathrm{ft}}\right]\mathbf{M}^\top\right). \tag{6}$$

Equality holds at $\mathbf{\Delta}_{\mathrm{lin}} = \mathbf{M}\mathbf{S}_{\mathrm{ft}}\left(\mathbf{S}_{\mathrm{ft}} + \mathbf{S}_{\mathrm{pt}}\right)^{-1}$*. The bound is positive if* $\mathbf{S}_{\mathrm{ft}}, \mathbf{S}_{\mathrm{pt}} \succ 0$ *and* $\mathbf{M} \neq \mathbf{0}$*.*

Proof. By Lemma 1, $\mathbf{\Delta}_{\text{lin}} = \mathbf{M}\mathbf{S}_{\text{ft}} \left(\mathbf{S}_{\text{ft}} + \mathbf{S}_{\text{pt}}\right)^{-1}$ minimizes the sum of the two losses. Hence,

$$\min_{\mathbf{\Delta}} \left\{\mathcal{L}_{\text{ft}}(\mathbf{\Delta}) + \mathcal{L}_{\text{pt}}(\mathbf{\Delta})\right\} = \mathcal{L}_{\text{ft}}(\mathbf{\Delta}_{\text{lin}}) + \mathcal{L}_{\text{pt}}(\mathbf{\Delta}_{\text{lin}}).$$

We compute the right-hand side using $\mathbb{E}\left[\|\mathbf{C}\mathbf{x}\|_2^2\right] = \operatorname{Tr}\left(\mathbf{C}\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]\mathbf{C}^\top\right)$ for a fixed matrix $\mathbf{C}$. Expanding the fine-tuning square gives

$$\begin{aligned}
&\mathcal{L}_{\text{ft}}(\mathbf{\Delta}_{\text{lin}}) + \mathcal{L}_{\text{pt}}(\mathbf{\Delta}_{\text{lin}}) \\
&= \frac{1}{2}\operatorname{Tr}\left(\left(\mathbf{\Delta}_{\text{lin}} - \mathbf{M}\right)\mathbf{S}_{\text{ft}}\left(\mathbf{\Delta}_{\text{lin}} - \mathbf{M}\right)^\top\right) + \frac{1}{2}\operatorname{Tr}\left(\mathbf{\Delta}_{\text{lin}}\mathbf{S}_{\text{pt}}\mathbf{\Delta}_{\text{lin}}^\top\right) \\
&= \frac{1}{2}\operatorname{Tr}\left(\mathbf{\Delta}_{\text{lin}}\left(\mathbf{S}_{\text{ft}} + \mathbf{S}_{\text{pt}}\right)\mathbf{\Delta}_{\text{lin}}^\top\right) - \operatorname{Tr}\left(\mathbf{\Delta}_{\text{lin}}\mathbf{S}_{\text{ft}}\mathbf{M}^\top\right) + \frac{1}{2}\operatorname{Tr}\left(\mathbf{M}\mathbf{S}_{\text{ft}}\mathbf{M}^\top\right).
\end{aligned}$$

Substituting the formula for $\mathbf{\Delta}_{\text{lin}}$ into the first two terms gives

$$\mathbf{\Delta}_{\text{lin}}\left(\mathbf{S}_{\text{ft}} + \mathbf{S}_{\text{pt}}\right)\mathbf{\Delta}_{\text{lin}}^\top = \mathbf{M}\mathbf{S}_{\text{ft}}\left(\mathbf{S}_{\text{ft}} + \mathbf{S}_{\text{pt}}\right)^{-1}\mathbf{S}_{\text{ft}}\mathbf{M}^\top = \mathbf{\Delta}_{\text{lin}}\mathbf{S}_{\text{ft}}\mathbf{M}^\top.$$

Thus the first two trace terms combine, leaving

$$\mathcal{L}_{\text{ft}}(\mathbf{\Delta}_{\text{lin}}) + \mathcal{L}_{\text{pt}}(\mathbf{\Delta}_{\text{lin}}) = \frac{1}{2}\operatorname{Tr}\left(\mathbf{M}\left[\mathbf{S}_{\text{ft}} - \mathbf{S}_{\text{ft}}\left(\mathbf{S}_{\text{ft}} + \mathbf{S}_{\text{pt}}\right)^{-1}\mathbf{S}_{\text{ft}}\right]\mathbf{M}^\top\right).$$

Every linear correction has total loss at least this minimum, proving (6) with equality at $\mathbf{\Delta}_{\text{lin}}$. If both second moments are positive definite, zero FT loss requires $\mathbf{\Delta}_{\text{lin}} = \mathbf{M}$, while zero PT loss requires $\mathbf{\Delta}_{\text{lin}} = \mathbf{0}$. Since both losses are nonnegative, their sum cannot be zero when $\mathbf{M} \neq \mathbf{0}$, proving strict positivity. □

Lemma 2 (MMSE correction and loss)**.** *Let p_{ft} and p_{pt} be densities with finite second moments. Under the losses in (4), the correction $f^\star(\mathbf{x}) = \frac{p_{\text{ft}}(\mathbf{x})}{p_{\text{ft}}(\mathbf{x}) + p_{\text{pt}}(\mathbf{x})}\mathbf{M}\mathbf{x}$ minimizes the total loss over all measurable corrections, and*

$$\mathcal{L}_{\text{MMSE}} = \min_f \left\{\mathcal{L}_{\text{ft}}(f) + \mathcal{L}_{\text{pt}}(f)\right\} = \frac{1}{2}\int \frac{p_{\text{ft}}(\mathbf{x})p_{\text{pt}}(\mathbf{x})}{p_{\text{ft}}(\mathbf{x}) + p_{\text{pt}}(\mathbf{x})}\|\mathbf{M}\mathbf{x}\|_2^2 \, d\mathbf{x}.$$

Fractions are defined as zero where both densities vanish.

Proof. We start from the definitions of the two losses and express their sum as one integral:

$$\mathcal{L}(f) = \frac{1}{2}\int \left[p_{\text{ft}}(\mathbf{x})\|f(\mathbf{x}) - \mathbf{M}\mathbf{x}\|_2^2 + p_{\text{pt}}(\mathbf{x})\|f(\mathbf{x})\|_2^2\right] d\mathbf{x}.$$

To minimize this loss, we minimize the contribution at each fixed input $\mathbf{x}$. Temporarily omit the argument $\mathbf{x}$ from the two densities, and suppose $p_{\text{ft}} + p_{\text{pt}} > 0$. Differentiating this contribution with respect to the output vector $f(\mathbf{x})$, not the input $\mathbf{x}$, and setting the gradient to zero gives

$$p_{\text{ft}}\left(f(\mathbf{x}) - \mathbf{M}\mathbf{x}\right) + p_{\text{pt}}f(\mathbf{x}) = \mathbf{0}.$$

Collecting the terms containing $f(\mathbf{x})$ and dividing by the positive density sum yields

$$\left(p_{\text{ft}} + p_{\text{pt}}\right)f(\mathbf{x}) = p_{\text{ft}}\mathbf{M}\mathbf{x}, \qquad f^\star(\mathbf{x}) = \frac{p_{\text{ft}}}{p_{\text{ft}} + p_{\text{pt}}}\mathbf{M}\mathbf{x}.$$

The Hessian of the contribution is $\left(p_{\text{ft}} + p_{\text{pt}}\right)\mathbf{I}_{d_y} \succ 0$, so this stationary point is its unique minimum. Where both densities vanish, the contribution is zero for any output, and we set $f^\star(\mathbf{x}) = \mathbf{0}$. We next compute the loss at this correction. Wherever the density sum is positive,

$$f^\star(\mathbf{x}) - \mathbf{M}\mathbf{x} = \left(\frac{p_{\text{ft}}}{p_{\text{ft}} + p_{\text{pt}}} - 1\right)\mathbf{M}\mathbf{x} = -\frac{p_{\text{pt}}}{p_{\text{ft}} + p_{\text{pt}}}\mathbf{M}\mathbf{x}.$$

Substituting into the two squared errors and canceling one factor of $p_{\text{ft}} + p_{\text{pt}}$ gives

$$\mathcal{L}(f^\star) = \frac{1}{2}\int \frac{p_{\text{ft}}p_{\text{pt}}^2 + p_{\text{pt}}p_{\text{ft}}^2}{\left(p_{\text{ft}} + p_{\text{pt}}\right)^2}\|\mathbf{M}\mathbf{x}\|_2^2 \, d\mathbf{x} = \frac{1}{2}\int \frac{p_{\text{ft}}(\mathbf{x})p_{\text{pt}}(\mathbf{x})}{p_{\text{ft}}(\mathbf{x}) + p_{\text{pt}}(\mathbf{x})}\|\mathbf{M}\mathbf{x}\|_2^2 \, d\mathbf{x}.$$

Since $f^\star$ minimizes the contribution at every input, integrating gives $\mathcal{L}(f) \geq \mathcal{L}(f^\star)$ for every measurable correction f. The correction $f^\star$ is measurable, and its loss is finite because $0 \leq \frac{p_{\text{ft}}p_{\text{pt}}}{p_{\text{ft}} + p_{\text{pt}}} \leq p_{\text{ft}}$ and the second moments are finite. This proves (9). □

A.2 Proof of Theorem 1

Theorem 1 (Distribution-free MMSE bounds). *Let p_{ft} and p_{pt} be arbitrary densities with finite second moments, define $q(\mathbf{x}) = \|\mathbf{M}\mathbf{x}\|_2^2$, and denote the average correction energy under their equal mixture by $\bar{q} = \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{mix}}}[q(\mathbf{x})]$. Assume $\bar{q} > 0$ and define the normalized q-weighted total variation*

$$\mathrm{TV}_q(p_{\mathrm{ft}}, p_{\mathrm{pt}}) = \frac{1}{2\bar{q}} \int q(\mathbf{x}) |p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})| d\mathbf{x}, \tag{10}$$

where $0 \le \mathrm{TV}_q(p_{\mathrm{ft}}, p_{\mathrm{pt}}) \le 1$. Then

$$\frac{1}{4}\big(1 - \mathrm{TV}_q(p_{\mathrm{ft}}, p_{\mathrm{pt}})\big)\bar{q} \le \mathcal{L}_{\mathrm{MMSE}} \le \frac{1 - \mathrm{TV}_q^2(p_{\mathrm{ft}}, p_{\mathrm{pt}})}{4}\bar{q} \le \frac{1 - \mathrm{TV}_q(p_{\mathrm{ft}}, p_{\mathrm{pt}})}{2}\bar{q}. \tag{11}$$

Proof. We start from the exact MMSE in (9). Using $q(\mathbf{x}) = \|\mathbf{M}\mathbf{x}\|_2^2$, it can be written as

$$\mathcal{L}_{\mathrm{MMSE}} = \frac{1}{2} \int \frac{p_{\mathrm{ft}}(\mathbf{x}) p_{\mathrm{pt}}(\mathbf{x})}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})} q(\mathbf{x})\, d\mathbf{x}. \tag{23}$$

To rewrite this quantity, define the q-weighted triangular discrimination

$$\Delta_q(p_{\mathrm{ft}}, p_{\mathrm{pt}}) = \int q(\mathbf{x}) \frac{(p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x}))^2}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})}\, d\mathbf{x}, \tag{24}$$

where the fraction is defined as zero when both densities are zero. The required algebra follows from the elementary identity

$$(a+b)^2 - (a-b)^2 = 4ab, \qquad a, b \ge 0.$$

Whenever $a + b > 0$, dividing this identity by $a + b$ gives

$$a + b - \frac{(a-b)^2}{a+b} = \frac{4ab}{a+b}.$$

We apply this pointwise with $a = p_{\mathrm{ft}}(\mathbf{x})$ and $b = p_{\mathrm{pt}}(\mathbf{x})$. Multiplying by $q(\mathbf{x})$ and integrating yields

$$\begin{aligned}\int q(\mathbf{x})\,(p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x}))\, d\mathbf{x} - \int q(\mathbf{x}) \frac{(p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x}))^2}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})}\, d\mathbf{x} \\ = 4 \int \frac{p_{\mathrm{ft}}(\mathbf{x}) p_{\mathrm{pt}}(\mathbf{x})}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})} q(\mathbf{x})\, d\mathbf{x}.\end{aligned}$$

The first integral on the left is

$$\int q(\mathbf{x})\,(p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x}))\, d\mathbf{x} = 2 \int q(\mathbf{x}) p_{\mathrm{mix}}(\mathbf{x})\, d\mathbf{x} = 2\bar{q}.$$

The second integral on the left is exactly $\Delta_q(p_{\mathrm{ft}}, p_{\mathrm{pt}})$ by (24). The integral on the right equals $2\mathcal{L}_{\mathrm{MMSE}}$ by (23), so the entire right-hand side equals $8\mathcal{L}_{\mathrm{MMSE}}$. Therefore, dividing the resulting identity by 8 gives

$$\mathcal{L}_{\mathrm{MMSE}} = \frac{1}{8}\left(2\bar{q} - \Delta_q(p_{\mathrm{ft}}, p_{\mathrm{pt}})\right). \tag{25}$$

This is a weighted form of the squared-loss identity in Reid & Williamson (2011, Section 5.5). The exact identity above reduces the proof to comparing $\Delta_q(p_{\mathrm{ft}}, p_{\mathrm{pt}})$ with the weighted total variation. Because Δ_q is subtracted from $2\bar{q}$, an upper bound on Δ_q gives a lower bound on $\mathcal{L}_{\mathrm{MMSE}}$, while a lower bound on Δ_q gives an upper bound on $\mathcal{L}_{\mathrm{MMSE}}$. We derive weighted versions of the comparisons in Topsoe (2000, Eq. (11)). We begin with the upper bound. Starting from the definition of Δ_q we have

$$\begin{aligned}\Delta_q(p_{\mathrm{ft}}, p_{\mathrm{pt}}) &= \int q(\mathbf{x}) \frac{|p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})|^2}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})}\, d\mathbf{x} \\ &= \int q(\mathbf{x}) |p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})| \frac{|p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})|}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})}\, d\mathbf{x}.\end{aligned}$$

Because p_{ft} and p_{pt} are probability densities, they are nonnegative, and hence $|p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})| \leq p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})$. The final fraction above is therefore at most one, which gives

$$\Delta_q\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right) \leq \int q(\mathbf{x}) |p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})| \, d\mathbf{x} = 2\bar{q}\, \mathrm{TV}_q\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right). \tag{26}$$

The same pointwise inequality and $q(\mathbf{x}) \geq 0$ also imply

$$0 \leq \int q(\mathbf{x}) |p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})| \, d\mathbf{x} \leq \int q(\mathbf{x}) \left(p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})\right) \, d\mathbf{x} = 2\bar{q}.$$

Dividing by $2\bar{q} > 0$ proves $0 \leq \mathrm{TV}_q\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right) \leq 1$. For the reverse comparison, observe that, whenever $p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x}) > 0$,

$$q(\mathbf{x}) |p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})| = \sqrt{q(\mathbf{x}) \left(p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})\right)} \sqrt{q(\mathbf{x}) \frac{\left(p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})\right)^2}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})}}.$$

Recall that the integral form of the Cauchy–Schwarz inequality states

$$\left(\int F(\mathbf{x}) G(\mathbf{x}) \, d\mathbf{x}\right)^2 \leq \left(\int F(\mathbf{x})^2 \, d\mathbf{x}\right) \left(\int G(\mathbf{x})^2 \, d\mathbf{x}\right).$$

Here we take

$$F(\mathbf{x}) = \sqrt{q(\mathbf{x}) \left(p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})\right)}, \qquad G(\mathbf{x}) = \sqrt{q(\mathbf{x}) \frac{\left(p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})\right)^2}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})}}.$$

Their product is $F(\mathbf{x})G(\mathbf{x}) = q(\mathbf{x})|p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})|$. Moreover,

$$\int F(\mathbf{x})^2 \, d\mathbf{x} = \int q(\mathbf{x}) \left(p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})\right) \, d\mathbf{x} = 2\bar{q}, \qquad \int G(\mathbf{x})^2 \, d\mathbf{x} = \Delta_q\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right).$$

Substituting these three identities into Cauchy–Schwarz gives

$$\begin{aligned}
&\left(\int q(\mathbf{x}) |p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})| \, d\mathbf{x}\right)^2 \\
&\leq \left(\int q(\mathbf{x}) \left(p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})\right) \, d\mathbf{x}\right) \left(\int q(\mathbf{x}) \frac{\left(p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})\right)^2}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})} \, d\mathbf{x}\right) \\
&= 2\bar{q} \Delta_q\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right).
\end{aligned}$$

Therefore,

$$\frac{\left(\int q(\mathbf{x}) |p_{\mathrm{ft}}(\mathbf{x}) - p_{\mathrm{pt}}(\mathbf{x})| \, d\mathbf{x}\right)^2}{2\bar{q}} \leq \Delta_q\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right). \tag{27}$$

Using the definition of TV_q, inequalities (26) and (27) become

$$2\bar{q} \left(\mathrm{TV}_q\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right)\right)^2 \leq \Delta_q\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right) \leq 2\bar{q}\, \mathrm{TV}_q\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right).$$

Substituting these two inequalities into (25) gives the first two bounds in (11). Since $0 \leq \mathrm{TV}_q\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right) \leq 1$, we also have $1 - \mathrm{TV}_q^2\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right) = \left(1 - \mathrm{TV}_q\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right)\right)\left(1 + \mathrm{TV}_q\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right)\right) \leq 2\left(1 - \mathrm{TV}_q\left(p_{\mathrm{ft}}, p_{\mathrm{pt}}\right)\right)$, which proves the final bound. Finally, if $\bar{q} = 0$, then $q(\mathbf{x}) = 0$ almost surely under both populations, and (23) gives $\mathcal{L}_{\mathrm{MMSE}} = 0$. □

B MISSING PROOFS FROM SECTION 4

B.1 PROOF OF THEOREM 2

Theorem 2 (MMSE likelihood-ratio gate)**.** *Let p_{ft} and p_{pt} be arbitrary population densities, with each population chosen with probability $\frac{1}{2}$. Wherever both densities are positive, define the log-likelihood ratio*

$$s(\mathbf{x}) = \log \frac{p_{\mathrm{ft}}(\mathbf{x})}{p_{\mathrm{pt}}(\mathbf{x})}. \tag{12}$$

Then the MMSE correction is

$$f^\star(\mathbf{x}) = \frac{p_{\mathrm{ft}}(\mathbf{x})}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})}\mathbf{M}\mathbf{x} = \sigma\left(s(\mathbf{x})\right)\mathbf{M}\mathbf{x}. \tag{13}$$

Proof. By (7), the MMSE correction is $f^\star(\mathbf{x}) = \mathbb{P}\left(Z = 1 \mid \mathbf{x}\right)\mathbf{M}\mathbf{x}$. It therefore remains to express the posterior probability in terms of the two densities. Under the equal mixture probabilities, Bayes' rule gives

$$\mathbb{P}\left(Z = 1 \mid \mathbf{x}\right) = \frac{p_{\mathrm{ft}}(\mathbf{x})}{p_{\mathrm{ft}}(\mathbf{x}) + p_{\mathrm{pt}}(\mathbf{x})} = \frac{1}{1 + \frac{p_{\mathrm{pt}}(\mathbf{x})}{p_{\mathrm{ft}}(\mathbf{x})}}.$$

At every point where both densities are positive, the log-likelihood ratio satisfies

$$s(\mathbf{x}) = \log \frac{p_{\mathrm{ft}}(\mathbf{x})}{p_{\mathrm{pt}}(\mathbf{x})}.$$

Then $\frac{p_{\mathrm{pt}}(\mathbf{x})}{p_{\mathrm{ft}}(\mathbf{x})} = e^{-s(\mathbf{x})}$, and hence

$$\mathbb{P}\left(Z = 1 \mid \mathbf{x}\right) = \frac{1}{1 + e^{-s(\mathbf{x})}} = \sigma\left(s(\mathbf{x})\right).$$

Multiplying this identity by $\mathbf{M}\mathbf{x}$ proves (13). □

B.2 PROOF OF COROLLARY 1

Corollary 1 (Gaussian likelihood-ratio gate)**.** *Suppose inputs are drawn with equal probability from $p_{\mathrm{ft}} = \mathcal{N}\left(\boldsymbol{\mu}_{\mathrm{ft}}, \boldsymbol{\Sigma}_{\mathrm{ft}}\right)$ and $p_{\mathrm{pt}} = \mathcal{N}\left(\boldsymbol{\mu}_{\mathrm{pt}}, \boldsymbol{\Sigma}_{\mathrm{pt}}\right)$, where $\boldsymbol{\Sigma}_{\mathrm{ft}}, \boldsymbol{\Sigma}_{\mathrm{pt}} \succ 0$. Define*

$$\mathbf{A} = \frac{1}{2}\left(\boldsymbol{\Sigma}_{\mathrm{pt}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{ft}}^{-1}\right), \qquad \mathbf{w} = \boldsymbol{\Sigma}_{\mathrm{ft}}^{-1}\boldsymbol{\mu}_{\mathrm{ft}} - \boldsymbol{\Sigma}_{\mathrm{pt}}^{-1}\boldsymbol{\mu}_{\mathrm{pt}},$$
$$c = \frac{1}{2}\left(\boldsymbol{\mu}_{\mathrm{pt}}^\top\boldsymbol{\Sigma}_{\mathrm{pt}}^{-1}\boldsymbol{\mu}_{\mathrm{pt}} - \boldsymbol{\mu}_{\mathrm{ft}}^\top\boldsymbol{\Sigma}_{\mathrm{ft}}^{-1}\boldsymbol{\mu}_{\mathrm{ft}}\right) + \frac{1}{2}\log\frac{\det\left(\boldsymbol{\Sigma}_{\mathrm{pt}}\right)}{\det\left(\boldsymbol{\Sigma}_{\mathrm{ft}}\right)}.$$

Then the MMSE correction is

$$f^\star(\mathbf{x}) = \sigma\left(\mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{w}^\top\mathbf{x} + c\right)\mathbf{M}\mathbf{x}. \tag{14}$$

Proof. By Theorem 2, it remains only to evaluate the Gaussian log-likelihood ratio $s(\mathbf{x})$. The density of $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d_x/2}\det\left(\boldsymbol{\Sigma}\right)^{1/2}}\exp\left(-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right)\right).$$

Taking the logarithm of the ratio of the two Gaussian densities cancels the common $(2\pi)^{-d_x/2}$ factor and gives

$$s(\mathbf{x}) = \frac{1}{2}\log\frac{\det\left(\boldsymbol{\Sigma}_{\mathrm{pt}}\right)}{\det\left(\boldsymbol{\Sigma}_{\mathrm{ft}}\right)} - \frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_{\mathrm{ft}}\right)^\top\boldsymbol{\Sigma}_{\mathrm{ft}}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_{\mathrm{ft}}\right) + \frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_{\mathrm{pt}}\right)^\top\boldsymbol{\Sigma}_{\mathrm{pt}}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_{\mathrm{pt}}\right).$$

For any symmetric matrix $\mathbf{\Sigma}^{-1}$,

$$(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x} - 2\boldsymbol{\mu}^\top\mathbf{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu}.$$

Applying this identity to both quadratic forms and grouping terms of the same order in $\mathbf{x}$ yields

$$\begin{aligned} s(\mathbf{x}) &= \frac{1}{2}\mathbf{x}^\top\left(\mathbf{\Sigma}_{\mathrm{pt}}^{-1} - \mathbf{\Sigma}_{\mathrm{ft}}^{-1}\right)\mathbf{x} + \left(\mathbf{\Sigma}_{\mathrm{ft}}^{-1}\boldsymbol{\mu}_{\mathrm{ft}} - \mathbf{\Sigma}_{\mathrm{pt}}^{-1}\boldsymbol{\mu}_{\mathrm{pt}}\right)^\top\mathbf{x} \\ &\quad + \frac{1}{2}\left(\boldsymbol{\mu}_{\mathrm{pt}}^\top\mathbf{\Sigma}_{\mathrm{pt}}^{-1}\boldsymbol{\mu}_{\mathrm{pt}} - \boldsymbol{\mu}_{\mathrm{ft}}^\top\mathbf{\Sigma}_{\mathrm{ft}}^{-1}\boldsymbol{\mu}_{\mathrm{ft}}\right) + \frac{1}{2}\log\frac{\det\left(\mathbf{\Sigma}_{\mathrm{pt}}\right)}{\det\left(\mathbf{\Sigma}_{\mathrm{ft}}\right)} \\ &= \mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{w}^\top\mathbf{x} + c. \end{aligned}$$

Substituting this expression into $\mathbb{P}(Z=1 \mid \mathbf{x}) = \sigma(s(\mathbf{x}))$, and then into (7), gives

$$f^\star(\mathbf{x}) = \sigma\left(\mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{w}^\top\mathbf{x} + c\right)\mathbf{M}\mathbf{x},$$

which proves (14). □

C Missing proofs from Section 5

Gaussian preliminaries. We first collect the Gaussian facts used throughout the proofs in this section. Stating them once keeps the later arguments focused on the gate and GF dynamics. Throughout these lemmas, $g \sim \mathcal{N}(0,1)$, and

$$\phi(t) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)$$

denotes the standard Gaussian density.

Lemma 3 (Linear images of a Gaussian vector)**.** *Let $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$. For every deterministic matrix $\mathbf{C}$ and vector $\mathbf{d}$ of compatible dimensions,*

$$\mathbf{C}\mathbf{x} + \mathbf{d} \sim \mathcal{N}\left(\mathbf{C}\boldsymbol{\mu} + \mathbf{d}, \mathbf{C}\mathbf{\Sigma}\mathbf{C}^\top\right).$$

In particular, let $\mathbf{u}$ be a unit vector, $\mathbf{P}_\perp = \mathbf{I} - \mathbf{u}\mathbf{u}^\top$, $\tau = \mathbf{u}^\top\mathbf{x}$, and $\mathbf{z} = \mathbf{P}_\perp\mathbf{x}$. Then $(\tau, \mathbf{z})$ is jointly Gaussian, and

$$\mathbb{E}[\tau] = \mathbf{u}^\top\boldsymbol{\mu}, \qquad \mathbb{E}[\mathbf{z}] = \mathbf{P}_\perp\boldsymbol{\mu},$$

$$\mathrm{Var}(\tau) = \mathbf{u}^\top\mathbf{\Sigma}\mathbf{u}, \qquad \mathrm{Cov}(\mathbf{z}) = \mathbf{P}_\perp\mathbf{\Sigma}\mathbf{P}_\perp, \qquad \mathrm{Cov}(\mathbf{z}, \tau) = \mathbf{P}_\perp\mathbf{\Sigma}\mathbf{u}.$$

Consequently, if $\mathbf{P}_\perp\mathbf{\Sigma}\mathbf{u} = \mathbf{0}$, then τ and $\mathbf{z}$ are independent.

Proof. We first verify that the affine transformation is Gaussian. Let $\mathbf{a}$ be any deterministic vector of the same dimension as $\mathbf{C}\mathbf{x} + \mathbf{d}$. Then

$$\mathbf{a}^\top(\mathbf{C}\mathbf{x} + \mathbf{d}) = \left(\mathbf{C}^\top\mathbf{a}\right)^\top\mathbf{x} + \mathbf{a}^\top\mathbf{d}.$$

Since $\mathbf{x}$ is a Gaussian vector, $\left(\mathbf{C}^\top\mathbf{a}\right)^\top\mathbf{x}$ is a scalar Gaussian random variable; adding the deterministic scalar $\mathbf{a}^\top\mathbf{d}$ preserves Gaussianity. Thus every linear projection of $\mathbf{C}\mathbf{x} + \mathbf{d}$ is Gaussian, which proves that $\mathbf{C}\mathbf{x} + \mathbf{d}$ is a Gaussian vector. Its mean follows from linearity of expectation:

$$\mathbb{E}[\mathbf{C}\mathbf{x} + \mathbf{d}] = \mathbf{C}\mathbb{E}[\mathbf{x}] + \mathbf{d} = \mathbf{C}\boldsymbol{\mu} + \mathbf{d}.$$

Its covariance is

$$\begin{aligned} \mathrm{Cov}(\mathbf{C}\mathbf{x} + \mathbf{d}) &= \mathbb{E}\left[\mathbf{C}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{C}^\top\right] \\ &= \mathbf{C}\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right]\mathbf{C}^\top = \mathbf{C}\mathbf{\Sigma}\mathbf{C}^\top. \end{aligned}$$

For the particular decomposition, define the stacked random vector

$$\begin{pmatrix}\tau\\ \mathbf{z}\end{pmatrix} = \begin{pmatrix}\mathbf{u}^\top\\ \mathbf{P}_\perp\end{pmatrix}\mathbf{x}.$$

This is a linear image of $\mathbf{x}$, so the result above shows that $(\tau, \mathbf{z})$ is jointly Gaussian. Its mean is

$$\mathbb{E}\left[\begin{pmatrix}\tau\\ \mathbf{z}\end{pmatrix}\right] = \begin{pmatrix}\mathbf{u}^\top\boldsymbol{\mu}\\ \mathbf{P}_\perp\boldsymbol{\mu}\end{pmatrix},$$

and its covariance is

$$\mathrm{Cov}\left(\begin{pmatrix}\tau\\ \mathbf{z}\end{pmatrix}\right) = \begin{pmatrix}\mathbf{u}^\top\\ \mathbf{P}_\perp\end{pmatrix}\boldsymbol{\Sigma}\begin{pmatrix}\mathbf{u} & \mathbf{P}_\perp\end{pmatrix} = \begin{pmatrix}\mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u} & \mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{P}_\perp\\ \mathbf{P}_\perp\boldsymbol{\Sigma}\mathbf{u} & \mathbf{P}_\perp\boldsymbol{\Sigma}\mathbf{P}_\perp\end{pmatrix},$$

where we used $\mathbf{P}_\perp^\top = \mathbf{P}_\perp$. The three covariance blocks can be verified separately from $\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right] = \boldsymbol{\Sigma}$:

$$\begin{aligned}\mathrm{Var}(\tau) &= \mathbb{E}\left[\left(\mathbf{u}^\top(\mathbf{x}-\boldsymbol{\mu})\right)^2\right] = \mathbf{u}^\top\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right]\mathbf{u} = \mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u},\\ \mathrm{Cov}(\mathbf{z}) &= \mathbb{E}\left[\mathbf{P}_\perp(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{P}_\perp^\top\right] = \mathbf{P}_\perp\boldsymbol{\Sigma}\mathbf{P}_\perp,\\ \mathrm{Cov}(\mathbf{z},\tau) &= \mathbb{E}\left[\mathbf{P}_\perp(\mathbf{x}-\boldsymbol{\mu})\left(\mathbf{u}^\top(\mathbf{x}-\boldsymbol{\mu})\right)\right] = \mathbf{P}_\perp\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right]\mathbf{u} = \mathbf{P}_\perp\boldsymbol{\Sigma}\mathbf{u}.\end{aligned}$$

If $\mathbf{P}_\perp\boldsymbol{\Sigma}\mathbf{u} = \mathbf{0}$, then the lower-left cross-covariance block is zero. Since $\boldsymbol{\Sigma}$ and $\mathbf{P}_\perp$ are symmetric, the upper-right block is its transpose and is also zero. The covariance of the jointly Gaussian pair $(\tau, \mathbf{z})$ is therefore block diagonal. The Gaussian characteristic function then factors into the product of the characteristic functions of τ and $\mathbf{z}$, which proves their independence. □

Lemma 4 (Exponentially tilted Gaussian quadratic moment)**.** *Let* $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{a} \in \mathbb{R}^{d_x}$, *and* $\mathbf{Q} \in \mathbb{R}^{d_x \times d_x}$. *Then*

$$\mathbb{E}\left[e^{\mathbf{a}^\top\mathbf{x}}\mathbf{x}^\top\mathbf{Q}\mathbf{x}\right] = \exp\left(\mathbf{a}^\top\boldsymbol{\mu} + \frac{1}{2}\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}\right)\left[(\boldsymbol{\mu}+\boldsymbol{\Sigma}\mathbf{a})^\top\mathbf{Q}(\boldsymbol{\mu}+\boldsymbol{\Sigma}\mathbf{a}) + \mathrm{Tr}(\mathbf{Q}\boldsymbol{\Sigma})\right]. \tag{28}$$

Proof. Let $\mathbf{C} = \boldsymbol{\Sigma}^{1/2}$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. By Lemma 3, $\mathbf{x}$ has the same distribution as $\boldsymbol{\mu} + \mathbf{C}\mathbf{z}$, including when $\boldsymbol{\Sigma}$ is singular. Set $\mathbf{b} = \mathbf{C}^\top\mathbf{a}$. In the Gaussian density of $\mathbf{z}$, completing the square gives

$$-\frac{1}{2}\|\mathbf{z}\|_2^2 + \mathbf{b}^\top\mathbf{z} = -\frac{1}{2}\|\mathbf{z}-\mathbf{b}\|_2^2 + \frac{1}{2}\|\mathbf{b}\|_2^2.$$

Moreover, $\|\mathbf{b}\|_2^2 = \mathbf{a}^\top\mathbf{C}\mathbf{C}^\top\mathbf{a} = \mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}$. Therefore multiplying the density by $e^{\mathbf{a}^\top\mathbf{x}} = e^{\mathbf{a}^\top\boldsymbol{\mu}}e^{\mathbf{b}^\top\mathbf{z}}$ changes the mean of $\mathbf{z}$ from $\mathbf{0}$ to $\mathbf{b}$ and contributes the factor

$$\exp\left(\mathbf{a}^\top\boldsymbol{\mu} + \frac{1}{2}\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}\right).$$

Consequently, if $\widetilde{\mathbf{x}} \sim \mathcal{N}(\boldsymbol{\mu}+\boldsymbol{\Sigma}\mathbf{a}, \boldsymbol{\Sigma})$, then

$$\mathbb{E}\left[e^{\mathbf{a}^\top\mathbf{x}}\mathbf{x}^\top\mathbf{Q}\mathbf{x}\right] = \exp\left(\mathbf{a}^\top\boldsymbol{\mu} + \frac{1}{2}\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}\right)\mathbb{E}\left[\widetilde{\mathbf{x}}^\top\mathbf{Q}\widetilde{\mathbf{x}}\right].$$

Writing $\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{a}$, the quadratic expectation is

$$\mathbb{E}\left[\widetilde{\mathbf{x}}^\top\mathbf{Q}\widetilde{\mathbf{x}}\right] = \widetilde{\boldsymbol{\mu}}^\top\mathbf{Q}\widetilde{\boldsymbol{\mu}} + \mathrm{Tr}(\mathbf{Q}\boldsymbol{\Sigma}),$$

because $\mathbb{E}[\widetilde{\mathbf{x}} - \widetilde{\boldsymbol{\mu}}] = \mathbf{0}$ and $\mathbb{E}\left[(\widetilde{\mathbf{x}}-\widetilde{\boldsymbol{\mu}})(\widetilde{\mathbf{x}}-\widetilde{\boldsymbol{\mu}})^\top\right] = \boldsymbol{\Sigma}$. Substituting this expression proves (28). □

Lemma 5 (Norm of a standard Gaussian vector)**.** *Let* $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x})$*. For every* $a \geq 0$*,*

$$\mathbb{P}\left(\|\mathbf{g}\|_2 \geq \sqrt{d_x} + a\right) \leq e^{-\frac{a^2}{2}}.$$

Proof. For $a = 0$, the right-hand side is one. Suppose $a > 0$. Introduce an exponential coefficient $t \in (0, 1/2)$, which we will choose by minimizing the resulting bound. Squaring both nonnegative sides and exponentiating preserve the event; Markov's inequality then gives

$$\begin{aligned}\mathbb{P}\left(\|\mathbf{g}\|_2 \geq \sqrt{d_x} + a\right) &= \mathbb{P}\left(\|\mathbf{g}\|_2^2 \geq (\sqrt{d_x} + a)^2\right) \\ &= \mathbb{P}\left(e^{t\|\mathbf{g}\|_2^2} \geq e^{t(\sqrt{d_x}+a)^2}\right) \\ &\leq e^{-t(\sqrt{d_x}+a)^2}\mathbb{E}\left[e^{t\|\mathbf{g}\|_2^2}\right].\end{aligned}$$

To evaluate the remaining expectation, use $\|\mathbf{g}\|_2^2 = \sum_{i=1}^{d_x} g_i^2$ and independence:

$$\mathbb{E}\left[e^{t\|\mathbf{g}\|_2^2}\right] = \prod_{i=1}^{d_x} \mathbb{E}\left[e^{tg_i^2}\right].$$

For each coordinate, integrating against its Gaussian density, with integration variable z, gives

$$\begin{aligned}\mathbb{E}\left[e^{tg_i^2}\right] &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(tz^2 - \frac{z^2}{2}\right) dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{(1-2t)z^2}{2}\right) dz.\end{aligned}$$

Introduce the integration variable $v = \sqrt{1-2t}\, z$, so $z = v/\sqrt{1-2t}$ and $dz = dv/\sqrt{1-2t}$. Since $1 - 2t > 0$, both variables range over $\mathbb{R}$. Substituting gives

$$\begin{aligned}\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{(1-2t)z^2}{2}}\, dz &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{v^2}{2}} \frac{dv}{\sqrt{1-2t}} \\ &= \frac{1}{\sqrt{1-2t}} \underbrace{\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{v^2}{2}}\, dv}_{=1} = (1-2t)^{-\frac{1}{2}}.\end{aligned}$$

Multiplying the d_x factors and substituting into the probability bound gives

$$\mathbb{P}\left(\|\mathbf{g}\|_2 \geq \sqrt{d_x} + a\right) \leq e^{-t(\sqrt{d_x}+a)^2}(1-2t)^{-\frac{d_x}{2}}.$$

This holds for every $t \in (0, 1/2)$, so we choose the value that minimizes the right-hand side. Applying the product and chain rules gives

$$\begin{aligned}&\frac{d}{dt}\left[e^{-t(\sqrt{d_x}+a)^2}(1-2t)^{-\frac{d_x}{2}}\right] \\ &\quad= -(\sqrt{d_x}+a)^2 e^{-t(\sqrt{d_x}+a)^2}(1-2t)^{-\frac{d_x}{2}} \\ &\qquad + e^{-t(\sqrt{d_x}+a)^2}\left(-\frac{d_x}{2}\right)(1-2t)^{-\frac{d_x}{2}-1}(-2) \\ &\quad= e^{-t(\sqrt{d_x}+a)^2}(1-2t)^{-\frac{d_x}{2}}\left[-(\sqrt{d_x}+a)^2 + \frac{d_x}{1-2t}\right].\end{aligned}$$

The factor outside the brackets is positive, so the derivative is zero exactly when

$$\frac{d_x}{1-2t} = (\sqrt{d_x}+a)^2 \quad\Longleftrightarrow\quad t = \frac{1}{2}\left(1 - \frac{d_x}{(\sqrt{d_x}+a)^2}\right).$$

Since $a > 0$, this value lies in $(0, 1/2)$. The bracket increases with t, so the derivative is negative before this value and positive afterward; hence it minimizes the bound. For this choice, the exponent and the power simplify separately:

$$-t(\sqrt{d_x}+a)^2 = -\frac{(\sqrt{d_x}+a)^2 - d_x}{2} = -a\sqrt{d_x} - \frac{a^2}{2},$$
$$(1-2t)^{-\frac{d_x}{2}} = \left(\frac{d_x}{(\sqrt{d_x}+a)^2}\right)^{-\frac{d_x}{2}} = \left(\frac{\sqrt{d_x}+a}{\sqrt{d_x}}\right)^{d_x} = \left(1+\frac{a}{\sqrt{d_x}}\right)^{d_x}.$$

Substituting into the probability bound gives

$$\mathbb{P}\left(\|\mathbf{g}\|_2 \geq \sqrt{d_x}+a\right) \leq e^{-a\sqrt{d_x}-\frac{a^2}{2}}\left(1+\frac{a}{\sqrt{d_x}}\right)^{d_x}.$$

To bound the remaining power, use $1+v \leq e^v$ for $v \geq 0$, which follows from $e^v = 1 + \int_0^v e^z\,dz \geq 1+v$. With $v = a/\sqrt{d_x}$, we obtain

$$\begin{aligned}\mathbb{P}\left(\|\mathbf{g}\|_2 \geq \sqrt{d_x}+a\right) &\leq e^{-a\sqrt{d_x}-\frac{a^2}{2}}\left(e^{\frac{a}{\sqrt{d_x}}}\right)^{d_x} \\ &= e^{-a\sqrt{d_x}-\frac{a^2}{2}}e^{a\sqrt{d_x}} = e^{-\frac{a^2}{2}}.\end{aligned}$$

□

C.1 Proof of Proposition 2

Proposition 2 (Anytime retention). *Suppose $p_{\mathrm{pt}} = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathrm{pt}})$ with $\mathbf{\Sigma}_{\mathrm{pt}} \preceq R\mathbf{I}$ for $R > 0$. For $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{i.i.d.}}{\sim} p_{\mathrm{ft}}$, assume $\widehat{\mathcal{L}}_\lambda(\mathbf{w}(t))$ is non-increasing and $\widehat{E}_0 := \widehat{\mathcal{L}}_\lambda(\mathbf{w}_0)$. If $2R\widehat{E}_0 \leq \lambda B$, then*

$$\sup_{t\geq 0}\mathcal{L}_{\mathrm{pt}}(\mathbf{w}(t)) \leq \frac{R(1+B)}{2}e^{-\frac{B}{2}}\|\mathbf{M}\|_{\mathrm{F}}^2. \tag{19}$$

A sufficient choice is $R\|\mathbf{w}_0\|_2^2 \leq B/2$ and $\lambda \geq \frac{4R}{Bn}\sum_{i=1}^n q(\mathbf{x}_i)$. The bound holds for every realization satisfying the initial-objective condition.

Proof. We first bound the pre-training loss for a fixed gate vector, then use the decrease of the empirical objective to control this bound throughout training. For a fixed $\mathbf{w}$, the pre-training loss is

$$\mathcal{L}_{\mathrm{pt}}(\mathbf{w}) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{pt}}}\left[\sigma^2\left(\mathbf{w}^\top\mathbf{x} - B\right)\mathbf{x}^\top\mathbf{Q}\mathbf{x}\right], \qquad \mathbf{Q} = \mathbf{M}^\top\mathbf{M}.$$

Since $0 < \sigma(s) < 1$ and $\sigma(s) = \frac{e^s}{1+e^s} \leq e^s$, we have $\sigma^2(s) \leq \sigma(s) \leq e^s$. Therefore

$$\mathcal{L}_{\mathrm{pt}}(\mathbf{w}) \leq \frac{e^{-B}}{2}\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{pt}}}\left[e^{\mathbf{w}^\top\mathbf{x}}\mathbf{x}^\top\mathbf{Q}\mathbf{x}\right]. \tag{29}$$

Apply Lemma 4 with $\boldsymbol{\mu} = \mathbf{0}$, $\mathbf{\Sigma} = \mathbf{\Sigma}_{\mathrm{pt}}$, and $\mathbf{a} = \mathbf{w}$. It gives

$$\mathbb{E}_{p_{\mathrm{pt}}}\left[e^{\mathbf{w}^\top\mathbf{x}}\mathbf{x}^\top\mathbf{Q}\mathbf{x}\right] = \exp\left(\frac{1}{2}\mathbf{w}^\top\mathbf{\Sigma}_{\mathrm{pt}}\mathbf{w}\right)\left[\mathrm{Tr}\left(\mathbf{Q}\mathbf{\Sigma}_{\mathrm{pt}}\right) + \left(\mathbf{\Sigma}_{\mathrm{pt}}\mathbf{w}\right)^\top\mathbf{Q}\left(\mathbf{\Sigma}_{\mathrm{pt}}\mathbf{w}\right)\right]. \tag{30}$$

The covariance bound $\mathbf{\Sigma}_{\mathrm{pt}} \preceq R\mathbf{I}$ implies $\mathbf{w}^\top\mathbf{\Sigma}_{\mathrm{pt}}\mathbf{w} \leq R\|\mathbf{w}\|_2^2$ and

$$\mathrm{Tr}\left(\mathbf{Q}\mathbf{\Sigma}_{\mathrm{pt}}\right) \leq R\,\mathrm{Tr}\left(\mathbf{Q}\right) = R\|\mathbf{M}\|_{\mathrm{F}}^2.$$

Moreover, $\|\mathbf{\Sigma}_{\mathrm{pt}}\|_{\mathrm{op}} \leq R$, $\|\mathbf{Q}\|_{\mathrm{op}} = \|\mathbf{M}\|_{\mathrm{op}}^2$, and $\|\mathbf{M}\|_{\mathrm{op}} \leq \|\mathbf{M}\|_{\mathrm{F}}$, so

$$\left(\mathbf{\Sigma}_{\mathrm{pt}}\mathbf{w}\right)^\top\mathbf{Q}\left(\mathbf{\Sigma}_{\mathrm{pt}}\mathbf{w}\right) \leq \|\mathbf{Q}\|_{\mathrm{op}}\|\mathbf{\Sigma}_{\mathrm{pt}}\mathbf{w}\|_2^2 \leq R^2\|\mathbf{M}\|_{\mathrm{F}}^2\|\mathbf{w}\|_2^2.$$

Substituting these bounds into (30) and then into (29) yields

$$\mathcal{L}_{\mathrm{pt}}(\mathbf{w}) \le \frac{R\,\|\mathbf{M}\|_{\mathrm{F}}^2}{2}\left(1 + R\,\|\mathbf{w}\|_2^2\right)\exp\left(-B + \frac{R}{2}\,\|\mathbf{w}\|_2^2\right). \tag{31}$$

To keep this bound small throughout training, it suffices to show that $R\,\|\mathbf{w}(t)\|_2^2 \le B$. We now obtain this from the initial objective, without requiring $\mathbf{w}_0 = \mathbf{0}$. Recall that

$$\widehat{E}_0 = \widehat{\mathcal{L}}_\lambda(\mathbf{w}_0) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \sigma(\mathbf{w}_0^\top \mathbf{x}_i - B)\right)^2 q(\mathbf{x}_i) + \frac{\lambda}{2}\,\|\mathbf{w}_0\|_2^2.$$

Every term in the empirical objective is nonnegative, and the objective is non-increasing by assumption. Hence, for every $t \ge 0$,

$$\frac{\lambda}{2}\,\|\mathbf{w}(t)\|_2^2 \le \widehat{\mathcal{L}}_\lambda(\mathbf{w}(t)) \le \widehat{\mathcal{L}}_\lambda(\mathbf{w}(0)) = \widehat{E}_0.$$

Consequently, $\|\mathbf{w}(t)\|_2^2 \le \frac{2\widehat{E}_0}{\lambda}$. The condition $2R\widehat{E}_0 \le \lambda B$ therefore gives

$$R\,\|\mathbf{w}(t)\|_2^2 \le \frac{2R\widehat{E}_0}{\lambda} \le B.$$

It follows that $1 + R\,\|\mathbf{w}(t)\|_2^2 \le 1 + B$ and $-B + \frac{R}{2}\,\|\mathbf{w}(t)\|_2^2 \le -\frac{B}{2}$. Applying these inequalities to (31) gives

$$\mathcal{L}_{\mathrm{pt}}(\mathbf{w}(t)) \le \frac{R(1+B)}{2} e^{-\frac{B}{2}}\,\|\mathbf{M}\|_{\mathrm{F}}^2 \qquad \text{for every } t \ge 0.$$

Taking the supremum over t proves (19) for every initialization satisfying the condition. To see how to ensure this condition, expand the initial objective and use $0 \le (1 - \sigma(s))^2 \le 1$:

$$\frac{2R\widehat{E}_0}{\lambda} = R\,\|\mathbf{w}_0\|_2^2 + \frac{2R}{\lambda n}\sum_{i=1}^{n}\left(1 - \sigma(\mathbf{w}_0^\top \mathbf{x}_i - B)\right)^2 q(\mathbf{x}_i) \le R\,\|\mathbf{w}_0\|_2^2 + \frac{2R}{\lambda n}\sum_{i=1}^{n} q(\mathbf{x}_i).$$

It is enough to make each term on the right at most $B/2$. This gives the sufficient choices stated in the proposition:

$$R\,\|\mathbf{w}_0\|_2^2 \le \frac{B}{2}, \qquad \lambda \ge \frac{4R}{Bn}\sum_{i=1}^{n} q(\mathbf{x}_i), \qquad \frac{2R\widehat{E}_0}{\lambda} \le \frac{B}{2} + \frac{B}{2} = B.$$

For Gaussian initialization, we ensure $\|\mathbf{w}_0\|_2 \le \sqrt{B/(2R)}$ with probability at least $1 - \delta$. Fix an allowed failure probability $\delta \in (0, 1)$ and draw independently of the training data

$$\mathbf{w}_0 = \frac{\omega}{\sqrt{d_x}}\mathbf{g} \sim \mathcal{N}\left(\mathbf{0}, \frac{\omega^2}{d_x}\mathbf{I}_{d_x}\right), \qquad \mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x}), \qquad \omega > 0.$$

Apply Lemma 5 with $a = \sqrt{2\log(1/\delta)}$, so $e^{-a^2/2} = \delta$. With probability at least $1 - \delta$,

$$\|\mathbf{w}_0\|_2 \le \frac{\omega}{\sqrt{d_x}}\left(\sqrt{d_x} + \sqrt{2\log(1/\delta)}\right) = \omega\left(1 + \sqrt{\frac{2\log(1/\delta)}{d_x}}\right).$$

The required norm condition therefore holds if

$$\omega\left(1 + \sqrt{\frac{2\log(1/\delta)}{d_x}}\right) \le \sqrt{\frac{B}{2R}}.$$

With this choice and $\lambda \ge \frac{4R}{Bn}\sum_{i=1}^{n} q(\mathbf{x}_i)$, retention holds simultaneously for all training times with probability at least $1 - \delta$ over initialization, conditional on any training dataset. Finally, empirical GF starts from $\mathbf{w}(0) = \mathbf{w}_0$ and satisfies $\dot{\mathbf{w}}(t) = -\nabla\widehat{\mathcal{L}}_\lambda(\mathbf{w}(t))$, so

$$\frac{d}{dt}\widehat{\mathcal{L}}_\lambda(\mathbf{w}(t)) = -\left\|\nabla\widehat{\mathcal{L}}_\lambda(\mathbf{w}(t))\right\|_2^2 \le 0.$$

Thus empirical GF satisfies the required descent assumption. The empirical objective is smooth, and its nonnegative quadratic penalty bounds the trajectory by its finite initial objective value, preventing finite-time escape. Hence this flow exists for all $t \ge 0$. □

C.2 Proof of Proposition 3

Proposition 3 (Gaussian adaptation and retention without PT data)**.** *Assume* $p_{\mathrm{ft}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{ft}}, \boldsymbol{\Sigma}_{\mathrm{ft}})$ *and* $p_{\mathrm{pt}} = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{pt}})$*, with* $\boldsymbol{\Sigma}_{\mathrm{ft}}, \boldsymbol{\Sigma}_{\mathrm{pt}} \succ 0$ *and* $\boldsymbol{\Sigma}_{\mathrm{pt}} \preceq R\mathbf{I}$ *for a known* $R > 0$*. Let* $\rho = \|\boldsymbol{\mu}_{\mathrm{ft}}\|_2 > 0$ *and* $\kappa = \boldsymbol{\mu}_{\mathrm{ft}}^\top \boldsymbol{\Sigma}_{\mathrm{ft}} \boldsymbol{\mu}_{\mathrm{ft}} / \rho^2$*. If* $B > 0$ *satisfies*

$$\rho^2 \geq \frac{(3R + \kappa)^2}{4R} B, \tag{20}$$

then $\widehat{\mathbf{w}} = \frac{1}{\rho}\sqrt{\frac{B}{R}}\,\boldsymbol{\mu}_{\mathrm{ft}}$*, requiring neither PT samples nor knowledge of* $\boldsymbol{\Sigma}_{\mathrm{pt}}$*, satisfies*

$$\mathcal{L}_{\mathrm{ft}}(\widehat{\mathbf{w}}) \leq \Theta(\widehat{\mathbf{w}}) e^{-\frac{B}{2}}, \qquad \mathcal{L}_{\mathrm{pt}}(\widehat{\mathbf{w}}) \leq C\,(B + R) e^{-\frac{B}{2}} \tag{21}$$

where $\Theta(\mathbf{w}) := \mathrm{Tr}(\mathbf{M}\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{M}^\top) + \|\mathbf{M}(\boldsymbol{\mu}_{\mathrm{ft}} - \boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w})\|_2^2$ *and* $C = \max\{R\,\|\mathbf{M}\|_{\mathrm{op}}^2, \|\mathbf{M}\|_{\mathrm{F}}^2\}$.

Proof. We bound each loss for an arbitrary gate and then evaluate the resulting bounds at the explicit gate $\widehat{\mathbf{w}}$. Since $f(\mathbf{x}) = \sigma(\mathbf{w}^\top\mathbf{x} - B)\mathbf{M}\mathbf{x}$, the only learnable parameter in this proposition is $\mathbf{w}$, so we write the two losses as functions of $\mathbf{w}$. We will repeatedly use

$$\|\mathbf{M}\mathbf{x}\|_2^2 = \mathbf{x}^\top \mathbf{M}^\top \mathbf{M}\mathbf{x}.$$

Both bounds below use the exponentially tilted Gaussian quadratic moment from Lemma 4, with $\mathbf{Q} = \mathbf{M}^\top\mathbf{M}$.

We first prove the adaptation bound. For every $u \in \mathbb{R}$,

$$(1 - \sigma(u))^2 \leq 1 - \sigma(u) = \frac{1}{1 + e^u} \leq e^{-u}.$$

Therefore,

$$\mathcal{L}_{\mathrm{ft}}(\mathbf{w}) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{ft}}}\left[\left(1 - \sigma(\mathbf{w}^\top\mathbf{x} - B)\right)^2 \|\mathbf{M}\mathbf{x}\|_2^2\right] \leq \frac{1}{2} e^{B} \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{ft}}}\left[e^{-\mathbf{w}^\top\mathbf{x}} \|\mathbf{M}\mathbf{x}\|_2^2\right]. \tag{32}$$

Apply Lemma 4 with $\boldsymbol{\mu} = \boldsymbol{\mu}_{\mathrm{ft}}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathrm{ft}}$, $\mathbf{a} = -\mathbf{w}$, and $\mathbf{Q} = \mathbf{M}^\top\mathbf{M}$. Using the definition of Θ, the lemma gives

$$\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{ft}}}\left[e^{-\mathbf{w}^\top\mathbf{x}} \|\mathbf{M}\mathbf{x}\|_2^2\right] = \exp\left(-\mathbf{w}^\top\boldsymbol{\mu}_{\mathrm{ft}} + \frac{1}{2}\mathbf{w}^\top\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}\right)\Theta(\mathbf{w}). \tag{33}$$

Combining (32) and (33), and using the definition of Θ, we get

$$\mathcal{L}_{\mathrm{ft}}(\mathbf{w}) \leq \frac{1}{2}\Theta(\mathbf{w}) \exp\left(B - \mathbf{w}^\top\boldsymbol{\mu}_{\mathrm{ft}} + \frac{1}{2}\mathbf{w}^\top\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}\right). \tag{34}$$

Lemma 4 also gives a useful forgetting bound for an arbitrary gate. Since $\sigma^2(s) \leq \sigma(s) \leq e^s$,

$$\mathcal{L}_{\mathrm{pt}}(\mathbf{w}) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{pt}}}\left[\sigma^2(\mathbf{w}^\top\mathbf{x} - B)\,\|\mathbf{M}\mathbf{x}\|_2^2\right] \leq \frac{1}{2} e^{-B} \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{pt}}}\left[e^{\mathbf{w}^\top\mathbf{x}} \|\mathbf{M}\mathbf{x}\|_2^2\right].$$

Applying Lemma 4 with $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathrm{pt}}$, $\mathbf{a} = \mathbf{w}$, and $\mathbf{Q} = \mathbf{M}^\top\mathbf{M}$, we obtain

$$\mathcal{L}_{\mathrm{pt}}(\mathbf{w}) \leq \frac{1}{2}\exp\left(-B + \frac{1}{2}\mathbf{w}^\top\boldsymbol{\Sigma}_{\mathrm{pt}}\mathbf{w}\right)\left(\|\mathbf{M}\boldsymbol{\Sigma}_{\mathrm{pt}}\mathbf{w}\|_2^2 + \mathrm{Tr}(\mathbf{M}\boldsymbol{\Sigma}_{\mathrm{pt}}\mathbf{M}^\top)\right). \tag{35}$$

For $\widehat{\mathbf{w}} = \frac{1}{\rho}\sqrt{\frac{B}{R}}\,\boldsymbol{\mu}_{\mathrm{ft}}$, we have

$$\widehat{\mathbf{w}}^\top\boldsymbol{\mu}_{\mathrm{ft}} = \rho\sqrt{\frac{B}{R}}, \qquad \|\widehat{\mathbf{w}}\|_2^2 = \frac{B}{R}, \qquad \widehat{\mathbf{w}}^\top\boldsymbol{\Sigma}_{\mathrm{ft}}\widehat{\mathbf{w}} = \frac{\kappa B}{R}.$$

Thus

$$B - \widehat{\mathbf{w}}^\top\boldsymbol{\mu}_{\mathrm{ft}} + \frac{1}{2}\widehat{\mathbf{w}}^\top\boldsymbol{\Sigma}_{\mathrm{ft}}\widehat{\mathbf{w}} = B - \rho\sqrt{\frac{B}{R}} + \frac{\kappa B}{2R}.$$

Condition (20) is equivalent to

$$\rho \geq \frac{3R+\kappa}{2}\sqrt{\frac{B}{R}}.$$

Multiplying this inequality by $\sqrt{\frac{B}{R}}$ gives

$$\rho\sqrt{\frac{B}{R}} \geq \frac{3B}{2} + \frac{\kappa B}{2R}.$$

Therefore

$$B - \rho\sqrt{\frac{B}{R}} + \frac{\kappa B}{2R} \leq -\frac{B}{2}.$$

Using this in (34) and dropping the harmless factor $\frac{1}{2}$ yields

$$\mathcal{L}_{\mathrm{ft}}(\widehat{\mathbf{w}}) \leq \Theta(\widehat{\mathbf{w}})e^{-B/2}. \tag{36}$$

To finish the forgetting estimate, evaluate (35) at $\widehat{\mathbf{w}}$. Using $\mathbf{\Sigma}_{\mathrm{pt}} \preceq R\mathbf{I}$ and $\|\widehat{\mathbf{w}}\|_2^2 = \frac{B}{R}$, the exponential term satisfies

$$\widehat{\mathbf{w}}^\top \mathbf{\Sigma}_{\mathrm{pt}} \widehat{\mathbf{w}} \leq R\,\|\widehat{\mathbf{w}}\|_2^2 = B.$$

The trace term is bounded by

$$\mathrm{Tr}(\mathbf{M}\mathbf{\Sigma}_{\mathrm{pt}}\mathbf{M}^\top) = \mathrm{Tr}(\mathbf{M}^\top\mathbf{M}\mathbf{\Sigma}_{\mathrm{pt}}) \leq R\,\mathrm{Tr}(\mathbf{M}^\top\mathbf{M}) = R\,\mathrm{Tr}(\mathbf{M}\mathbf{M}^\top).$$

For the mean term after tilting, use $\|\mathbf{\Sigma}_{\mathrm{pt}}\widehat{\mathbf{w}}\|_2 \leq R\,\|\widehat{\mathbf{w}}\|_2$, which follows from $\mathbf{\Sigma}_{\mathrm{pt}} \preceq R\mathbf{I}$. Then

$$\|\mathbf{M}\mathbf{\Sigma}_{\mathrm{pt}}\widehat{\mathbf{w}}\|_2^2 \leq \|\mathbf{M}\|_{\mathrm{op}}^2\,\|\mathbf{\Sigma}_{\mathrm{pt}}\widehat{\mathbf{w}}\|_2^2 \leq R^2\,\|\mathbf{M}\|_{\mathrm{op}}^2\,\|\widehat{\mathbf{w}}\|_2^2 = R\,\|\mathbf{M}\|_{\mathrm{op}}^2\,B.$$

Combining these estimates with (35), we obtain

$$\mathcal{L}_{\mathrm{pt}}(\widehat{\mathbf{w}}) \leq \frac{1}{2}e^{-B}e^{B/2}\left(R\,\mathrm{Tr}(\mathbf{M}\mathbf{M}^\top) + R\,\|\mathbf{M}\|_{\mathrm{op}}^2\,B\right) = \frac{1}{2}e^{-B/2}\left(R\,\mathrm{Tr}(\mathbf{M}\mathbf{M}^\top) + R\,\|\mathbf{M}\|_{\mathrm{op}}^2\,B\right).$$

Let $C = \max\{R\,\|\mathbf{M}\|_{\mathrm{op}}^2, \mathrm{Tr}(\mathbf{M}\mathbf{M}^\top)\}$. Then

$$R\,\mathrm{Tr}(\mathbf{M}\mathbf{M}^\top) \leq RC, \qquad R\,\|\mathbf{M}\|_{\mathrm{op}}^2\,B \leq BC.$$

Dropping the harmless factor $\frac{1}{2}$, we get

$$\mathcal{L}_{\mathrm{pt}}(\widehat{\mathbf{w}}) \leq C(B+R)e^{-B/2}. \tag{37}$$

Combining (36) and (37) proves (21). □

We first establish two Gaussian identities used in the proof of Theorem 3.

Lemma 6 (Gaussian integration by parts)**.** *Let $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, with $\mathbf{\Sigma} \succeq 0$, and let $\psi : \mathbb{R}^{d_x} \to \mathbb{R}$ be continuously differentiable. Suppose $|\psi(\mathbf{x})| + \|\nabla\psi(\mathbf{x})\|_2 \leq C(1+\|\mathbf{x}\|_2)^k$ for every $\mathbf{x} \in \mathbb{R}^{d_x}$, with constants $C > 0$ and $k \geq 0$. Then, including when $\mathbf{\Sigma}$ is singular,*

$$\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})\psi(\mathbf{x})\right] = \mathbf{\Sigma}\mathbb{E}\left[\nabla_{\mathbf{x}}\psi(\mathbf{x})\right]. \tag{38}$$

Proof. By Lemma 3, we may write $\mathbf{x} = \boldsymbol{\mu} + \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{g}$, where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x})$. Therefore

$$\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})\psi(\mathbf{x})\right] = \mathbf{\Sigma}^{\frac{1}{2}}\mathbb{E}\left[\mathbf{g}\psi\left(\boldsymbol{\mu} + \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{g}\right)\right].$$

The vector expectation on the right is computed one coordinate at a time. Fix a coordinate g_i, and temporarily hold all other coordinates fixed. Since $\phi'(g_i) = -g_i\phi(g_i)$, the one-dimensional rule $\int ab' = [ab] - \int a'b$, with $a = \psi\left(\boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{g}\right)$ and $b = \phi(g_i)$, gives

$$\begin{aligned}\int_{\mathbb{R}} g_i\psi(\mathbf{x})\phi(g_i)\,dg_i &= -\int_{\mathbb{R}} \psi(\mathbf{x})\phi'(g_i)\,dg_i \\ &= -\left[\psi(\mathbf{x})\phi(g_i)\right]_{g_i=-\infty}^{g_i=+\infty} + \int_{\mathbb{R}} \frac{\partial\psi(\mathbf{x})}{\partial g_i}\phi(g_i)\,dg_i \\ &= \int_{\mathbb{R}} \frac{\partial\psi(\mathbf{x})}{\partial g_i}\phi(g_i)\,dg_i.\end{aligned}$$

The boundary term vanishes because ψ grows at most polynomially, whereas the Gaussian density decays faster than any polynomial. The coordinates of $\mathbf{g}$ are independent, so averaging this equality over the remaining coordinates gives

$$\mathbb{E}\left[g_i\psi(\mathbf{x})\right] = \mathbb{E}\left[\frac{\partial\psi(\mathbf{x})}{\partial g_i}\right].$$

Both expectations now include every coordinate. The growth assumption makes all these integrals absolutely convergent, so the order of integration is valid. To compute the derivative, write $x_j = \mu_j + \sum_{\ell=1}^{d_x}(\boldsymbol{\Sigma}^{\frac{1}{2}})_{j\ell}g_\ell$, so $\frac{\partial x_j}{\partial g_i} = (\boldsymbol{\Sigma}^{\frac{1}{2}})_{ji}$. The chain rule then gives

$$\frac{\partial\psi(\mathbf{x})}{\partial g_i} = \sum_{j=1}^{d_x} \frac{\partial\psi(\mathbf{x})}{\partial x_j}(\boldsymbol{\Sigma}^{\frac{1}{2}})_{ji} = \left(\boldsymbol{\Sigma}^{\frac{1}{2}}\nabla_{\mathbf{x}}\psi(\mathbf{x})\right)_i,$$

where the last equality uses the symmetric square root. Collecting the coordinate identities and taking the fixed matrix outside the expectation yields

$$\mathbb{E}\left[\mathbf{g}\psi(\mathbf{x})\right] = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbb{E}\left[\nabla_{\mathbf{x}}\psi(\mathbf{x})\right].$$

Substitution into the first display, using $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{\Sigma}$, proves the result. No inverse of $\boldsymbol{\Sigma}$ is needed. □

Lemma 7 (Gaussian product moment)**.** *Let $\mathbf{z} \in \mathbb{R}^{d_y}$ and $Y \in \mathbb{R}$ be jointly Gaussian, with arbitrary means and possibly singular covariance. Then*

$$\mathbb{E}\left[\|\mathbf{z}\|_2^2\, Y^2\right] \le 3\mathbb{E}\left[\|\mathbf{z}\|_2^2\right]\mathbb{E}\left[Y^2\right].$$

Proof. First consider a scalar Gaussian $U = m + \sqrt{b}\,g$, where $b \ge 0$ and $g \sim \mathcal{N}(0,1)$. Its odd centered moments vanish by symmetry, $\mathbb{E}\left[g^2\right] = 1$, and Lemma 6 with $\psi(g) = g^3$ gives $\mathbb{E}\left[g^4\right] = 3\mathbb{E}\left[g^2\right] = 3$. Expanding the fourth power therefore gives

$$\mathbb{E}\left[U^4\right] = m^4 + 6m^2b + 3b^2 \le 3(m^2+b)^2 = 3\left(\mathbb{E}\left[U^2\right]\right)^2.$$

Each coordinate z_i of $\mathbf{z}$, and also Y, is scalar Gaussian. Expanding the squared norm and applying Cauchy–Schwarz to each term yields

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{z}\|_2^2\, Y^2\right] &= \sum_{i=1}^{d_y}\mathbb{E}\left[z_i^2Y^2\right] \le \sum_{i=1}^{d_y}\sqrt{\mathbb{E}\left[z_i^4\right]\mathbb{E}\left[Y^4\right]} \\ &\le 3\sum_{i=1}^{d_y}\mathbb{E}\left[z_i^2\right]\mathbb{E}\left[Y^2\right] = 3\mathbb{E}\left[\|\mathbf{z}\|_2^2\right]\mathbb{E}\left[Y^2\right].\end{aligned}$$

No independence or positive-definiteness assumption is needed. □

C.3 Proof of Theorem 3

Theorem 3 (Adaptation and retention under GF)**.** *Suppose* $p_{\mathrm{ft}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{ft}}, \boldsymbol{\Sigma}_{\mathrm{ft}})$ *and* $p_{\mathrm{pt}} = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{pt}})$*, with* $\boldsymbol{\Sigma}_{\mathrm{ft}} \succeq 0$ *and* $\mathbf{0} \preceq \boldsymbol{\Sigma}_{\mathrm{pt}} \preceq R\mathbf{I}$ *for* $R > 0$*. Let* $\rho = \|\boldsymbol{\mu}_{\mathrm{ft}}\|_2 > 0$ *and assume* $\mathbb{E}_{p_{\mathrm{ft}}}[q(\mathbf{x})] > 0$*,* $B \geq 10$*,* $\rho^2 \geq 360B\max\{\|\boldsymbol{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}, R\}$*, and* $\|\mathbf{w}_0\|_2 \leq \frac{1}{20\rho}$*. Define*

$$\lambda_{\min} = \frac{\rho^2(B+10)}{4B}\mathbb{E}_{p_{\mathrm{ft}}}[q(\mathbf{x})]\, e^{-\frac{39B}{20}}, \qquad \lambda_{\max} = \frac{\rho^2}{8B}\mathbb{E}_{p_{\mathrm{ft}}}[q(\mathbf{x})]\, e^{-B-\frac{1}{20}-\sqrt{\frac{3B}{10}}}.$$

Then $0 < \lambda_{\min} < \lambda_{\max}$*. For every* $\lambda \in (\lambda_{\min}, \lambda_{\max})$*, the population flow* (18) *exists for all* $t \geq 0$ *and converges to a stationary point* $\mathbf{w}_\infty$*. The time at which its mean FT score first reaches* $B/2$ *satisfies*

$$T := \inf\left\{t \geq 0 : \boldsymbol{\mu}_{\mathrm{ft}}^\top \mathbf{w}(t) - B \geq \frac{B}{2}\right\} \leq \frac{1+\frac{1}{30B}}{\lambda_{\max}-\lambda} < \infty.$$

The population losses satisfy

$$\begin{aligned}\sup_{t\geq T}\mathcal{L}_{\mathrm{ft}}(\mathbf{w}(t)) &\leq \left(1+\frac{B}{10}\right)\mathbb{E}_{p_{\mathrm{ft}}}[q(\mathbf{x})]\, e^{-\frac{19B}{20}},\\ \sup_{t\geq 0}\mathcal{L}_{\mathrm{pt}}(\mathbf{w}(t)) &\leq \frac{R}{2}\left(1+\frac{B}{10}\right)\|\mathbf{M}\|_{\mathrm{F}}^2\, e^{-\frac{39B}{20}}.\end{aligned} \tag{22}$$

Moreover, $\|\mathbf{w}(t) - \mathbf{w}_\infty\|_2 \leq \frac{6B}{\rho}e^{-\frac{\lambda}{2}(t-T)}$ *for every* $t \geq T$*. For Gaussian initialization, fix* $\delta \in (0,1)$ *and choose* $\omega > 0$ *so that*

$$\mathbf{w}_0 \sim \mathcal{N}\left(\mathbf{0}, \frac{\omega^2}{d_x}\mathbf{I}_{d_x}\right), \qquad \omega\sqrt{2+\frac{4}{d_x}\log\frac{1}{\delta}} \leq \frac{1}{20\rho}.$$

Then the required initial norm bound, and hence all the conclusions above, hold with probability at least $1-\delta$ *over initialization.*

Proof. We first study the direction and size of the fine-tuning force. These estimates will let us find a region that contains the training path and show that the gate opens enough on fine-tuning inputs. We then bound both population losses, prove convergence, and check when a small random initialization satisfies the required conditions.

C.3.1 Bound the fine-tuning force

We start by identifying the part of the gradient that comes from fine-tuning. After dividing out its average weight, this force becomes a weighted average of the inputs. We bound its distance from the fine-tuning mean; this will let us control motion away from the mean direction later.

Let $p_{\mathrm{ft}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{ft}}, \boldsymbol{\Sigma}_{\mathrm{ft}})$, where $\boldsymbol{\Sigma}_{\mathrm{ft}} \succeq 0$, and write $\rho = \|\boldsymbol{\mu}_{\mathrm{ft}}\|_2 > 0$ and $\mathbf{u} = \boldsymbol{\mu}_{\mathrm{ft}}/\rho$. Recall that $q(\mathbf{x}) = \|\mathbf{M}\mathbf{x}\|_2^2 = \mathbf{x}^\top\mathbf{Q}\mathbf{x}$, where $\mathbf{Q} = \mathbf{M}^\top\mathbf{M} \succeq 0$. We keep $\mathbf{M}$ and $B > 0$ fixed and train only $\mathbf{w}$ using the objective

$$\mathcal{L}_\lambda(\mathbf{w}) = \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{ft}}}\left[\left(1-\sigma(\mathbf{w}^\top\mathbf{x}-B)\right)^2 q(\mathbf{x})\right] + \frac{\lambda}{2}\|\mathbf{w}\|_2^2, \qquad \lambda > 0.$$

All expectations below are over $\mathbf{x} \sim p_{\mathrm{ft}}$. We assume $\mathbb{E}[q(\mathbf{x})] > 0$; otherwise $q(\mathbf{x}) = 0$ almost surely, so there is no fine-tuning correction to learn. Lemma 4, with its vector parameter $\mathbf{a} = \mathbf{0}$, gives the explicit expression

$$\mathbb{E}[q(\mathbf{x})] = \mathrm{Tr}(\mathbf{Q}\boldsymbol{\Sigma}_{\mathrm{ft}}) + \boldsymbol{\mu}_{\mathrm{ft}}^\top\mathbf{Q}\boldsymbol{\mu}_{\mathrm{ft}} > 0.$$

To determine how the gate moves, we first differentiate this objective with respect to $\mathbf{w}$. For a fixed input $\mathbf{x}$, write $s = \mathbf{w}^\top\mathbf{x} - B$, so $\nabla_{\mathbf{w}}s = \mathbf{x}$. The factor $q(\mathbf{x})$ does not depend on $\mathbf{w}$. Using $\sigma'(s) = \sigma(s)(1-\sigma(s))$, the chain rule gives

$$\nabla_{\mathbf{w}}\left[\left(1-\sigma(s)\right)^2 q(\mathbf{x})\right] = -2\left(1-\sigma(s)\right)\sigma'(s)q(\mathbf{x})\mathbf{x} = -2\sigma(s)\left(1-\sigma(s)\right)^2 q(\mathbf{x})\mathbf{x}.$$

Since $0 < \sigma(s) < 1$, the norm of this derivative is at most $2q(\mathbf{x}) \|\mathbf{x}\|_2 \le 2 \|\mathbf{M}\|_{\text{op}}^2 \|\mathbf{x}\|_2^3$, which has finite expectation under the Gaussian distribution and does not depend on $\mathbf{w}$. We may therefore differentiate under the expectation. The regularizer contributes $\nabla_{\mathbf{w}} \left(\frac{\lambda}{2} \|\mathbf{w}\|_2^2 \right) = \lambda \mathbf{w}$, so

$$\nabla \mathcal{L}_\lambda(\mathbf{w}) = -2\mathbb{E}\left[\sigma(\mathbf{w}^\top \mathbf{x} - B) \left(1 - \sigma(\mathbf{w}^\top \mathbf{x} - B)\right)^2 q(\mathbf{x}) \mathbf{x} \right] + \lambda \mathbf{w}.$$

GF follows $\dot{\mathbf{w}}(t) = -\nabla \mathcal{L}_\lambda \left(\mathbf{w}(t)\right)$. To separate its fine-tuning contribution from the regularizer, denote the repeated scalar factor by $h(s) = \sigma(s) \left(1 - \sigma(s)\right)^2$ and define the fine-tuning force by

$$\mathbf{F}_{\text{ft}}(\mathbf{w}) = 2\mathbb{E}\left[h(\mathbf{w}^\top \mathbf{x} - B) q(\mathbf{x}) \mathbf{x} \right], \qquad \dot{\mathbf{w}}(t) = \mathbf{F}_{\text{ft}}\left(\mathbf{w}(t)\right) - \lambda \mathbf{w}(t). \tag{39}$$

Fix an arbitrary $\mathbf{w}$. Since $h(s) > 0$, $h(s) \le 1$, and $0 < \mathbb{E}\left[q(\mathbf{x})\right] < \infty$, the average weight $\mathbb{E}\left[h(\mathbf{w}^\top \mathbf{x} - B) q(\mathbf{x})\right]$ is positive and finite. Dividing the force by twice this weight gives

$$\frac{\mathbf{F}_{\text{ft}}(\mathbf{w})}{2\mathbb{E}\left[h(\mathbf{w}^\top \mathbf{x} - B) q(\mathbf{x})\right]} = \frac{\mathbb{E}\left[h(\mathbf{w}^\top \mathbf{x} - B) q(\mathbf{x}) \mathbf{x}\right]}{\mathbb{E}\left[h(\mathbf{w}^\top \mathbf{x} - B) q(\mathbf{x})\right]}.$$

This is the weighted average of the inputs: dividing by a positive scalar preserves the force direction and allows us to compare it directly with $\boldsymbol{\mu}_{\text{ft}}$. With $s = \mathbf{w}^\top \mathbf{x} - B$ as above, substituting $\mathbf{x} = \boldsymbol{\mu}_{\text{ft}} + (\mathbf{x} - \boldsymbol{\mu}_{\text{ft}})$ into the numerator and using linearity of expectation gives

$$\mathbb{E}\left[h(s) q(\mathbf{x}) \mathbf{x}\right] = \mathbb{E}\left[h(s) q(\mathbf{x})\right] \boldsymbol{\mu}_{\text{ft}} + \mathbb{E}\left[h(s) q(\mathbf{x}) \left(\mathbf{x} - \boldsymbol{\mu}_{\text{ft}}\right)\right].$$

Dividing by $\mathbb{E}\left[h(s) q(\mathbf{x})\right]$ and subtracting $\boldsymbol{\mu}_{\text{ft}}$ therefore yields

$$\frac{\mathbf{F}_{\text{ft}}(\mathbf{w})}{2\mathbb{E}\left[h(s) q(\mathbf{x})\right]} - \boldsymbol{\mu}_{\text{ft}} = \frac{\mathbb{E}\left[h(s) q(\mathbf{x}) \left(\mathbf{x} - \boldsymbol{\mu}_{\text{ft}}\right)\right]}{\mathbb{E}\left[h(s) q(\mathbf{x})\right]}. \tag{40}$$

Thus the quantity we need to control is exactly

$$\left\| \frac{\mathbf{F}_{\text{ft}}(\mathbf{w})}{2\mathbb{E}\left[h(s) q(\mathbf{x})\right]} - \rho \mathbf{u} \right\|_2 = \frac{\left\| \mathbb{E}\left[h(s) q(\mathbf{x}) \left(\mathbf{x} - \boldsymbol{\mu}_{\text{ft}}\right)\right] \right\|_2}{\mathbb{E}\left[h(s) q(\mathbf{x})\right]}.$$

If this ratio is small compared with ρ, the force points mostly along $\mathbf{u}$, although its orthogonal component need not vanish. To compute the numerator, apply Lemma 6 with $\boldsymbol{\mu} = \boldsymbol{\mu}_{\text{ft}}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\text{ft}}$, and $\psi(\mathbf{x}) = q(\mathbf{x}) h(\mathbf{w}^\top \mathbf{x} - B)$. Since q is quadratic and h, h' are bounded, the lemma's polynomial-growth condition holds. Thus

$$\mathbb{E}\left[h(s) q(\mathbf{x}) (\mathbf{x} - \boldsymbol{\mu}_{\text{ft}})\right] = \boldsymbol{\Sigma}_{\text{ft}} \mathbb{E}\left[\nabla_{\mathbf{x}} \psi(\mathbf{x})\right].$$

We now evaluate $\nabla_{\mathbf{x}} \psi(\mathbf{x})$ for our weight $\psi(\mathbf{x}) = q(\mathbf{x}) h(s)$. Since $\mathbf{Q}$ is symmetric, $\nabla_{\mathbf{x}} q(\mathbf{x}) = 2\mathbf{Q}\mathbf{x}$, while $\nabla_{\mathbf{x}} s = \mathbf{w}$. Using $\sigma'(s) = \sigma(s) \left(1 - \sigma(s)\right)$, the scalar derivative simplifies to

$$h'(s) = \sigma'(s) \left(1 - \sigma(s)\right)^2 - 2\sigma(s) \left(1 - \sigma(s)\right) \sigma'(s) = h(s) \left(1 - 3\sigma(s)\right).$$

The product and chain rules therefore give

$$\nabla_{\mathbf{x}} \psi(\mathbf{x}) = h(s) \nabla_{\mathbf{x}} q(\mathbf{x}) + q(\mathbf{x}) h'(s) \nabla_{\mathbf{x}} s = 2h(s) \mathbf{Q}\mathbf{x} + q(\mathbf{x}) h(s) \left(1 - 3\sigma(s)\right) \mathbf{w}.$$

Substituting this derivative into the Gaussian identity and taking the fixed matrix $\mathbf{Q}$ and vector $\mathbf{w}$ outside their expectations gives

$$\mathbb{E}\left[h(s) q(\mathbf{x}) \left(\mathbf{x} - \boldsymbol{\mu}_{\text{ft}}\right)\right] = 2\boldsymbol{\Sigma}_{\text{ft}} \mathbf{Q} \mathbb{E}\left[h(s) \mathbf{x}\right] + \boldsymbol{\Sigma}_{\text{ft}} \mathbf{w} \mathbb{E}\left[q(\mathbf{x}) h(s) \left(1 - 3\sigma(s)\right)\right]. \tag{41}$$

The two terms come from differentiating $q(\mathbf{x})$ and the gate factor, respectively. The triangle inequality now bounds the norm of their sum:

$$\begin{aligned} \left\| \mathbb{E}\left[h(s) q(\mathbf{x}) \left(\mathbf{x} - \boldsymbol{\mu}_{\text{ft}}\right)\right] \right\|_2 &= \left\| 2\boldsymbol{\Sigma}_{\text{ft}} \mathbf{Q} \mathbb{E}\left[h(s) \mathbf{x}\right] + \boldsymbol{\Sigma}_{\text{ft}} \mathbf{w} \mathbb{E}\left[q(\mathbf{x}) h(s) \left(1 - 3\sigma(s)\right)\right] \right\|_2 \\ &\le 2 \left\| \boldsymbol{\Sigma}_{\text{ft}} \mathbf{Q} \mathbb{E}\left[h(s) \mathbf{x}\right] \right\|_2 + \left\| \boldsymbol{\Sigma}_{\text{ft}} \mathbf{w} \right\|_2 \left| \mathbb{E}\left[q(\mathbf{x}) h(s) \left(1 - 3\sigma(s)\right)\right] \right|. \end{aligned}$$

In the second term, the expectation is a scalar and differs from our denominator only by the factor $1 - 3\sigma(s)$. Since $0 < \sigma(s) < 1$, we have $|1 - 3\sigma(s)| \le 2$. Using $|\mathbb{E}[Y]| \le \mathbb{E}[|Y|]$ and $q(\mathbf{x}) h(s) \ge 0$ gives

$$\begin{aligned} \left| \mathbb{E}\left[q(\mathbf{x}) h(s) \left(1 - 3\sigma(s)\right)\right] \right| &\le \mathbb{E}\left[q(\mathbf{x}) h(s) |1 - 3\sigma(s)|\right] \\ &\le 2\mathbb{E}\left[q(\mathbf{x}) h(s)\right]. \end{aligned}$$

The positive expectation on the right is exactly the denominator of our ratio. Dividing therefore bounds the contribution of this second term by

$$\frac{\|\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{w}\|_2 \left|\mathbb{E}\left[q(\mathbf{x})h(s)\left(1-3\sigma(s)\right)\right]\right|}{\mathbb{E}\left[h(s)q(\mathbf{x})\right]} \leq \frac{2\|\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{w}\|_2 \mathbb{E}\left[h(s)q(\mathbf{x})\right]}{\mathbb{E}\left[h(s)q(\mathbf{x})\right]} = 2\|\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{w}\|_2 .$$

Substituting this bound into the triangle inequality above and recalling the expression for the difference between the weighted average and $\rho\mathbf{u}$, we obtain

$$\left\| \frac{\mathbf{F}_{\mathrm{ft}}(\mathbf{w})}{2\mathbb{E}\left[h(s)q(\mathbf{x})\right]} - \rho\mathbf{u} \right\|_2 \leq 2\|\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{w}\|_2 + \frac{2\|\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{Q}\mathbb{E}\left[h(s)\mathbf{x}\right]\|_2}{\mathbb{E}\left[h(s)q(\mathbf{x})\right]}. \tag{42}$$

It remains to bound the fraction on the right. The factor $h(s)$ cannot be cancelled across different expectations. Instead, divide the numerator and denominator by the same positive scalar $\mathbb{E}\left[h(s)\right]$:

$$\frac{2\|\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{Q}\mathbb{E}\left[h(s)\mathbf{x}\right]\|_2}{\mathbb{E}\left[h(s)q(\mathbf{x})\right]} = \frac{2\left\|\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{Q}\frac{\mathbb{E}[h(s)\mathbf{x}]}{\mathbb{E}[h(s)]}\right\|_2}{\frac{\mathbb{E}[h(s)q(\mathbf{x})]}{\mathbb{E}[h(s)]}}.$$

The weights $h(s)/\mathbb{E}\left[h(s)\right]$ are positive and average to one, so they define a probability law. For any function φ with finite weighted expectation, write

$$\mathbb{E}_h\left[\varphi(\mathbf{x})\right] = \frac{\mathbb{E}\left[h(s)\varphi(\mathbf{x})\right]}{\mathbb{E}\left[h(s)\right]}, \qquad s = \mathbf{w}^\top\mathbf{x} - B.$$

The subscript h denotes weighting at this fixed $\mathbf{w}$; the resulting law need not be Gaussian. With this notation, the fraction becomes

$$\frac{2\|\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{Q}\mathbb{E}\left[h(s)\mathbf{x}\right]\|_2}{\mathbb{E}\left[h(s)q(\mathbf{x})\right]} = \frac{2\|\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{Q}\mathbb{E}_h\left[\mathbf{x}\right]\|_2}{\mathbb{E}_h\left[q(\mathbf{x})\right]}.$$

Write $\boldsymbol{\mu}_h = \mathbb{E}_h\left[\mathbf{x}\right]$ for the weighted mean in the numerator. To express the denominator using the same mean, substitute $\mathbf{x} = \boldsymbol{\mu}_h + (\mathbf{x} - \boldsymbol{\mu}_h)$ into $q(\mathbf{x}) = \mathbf{x}^\top\mathbf{Q}\mathbf{x}$ and expand:

$$\begin{aligned}\mathbb{E}_h\left[q(\mathbf{x})\right] &= \boldsymbol{\mu}_h^\top\mathbf{Q}\boldsymbol{\mu}_h + 2\boldsymbol{\mu}_h^\top\mathbf{Q}\mathbb{E}_h\left[\mathbf{x}-\boldsymbol{\mu}_h\right] + \mathbb{E}_h\left[\left(\mathbf{x}-\boldsymbol{\mu}_h\right)^\top\mathbf{Q}\left(\mathbf{x}-\boldsymbol{\mu}_h\right)\right] \\ &= \|\mathbf{M}\boldsymbol{\mu}_h\|_2^2 + \mathbb{E}_h\left[\|\mathbf{M}\left(\mathbf{x}-\boldsymbol{\mu}_h\right)\|_2^2\right].\end{aligned}$$

Here the middle term vanishes because $\mathbb{E}_h\left[\mathbf{x}-\boldsymbol{\mu}_h\right] = \mathbf{0}$, and the last equality uses $\mathbf{Q} = \mathbf{M}^\top\mathbf{M}$. All these moments are finite because h is bounded, $\mathbb{E}\left[h(s)\right] > 0$, and the original Gaussian has finite second moments. Our fraction therefore has the exact form

$$\frac{2\|\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{Q}\mathbb{E}\left[h(s)\mathbf{x}\right]\|_2}{\mathbb{E}\left[h(s)q(\mathbf{x})\right]} = \frac{2\|\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{Q}\boldsymbol{\mu}_h\|_2}{\|\mathbf{M}\boldsymbol{\mu}_h\|_2^2 + \mathbb{E}_h\left[\|\mathbf{M}\left(\mathbf{x}-\boldsymbol{\mu}_h\right)\|_2^2\right]}. \tag{43}$$

We first bound the numerator using $\mathbf{Q} = \mathbf{M}^\top\mathbf{M}$ and the symmetric square root of $\mathbf{\Sigma}_{\mathrm{ft}}$. Applying $\|\mathbf{A}\mathbf{v}\|_2 \leq \|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{v}\|_2$ and then bounding the operator norm by the Frobenius norm gives

$$\begin{aligned}\|\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{Q}\boldsymbol{\mu}_h\|_2^2 &= \left\|\mathbf{\Sigma}_{\mathrm{ft}}^{\frac{1}{2}}\left(\mathbf{\Sigma}_{\mathrm{ft}}^{\frac{1}{2}}\mathbf{M}^\top\right)\mathbf{M}\boldsymbol{\mu}_h\right\|_2^2 \\ &\leq \left\|\mathbf{\Sigma}_{\mathrm{ft}}^{\frac{1}{2}}\right\|_{\mathrm{op}}^2 \left\|\mathbf{\Sigma}_{\mathrm{ft}}^{\frac{1}{2}}\mathbf{M}^\top\right\|_{\mathrm{F}}^2 \|\mathbf{M}\boldsymbol{\mu}_h\|_2^2 \\ &= \|\mathbf{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}} \operatorname{Tr}\left(\mathbf{Q}\mathbf{\Sigma}_{\mathrm{ft}}\right)\|\mathbf{M}\boldsymbol{\mu}_h\|_2^2 .\end{aligned}$$

The last equality uses $\left\|\mathbf{\Sigma}_{\mathrm{ft}}^{\frac{1}{2}}\right\|_{\mathrm{op}}^2 = \|\mathbf{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}$ and $\left\|\mathbf{\Sigma}_{\mathrm{ft}}^{\frac{1}{2}}\mathbf{M}^\top\right\|_{\mathrm{F}}^2 = \operatorname{Tr}(\mathbf{M}\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{M}^\top) = \operatorname{Tr}(\mathbf{Q}\mathbf{\Sigma}_{\mathrm{ft}})$. The factor $\|\mathbf{M}\boldsymbol{\mu}_h\|_2^2$ already occurs in our denominator. To control the fraction, we therefore compare its other term, $\mathbb{E}_h\left[\|\mathbf{M}(\mathbf{x}-\boldsymbol{\mu}_h)\|_2^2\right]$, with $\operatorname{Tr}(\mathbf{Q}\mathbf{\Sigma}_{\mathrm{ft}}) = \mathbb{E}\left[\|\mathbf{M}(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ft}})\|_2^2\right]$. To lower bound this expected square, we use the fact that the weight depends only on the scalar score $s = \mathbf{w}^\top\mathbf{x} - B$. By Lemma 3, its original Gaussian distribution has mean and variance

$$\bar{s} = \mathbf{w}^\top\boldsymbol{\mu}_{\mathrm{ft}} - B, \qquad v = \mathbf{w}^\top\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{w}, \qquad s \sim \mathcal{N}(\bar{s}, v).$$

If $v = 0$, the score and its weight are constant almost surely, so $\boldsymbol{\mu}_h = \boldsymbol{\mu}_{\mathrm{ft}}$ and $\mathbb{E}_h\left[\|\mathbf{M}(\mathbf{x}-\boldsymbol{\mu}_h)\|_2^2\right] = \mathrm{Tr}\left(\mathbf{Q}\boldsymbol{\Sigma}_{\mathrm{ft}}\right)$. Suppose $v > 0$. We separate the input into the part correlated with the score and the remaining part by defining

$$\mathbf{z} = \mathbf{x} - \boldsymbol{\mu}_{\mathrm{ft}} - \frac{\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}}{v}(s-\bar{s}), \qquad \mathbf{x} = \boldsymbol{\mu}_{\mathrm{ft}} + \frac{\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}}{v}(s-\bar{s}) + \mathbf{z}.$$

The coefficient is chosen to make the residual uncorrelated with the score. Indeed, these are affine functions of a Gaussian input, so Lemma 3 gives joint Gaussianity and

$$\mathbb{E}\left[\mathbf{z}\right] = \mathbf{0}, \qquad \mathrm{Cov}(\mathbf{z}, s) = \boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w} - \frac{\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}}{v}v = \mathbf{0}.$$

Thus $\mathbf{z}$ and s are independent. For bounded functions a, b, $\mathbb{E}_h\left[a(s)b(\mathbf{z})\right] = \mathbb{E}\left[h(s)a(s)b(\mathbf{z})\right]/\mathbb{E}\left[h(s)\right] = \mathbb{E}_h\left[a(s)\right]\mathbb{E}\left[b(\mathbf{z})\right]$. Thus weighting changes only the score distribution: the residual remains independent and keeps its original law. Taking the weighted mean gives $\boldsymbol{\mu}_h = \boldsymbol{\mu}_{\mathrm{ft}} + (\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}/v)(\mathbb{E}_h\left[s\right] - \bar{s})$, and subtracting it gives

$$\mathbf{x} - \boldsymbol{\mu}_h = \frac{\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}}{v}\left(s - \mathbb{E}_h\left[s\right]\right) + \mathbf{z}.$$

We can now compute the expected square we want to bound. Applying $\mathbf{M}$, expanding the squared norm, and using independence and $\mathbb{E}\left[\mathbf{z}\right] = \mathbf{0}$ to remove the cross term gives

$$\mathbb{E}_h\left[\|\mathbf{M}\left(\mathbf{x}-\boldsymbol{\mu}_h\right)\|_2^2\right] = \frac{\|\mathbf{M}\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}\|_2^2}{v^2}\,\mathrm{Var}_h(s) + \mathbb{E}\left[\|\mathbf{M}\mathbf{z}\|_2^2\right].$$

In particular, the cross term contains $\mathbb{E}_h\left[(s-\mathbb{E}_h\left[s\right])\mathbf{z}\right] = \mathbb{E}_h\left[s-\mathbb{E}_h\left[s\right]\right]\mathbb{E}\left[\mathbf{z}\right] = \mathbf{0}$. The same expansion under the original law, using $\mathrm{Var}(s) = v$, gives

$$\mathrm{Tr}\left(\mathbf{Q}\boldsymbol{\Sigma}_{\mathrm{ft}}\right) = \mathbb{E}\left[\|\mathbf{M}\left(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ft}}\right)\|_2^2\right] = \frac{\|\mathbf{M}\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}\|_2^2}{v} + \mathbb{E}\left[\|\mathbf{M}\mathbf{z}\|_2^2\right].$$

Only the score variance changes between these expressions. To lower bound it, apply Lemma 6 to $s \sim \mathcal{N}(\bar{s}, v)$ with $\psi(s) = (s-\mathbb{E}_h\left[s\right])h(s)$, treating the weighted mean as a constant. Since h, h' are bounded, ψ and $\psi'(s) = h(s) + (s-\mathbb{E}_h\left[s\right])h'(s)$ grow at most linearly. The lemma gives

$$\mathbb{E}\left[(s-\bar{s})(s-\mathbb{E}_h\left[s\right])h(s)\right] = v\mathbb{E}\left[h(s)\right] + v\mathbb{E}\left[(s-\mathbb{E}_h\left[s\right])h'(s)\right].$$

Divide by $\mathbb{E}\left[h(s)\right] > 0$. On the left, adding and subtracting the weighted mean shows that

$$\mathbb{E}_h\left[(s-\bar{s})(s-\mathbb{E}_h\left[s\right])\right] = \mathrm{Var}_h(s) + (\mathbb{E}_h\left[s\right]-\bar{s})\mathbb{E}_h\left[s-\mathbb{E}_h\left[s\right]\right] = \mathrm{Var}_h(s).$$

On the right, use the identity $h'(s)/h(s) = 1 - 3\sigma(s)$ already derived above. The centered factor has weighted mean zero, so

$$\mathrm{Var}_h(s) = v + v\mathbb{E}_h\left[(s-\mathbb{E}_h\left[s\right])(1-3\sigma(s))\right] = v - 3v\,\mathrm{Cov}_h\left(s, \sigma(s)\right).$$

We need an upper bound on this covariance because it is subtracted. Since $\mathbb{E}_h\left[s-\mathbb{E}_h\left[s\right]\right] = 0$, subtracting the constant $\sigma(\mathbb{E}_h\left[s\right])$ from its other factor does not change the covariance. Also, $0 \le \sigma'(s) = \sigma(s)(1-\sigma(s)) \le 1/4$, so $|\sigma(s) - \sigma(\mathbb{E}_h\left[s\right])| \le |s - \mathbb{E}_h\left[s\right]|/4$. Therefore

$$\begin{aligned}\mathrm{Cov}_h\left(s, \sigma(s)\right) &= \mathbb{E}_h\left[\left(s-\mathbb{E}_h\left[s\right]\right)\left(\sigma(s)-\sigma(\mathbb{E}_h\left[s\right])\right)\right] \\ &\le \frac{1}{4}\mathbb{E}_h\left[\left(s-\mathbb{E}_h\left[s\right]\right)^2\right] = \frac{1}{4}\,\mathrm{Var}_h(s).\end{aligned}$$

Substitution into the variance identity and rearrangement yield

$$\mathrm{Var}_h(s) \ge v - \frac{3v}{4}\,\mathrm{Var}_h(s), \qquad \mathrm{Var}_h(s) \ge \frac{v}{1+\frac{3}{4}v}.$$

We now substitute this bound into the expected square. Since the unchanged residual term is nonnegative and $1 + \frac{3}{4}v \ge 1$,

$$\begin{aligned}\mathbb{E}_h\left[\|\mathbf{M}\left(\mathbf{x}-\boldsymbol{\mu}_h\right)\|_2^2\right] &\ge \frac{\|\mathbf{M}\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}\|_2^2}{v(1+\frac{3}{4}v)} + \mathbb{E}\left[\|\mathbf{M}\mathbf{z}\|_2^2\right] \\ &\ge \frac{\|\mathbf{M}\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}\|_2^2/v + \mathbb{E}\left[\|\mathbf{M}\mathbf{z}\|_2^2\right]}{1+\frac{3}{4}v} = \frac{\mathrm{Tr}\left(\mathbf{Q}\boldsymbol{\Sigma}_{\mathrm{ft}}\right)}{1+\frac{3}{4}v}.\end{aligned}$$

Together with the case $v = 0$, this gives exactly the denominator comparison we need, for every $\mathbf{w}$, including singular $\mathbf{\Sigma}_{\text{ft}}$:

$$\mathbb{E}_h\left[\|\mathbf{M}(\mathbf{x}-\boldsymbol{\mu}_h)\|_2^2\right] \geq \frac{\operatorname{Tr}(\mathbf{Q}\mathbf{\Sigma}_{\text{ft}})}{1+\frac{3}{4}\mathbf{w}^\top\mathbf{\Sigma}_{\text{ft}}\mathbf{w}}. \tag{44}$$

Equivalently, $\operatorname{Tr}(\mathbf{Q}\mathbf{\Sigma}_{\text{ft}}) \leq (1+\frac{3}{4}v)\mathbb{E}_h\left[\|\mathbf{M}(\mathbf{x}-\boldsymbol{\mu}_h)\|_2^2\right]$. Substituting into the numerator bound obtained after (43) gives

$$\|\mathbf{\Sigma}_{\text{ft}}\mathbf{Q}\boldsymbol{\mu}_h\|_2^2 \leq \|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}\left(1+\frac{3}{4}v\right)\mathbb{E}_h\left[\|\mathbf{M}(\mathbf{x}-\boldsymbol{\mu}_h)\|_2^2\right]\|\mathbf{M}\boldsymbol{\mu}_h\|_2^2.$$

The last two factors are exactly the two terms in our denominator. Both are nonnegative. Taking square roots and using $2\sqrt{ab} \leq a+b$, which follows from $\left(\sqrt{a}-\sqrt{b}\right)^2 \geq 0$, with $a = \mathbb{E}_h\left[\|\mathbf{M}(\mathbf{x}-\boldsymbol{\mu}_h)\|_2^2\right]$ and $b = \|\mathbf{M}\boldsymbol{\mu}_h\|_2^2$, gives

$$2\|\mathbf{\Sigma}_{\text{ft}}\mathbf{Q}\boldsymbol{\mu}_h\|_2 \leq \sqrt{\|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}\left(1+\frac{3}{4}v\right)}\left(\|\mathbf{M}\boldsymbol{\mu}_h\|_2^2 + \mathbb{E}_h\left[\|\mathbf{M}(\mathbf{x}-\boldsymbol{\mu}_h)\|_2^2\right]\right).$$

The last parentheses equal $\mathbb{E}_h[q(\mathbf{x})] = \mathbb{E}[h(s)q(\mathbf{x})]/\mathbb{E}[h(s)] > 0$, so we may divide by them. Recalling $v = \mathbf{w}^\top\mathbf{\Sigma}_{\text{ft}}\mathbf{w}$, our remaining fraction is therefore bounded by

$$\frac{2\|\mathbf{\Sigma}_{\text{ft}}\mathbf{Q}\boldsymbol{\mu}_h\|_2}{\mathbb{E}_h[q(\mathbf{x})]} \leq \sqrt{\|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}\left(1+\frac{3}{4}\mathbf{w}^\top\mathbf{\Sigma}_{\text{ft}}\mathbf{w}\right)}.$$

Substituting this into the force inequality derived earlier completes the two bounds we needed:

$$\left\|\frac{\mathbf{F}_{\text{ft}}(\mathbf{w})}{2\mathbb{E}[h(s)q(\mathbf{x})]} - \rho\mathbf{u}\right\|_2 \leq 2\|\mathbf{\Sigma}_{\text{ft}}\mathbf{w}\|_2 + \sqrt{\|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}\left(1+\frac{3}{4}\mathbf{w}^\top\mathbf{\Sigma}_{\text{ft}}\mathbf{w}\right)}. \tag{45}$$

For a simpler expression depending only on the covariance norm and the gate norm, use $\|\mathbf{\Sigma}_{\text{ft}}\mathbf{w}\|_2 \leq \|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}\|\mathbf{w}\|_2$, $\mathbf{w}^\top\mathbf{\Sigma}_{\text{ft}}\mathbf{w} \leq \|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}\|\mathbf{w}\|_2^2$, and $\sqrt{1+a} \leq 1+\sqrt{a}$ for $a \geq 0$. These give

$$\begin{aligned}2\|\mathbf{\Sigma}_{\text{ft}}\mathbf{w}\|_2 + \sqrt{\|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}\left(1+\frac{3}{4}\mathbf{w}^\top\mathbf{\Sigma}_{\text{ft}}\mathbf{w}\right)} &\leq \sqrt{\|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}} + \left(2+\frac{\sqrt{3}}{2}\right)\|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}\|\mathbf{w}\|_2 \\ &\leq \sqrt{\|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}} + 3\|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}\|\mathbf{w}\|_2.\end{aligned}$$

Consequently, recalling $s = \mathbf{w}^\top\mathbf{x} - B$, the final force estimate is

$$\boxed{\left\|\frac{\mathbf{F}_{\text{ft}}(\mathbf{w})}{2\mathbb{E}[h(s)q(\mathbf{x})]} - \rho\mathbf{u}\right\|_2 \leq \sqrt{\|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}} + 3\|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}\|\mathbf{w}\|_2.} \tag{46}$$

C.3.2 Control deviation from the mean direction

We now use the force estimate along the training path to bound the perpendicular component in terms of progress along the mean. This comparison holds while the force is sufficiently close to the mean direction.

Boundedness of h' and $\mathbb{E}\left[q(\mathbf{x})\|\mathbf{x}\|_2^2\right] < \infty$ make the vector field locally Lipschitz, so every finite $\mathbf{w}(0)$ has a unique local solution. Along this solution, the chain rule gives

$$\frac{d}{dt}\mathcal{L}_\lambda(\mathbf{w}(t)) = \nabla\mathcal{L}_\lambda(\mathbf{w}(t))^\top\dot{\mathbf{w}}(t) = -\|\nabla\mathcal{L}_\lambda(\mathbf{w}(t))\|_2^2 \leq 0.$$

Thus the objective cannot exceed its initial value. Its fine-tuning term is nonnegative, so the regularization term alone satisfies

$$\frac{\lambda}{2}\|\mathbf{w}(t)\|_2^2 \leq \mathcal{L}_\lambda(\mathbf{w}(t)) \leq \mathcal{L}_\lambda(\mathbf{w}(0)).$$

Since $\lambda > 0$, dividing and taking square roots gives

$$\boxed{\|\mathbf{w}(t)\|_2 \le \sqrt{\frac{2\mathcal{L}_\lambda\left(\mathbf{w}(0)\right)}{\lambda}}, \qquad t \ge 0.} \tag{47}$$

Initially this bound holds on the solution's interval of existence. It prevents escape to infinity in finite time, so the solution extends to all $t \ge 0$. The bound can be loose when λ is small; we seek a sharper one by controlling the components along and perpendicular to the mean. Recall $\mathbf{u} = \boldsymbol{\mu}_{\mathrm{ft}}/\rho$ and $\|\mathbf{u}\|_2 = 1$. With $\mathbf{P}_\perp = \mathbf{I} - \mathbf{u}\mathbf{u}^\top$, define

$$c(t) = \mathbf{u}^\top \mathbf{w}(t), \qquad \mathbf{r}(t) = \mathbf{P}_\perp \mathbf{w}(t) = \mathbf{w}(t) - c(t)\mathbf{u}.$$

To see how the two components evolve, recall the flow equation and its fine-tuning force:

$$\dot{\mathbf{w}}(t) = -\nabla\mathcal{L}_\lambda\left(\mathbf{w}(t)\right) = \underbrace{2\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{ft}}}\left[h\left(\mathbf{w}(t)^\top\mathbf{x} - B\right)q(\mathbf{x})\mathbf{x}\right]}_{\mathbf{F}_{\mathrm{ft}}(\mathbf{w}(t))} - \lambda\mathbf{w}(t).$$

Here $h(s) = \sigma(s)\left(1-\sigma(s)\right)^2$ and $q(\mathbf{x}) = \|\mathbf{M}\mathbf{x}\|_2^2$. The vector $\mathbf{u}$ and the matrix $\mathbf{P}_\perp$ are fixed in time, so differentiating the definitions above gives

$$\begin{aligned}\dot{c}(t) &= \mathbf{u}^\top\dot{\mathbf{w}}(t) = \mathbf{u}^\top\mathbf{F}_{\mathrm{ft}}\left(\mathbf{w}(t)\right) - \lambda c(t),\\ \dot{\mathbf{r}}(t) &= \mathbf{P}_\perp\dot{\mathbf{w}}(t) = \mathbf{P}_\perp\mathbf{F}_{\mathrm{ft}}\left(\mathbf{w}(t)\right) - \lambda\mathbf{r}(t).\end{aligned}$$

We first lower bound $\dot{c}(t)$ to control progress along the mean. We will then bound the fine-tuning term in the equation for $\dot{\mathbf{r}}(t)$, so that integrating this equation controls $\|\mathbf{r}(t)\|_2$. Recall $s(t) = \mathbf{w}(t)^\top\mathbf{x} - B$ and $\boldsymbol{\mu}_{\mathrm{ft}} = \rho\mathbf{u}$. Starting from the equation for $\dot{c}(t)$, substituting the definition of the force, and writing $\mathbf{x} = \boldsymbol{\mu}_{\mathrm{ft}} + (\mathbf{x} - \boldsymbol{\mu}_{\mathrm{ft}})$ gives

$$\begin{aligned}\dot{c}(t) &= \mathbf{u}^\top\mathbf{F}_{\mathrm{ft}}\left(\mathbf{w}(t)\right) - \lambda c(t)\\ &= 2\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})\mathbf{u}^\top\mathbf{x}\right] - \lambda c(t)\\ &= 2\rho\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})\right] + 2\mathbf{u}^\top\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ft}})\right] - \lambda c(t).\end{aligned}$$

Here $\mathbf{u}^\top\boldsymbol{\mu}_{\mathrm{ft}} = \rho$, and the fixed vector $\mathbf{u}$ comes outside the expectation. The first term is positive, but the second may be negative. Cauchy–Schwarz gives $|\mathbf{u}^\top\mathbf{a}| \le \|\mathbf{u}\|_2\|\mathbf{a}\|_2 = \|\mathbf{a}\|_2$, hence $\mathbf{u}^\top\mathbf{a} \ge -\|\mathbf{a}\|_2$. Applying this to the vector expectation yields

$$\dot{c}(t) \ge 2\rho\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})\right] - 2\left\|\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ft}})\right]\right\|_2 - \lambda c(t).$$

We now need an upper bound on the norm being subtracted. This is exactly the centered numerator in (40). Dividing by the positive average weight and using (46) gives

$$\begin{aligned}\frac{\left\|\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ft}})\right]\right\|_2}{\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})\right]} &= \left\|\frac{\mathbf{F}_{\mathrm{ft}}\left(\mathbf{w}(t)\right)}{2\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})\right]} - \rho\mathbf{u}\right\|_2\\ &\le \sqrt{\|\boldsymbol{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}} + 3\|\boldsymbol{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}\|\mathbf{w}(t)\|_2.\end{aligned}$$

Multiplying back by the average weight and substituting into the lower bound for $\dot{c}(t)$ gives

$$\dot{c}(t) \ge 2\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})\right]\left(\rho - \sqrt{\|\boldsymbol{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}} - 3\|\boldsymbol{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}\|\mathbf{w}(t)\|_2\right) - \lambda c(t).$$

Next, we bound the fine-tuning contribution to the orthogonal component. Its equation gives $\dot{\mathbf{r}}(t) + \lambda\mathbf{r}(t) = \mathbf{P}_\perp\mathbf{F}_{\mathrm{ft}}(\mathbf{w}(t))$. Substituting the force definition, using $\mathbf{P}_\perp\boldsymbol{\mu}_{\mathrm{ft}} = \mathbf{0}$, and then using that projection cannot increase a norm yields

$$\begin{aligned}\|\dot{\mathbf{r}}(t) + \lambda\mathbf{r}(t)\|_2 = \|\mathbf{P}_\perp\mathbf{F}_{\mathrm{ft}}\left(\mathbf{w}(t)\right)\|_2 &= 2\|\mathbf{P}_\perp\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ft}})\right]\|_2\\ &\le 2\left\|\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ft}})\right]\right\|_2\\ &\le 2\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})\right]\left(\sqrt{\|\boldsymbol{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}} + 3\|\boldsymbol{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}\|\mathbf{w}(t)\|_2\right).\end{aligned}$$

The same error estimate reduces our lower bound on the forward force and increases our upper bound on the orthogonal force. We therefore require it to be small compared with ρ. Choosing it at most $\rho/4$ leaves at least $3\rho/4$ in the forward coefficient and at most $\rho/4$ in the orthogonal one. This gives

a convenient one-to-three comparison; the fraction $1/4$ is not a necessary threshold. We will use the sufficient condition

$$\sqrt{\|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}} + 3\,\|\mathbf{\Sigma}_{\text{ft}}\|_{\text{op}}\,\|\mathbf{w}(t)\|_2 \le \frac{\rho}{4}. \tag{48}$$

For now, work on an interval starting at zero where this condition holds; later we will verify it along the whole trajectory. Substituting into the two bounds just derived gives

$$\dot{c}(t) \ge \frac{3\rho}{2}\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})\right] - \lambda c(t), \qquad \|\mathbf{P}_{\perp}\mathbf{F}_{\text{ft}}\left(\mathbf{w}(t)\right)\|_2 \le \frac{\rho}{2}\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})\right].$$

To turn these force bounds into bounds on $c(t)$ and $\|\mathbf{r}(t)\|_2$, we integrate their flow equations from initialization to time t. First move $-\lambda c(t)$ to the left in the inequality for $\dot{c}(t)$. Multiplication by $e^{\lambda t}$ makes that left-hand side a single derivative, since the product rule gives

$$\frac{d}{dt}\left(e^{\lambda t}c(t)\right) = e^{\lambda t}\left(\dot{c}(t) + \lambda c(t)\right) \ge \frac{3\rho}{2}e^{\lambda t}\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})\right].$$

Integrating from 0 to t, with τ denoting time inside the integral, gives

$$e^{\lambda t}c(t) - c(0) \ge \frac{3\rho}{2}\int_0^t e^{\lambda\tau}\mathbb{E}\left[h\left(s(\tau)\right)q(\mathbf{x})\right]\,d\tau.$$

Multiplying by $e^{-\lambda t} > 0$ therefore yields

$$c(t) \ge e^{-\lambda t}c(0) + \frac{3\rho}{2}\int_0^t e^{-\lambda(t-\tau)}\mathbb{E}\left[h\left(s(\tau)\right)q(\mathbf{x})\right]\,d\tau. \tag{49}$$

For the orthogonal component, we integrate its vector equation directly, so the argument also covers times when $\mathbf{r}(t) = \mathbf{0}$. The same product rule, applied to each coordinate, gives

$$\frac{d}{dt}\left(e^{\lambda t}\mathbf{r}(t)\right) = e^{\lambda t}\left(\dot{\mathbf{r}}(t) + \lambda\mathbf{r}(t)\right) = e^{\lambda t}\mathbf{P}_{\perp}\mathbf{F}_{\text{ft}}\left(\mathbf{w}(t)\right).$$

Integrating and multiplying by $e^{-\lambda t}$ gives the exact expression

$$\mathbf{r}(t) = e^{-\lambda t}\mathbf{r}(0) + \int_0^t e^{-\lambda(t-\tau)}\mathbf{P}_{\perp}\mathbf{F}_{\text{ft}}\left(\mathbf{w}(\tau)\right)\,d\tau.$$

Taking norms, using the triangle inequality $\left\|\int \mathbf{v}(\tau)\,d\tau\right\|_2 \le \int \|\mathbf{v}(\tau)\|_2\,d\tau$, and applying the orthogonal force bound derived above yields

$$\begin{aligned}\|\mathbf{r}(t)\|_2 &\le e^{-\lambda t}\|\mathbf{r}(0)\|_2 + \int_0^t e^{-\lambda(t-\tau)}\|\mathbf{P}_{\perp}\mathbf{F}_{\text{ft}}\left(\mathbf{w}(\tau)\right)\|_2\,d\tau \\ &\le e^{-\lambda t}\|\mathbf{r}(0)\|_2 + \frac{\rho}{2}\int_0^t e^{-\lambda(t-\tau)}\mathbb{E}\left[h\left(s(\tau)\right)q(\mathbf{x})\right]\,d\tau. \end{aligned}\tag{50}$$

The same integral appears in both bounds. Subtracting $e^{-\lambda t}c(0)$ from (49) and dividing by three therefore bounds the integral term in (50):

$$\frac{\rho}{2}\int_0^t e^{-\lambda(t-\tau)}\mathbb{E}\left[h\left(s(\tau)\right)q(\mathbf{x})\right]\,d\tau \le \frac{c(t) - e^{-\lambda t}c(0)}{3}.$$

The right-hand side is nonnegative by the same integrated bound, even if $c(0) < 0$. Substituting this upper bound gives

$$\begin{aligned}\|\mathbf{r}(t)\|_2 &\le e^{-\lambda t}\|\mathbf{r}(0)\|_2 + \frac{c(t) - e^{-\lambda t}c(0)}{3} \\ &= \frac{c(t)}{3} + e^{-\lambda t}\left(\|\mathbf{r}(0)\|_2 - \frac{c(0)}{3}\right).\end{aligned}$$

Let $\varepsilon_0 > 0$ satisfy $\|\mathbf{w}(0)\|_2 \le \varepsilon_0$, so $|c(0)|, \|\mathbf{r}(0)\|_2 \le \varepsilon_0$. The nonnegative integral in (49) gives $c(t) \ge e^{-\lambda t}c(0) \ge -\varepsilon_0$, and $e^{-\lambda t}(\|\mathbf{r}(0)\|_2 - c(0)/3) \le 4\varepsilon_0/3$. Thus

$$\boxed{c(t) \ge -\varepsilon_0, \qquad \|\mathbf{r}(t)\|_2 \le \frac{c(t)}{3} + e^{-\lambda t}\left(\|\mathbf{r}(0)\|_2 - \frac{c(0)}{3}\right) \le \frac{c(t)}{3} + \frac{4\varepsilon_0}{3}.} \tag{51}$$

This comparison holds while (48) holds, also for a random initialization whose realized norm is at most ε_0. It does not imply $\mathbf{r}(t) \to 0$ or guarantee sufficient opening.

C.3.3 LOWER BOUND THE FORCE THAT OPENS THE GATE

Controlling deviation from the mean is not enough: the gate must also open on fine-tuning inputs. We therefore need a lower bound on the positive term driving motion along the mean, so that we can compare it with the opposing regularization term. Return to the bound already obtained:

$$\dot{c}(t) \geq \frac{3\rho}{2}\mathbb{E}\left[h\left(s(t)\right)q(\mathbf{x})\right] - \lambda c(t).$$

To guarantee $\dot{c}(t) > 0$, the positive term must exceed $\lambda c(t)$. We first lower bound its expectation at an arbitrary fixed $\mathbf{w}$. With $s = \mathbf{w}^\top \mathbf{x} - B$, $c = \mathbf{u}^\top \mathbf{w}$, and $\boldsymbol{\mu}_{\mathrm{ft}} = \rho\mathbf{u}$,

$$\mathbb{E}_{\mathrm{ft}}\left[s\right] = \rho c - B, \qquad s - (\rho c - B) = \mathbf{w}^\top\left(\mathbf{x} - \boldsymbol{\mu}_{\mathrm{ft}}\right).$$

Write $\bar{s} = \rho c - B$. We compare $h(s)$ with its positive value at this average score. To bound their ratio, we use its logarithm: the identity $h'(s) = h(s)(1 - 3\sigma(s))$ already proved above gives $(\log h)'(s) = 1 - 3\sigma(s)$, whose absolute value is at most two. Integrating between the two scores yields

$$\log\frac{h(s)}{h(\bar{s})} = \int_{\bar{s}}^{s}\left(1 - 3\sigma(a)\right)\,da \geq -2|s - \bar{s}|.$$

Exponentiating gives $h(s) \geq h(\bar{s})e^{-2|s-\bar{s}|}$. Multiplying by $q(\mathbf{x}) \geq 0$, taking expectations, and substituting the score difference gives

$$\mathbb{E}\left[h(s)q(\mathbf{x})\right] \geq h(\rho c - B)\mathbb{E}\left[q(\mathbf{x})\exp\left(-2\left|\mathbf{w}^\top\left(\mathbf{x} - \boldsymbol{\mu}_{\mathrm{ft}}\right)\right|\right)\right]. \tag{52}$$

The remaining expectation accounts for variation around the average score. Its two factors depend on the same input, so they cannot be separated. Instead, we treat $q(\mathbf{x})$ as a weight and normalize it:

$$\frac{q(\mathbf{x})}{\mathbb{E}\left[q(\mathbf{x})\right]} \geq 0, \qquad \mathbb{E}\left[\frac{q(\mathbf{x})}{\mathbb{E}\left[q(\mathbf{x})\right]}\right] = \frac{\mathbb{E}\left[q(\mathbf{x})\right]}{\mathbb{E}\left[q(\mathbf{x})\right]} = 1.$$

Under these weights, the average score difference is $\mathbb{E}\left[q(\mathbf{x})|s - \bar{s}|\right]/\mathbb{E}\left[q(\mathbf{x})\right]$, finite by Gaussian third moments. Since $a \mapsto e^{-2a}$ has positive second derivative, Jensen's inequality bounds its average below by its value at that weighted average. Applying this inequality and multiplying by $\mathbb{E}\left[q(\mathbf{x})\right] > 0$ gives

$$\mathbb{E}\left[q(\mathbf{x})e^{-2|s-\bar{s}|}\right] \geq \mathbb{E}\left[q(\mathbf{x})\right]\exp\left(-2\frac{\mathbb{E}\left[q(\mathbf{x})|s - \bar{s}|\right]}{\mathbb{E}\left[q(\mathbf{x})\right]}\right). \tag{53}$$

Because the average is multiplied by -2, we need to upper bound it. Cauchy–Schwarz replaces the absolute difference by a squared difference, whose Gaussian moments we can compute. Apply it to $\sqrt{q(\mathbf{x})}$ and $\sqrt{q(\mathbf{x})}|s - \bar{s}|$:

$$\begin{aligned}\mathbb{E}\left[q(\mathbf{x})|s - \bar{s}|\right] &= \mathbb{E}\left[\sqrt{q(\mathbf{x})}\left(\sqrt{q(\mathbf{x})}|s - \bar{s}|\right)\right] \\ &\leq \sqrt{\mathbb{E}\left[q(\mathbf{x})\right]}\sqrt{\mathbb{E}\left[q(\mathbf{x})(s - \bar{s})^2\right]}.\end{aligned}$$

Dividing by $\mathbb{E}\left[q(\mathbf{x})\right] > 0$ and simplifying the square roots yields

$$\frac{\mathbb{E}\left[q(\mathbf{x})|s - \bar{s}|\right]}{\mathbb{E}\left[q(\mathbf{x})\right]} \leq \frac{\sqrt{\mathbb{E}\left[q(\mathbf{x})\right]}\sqrt{\mathbb{E}\left[q(\mathbf{x})(s - \bar{s})^2\right]}}{\mathbb{E}\left[q(\mathbf{x})\right]} = \sqrt{\frac{\mathbb{E}\left[q(\mathbf{x})(s - \bar{s})^2\right]}{\mathbb{E}\left[q(\mathbf{x})\right]}}.$$

Multiplying by -2 reverses the inequality. Substitution into (53) and then (52) therefore gives

$$\mathbb{E}\left[h(s)q(\mathbf{x})\right] \geq h(\rho c - B)\mathbb{E}\left[q(\mathbf{x})\right]\exp\left(-2\sqrt{\frac{\mathbb{E}\left[q(\mathbf{x})(s - \bar{s})^2\right]}{\mathbb{E}\left[q(\mathbf{x})\right]}}\right). \tag{54}$$

To bound the remaining numerator, apply Lemma 7 with $\mathbf{z} = \mathbf{M}\mathbf{x}$ and $Y = s - \bar{s} = \mathbf{w}^\top(\mathbf{x} - \boldsymbol{\mu}_{\mathrm{ft}})$. They are jointly Gaussian because both are affine functions of $\mathbf{x}$. Also, $\|\mathbf{z}\|_2^2 = q(\mathbf{x})$ and $\mathbb{E}\left[Y^2\right] = \mathbf{w}^\top\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}$. The lemma therefore gives

$$\mathbb{E}\left[q(\mathbf{x})(s - \bar{s})^2\right] \leq 3\mathbb{E}\left[q(\mathbf{x})\right]\,\mathbf{w}^\top\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}, \qquad \frac{\mathbb{E}\left[q(\mathbf{x})(s - \bar{s})^2\right]}{\mathbb{E}\left[q(\mathbf{x})\right]} \leq 3\mathbf{w}^\top\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}.$$

Substituting this upper bound into the negative exponent in (54), and recalling the explicit expression for $\mathbb{E}\left[q(\mathbf{x})\right]$, gives the desired lower bound:

$$\boxed{\mathbb{E}\left[h(s)q(\mathbf{x})\right] \geq h(\rho c - B)\left(\mathrm{Tr}\left(\mathbf{Q}\boldsymbol{\Sigma}_{\mathrm{ft}}\right) + \boldsymbol{\mu}_{\mathrm{ft}}^\top\mathbf{Q}\boldsymbol{\mu}_{\mathrm{ft}}\right)\exp\left(-2\sqrt{3\mathbf{w}^\top\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}}\right).} \tag{55}$$

This estimate holds for every $\mathbf{w}$, including singular $\boldsymbol{\Sigma}_{\mathrm{ft}}$, without a separation assumption.

C.3.4 CHOOSE THE REGION AND REGULARIZATION INTERVAL

The opening-force bound depends on the average score and its variance. We now choose a bounded region that controls both and keeps the earlier direction estimate valid. We then find a nonempty regularization interval: the penalty must allow the gate to open, but also prevent the parameters from growing beyond this region.

First, to make the gate close to one at the fine-tuning mean, we ask for $\rho c(t) - B \geq B/2$, which gives $1 - \sigma(\rho c(t) - B) \leq e^{-B/2}$. Solving for $c(t)$ gives the target

$$\rho c(t) - B \geq \frac{B}{2} \quad \Longleftrightarrow \quad c(t) \geq \frac{3B}{2\rho}, \qquad c_{\text{open}} := \frac{3B}{2\rho}.$$

Opening at the mean alone does not guarantee small fine-tuning loss; the score variance must also be controlled. First consider the mean-score factor in (55). For $0 \leq c \leq c_{\text{open}}$, the average score lies in $[-B, B/2]$. For $-B \leq s \leq 0$, we have $1 + e^s \leq 2$, while for $0 \leq s \leq B/2$, we have $1 + e^{-s} \leq 2$. Thus

$$h(s) = \frac{e^s}{(1+e^s)^3} \geq \frac{e^s}{8} \geq \frac{e^{-B}}{8}, \qquad -B \leq s \leq 0,$$
$$h(s) = \frac{e^{-2s}}{(1+e^{-s})^3} \geq \frac{e^{-2s}}{8} \geq \frac{e^{-B}}{8}, \qquad 0 \leq s \leq B/2.$$

Thus $h(\rho c - B) \geq e^{-B}/8$ on this range. To control the score variance, use $\mathbf{w}^\top \boldsymbol{\Sigma}_{\text{ft}} \mathbf{w} \leq \|\boldsymbol{\Sigma}_{\text{ft}}\|_{\text{op}} (c^2 + \|\mathbf{r}\|_2^2)$. We choose the sideways boundary $\|\mathbf{r}\|_2 \leq c/2 + 2\varepsilon_0$ to leave a positive gap beyond (51): for $c \geq -\varepsilon_0$,

$$\left(\frac{c}{2} + 2\varepsilon_0\right) - \left(\frac{c}{3} + \frac{4\varepsilon_0}{3}\right) = \frac{c}{6} + \frac{2\varepsilon_0}{3} \geq \frac{\varepsilon_0}{2} > 0.$$

This keeps the trajectory below the sideways boundary while (48) holds. We still need an upper boundary for c to bound the full norm. It should lie beyond c_{open}, with an opening force small enough for regularization to oppose further growth. Recall that

$$h(s) = \frac{e^{-2s}}{(1+e^{-s})^3} \leq e^{-2s}.$$

At average score B, $h(B) \leq e^{-2B}$, smaller than the order e^{-B} used before the opening target. Solving $\rho c - B = B$ suggests the upper boundary $c = 2B/\rho$. We must still check the actual expected force there, not just its value at the average score. Combining the proposed boundaries gives

$$\mathcal{D} = \left\{ \mathbf{w} = c\mathbf{u} + \mathbf{r} : \ \mathbf{u}^\top \mathbf{r} = 0, \quad -\varepsilon_0 \leq c \leq \frac{2B}{\rho}, \quad \|\mathbf{r}\|_2 \leq \frac{c}{2} + 2\varepsilon_0 \right\}. \tag{56}$$

First we check that it contains the initialization. Since $|c(0)| \leq \varepsilon_0$, the upper boundary includes every allowed initialization if $\varepsilon_0 \leq 2B/\rho$, which we require from now on. The sideways boundary also contains it, because

$$\|\mathbf{r}(0)\|_2 \leq \varepsilon_0 < \frac{3\varepsilon_0}{2} \leq \frac{c(0)}{2} + 2\varepsilon_0.$$

Next consider any fixed $\mathbf{w} \in \mathcal{D}$, without assuming that the trajectory stays there. The restrictions on c and ε_0 imply $|c| \leq 2B/\rho$, while the sideways restriction gives $\|\mathbf{r}\|_2 \leq B/\rho + 2\varepsilon_0$. Orthogonality therefore yields

$$\|\mathbf{w}\|_2^2 = c^2 + \|\mathbf{r}\|_2^2 \leq \frac{4B^2}{\rho^2} + \left(\frac{B}{\rho} + 2\varepsilon_0\right)^2, \qquad \mathbf{w} \in \mathcal{D}. \tag{57}$$

This norm bound does not depend on λ. Substitution into (46) gives a sufficient condition for (48) throughout $\mathcal{D}$:

$$\boxed{\sqrt{\|\boldsymbol{\Sigma}_{\text{ft}}\|_{\text{op}}} + 3\,\|\boldsymbol{\Sigma}_{\text{ft}}\|_{\text{op}} \sqrt{\frac{4B^2}{\rho^2} + \left(\frac{B}{\rho} + 2\varepsilon_0\right)^2} \leq \frac{\rho}{4}.} \tag{58}$$

We still need to prove that the trajectory stays in $\mathcal{D}$. This requires $\dot{c} < 0$ at its upper boundary, while $\dot{c} > 0$ before the opening target; these requirements will determine the allowed λ interval. At every

point of $\mathcal{D}$, the normalized force is within $\rho/4$ of $\rho\mathbf{u}$. Applying both sides of the projection inequality used earlier gives

$$\frac{3\rho}{4} \le \frac{\mathbf{u}^\top \mathbf{F}_{\mathrm{ft}}(\mathbf{w})}{2\mathbb{E}\left[h(s)q(\mathbf{x})\right]} \le \frac{5\rho}{4}, \qquad \frac{3\rho}{2}\mathbb{E}\left[h(s)q(\mathbf{x})\right] - \lambda c \le \dot{c} \le \frac{5\rho}{2}\mathbb{E}\left[h(s)q(\mathbf{x})\right] - \lambda c.$$

For $-\varepsilon_0 \le c \le c_{\mathrm{open}}$, the mean score lies in $[-B - \rho\varepsilon_0, B/2]$. The earlier two cases give $h(\rho c - B) \ge e^{\rho c - B}/8 \ge e^{-B-\rho\varepsilon_0}/8$ when the mean score is nonpositive, and $h(\rho c - B) \ge e^{-B}/8$ otherwise. Thus the weaker bound $e^{-B-\rho\varepsilon_0}/8$ covers both, including negative initial $c(0)$. Writing $v = \mathbf{w}^\top \mathbf{\Sigma}_{\mathrm{ft}}\mathbf{w}$, substituting into (55), and using $-\lambda c \ge -\lambda c_{\mathrm{open}}$ gives

$$\dot{c} \ge \frac{3\rho}{16}\mathbb{E}\left[q(\mathbf{x})\right] \exp\left(-B - \rho\varepsilon_0 - 2\sqrt{3v}\right) - \lambda\frac{3B}{2\rho}, \qquad -\varepsilon_0 \le c \le c_{\mathrm{open}}. \tag{59}$$

Thus an upper restriction on λ will keep this lower bound positive. To find the lower restriction on λ, we next upper bound the expected force. Since $h(s) \le e^{-2s}$, it suffices to bound $\mathbb{E}\left[e^{-2s}q(\mathbf{x})\right]$. We do this for an arbitrary $\mathbf{w}$, then evaluate the bound at $c = 2B/\rho$. Lemma 4, with $\mathbf{a} = -2\mathbf{w}$, $\boldsymbol{\mu} = \boldsymbol{\mu}_{\mathrm{ft}}$, $\mathbf{\Sigma} = \mathbf{\Sigma}_{\mathrm{ft}}$, and $\mathbf{Q} = \mathbf{M}^\top\mathbf{M}$, gives

$$\mathbb{E}\left[e^{-2s}q(\mathbf{x})\right] = e^{2B}\mathbb{E}\left[e^{-2\mathbf{w}^\top\mathbf{x}}\mathbf{x}^\top\mathbf{Q}\mathbf{x}\right] = e^{-2(\rho c - B)+2v}\left[\mathrm{Tr}\left(\mathbf{Q}\mathbf{\Sigma}_{\mathrm{ft}}\right) + \left\|\mathbf{M}\left(\boldsymbol{\mu}_{\mathrm{ft}} - 2\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{w}\right)\right\|_2^2\right].$$

Here $c = \mathbf{u}^\top\mathbf{w}$ and $v = \mathbf{w}^\top\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{w}$. To control the shifted mean, first apply the matrix bound $\|\mathbf{A}\mathbf{b}\|_2 \le \|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{b}\|_2$:

$$\begin{aligned}\left\|\mathbf{M}\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{w}\right\|_2^2 &= \left\|\left(\mathbf{M}\mathbf{\Sigma}_{\mathrm{ft}}^{\frac{1}{2}}\right)\left(\mathbf{\Sigma}_{\mathrm{ft}}^{\frac{1}{2}}\mathbf{w}\right)\right\|_2^2 \\ &\le \left\|\mathbf{M}\mathbf{\Sigma}_{\mathrm{ft}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 \left\|\mathbf{\Sigma}_{\mathrm{ft}}^{\frac{1}{2}}\mathbf{w}\right\|_2^2 = v\,\mathrm{Tr}\left(\mathbf{Q}\mathbf{\Sigma}_{\mathrm{ft}}\right) \le v\mathbb{E}\left[q(\mathbf{x})\right].\end{aligned}$$

The last inequality uses $\mathbb{E}\left[q(\mathbf{x})\right] = \mathrm{Tr}(\mathbf{Q}\mathbf{\Sigma}_{\mathrm{ft}}) + \|\mathbf{M}\boldsymbol{\mu}_{\mathrm{ft}}\|_2^2$. Applying $\|\mathbf{a}+\mathbf{b}\|_2^2 \le 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$ with $\mathbf{a} = \mathbf{M}\boldsymbol{\mu}_{\mathrm{ft}}$ and $\mathbf{b} = -2\mathbf{M}\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{w}$ now gives

$$\begin{aligned}\mathrm{Tr}\left(\mathbf{Q}\mathbf{\Sigma}_{\mathrm{ft}}\right) + \left\|\mathbf{M}\left(\boldsymbol{\mu}_{\mathrm{ft}} - 2\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{w}\right)\right\|_2^2 &\le \mathrm{Tr}\left(\mathbf{Q}\mathbf{\Sigma}_{\mathrm{ft}}\right) + 2\left\|\mathbf{M}\boldsymbol{\mu}_{\mathrm{ft}}\right\|_2^2 + 8\left\|\mathbf{M}\mathbf{\Sigma}_{\mathrm{ft}}\mathbf{w}\right\|_2^2 \\ &\le (2+8v)\mathbb{E}\left[q(\mathbf{x})\right].\end{aligned}$$

Consequently, for every $\mathbf{w}$,

$$\mathbb{E}\left[e^{-2s}q(\mathbf{x})\right] \le (2+8v)\mathbb{E}\left[q(\mathbf{x})\right]e^{-2(\rho c - B)+2v}, \qquad s = \mathbf{w}^\top\mathbf{x} - B. \tag{60}$$

At $c = 2B/\rho$, we have $\rho c - B = B$. Using $h(s) \le e^{-2s}$ and the upper bound on $\dot{c}$ therefore yields

$$\mathbb{E}\left[h(s)q(\mathbf{x})\right] \le (2+8v)\mathbb{E}\left[q(\mathbf{x})\right]e^{-2B+2v}, \qquad \dot{c} \le \frac{5\rho}{2}(2+8v)\mathbb{E}\left[q(\mathbf{x})\right]e^{-2B+2v} - \lambda\frac{2B}{\rho}. \tag{61}$$

The positive terms contain e^{-B} before the target and e^{-2B} at the upper boundary. To preserve this difference, it suffices to keep $v \le B/40$, making the latter exponential at most $e^{-39B/20}$. We also require $\rho\varepsilon_0 \le 1/20$ to limit the initialization contribution. These are sufficient margins, not necessary thresholds. For $B \ge 1$ and $\varepsilon_0 \le 1/(20\rho)$, (57) gives

$$\|\mathbf{w}\|_2^2 \le \frac{4B^2 + (B + \frac{1}{10})^2}{\rho^2} \le \frac{9B^2}{\rho^2}, \qquad v \le \|\mathbf{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}\|\mathbf{w}\|_2^2 \le \frac{9B^2\|\mathbf{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}}{\rho^2}.$$

Here $B + 1/10 \le 2B$, so the numerator in the first bound is at most $8B^2 \le 9B^2$. To make the last expression at most $B/40$, it is sufficient that $9B^2\|\mathbf{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}/\rho^2 \le B/40$, or equivalently $\rho^2 \ge 360B\|\mathbf{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}$. We therefore impose

$$\boxed{B \ge 1, \qquad 0 < \varepsilon_0 \le \frac{1}{20\rho}, \qquad \rho^2 \ge 360B\|\mathbf{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}.} \tag{62}$$

These conditions also imply the earlier force requirement (58): throughout $\mathcal{D}$,

$$\sqrt{\|\mathbf{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}} + 3\|\mathbf{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}\|\mathbf{w}\|_2 \le \frac{\rho}{\sqrt{360B}} + \frac{9B\|\mathbf{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}}{\rho} \le \rho\left(\frac{1}{\sqrt{360B}} + \frac{1}{40}\right) < \frac{\rho}{4}.$$

Thus both projected-force estimates apply throughout this region, and $v \le B/40$ there. In (59), we now have $\rho\varepsilon_0 \le 1/20$ and $2\sqrt{3v} \le \sqrt{3B/10}$. To make the resulting lower bound for $\dot{c}$ positive for every $-\varepsilon_0 \le c \le c_{\text{open}}$, we require

$$\lambda \frac{3B}{2\rho} < \frac{3\rho}{16} \mathbb{E}\left[q(\mathbf{x})\right] \exp\left(-B - \frac{1}{20} - \sqrt{\frac{3B}{10}}\right).$$

Dividing by $3B/(2\rho) > 0$ gives the upper endpoint of the allowed interval:

$$\boxed{\lambda < \lambda_{\max} := \frac{\rho^2}{8B} \mathbb{E}\left[q(\mathbf{x})\right] \exp\left(-B - \frac{1}{20} - \sqrt{\frac{3B}{10}}\right).} \tag{63}$$

In particular, $\dot{c} \ge (\lambda_{\max} - \lambda) c_{\text{open}} > 0$ at every point of $\mathcal{D}$ with $c \le c_{\text{open}}$. At the upper boundary, $2 + 8v \le 2 + B/5$ and $-2B + 2v \le -39B/20$. To make the upper bound in (61) negative, we require

$$\lambda \frac{2B}{\rho} > \frac{5\rho}{2} \left(2 + \frac{B}{5}\right) \mathbb{E}\left[q(\mathbf{x})\right] e^{-\frac{39B}{20}}.$$

Dividing by $2B/\rho > 0$, and using $5(2 + B/5) = B + 10$, gives the lower endpoint:

$$\boxed{\lambda > \lambda_{\min} := \frac{\rho^2 (B+10)}{4B} \mathbb{E}\left[q(\mathbf{x})\right] e^{-\frac{39B}{20}}.} \tag{64}$$

It follows that $\dot{c} \le (\lambda_{\min} - \lambda) 2B/\rho < 0$ at every point of $\mathcal{D}$ with $c = 2B/\rho$. We must still check that the two restrictions on λ are compatible. Both endpoints are positive because $\mathbb{E}\left[q(\mathbf{x})\right] > 0$. Cancelling their common positive factors gives

$$\frac{\lambda_{\min}}{\lambda_{\max}} = 2(B+10) \exp\left(-\frac{19B}{20} + \frac{1}{20} + \sqrt{\frac{3B}{10}}\right).$$

We need this ratio to be less than one. For $B \ge 10$, $\sqrt{3B/10} \le B/5$, $1/20 \le B/20$, and $B + 10 \le 2B$, so

$$\frac{\lambda_{\min}}{\lambda_{\max}} \le 4B \exp\left(-\frac{19B}{20} + \frac{B}{20} + \frac{B}{5}\right) = 4Be^{-\frac{7B}{10}} \le 40e^{-7} < 1.$$

The square-root bound follows by squaring both nonnegative sides; the last bound uses $\frac{d}{dB}(Be^{-7B/10}) = e^{-7B/10}(1 - 7B/10) < 0$ for $B \ge 10$. Thus strengthening the bias condition in (62) to $B \ge 10$ guarantees a nonempty interval:

$$\boxed{B \ge 10, \qquad 0 < \lambda_{\min} < \lambda < \lambda_{\max}.} \tag{65}$$

C.3.5 Keep the trajectory controlled and reach the target

We have bounds inside the proposed region, but have not yet shown that the trajectory stays there. We first rule out escape through each boundary. We then use the positive lower bound on motion along the mean to show that the opening target is reached in finite time and that the trajectory never falls below it afterward.

Initially, $c(0) < 2B/\rho$ and $\|\mathbf{r}(0)\|_2 < c(0)/2 + 2\varepsilon_0$, as checked above. If $c(0) = -\varepsilon_0$, the bound following (63) gives $\dot{c}(0) > 0$, so the trajectory enters the interior immediately. Otherwise it starts there. Suppose it later reaches a boundary, and let $t_{\text{exit}} > 0$ be its first such time. The force and component bounds apply up to this time because (48) holds throughout $\mathcal{D}$. The lower boundary is ruled out by the nonnegative integral in (49):

$$c(t_{\text{exit}}) \ge e^{-\lambda t_{\text{exit}}} c(0) \ge -\varepsilon_0 e^{-\lambda t_{\text{exit}}} > -\varepsilon_0.$$

Strictness follows from $\varepsilon_0, \lambda, t_{\text{exit}} > 0$. At the upper boundary, c would arrive from below, giving $\dot{c}(t_{\text{exit}}) \ge 0$ by taking the limit of $(c(t_{\text{exit}}) - c(t_{\text{exit}} - \delta))/\delta$ as $\delta \downarrow 0$. But the estimate following (64) gives the opposite sign:

$$\dot{c}(t_{\text{exit}}) \le (\lambda_{\min} - \lambda) \frac{2B}{\rho} < 0.$$

Finally, the sideways boundary cannot be reached either. Applying (51) at t_{exit}, and using $c(t_{\text{exit}}) \geq -\varepsilon_0$, gives a strictly positive gap to that boundary:

$$\frac{c(t_{\text{exit}})}{2} + 2\varepsilon_0 - \|\mathbf{r}(t_{\text{exit}})\|_2 \geq \frac{c(t_{\text{exit}})}{6} + \frac{2\varepsilon_0}{3} \geq \frac{\varepsilon_0}{2} > 0.$$

Thus none of the three boundaries can be reached after entering the interior. Since the flow exists for all time, $\mathcal{D}$ contains the whole trajectory, and the norm and variance bounds established there now hold throughout training:

$$\boxed{\mathbf{w}(t) \in \mathcal{D}, \qquad \|\mathbf{w}(t)\|_2 \leq \frac{3B}{\rho}, \qquad \mathbf{w}(t)^\top \mathbf{\Sigma}_{\text{ft}} \mathbf{w}(t) \leq \frac{B}{40}, \qquad t \geq 0.} \tag{66}$$

The component comparison (51) and the bound on $\dot{c}$ now apply for all time in their stated ranges. We next show that c reaches $c_{\text{open}} = 3B/(2\rho)$, where the average score is $B/2$. Since $c(0) \leq \varepsilon_0 < c_{\text{open}}$, define the first reaching time by

$$T := \inf\{t \geq 0 : c(t) \geq c_{\text{open}}\},$$

with $T = \infty$ if the target is never reached. For $0 \leq t < T$, the bound following (63) gives $\dot{c}(t) \geq (\lambda_{\max} - \lambda) c_{\text{open}} > 0$. Integrating, with a denoting time inside the integral, gives

$$c(t) - c(0) = \int_0^t \dot{c}(a)\, da \geq \int_0^t (\lambda_{\max} - \lambda) c_{\text{open}}\, da = (\lambda_{\max} - \lambda) c_{\text{open}} t, \qquad 0 \leq t < T,$$

Since $c(t) < c_{\text{open}}$ before T, this linear lower bound forces $t < (c_{\text{open}} - c(0))/((\lambda_{\max} - \lambda) c_{\text{open}})$ for every $t < T$. Thus $T < \infty$, and continuity gives $c(T) = c_{\text{open}}$. Using $c(0) \geq -\varepsilon_0$ and $\varepsilon_0/c_{\text{open}} \leq 1/(30B)$ yields

$$\boxed{T \leq \frac{c_{\text{open}} - c(0)}{(\lambda_{\max} - \lambda) c_{\text{open}}} \leq \frac{1 + \varepsilon_0/c_{\text{open}}}{\lambda_{\max} - \lambda} \leq \frac{1 + \frac{1}{30B}}{\lambda_{\max} - \lambda}.} \tag{67}$$

Opening also persists: whenever $c = c_{\text{open}}$, we have $\dot{c} \geq (\lambda_{\max} - \lambda) c_{\text{open}} > 0$. Hence c rises above the target just after T. A first return from above would require $\dot{c} \leq 0$, a contradiction. Combining $c(t) \geq c_{\text{open}}$ with the upper boundary of $\mathcal{D}$ and $\mathbb{E}_{\text{ft}}[s(t)] = \rho c(t) - B$ gives

$$\boxed{\frac{3B}{2\rho} \leq c(t) \leq \frac{2B}{\rho}, \qquad \frac{B}{2} \leq \mathbb{E}_{\text{ft}}[s(t)] \leq B, \qquad t \geq T.} \tag{68}$$

C.3.6 Bound adaptation error and forgetting

We now translate the trajectory bounds into the two losses of interest. After time T, the average fine-tuning score is at least $B/2$ and its variance is at most $B/40$, which will give small fine-tuning error. The parameter norm is controlled from the start, allowing us to bound forgetting throughout training under the stated pre-training conditions below.

Recall the fine-tuning loss for the learned gate:

$$\mathcal{L}_{\text{ft}}(\mathbf{w}(t)) = \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim p_{\text{ft}}} \left[(1 - \sigma(s(t)))^2 q(\mathbf{x}) \right], \qquad s(t) = \mathbf{w}(t)^\top \mathbf{x} - B.$$

Fix any $t \geq T$. For every real s, $1 - \sigma(s) = 1/(1 + e^s) \leq e^{-s}$, so $(1 - \sigma(s))^2 \leq e^{-2s}$, including when the individual score is negative. The exponential expectation is exactly the one bounded in (60). Applying that bound with $\mathbf{w} = \mathbf{w}(t)$ and $v = \mathbf{w}(t)^\top \mathbf{\Sigma}_{\text{ft}} \mathbf{w}(t)$ gives

$$\mathcal{L}_{\text{ft}}(\mathbf{w}(t)) \leq \frac{1}{2} \mathbb{E}\left[e^{-2s(t)} q(\mathbf{x}) \right] \leq (1 + 4v) \mathbb{E}[q(\mathbf{x})]\, e^{-2(\rho c(t) - B) + 2v}.$$

Finally, (68) gives $\rho c(t) - B \geq B/2$, and (66) gives $v \leq B/40$. Thus $1 + 4v \leq 1 + B/10$ and $-2(\rho c(t) - B) + 2v \leq -B + B/20 = -19B/20$. Since $t \geq T$ was arbitrary, we obtain

$$\boxed{\sup_{t \geq T} \mathcal{L}_{\text{ft}}(\mathbf{w}(t)) \leq \left(1 + \frac{B}{10}\right) \left(\text{Tr}(\mathbf{Q} \mathbf{\Sigma}_{\text{ft}}) + \|\mathbf{M} \boldsymbol{\mu}_{\text{ft}}\|_2^2 \right) e^{-\frac{19B}{20}}.} \tag{69}$$

The factor multiplying $\mathbb{E}[q(\mathbf{x})]$ decreases exponentially with B. This bounds the loss on the whole fine-tuning population at every time after T, without assuming convergence to an endpoint. We now control the pre-training loss using the all-time gate-norm bound already established. For this step, suppose $p_{\mathrm{pt}} = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathrm{pt}})$, where $0 \preceq \mathbf{\Sigma}_{\mathrm{pt}} \preceq R\mathbf{I}$ and $R > 0$. Recall that the desired correction on pre-training inputs is zero, so

$$\mathcal{L}_{\mathrm{pt}}(\mathbf{w}(t)) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{pt}}}\left[\sigma\left(\mathbf{w}(t)^\top \mathbf{x} - B\right)^2 q(\mathbf{x})\right].$$

Fix any $t \geq 0$. For every real s, $\sigma(s) = e^s/(1+e^s) \leq e^s$, hence $\sigma(s)^2 \leq e^{2s}$. Since $q(\mathbf{x}) \geq 0$, we obtain

$$\mathcal{L}_{\mathrm{pt}}(\mathbf{w}(t)) \leq \frac{e^{-2B}}{2}\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{pt}}}\left[e^{2\mathbf{w}(t)^\top \mathbf{x}}\mathbf{x}^\top \mathbf{Q}\mathbf{x}\right].$$

To evaluate this expectation, apply Lemma 4 with $\mathbf{a} = 2\mathbf{w}(t)$, $\boldsymbol{\mu} = \mathbf{0}$, $\mathbf{\Sigma} = \mathbf{\Sigma}_{\mathrm{pt}}$, and $\mathbf{Q} = \mathbf{M}^\top\mathbf{M}$. The mean is zero, so the lemma gives

$$\mathcal{L}_{\mathrm{pt}}(\mathbf{w}(t)) \leq \frac{1}{2}\exp\left(-2B + 2\mathbf{w}(t)^\top \mathbf{\Sigma}_{\mathrm{pt}}\mathbf{w}(t)\right)\left[\mathrm{Tr}\left(\mathbf{Q}\mathbf{\Sigma}_{\mathrm{pt}}\right) + 4\left\|\mathbf{M}\mathbf{\Sigma}_{\mathrm{pt}}\mathbf{w}(t)\right\|_2^2\right].$$

We next bound both occurrences of the pre-training covariance using $\mathbf{\Sigma}_{\mathrm{pt}} \preceq R\mathbf{I}$. First, $\mathbf{w}(t)^\top \mathbf{\Sigma}_{\mathrm{pt}}\mathbf{w}(t) \leq R\left\|\mathbf{w}(t)\right\|_2^2$. Also, $\mathrm{Tr}(\mathbf{Q}) = \left\|\mathbf{M}\right\|_{\mathrm{F}}^2$, $\left\|\mathbf{\Sigma}_{\mathrm{pt}}\right\|_{\mathrm{op}} \leq R$, and $\left\|\mathbf{M}\right\|_{\mathrm{op}} \leq \left\|\mathbf{M}\right\|_{\mathrm{F}}$, so

$$\mathrm{Tr}\left(\mathbf{Q}\mathbf{\Sigma}_{\mathrm{pt}}\right) \leq R\,\mathrm{Tr}\left(\mathbf{Q}\right) = R\left\|\mathbf{M}\right\|_{\mathrm{F}}^2,$$
$$\left\|\mathbf{M}\mathbf{\Sigma}_{\mathrm{pt}}\mathbf{w}(t)\right\|_2^2 \leq \left\|\mathbf{M}\right\|_{\mathrm{op}}^2\left\|\mathbf{\Sigma}_{\mathrm{pt}}\right\|_{\mathrm{op}}^2\left\|\mathbf{w}(t)\right\|_2^2 \leq R^2\left\|\mathbf{M}\right\|_{\mathrm{F}}^2\left\|\mathbf{w}(t)\right\|_2^2.$$

Substituting these bounds yields

$$\mathcal{L}_{\mathrm{pt}}(\mathbf{w}(t)) \leq \frac{R}{2}\left\|\mathbf{M}\right\|_{\mathrm{F}}^2\left(1 + 4R\left\|\mathbf{w}(t)\right\|_2^2\right)\exp\left(-2B + 2R\left\|\mathbf{w}(t)\right\|_2^2\right).$$

Thus it remains to ensure that the positive term in the exponent does not cancel the negative bias. By (66), $R\left\|\mathbf{w}(t)\right\|_2^2 \leq 9RB^2/\rho^2$ for every $t \geq 0$. As for the fine-tuning variance, it is sufficient to keep this quantity at most $B/40$. Solving $9RB^2/\rho^2 \leq B/40$ gives the additional condition

$$\boxed{\rho^2 \geq 360RB.} \tag{70}$$

Together with the earlier fine-tuning condition, this amounts to $\rho^2 \geq 360B\max\{\left\|\mathbf{\Sigma}_{\mathrm{ft}}\right\|_{\mathrm{op}}, R\}$. We then have $1 + 4R\left\|\mathbf{w}(t)\right\|_2^2 \leq 1 + B/10$ and $-2B + 2R\left\|\mathbf{w}(t)\right\|_2^2 \leq -39B/20$. Since $t \geq 0$ was arbitrary,

$$\boxed{\sup_{t\geq 0}\mathcal{L}_{\mathrm{pt}}(\mathbf{w}(t)) \leq \frac{R}{2}\left(1 + \frac{B}{10}\right)\left\|\mathbf{M}\right\|_{\mathrm{F}}^2 e^{-\frac{39B}{20}}.} \tag{71}$$

For fixed R and $\mathbf{M}$, this bound decreases exponentially with B. No pre-training samples enter the training objective; the pre-training assumptions are used only to bound its population loss. Together with (69), this proves small pre-training loss throughout training and small fine-tuning loss at every time after T.

C.3.7 Prove exponential convergence

The loss bounds do not yet show that the trajectory converges to a single endpoint. We prove this by showing positive curvature in every direction throughout the region reached after T, not throughout the whole parameter space. This will let us identify a stationary point and prove that the distance to it decreases exponentially.

The relevant region is the part of $\mathcal{D}$ where the gate has opened:

$$\mathcal{D}_{\mathrm{open}} = \left\{\mathbf{w} \in \mathcal{D} : \mathbf{u}^\top\mathbf{w} \geq c_{\mathrm{open}}\right\}, \qquad c_{\mathrm{open}} = \frac{3B}{2\rho}.$$

It is closed, bounded, and convex: its restrictions on c are linear, and the triangle inequality preserves $\left\|\mathbf{r}\right\|_2 \leq c/2 + 2\varepsilon_0$ under convex combinations. By (68), $\mathbf{w}(t) \in \mathcal{D}_{\mathrm{open}}$ for every $t \geq T$. Fix any $\mathbf{w}$ in this region. Differentiating the gradient computed at the start of the proof gives

$$\nabla^2\mathcal{L}_\lambda(\mathbf{w}) = \lambda\mathbf{I} - 2\mathbb{E}\left[h(s)\left(1 - 3\sigma(s)\right)q(\mathbf{x})\mathbf{x}\mathbf{x}^\top\right], \qquad s = \mathbf{w}^\top\mathbf{x} - B.$$

This differentiation is valid because h' is bounded and $\mathbb{E}\left[q(\mathbf{x})\left\|\mathbf{x}\right\|_2^2\right] < \infty$. To lower bound the Hessian, take any unit vector $\mathbf{v}$. Only the positive values of $h(s)(1-3\sigma(s))$ can reduce its quadratic form. Writing $[a]_+ = \max\{a, 0\}$, we have

$$\begin{aligned}\mathbf{v}^\top \nabla^2 \mathcal{L}_\lambda(\mathbf{w})\mathbf{v} &= \lambda - 2\mathbb{E}\left[h(s)\left(1-3\sigma(s)\right)q(\mathbf{x})(\mathbf{v}^\top\mathbf{x})^2\right] \\ &\geq \lambda - 2\mathbb{E}\left[\left[h(s)\left(1-3\sigma(s)\right)\right]_+ q(\mathbf{x})(\mathbf{v}^\top\mathbf{x})^2\right].\end{aligned}$$

The positive part is nonzero only if $\sigma(s) < 1/3$, or $s < -\log 2 < 0$, and is at most one. Thus, for any scalar $a > 0$, it is at most e^{-as}: on that negative-score event $e^{-as} \geq 1$, and elsewhere the positive part is zero. Consequently,

$$\mathbf{v}^\top \nabla^2 \mathcal{L}_\lambda(\mathbf{w})\mathbf{v} \geq \lambda - 2\mathbb{E}\left[e^{-as}q(\mathbf{x})(\mathbf{v}^\top\mathbf{x})^2\right], \qquad a > 0. \tag{72}$$

We need to make the subtracted expectation smaller than the positive regularization term. As in the loss estimates, Gaussian square completion handles the exponential. Write $v = \mathbf{w}^\top\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}$. The calculation in the proof of Lemma 4 applies also with the polynomial $q(\mathbf{x})(\mathbf{v}^\top\mathbf{x})^2$ inside the expectation. Explicitly, writing $\mathbf{x} = \boldsymbol{\mu}_{\mathrm{ft}} + \boldsymbol{\Sigma}_{\mathrm{ft}}^{\frac{1}{2}}\mathbf{g}$ with standard Gaussian $\mathbf{g}$, its density exponent becomes

$$-\frac{1}{2}\left\|\mathbf{g}\right\|_2^2 - a\mathbf{w}^\top\boldsymbol{\Sigma}_{\mathrm{ft}}^{\frac{1}{2}}\mathbf{g} = -\frac{1}{2}\left\|\mathbf{g} + a\boldsymbol{\Sigma}_{\mathrm{ft}}^{\frac{1}{2}}\mathbf{w}\right\|_2^2 + \frac{a^2 v}{2}.$$

Therefore, with $\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\mathrm{ft}} - a\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}$ and $\widetilde{\mathbf{x}} \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\mathrm{ft}})$,

$$\mathbb{E}\left[e^{-as}q(\mathbf{x})(\mathbf{v}^\top\mathbf{x})^2\right] = e^{-a(\rho c - B) + \frac{a^2 v}{2}}\mathbb{E}\left[\left\|\mathbf{M}\widetilde{\mathbf{x}}\right\|_2^2(\mathbf{v}^\top\widetilde{\mathbf{x}})^2\right].$$

This argument also covers singular $\boldsymbol{\Sigma}_{\mathrm{ft}}$, because the square completion is in the standard Gaussian vector $\mathbf{g}$. Apply Lemma 7 with $\mathbf{z} = \mathbf{M}\widetilde{\mathbf{x}}$ and $Y = \mathbf{v}^\top\widetilde{\mathbf{x}}$, which are jointly Gaussian. The remaining fourth moment is bounded by

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{M}\widetilde{\mathbf{x}}\right\|_2^2(\mathbf{v}^\top\widetilde{\mathbf{x}})^2\right] &\leq 3\mathbb{E}\left[\left\|\mathbf{M}\widetilde{\mathbf{x}}\right\|_2^2\right]\mathbb{E}\left[(\mathbf{v}^\top\widetilde{\mathbf{x}})^2\right] \\ &= 3\left(\mathrm{Tr}\left(\mathbf{Q}\boldsymbol{\Sigma}_{\mathrm{ft}}\right) + \left\|\mathbf{M}\widetilde{\boldsymbol{\mu}}\right\|_2^2\right)\left(\mathbf{v}^\top\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{v} + (\mathbf{v}^\top\widetilde{\boldsymbol{\mu}})^2\right).\end{aligned}$$

Since $\left\|\mathbf{v}\right\|_2 = 1$, the last factor is at most $\left\|\boldsymbol{\Sigma}_{\mathrm{ft}}\right\|_{\mathrm{op}} + \left\|\widetilde{\boldsymbol{\mu}}\right\|_2^2$. Also, the same squared-norm inequality used in the loss bounds gives

$$\begin{aligned}\mathrm{Tr}\left(\mathbf{Q}\boldsymbol{\Sigma}_{\mathrm{ft}}\right) + \left\|\mathbf{M}\widetilde{\boldsymbol{\mu}}\right\|_2^2 &\leq \mathrm{Tr}\left(\mathbf{Q}\boldsymbol{\Sigma}_{\mathrm{ft}}\right) + 2\left\|\mathbf{M}\boldsymbol{\mu}_{\mathrm{ft}}\right\|_2^2 + 2a^2\left\|\mathbf{M}\boldsymbol{\Sigma}_{\mathrm{ft}}\mathbf{w}\right\|_2^2 \\ &\leq (2 + 2a^2 v)\mathbb{E}\left[q(\mathbf{x})\right], \\ \left\|\widetilde{\boldsymbol{\mu}}\right\|_2 &\leq \rho + a\left\|\boldsymbol{\Sigma}_{\mathrm{ft}}\right\|_{\mathrm{op}}\left\|\mathbf{w}\right\|_2.\end{aligned}$$

The exponential factor is small throughout $\mathcal{D}_{\mathrm{open}}$, because $\rho c - B \geq B/2$ and $v \leq B/40$. In particular,

$$-a(\rho c - B) + \frac{a^2 v}{2} \leq \left(-\frac{a}{2} + \frac{a^2}{80}\right)B.$$

We choose $a = 10$, making this exponent at most $-15B/4$, which decays faster than the factor $e^{-39B/20}$ in $\lambda_{\min}$. This is only a choice in the estimate, not an additional assumption. Using the existing bounds $\left\|\mathbf{w}\right\|_2 \leq 3B/\rho$ and $\left\|\boldsymbol{\Sigma}_{\mathrm{ft}}\right\|_{\mathrm{op}} \leq \rho^2/(360B)$, we obtain

$$2 + 200v \leq 2 + 5B, \quad \left\|\widetilde{\boldsymbol{\mu}}\right\|_2 \leq \rho + \frac{\rho}{12} = \frac{13\rho}{12}, \quad \left\|\boldsymbol{\Sigma}_{\mathrm{ft}}\right\|_{\mathrm{op}} + \left\|\widetilde{\boldsymbol{\mu}}\right\|_2^2 \leq \left(\frac{1}{360B} + \frac{169}{144}\right)\rho^2 \leq 2\rho^2.$$

Substituting these estimates into the fourth-moment bound and then (72) yields

$$\mathbf{v}^\top \nabla^2 \mathcal{L}_\lambda(\mathbf{w})\mathbf{v} \geq \lambda - 12\rho^2(2 + 5B)\mathbb{E}\left[q(\mathbf{x})\right]e^{-\frac{15B}{4}}.$$

To compare the subtracted term with λ, recall $\lambda > \lambda_{\min} = \rho^2(B+10)\mathbb{E}\left[q(\mathbf{x})\right]e^{-39B/20}/(4B)$. Their ratio satisfies

$$\frac{12\rho^2(2+5B)\mathbb{E}\left[q(\mathbf{x})\right]e^{-\frac{15B}{4}}}{\lambda_{\min}} = \frac{48B(2+5B)}{B+10}e^{-\frac{9B}{5}} \leq 288Be^{-\frac{9B}{5}} \leq 2880e^{-18} < \frac{1}{2}.$$

Here $(2+5B)/(B+10) \le 6$, and $Be^{-9B/5}$ decreases for $B \ge 10$. Thus the possible negative curvature is smaller than $\lambda_{\min}/2 < \lambda/2$. Since $\mathbf{v}$ was an arbitrary unit vector,

$$\boxed{\nabla^2 \mathcal{L}_\lambda(\mathbf{w}) \succeq \frac{\lambda}{2}\mathbf{I}, \qquad \mathbf{w} \in \mathcal{D}_{\text{open}}.} \tag{73}$$

This proves strong convexity on the region reached after T, not on the entire parameter space. Before using it to prove convergence, we first identify a stationary point in this region. The descent identity proved at the start, together with $\mathcal{L}_\lambda \ge 0$, gives

$$\int_T^\infty \|\nabla \mathcal{L}_\lambda(\mathbf{w}(t))\|_2^2 \, dt \le \mathcal{L}_\lambda(\mathbf{w}(T)) < \infty.$$

There must therefore be times $t_k \to \infty$ with $\|\nabla \mathcal{L}_\lambda(\mathbf{w}(t_k))\|_2 \to 0$; otherwise this norm would stay bounded away from zero for all sufficiently large times, making the integral infinite. Since $\mathcal{D}_{\text{open}}$ is closed and bounded, a subsequence of $\mathbf{w}(t_k)$ converges to some $\mathbf{w}_\infty \in \mathcal{D}_{\text{open}}$. Continuity of the gradient gives $\nabla \mathcal{L}_\lambda(\mathbf{w}_\infty) = \mathbf{0}$. We now show that the whole trajectory converges to this point. Convexity of $\mathcal{D}_{\text{open}}$ keeps the segment between $\mathbf{w}(t)$ and $\mathbf{w}_\infty$ inside that region. Integrating the Hessian along this segment and using (73) gives

$$\begin{aligned}
&(\mathbf{w}(t) - \mathbf{w}_\infty)^\top \left(\nabla \mathcal{L}_\lambda(\mathbf{w}(t)) - \nabla \mathcal{L}_\lambda(\mathbf{w}_\infty)\right) \\
&\quad = \int_0^1 (\mathbf{w}(t) - \mathbf{w}_\infty)^\top \nabla^2 \mathcal{L}_\lambda \left(\mathbf{w}_\infty + a(\mathbf{w}(t) - \mathbf{w}_\infty)\right) (\mathbf{w}(t) - \mathbf{w}_\infty) \, da \\
&\quad \ge \frac{\lambda}{2} \|\mathbf{w}(t) - \mathbf{w}_\infty\|_2^2 .
\end{aligned}$$

Since $\dot{\mathbf{w}}(t) = -\nabla \mathcal{L}_\lambda(\mathbf{w}(t))$ and $\nabla \mathcal{L}_\lambda(\mathbf{w}_\infty) = \mathbf{0}$,

$$\frac{d}{dt} \|\mathbf{w}(t) - \mathbf{w}_\infty\|_2^2 = -2(\mathbf{w}(t) - \mathbf{w}_\infty)^\top \nabla \mathcal{L}_\lambda(\mathbf{w}(t)) \le -\lambda \|\mathbf{w}(t) - \mathbf{w}_\infty\|_2^2 .$$

Multiplying by $e^{\lambda(t-T)}$ shows that the corresponding weighted squared distance is non-increasing. Evaluating at T and taking square roots yields

$$\boxed{\|\mathbf{w}(t) - \mathbf{w}_\infty\|_2 \le \|\mathbf{w}(T) - \mathbf{w}_\infty\|_2 \, e^{-\frac{\lambda}{2}(t-T)} \le \frac{6B}{\rho} e^{-\frac{\lambda}{2}(t-T)}, \qquad t \ge T.} \tag{74}$$

The last inequality uses the norm bound $3B/\rho$ for both points. Together with the explicit upper bound on T in (67), this controls both the time to reach the open region and convergence thereafter. For example, for any target distance $\delta > 0$, it suffices that

$$t \ge \frac{1 + \frac{1}{30B}}{\lambda_{\max} - \lambda} + \frac{2}{\lambda}\left[\log \frac{6B}{\rho\delta}\right]_+ \quad \Longrightarrow \quad \|\mathbf{w}(t) - \mathbf{w}_\infty\|_2 \le \delta.$$

The same Hessian bound also implies that $\mathbf{w}_\infty$ is the unique stationary point in $\mathcal{D}_{\text{open}}$ and its unique minimizer there. Indeed, integrating the Hessian twice along a segment gives, for any $\mathbf{w} \in \mathcal{D}_{\text{open}}$,

$$\mathcal{L}_\lambda(\mathbf{w}) \ge \mathcal{L}_\lambda(\mathbf{w}_\infty) + \nabla \mathcal{L}_\lambda(\mathbf{w}_\infty)^\top (\mathbf{w} - \mathbf{w}_\infty) + \frac{\lambda}{4} \|\mathbf{w} - \mathbf{w}_\infty\|_2^2 = \mathcal{L}_\lambda(\mathbf{w}_\infty) + \frac{\lambda}{4} \|\mathbf{w} - \mathbf{w}_\infty\|_2^2 .$$

If another point in this region were stationary, the preceding gradient-difference inequality between the two stationary points would force their distance to be zero. These are claims only about $\mathcal{D}_{\text{open}}$; no global optimality outside this region is needed. For completeness, the regularized objective gap also decreases exponentially after T. Applying the same strong-convexity inequality with its expansion based at $\mathbf{w}$, and completing a square in $\mathbf{w} - \mathbf{w}_\infty$, gives

$$\begin{aligned}
0 \le \mathcal{L}_\lambda(\mathbf{w}) - \mathcal{L}_\lambda(\mathbf{w}_\infty) &\le \nabla \mathcal{L}_\lambda(\mathbf{w})^\top (\mathbf{w} - \mathbf{w}_\infty) - \frac{\lambda}{4} \|\mathbf{w} - \mathbf{w}_\infty\|_2^2 \\
&\le \frac{1}{\lambda} \|\nabla \mathcal{L}_\lambda(\mathbf{w})\|_2^2 .
\end{aligned}$$

Along the flow, the derivative of this gap is $-\|\nabla \mathcal{L}_\lambda(\mathbf{w}(t))\|_2^2$, so integration yields

$$0 \le \mathcal{L}_\lambda(\mathbf{w}(t)) - \mathcal{L}_\lambda(\mathbf{w}_\infty) \le \left(\mathcal{L}_\lambda(\mathbf{w}(T)) - \mathcal{L}_\lambda(\mathbf{w}_\infty)\right) e^{-\lambda(t-T)}, \qquad t \ge T. \tag{75}$$

Finally, both population losses are continuous in $\mathbf{w}$: their integrands are continuous and bounded by $q(\mathbf{x})/2$, which has finite expectation under the respective Gaussian laws. Passing to the limit in the already proved loss bounds therefore gives

$$\boxed{\begin{aligned}\mathcal{L}_{\mathrm{ft}}(\mathbf{w}_\infty) &\le \left(1+\frac{B}{10}\right)\left(\mathrm{Tr}\left(\mathbf{Q}\boldsymbol{\Sigma}_{\mathrm{ft}}\right)+\|\mathbf{M}\boldsymbol{\mu}_{\mathrm{ft}}\|_2^2\right)e^{-\frac{19B}{20}},\\ \mathcal{L}_{\mathrm{pt}}(\mathbf{w}_\infty) &\le \sup_{t\ge 0}\mathcal{L}_{\mathrm{pt}}(\mathbf{w}(t)) \le \frac{R}{2}\left(1+\frac{B}{10}\right)\|\mathbf{M}\|_{\mathrm{F}}^2\, e^{-\frac{39B}{20}}.\end{aligned}} \tag{76}$$

Writing $c_\infty = \mathbf{u}^\top \mathbf{w}_\infty$ and $\mathbf{r}_\infty = \mathbf{P}_\perp \mathbf{w}_\infty$, the strict signs of $\dot{c}$ at c_{open} and $2B/\rho$ exclude a stationary endpoint on either boundary. Taking limits in (51) also removes its decaying initialization term. Hence

$$\mathbf{w}_\infty = c_\infty \mathbf{u} + \mathbf{r}_\infty, \qquad \frac{3B}{2\rho} < c_\infty < \frac{2B}{\rho}, \qquad \mathbf{u}^\top \mathbf{r}_\infty = 0, \qquad \|\mathbf{r}_\infty\|_2 \le \frac{c_\infty}{3}.$$

In particular, the proof does not require $\mathbf{r}_\infty = \mathbf{0}$. The conclusions use only $B \ge 10$, $\rho^2 \ge 360B\max\{\|\boldsymbol{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}, R\}$, $0 < \mathbb{E}_{\mathrm{ft}}[q(\mathbf{x})]$, $\lambda \in (\lambda_{\min}, \lambda_{\max})$, and $\|\mathbf{w}(0)\|_2 \le \varepsilon_0 \le 1/(20\rho)$, together with the stated Gaussian populations. No alignment between $\boldsymbol{\mu}_{\mathrm{ft}}$, $\boldsymbol{\Sigma}_{\mathrm{ft}}$, and $\mathbf{Q}$ has been assumed.

C.3.8 Include small random initialization

The proof applies to every initial vector in the stated ball, not only the zero vector. For a random start, it therefore remains to check how likely it is to lie in that ball. We give an explicit Gaussian initialization scale for which all the preceding conclusions hold with high probability.

Let $\mathbf{w}(0) = \omega\mathbf{g}/\sqrt{d_x}$, where $\omega > 0$ and $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x})$, and fix a failure probability $\delta \in (0,1)$. Applying Lemma 5 with $a = \sqrt{2\log(1/\delta)}$ gives, with probability at least $1-\delta$,

$$\|\mathbf{w}(0)\|_2 \le \omega\left(1+\sqrt{\frac{2\log(1/\delta)}{d_x}}\right) \le \omega\sqrt{2+\frac{4}{d_x}\log\frac{1}{\delta}}.$$

The second inequality follows by expanding $(1-\sqrt{2\log(1/\delta)/d_x})^2 \ge 0$. Thus the required norm bound holds if

$$\omega\sqrt{2+\frac{4}{d_x}\log\frac{1}{\delta}} \le \varepsilon_0.$$

On this event, $\|\mathbf{w}(0)\|_2 \le \varepsilon_0$, so the entire deterministic argument applies. This completes the convergence and loss guarantees, including small random initialization. □

D Population GF Experiments

Setup and training. We use $d_x = d_y = 2$, $\mathbf{u} = (3, 1.4)^\top/\sqrt{3^2+1.4^2}$, $p_{\mathrm{ft}} = \mathcal{N}(\rho\mathbf{u}, \boldsymbol{\Sigma}_{\mathrm{ft}})$, and $p_{\mathrm{pt}} = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{pt}})$, with

$$\boldsymbol{\Sigma}_{\mathrm{ft}} = \begin{pmatrix} 1.8 & -0.65 \\ -0.65 & 0.65 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_{\mathrm{pt}} = \begin{pmatrix} 0.7 & 0.22 \\ 0.22 & 1.1 \end{pmatrix}, \qquad \mathbf{M} = \begin{pmatrix} 1.2 & 0.45 \\ -0.35 & 0.8 \end{pmatrix}.$$

The covariances are positive definite and unequal. We fix these matrices and $B = 10$, vary ρ over $1, 1.5, \ldots, 10$, and train only $\mathbf{w}$. Figure 1(a) displays the sweep up to $\rho = 8$. We reuse five draws from $\mathcal{N}(\mathbf{0}, \frac{1}{2}\mathbf{I}_2)$ at every separation. For each ρ, we set $\lambda = \sqrt{\lambda_{\min}\lambda_{\max}}$ using (64) and (63). These runs satisfy the bias condition but are outside the proof's sufficient separation and initialization conditions.

We fix $B = 10$, the theorem's stated threshold, without a performance-based search. The rule for λ, initialization, and stopping use no PT samples, statistics, or losses. All trajectories are saved before PT evaluation begins. The linear and MMSE benchmarks use both populations only for comparison.

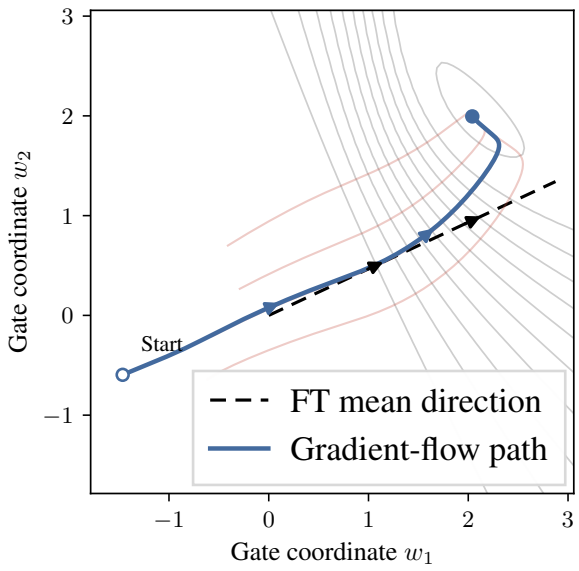


Figure 4: Trajectories at $\rho = 8$ and $B = 10$, with the first random start from Figure 1 highlighted and the other four faint. Gray contours show the regularized FT objective. Black arrows mark the FT mean direction, and blue arrows show the direction of training. The endpoint is numerically stationary and need not lie on the mean line.

Numerical integration. We compute population expectations by conditioning on the scalar score and using split Gauss–Legendre quadrature of order 24, without a sampled training set. We integrate the full two-dimensional flow by adaptive fourth-order Runge–Kutta with relative tolerance $5 \cdot 10^{-10}$, stopping when $\|\nabla \mathcal{L}_\lambda(\mathbf{w})\|_2 / \mathbb{E}_{\mathrm{ft}}[q(\mathbf{x})] < 10^{-10}$. We check selected runs with quadrature order 48 and tighter tolerances. The displayed time window is chosen using FT losses only.

The linear benchmark's losses are computed from second moments, and the MMSE loss is evaluated by numerical integration. Both benchmarks use both populations. Reported losses exclude regularization.

Figure 4 shows why the FT mean is a reference direction rather than a prescribed training path. The gradient also depends on the covariance and the required correction, so the path can bend away from this line. The endpoint still achieves the small FT and PT losses shown in Figure 1.

Effect of initialization size. For the comparison in Figure 5, we keep $B = 10$ and pair each of the five initial vectors above with a copy scaled by 0.01. All other settings, including regularization, agree within each pair.

Figure 5 shows that initialization affects training speed more than final performance in these runs. Both scales reach nearly the same final loss (left), but larger starts can delay the decrease in FT loss (right). In the highlighted pair, the larger start gives a more negative score at the FT mean, keeping the gate more strongly closed and weakening the initial FT learning signal.

Check of the theoretical bounds. We keep the same matrices, set $B = 10$ and $R = \|\mathbf{\Sigma}_{\mathrm{pt}}\|_{\mathrm{op}}$, and choose $\rho = 1.05\sqrt{360B \max\{\|\mathbf{\Sigma}_{\mathrm{ft}}\|_{\mathrm{op}}, R\}} \simeq 91.14$. The regularization rule gives $\lambda \simeq 1.436$, inside the admissible interval $(0.1915, 10.7633)$. We use zero and four draws from $\mathcal{N}(\mathbf{0}, \frac{0.01^2}{2\rho^2}\mathbf{I}_2)$, checking that each satisfies $\|\mathbf{w}(0)\|_2 \le 1/(20\rho)$. All five runs reach the target $c \ge 3B/(2\rho)$ by $1.50 \cdot 10^{-4}$, below the time bound 0.108. The largest recorded PT loss is $2.63 \cdot 10^{-9}$, below its bound $9.79 \cdot 10^{-9}$. The region, force, and post-entry FT-loss bounds also hold at every recorded point.

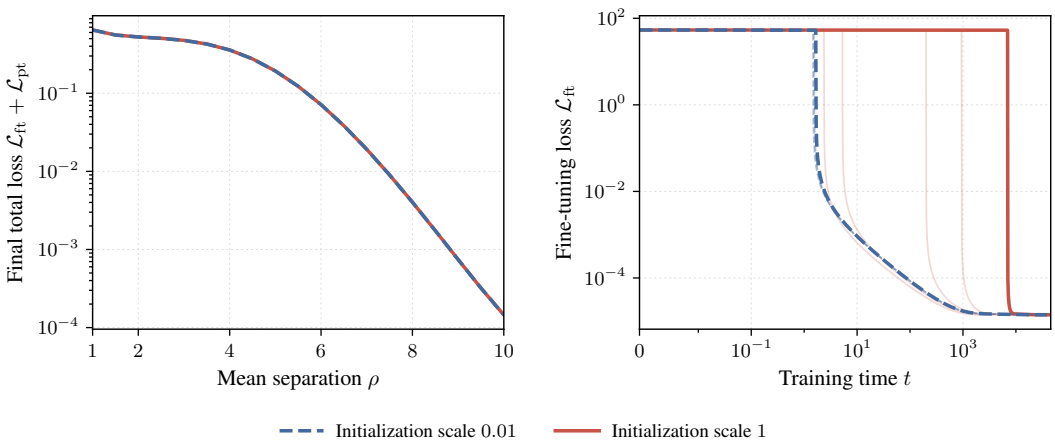


Figure 5: **Effect of Initialization Size.** Five paired starts at $B = 10$, with each smaller start equal to 0.01 times its larger counterpart. **Left:** mean final total loss with shaded ± 1 sample standard deviation across the five starts. **Right:** FT loss over time at $\rho = 8$. The first pair is highlighted, and faint curves show the other four pairs.