Compute-Optimal Pretrain–Fine-tune in Ridge Gradient Descent

Alex Buna
Department of Statistics
University of Oxford, UK

Fanghui Liu
School of Mathematical Sciences
Institute of Natural Sciences and MOE-LSC
Shanghai Jiao Tong University, China

Patrick Rebeschini
Department of Statistics
University of Oxford, UK

Abstract

Pretraining followed by fine-tuning introduces a compute-allocation problem: under a fixed training budget, compute spent improving the upstream objective reduces the compute available for downstream adaptation. Despite its practical importance, the theoretical understanding of this trade-off remains limited, even in simple models. In this paper, we cast this allocation as a compute-split problem under a two-stage pretrain–fine-tune procedure with fixed total optimisation budget, using regularised least squares trained by gradient descent as a tractable setting. We characterise the optimal split under data-dependent evaluation geometries induced by the fine-tuning problem. Our results show that the allocation depends on how pretraining directions affect fine-tuning predictions and how fine-tuning shifts are seen through downstream data geometry. In particular, the relevant quantities are determined by prediction-relevant spectral components of the pretraining and fine-tuning empirical covariances. Technically, the analysis relies on a basis-invariant, eigenspace-level spectral decomposition, together with perturbative control of the non-commuting pretraining and fine-tuning dynamics.

1 Introduction

A central paradigm in modern machine learning is pretraining followed by fine-tuning: a model is first trained on an upstream dataset and subsequently adapted to a downstream task [18, 5]. This two-phase procedure underlies many successful systems in language [6, 28] and vision [18], and fine-tuning may either update all parameters or use parameter-efficient mechanisms such as adapters and LoRA [13, 14]. Computationally, however, the two stages are coupled: for a fixed training budget, compute spent improving the upstream objective is compute not spent adapting to the downstream task.

This trade-off is procedural rather than purely statistical. Existing theory of pretraining and fine-tuning largely asks when upstream data helps downstream prediction, how task similarity affects transfer, or how pretraining changes generalisation [24, 8, 29]. We instead fix the two empirical objectives and ask how a finite optimisation budget should be allocated between them. A brute-force approach would save multiple pretraining checkpoints, fine-tune each one, and select the best downstream result; a similar intermediate-checkpoint fine-tuning approach is used in recent empirical studies to probe the relation between pretraining and fine-tuning [26]. Such a sweep is informative but expensive, since

Email addresses of corresponding authors: Alex Buna (alex.bunamarginean@spc.ox.ac.uk), Fanghui Liu (fanghui.liu@sjtu.edu.cn), Patrick Rebeschini (patrick.rebeschini@stats.ox.ac.uk).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

every candidate checkpoint requires a separate downstream training run. Our objective is to examine the allocation problem within a paradigmatic, analytically tractable framework, and to identify which upstream–downstream geometric quantities determine the optimal split.

To this end, we study this trade-off in ridge-regularised linear regression: suppose the upstream and downstream data are $(\boldsymbol{X}, \boldsymbol{y}) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$ and $(\boldsymbol{X}', \boldsymbol{y}') \in \mathbb{R}^{n'\times d} \times \mathbb{R}^{n'}$, respectively. For any ridge parameters $\gamma, \gamma' > 0$, we consider the pretraining $R_{\mathrm{pre}}(\boldsymbol{w}) = \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \frac{\gamma}{2}\|\boldsymbol{w}\|_2^2$ and fine-tuning $R_{\mathrm{fine}}(\boldsymbol{w}) = \frac{1}{2n'}\|\boldsymbol{X}'\boldsymbol{w} - \boldsymbol{y}'\|_2^2 + \frac{\gamma'}{2}\|\boldsymbol{w}\|_2^2$ empirical objectives, with corresponding minimisers $\widehat{\boldsymbol{w}}$ and $\widehat{\boldsymbol{w}}'$, respectively. For fixed per-iteration costs $c_{\mathrm{p}} > 0$ and $c_{\mathrm{f}} > 0$ of pretraining and fine-tuning, respectively, running gradient descent (GD) on R_{pre} for k iterations and then on R_{fine} for an additional ℓ iterations incurs a computational cost of $c_{\mathrm{p}}k + c_{\mathrm{f}}\ell$, which we constrain to equal a fixed total compute budget B. This gives the computational question studied in this paper:

Given fixed total compute budget B and per-iteration costs c_{p} , c_{f} ,
when should we switch from pretraining to fine-tuning during optimisation?

To avoid non-integrality issues, we assume that, after choosing a sufficiently fine compute unit, c_{p} and c_{f} are positive integers. For simplicity, we focus on budgets B divisible by both c_{p} and c_{f}, so that $k = 0$ and $\ell = 0$ are feasible allocations. Thus, there exists a positive integer T such that

$$B = TL, \quad \text{where } L := \mathrm{lcm}(c_{\mathrm{p}}, c_{\mathrm{f}}),$$

where $\mathrm{lcm}(\cdot,\cdot)$ denotes the least common multiple. Then, every exact-budget schedule (k, ℓ) with $c_{\mathrm{p}}k + c_{\mathrm{f}}\ell = B$ is uniquely represented by a feasible pretraining-compute allocation

$$t \in [T] := \{0, 1, 2, \ldots, T\}, \quad \text{with} \quad (k(t), \ell(t)) = \left(\frac{L}{c_{\mathrm{p}}}t, \frac{L}{c_{\mathrm{f}}}(T - t)\right). \tag{1}$$

Many practical adaptation methods are explicitly represented as a task-specific update on top of a shared base model: adapters, LoRA, delta tuning, and task vectors all adopt this viewpoint [13, 14, 7, 15]. This motivates using

$$\boldsymbol{\Delta} := \widehat{\boldsymbol{w}}' - \widehat{\boldsymbol{w}}$$

as the schedule-independent transfer update from the pretraining optimum to the fine-tuning optimum, and as a natural objective for revealing the trade-off we study in this paper.

A split $(k(t), \ell(t))$ with $t \in [T]$ does not produce $\boldsymbol{\Delta}$ exactly. It first reaches after k iterations the pretraining checkpoint $\boldsymbol{w}_{k(t)}$ and then fine-tunes from that checkpoint for ℓ iterations, reaching $\boldsymbol{w}'_{\ell(t)}$. The update induced by this finite-budget two-stage algorithm is therefore

$$\boldsymbol{\Delta}_t := \boldsymbol{w}'_{\ell(t)} - \boldsymbol{w}_{k(t)}.$$

We aim to identify the allocation t that allows $\boldsymbol{\Delta}_t$ to best approximate the benchmark update $\boldsymbol{\Delta}$.

Necessity of a fixed anchor: We use $\widehat{\boldsymbol{w}}$ as an anchor in order to compare all allocations against the same transfer update $\boldsymbol{\Delta}$. If the checkpoint $\boldsymbol{w}_{k(t)}$ were used as the anchor instead, the ideal update would be $\widehat{\boldsymbol{w}}' - \boldsymbol{w}_{k(t)}$, and the mismatch reduces to $\boldsymbol{\Delta}_t - (\widehat{\boldsymbol{w}}' - \boldsymbol{w}_{k(t)}) = \boldsymbol{w}'_{\ell(t)} - \widehat{\boldsymbol{w}}'$, measuring only how close the fine-tuning iterate is to the fine-tuning optimum. It would no longer compare the quality of the transfer update across different pretrain–fine-tune splits. Anchoring at $\widehat{\boldsymbol{w}}$ thus gives a fixed comparator that is independent of the split.

Once updates are anchored at $\widehat{\boldsymbol{w}}$, the natural way to evaluate a split $(k(t), \ell(t))$ is through the empirical fine-tuning risk of the transferred model $\widehat{\boldsymbol{w}} + \boldsymbol{\Delta}_t$, which can be expressed (see Lemma 1) as

$$R_{\mathrm{fine}}(\widehat{\boldsymbol{w}} + \boldsymbol{\Delta}_t) - R_{\mathrm{fine}}(\widehat{\boldsymbol{w}}') = \frac{1}{2}\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|_{\widehat{\boldsymbol{\Sigma}}'}^2 + \frac{\gamma'}{2}\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|_2^2 = \frac{1}{2}\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|_{\widehat{\boldsymbol{\Sigma}}' + \gamma'\boldsymbol{I}}^2 ,$$

where $\widehat{\boldsymbol{\Sigma}}' = \frac{1}{n'}(\boldsymbol{X}')^\top \boldsymbol{X}' \in \mathbb{R}^{d\times d}$ denotes the downstream empirical covariance. This identity motivates evaluating the *update mismatch* $\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|_{\mathtt{M}}^2$ under two different geometries $\mathtt{M} \in \{\boldsymbol{H}', \widehat{\boldsymbol{\Sigma}}'\}$, where $\boldsymbol{H}' := \widehat{\boldsymbol{\Sigma}}' + \gamma'\boldsymbol{I}$. The *norm* $\|\cdot\|_{\boldsymbol{H}'}$, induced by the positive definite fine-tuning Hessian $\boldsymbol{H}'$, exactly controls the empirical excess risk above. The *seminorm* $\|\cdot\|_{\widehat{\boldsymbol{\Sigma}}'}$ ignores directions in $\ker(\boldsymbol{X}')$, which are unidentifiable from the downstream sample and do not affect its empirical predictions. Consequently, we formalise the allocation problem as the minimisation of

$$f_{\mathtt{M}} : [T] \to [0, \infty), \quad f_{\mathtt{M}}(t) = \|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|_{\mathtt{M}}^2, \quad \mathtt{M} \in \{\boldsymbol{H}', \widehat{\boldsymbol{\Sigma}}'\}.$$

1.1 Our contributions and technical challenges

Localisation of the optimal split. Our main results (Theorem 6 and 12) show that, for the reparametrisation (1), for sufficiently large T, any global minimiser t^* of $f_{\mathtt{M}}$ satisfies

$$|t^* - (s^* + \theta^* T)| \le 1 + O(e^{-\delta T}), \qquad \text{where} \quad \theta^* := \frac{\frac{1}{c_{\mathrm{f}}} \ln(1 - \eta' \lambda'_{j^*})}{\frac{1}{c_{\mathrm{p}}} \ln(1 - \eta \lambda_{i^*}) + \frac{1}{c_{\mathrm{f}}} \ln(1 - \eta' \lambda'_{j^*})}. \quad (2)$$

The *leading fraction* θ^* depends on the step-sizes $\eta, \eta' > 0$ of GD in each of the two phases, the normalisation with respect to the two costs $c_{\mathrm{p}}, c_{\mathrm{f}}$, and two active bottlenecks $\lambda_{i^*}, \lambda'_{j^*}$. The pretraining bottleneck λ_{i^*} is the smallest eigenvalue among those whose upstream eigenspaces have a non-zero (visible) projection of $\widehat{\boldsymbol{w}}$ in $\mathtt{M}$-geometry, while the fine-tuning bottleneck λ'_{j^*} is the smallest eigenvalue among those whose downstream eigenspaces have a non-zero (visible) projection of $\boldsymbol{\Delta}$ in $\mathtt{M}$-geometry. The offset s^* is the next-order, T-independent correction. It depends on the scale and alignment of these smallest upstream and downstream visible projections. Furthermore, in Corollary 9 we obtain the corresponding expressions for the optimal split

$$(k(t^*), \ell(t^*)) \approx \left(\theta^* \frac{B}{c_{\mathrm{p}}}, (1 - \theta^*) \frac{B}{c_{\mathrm{f}}} \right).$$

Thus, the same upstream task can have different optimal allocations for different downstream tasks, and vice versa. This is illustrated in Fig. 1, where changing only the sample sizes moves the predicted optimal compute allocation from early to late pretraining.

Complementing these results, Theorem 7 shows that any fixed interior fraction $\widehat{\theta} \in (0, 1)$ gives an update mismatch exponentially smaller in T than allocations with only a constant amount of pretraining or fine-tuning, and the guaranteed improvement is maximised at $\widehat{\theta} = \theta^*$. Importantly, computing the offset s^* is not required for this exponential improvement, and this motivates the estimation approach in Section 3.4, also illustrated in Fig. 1. While evaluating $\theta^* T + s^*$ requires solution-level information, the leading fraction θ^* depends only on λ_{i^*} and λ'_{j^*}. We estimate these quantities using matrix-free Lanczos iterations, without computing the ridge solutions or full spectral decompositions. With a fixed number of Lanczos iterations, estimating these bottlenecks costs $O((n + n')d)$. This provides a potential alternative to a checkpoint sweep over many candidate splits.

An eigenspace-level perturbation analysis for non-commuting objectives. Technically, the proof uses spectral projectors associated with distinct eigenvalues, rather than an arbitrary eigenvector basis. This makes the active sets basis-invariant when eigenvalues have multiplicity and allows us to control the non-commuting cross-terms between the pretraining and fine-tuning ridge Hessians. Under the downstream prediction seminorm $\|\cdot\|_{\widehat{\boldsymbol{\Sigma}}'}$, the result requires a 'no-kernel-hitting' condition because $\widehat{\boldsymbol{\Sigma}}'$ can vanish on non-predictive directions. Proposition 13 shows why such a condition is necessary, while Propositions 14 and 15 give common settings where it holds.

1.2 Related work

Statistical theory of pretraining and fine-tuning. Most theoretical work on pretraining and fine-tuning studies statistical transfer: when upstream data improves downstream prediction, how task similarity affects sample complexity, or how covariate shift changes the value of pretraining [24, 8, 29]. The closest works to ours are linear-model analyses. In linear-teacher settings, Shachaf et al. [24] show that the relevant similarity notion depends on the source task, target task, and covariance structure of the target data. For linear regression under covariate shift, Wu et al. [29] derive instance-dependent guarantees for pretraining followed by fine-tuning with online SGD. These works identify when upstream data can statistically help downstream prediction. Our work is complementary: we fix the two empirical objectives and ask where a finite optimisation budget should switch from one objective to the other.

Compute allocation. Our question is also connected to compute-optimal training. Scaling-law works study how finite compute should be allocated across axes such as model size, data, and optimisation time [3, 12, 20]. Transfer settings introduce another allocation axis: upstream training versus downstream adaptation. For instance, for protein language models, Cheng et al. [4] find a non-trivial switching trade-off, allocating approximately 10–20% of the fixed compute budget to causal language modelling before switching to masked language modelling. Other empirical

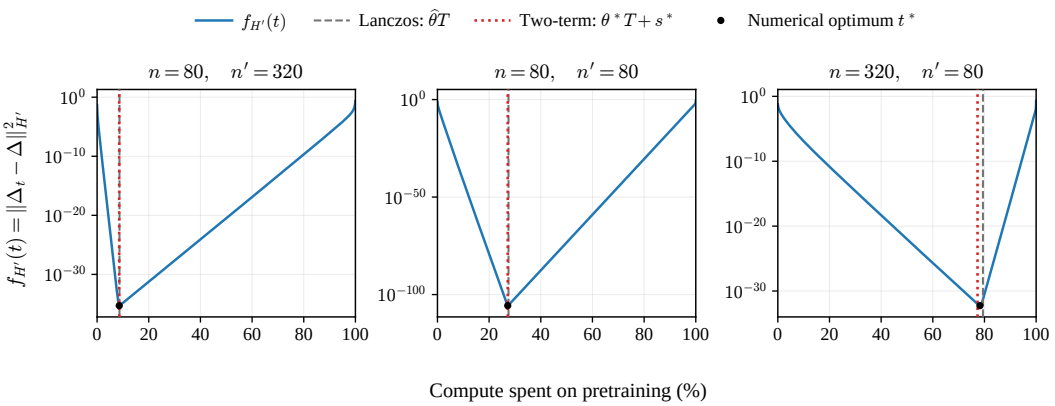

Figure 1: We generate random-design linear regression examples in dimension $d = 400$, with $(n, n') = (80, 320), (80, 80), (320, 80)$ pretraining and fine-tuning samples, respectively (left/middle/right). For each feasible allocation $t \in \{0, \dots, T\}$, we evaluate the closed-form full-batch GD dynamics for $k(t)$ pretraining iterations followed by $\ell(t)$ fine-tuning iterations, as in representation (1). Let $\boldsymbol{\Delta}_t$ be the algorithm-induced update and $\boldsymbol{\Delta} = \widehat{\boldsymbol{w}}' - \widehat{\boldsymbol{w}}$ the ideal transfer update. The blue curve shows $f_{\boldsymbol{H}'}(t) = \|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|^2_{\boldsymbol{H}'}$ against the percentage of compute spent on pretraining, $100t/T$. The panels illustrate how sample sizes and the resulting empirical geometry affect the optimal allocation (c.f. Remark 11). The black dot marks the numerical optimum found by evaluating every feasible allocation. The red dotted line marks the predicted $\theta^* T + s^*$ in (2), computed using exact spectral and solution-level quantities according to Theorem 6. The grey dashed line marks the estimated leading prediction $\widehat{\theta}T$, based on Lanczos iterations and described in Section 3.4. This estimate $\widehat{\theta}$ is computed without the ridge solutions or full spectral decompositions. For reproducibility, further experimental details have been included in Appendix D.

transfer-scaling studies show that downstream behaviour depends on the amount and alignment of pretraining and fine-tuning data [11, 16], and LLM fine-tuning studies find that performance depends on model size, pretraining and fine-tuning data, and the fine-tuning method [31]. Such observations motivate a theory in which the optimal split depends on both upstream and downstream geometry, rather than on upstream loss alone.

Checkpoint selection and update-based adaptation. A practical way to study the pretrain–fine-tune interface is to fine-tune several intermediate pretraining checkpoints. The work of [26] takes this approach for LLMs and shows that downstream behaviour after supervised fine-tuning can reveal effects of pretraining that are not apparent from the pretraining checkpoint alone. Our ridge model abstracts this situation into a discrete compute-allocation problem. The update-based formulation is also motivated by parameter-efficient and arithmetic views of adaptation: adapters and LoRA learn small task-specific modifications [13, 14], delta tuning frames adaptation as a parameter delta [7], and task arithmetic manipulates differences between base and fine-tuned weights [15]. This motivates evaluating the quality of the finite-budget update $\boldsymbol{\Delta}_t$, not only the final fine-tuned iterate.

Ridge dynamics. Ridge-regularised least squares yields closed-form GD trajectories through the spectrum of the empirical covariance [10, 2]. Spectral analyses have been used to characterise mode-wise learning dynamics in linear and deep-linear models [23, 27]. Related work connects optimisation paths to explicit ℓ_2 regularisation and spectral filtering [21, 9, 25, 32, 1, 30]. Recent advances also show that fixed-feature models, similar to the setting under consideration, can remain informative beyond that formal setting via spectral predictions. For instance, Li et al. [19] derive a functional scaling law under power-law spectral assumptions and verify empirically that it predicts LLM loss trajectories under unseen schedules; and Karkada et al. [17] empirically show that covariance-derived kernel eigenvalues predict the order in which feature-learning MLPs learn Hermite-polynomial targets.

1.3 Notation and organisation of the paper

For an integer $m \geq 0$, let $[m] := \{0, 1, \dots, m\}$, with the convention that $[-1] = \emptyset$. We denote the zero vector by $\mathbf{0}$ and the identity matrix by $\boldsymbol{I}$. For vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and a symmetric positive

semidefinite matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, we use the notations $\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\boldsymbol{A}} = \boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{v}$ and $\|\boldsymbol{u}\|_A = \langle \boldsymbol{u}, \boldsymbol{u}\rangle_{\boldsymbol{A}}^{1/2}$. The Euclidean and operator norms will be denoted by $\|\boldsymbol{v}\|_2$ and $\|\boldsymbol{A}\|_2$, respectively.

Section 2 formally introduces the model, algorithm, and evaluation metric. Section 3 presents our main theoretical findings on the localisation of the optimal allocation, with a proof sketch of one of these results in Appendix B. All missing proofs, deferred to Appendix C, rely on additional notation in Appendix A. Experimental details for Fig. 1 can be found in Appendix D.

2 Problem settings

This section presents our framework: ridge regression optimised by GD, and the evaluation metrics.

Ridge regression and gradient descent. Let $(\boldsymbol{X}, \boldsymbol{y}) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$ and $(\boldsymbol{X}', \boldsymbol{y}') \in \mathbb{R}^{n'\times d} \times \mathbb{R}^{n'}$ denote the pretraining and fine-tuning datasets, respectively. For fixed ridge parameters $\gamma, \gamma' > 0$, define the ridge-regularised empirical risks

$$R_{\mathrm{pre}}(\boldsymbol{w}) = \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \frac{\gamma}{2}\|\boldsymbol{w}\|_2^2, \quad R_{\mathrm{fine}}(\boldsymbol{w}) = \frac{1}{2n'}\|\boldsymbol{X}'\boldsymbol{w} - \boldsymbol{y}'\|_2^2 + \frac{\gamma'}{2}\|\boldsymbol{w}\|_2^2.$$

Consider the associated empirical covariances and positive definite ridge Hessians

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}, \quad \boldsymbol{H} = \widehat{\boldsymbol{\Sigma}} + \gamma\boldsymbol{I}, \quad \widehat{\boldsymbol{\Sigma}}' = \frac{1}{n'}\boldsymbol{X}'^\top\boldsymbol{X}', \quad \boldsymbol{H}' = \widehat{\boldsymbol{\Sigma}}' + \gamma'\boldsymbol{I}.$$

The pretraining and fine-tuning unique global minimisers are given by

$$\widehat{\boldsymbol{w}} := \arg\min_{\boldsymbol{w}} R_{\mathrm{pre}}(\boldsymbol{w}) = \frac{1}{n}\boldsymbol{H}^{-1}\boldsymbol{X}^\top\boldsymbol{y}, \quad \widehat{\boldsymbol{w}}' := \arg\min_{\boldsymbol{w}} R_{\mathrm{fine}}(\boldsymbol{w}) = \frac{1}{n'}(\boldsymbol{H}')^{-1}(\boldsymbol{X}')^\top\boldsymbol{y}'.$$

Recall that we aim to recover the displacement $\boldsymbol{\Delta} = \widehat{\boldsymbol{w}}' - \widehat{\boldsymbol{w}}$. For this, consider a fixed computational budget $B > 0$ and step-sizes $0 < \eta < \|\boldsymbol{H}\|_2^{-1}$, $0 < \eta' < \|\boldsymbol{H}'\|_2^{-1}$. For any split (k, ℓ) with $c_{\mathrm{p}}k + c_{\mathrm{f}}\ell = B$, define the following two-phase GD procedure:

Pretraining phase: Initialise $\boldsymbol{w}_0 = \boldsymbol{0}$ and run $\boldsymbol{w}_{s+1} = \boldsymbol{w}_s - \eta\nabla R_{\mathrm{pre}}(\boldsymbol{w}_s)$ for $s \in [k-1]$.

Fine-tuning phase: Initialise $\boldsymbol{w}'_0 = \boldsymbol{w}_k$ and run $\boldsymbol{w}'_{s+1} = \boldsymbol{w}'_s - \eta'\nabla R_{\mathrm{fine}}(\boldsymbol{w}'_s)$ for $s \in [\ell-1]$.

Accordingly, under the representation (1), a natural estimator for $\boldsymbol{\Delta}$ is the update $\boldsymbol{\Delta}_t = \boldsymbol{w}'_{\ell(t)} - \boldsymbol{w}_{k(t)}$.

Evaluation metrics. Recalling the discussion from Section 1 on the necessity of a canonical anchor, once one evaluates such an update anchored at $\widehat{\boldsymbol{w}}$, the natural performance criterion is the empirical fine-tuning risk of the transferred model $\widehat{\boldsymbol{w}} + \boldsymbol{\Delta}_t$. In ridge linear regression, this connection is exact:

Lemma 1. *The following fine-tuning empirical excess risk identity holds:*

$$R_{\mathrm{fine}}(\widehat{\boldsymbol{w}} + \boldsymbol{\Delta}_t) - R_{\mathrm{fine}}(\widehat{\boldsymbol{w}}') = \frac{1}{2}\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|_{\widehat{\boldsymbol{\Sigma}}'}^2 + \frac{\gamma'}{2}\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|_2^2 = \frac{1}{2}\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|_{\boldsymbol{H}'}^2.$$

One goal of our work is to understand how one can choose t so that the anchored update fine-tuning risk $R_{\mathrm{fine}}(\widehat{\boldsymbol{w}} + \boldsymbol{\Delta}_t)$ is as close as possible to $R_{\mathrm{fine}}(\widehat{\boldsymbol{w}}')$. With the viewpoint of anchored update given by Lemma 1, we measure the quality of this approximation via a parameter-space distance

$$f_{\boldsymbol{H}'}(t) := \|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|_{\boldsymbol{H}'}^2 \quad \text{and} \quad f_{\widehat{\boldsymbol{\Sigma}}'}(t) := \|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|_{\widehat{\boldsymbol{\Sigma}}'}^2 = \frac{1}{n'}\|\boldsymbol{X}'(\boldsymbol{\Delta}_t - \boldsymbol{\Delta})\|_2^2.$$

In general, for $\mathtt{M} \in \{\boldsymbol{H}', \widehat{\boldsymbol{\Sigma}}'\}$ we write $f_{\mathtt{M}}(t) = \|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|_{\mathtt{M}}^2$. The set of minimisers

$$\mathcal{T}^* := \arg\min_{t\in[T]} f_{\mathtt{M}}(t)$$

represents the optimal allocation of compute between pretraining and fine-tuning that best approximates the ideal transfer shift under the fine-tuning geometry or the $\boldsymbol{H}'$-induced geometry.

3 Localisation of the optimal pretrain/fine-tune split

Before delivering our results, we give an alternative formulation for $f_{\mathtt{M}}$ involving $\widehat{\boldsymbol{w}}$, $\boldsymbol{\Delta}$, $\boldsymbol{H}$, $\boldsymbol{H}'$:

Lemma 2. *For every split $(k(t), \ell(t))$, the pretraining and fine-tuning GD iterates satisfy*

$$\boldsymbol{w}_{k(t)} = (\boldsymbol{I} - (\boldsymbol{I} - \eta \boldsymbol{H})^{k(t)})\widehat{\boldsymbol{w}}, \quad \boldsymbol{w}'_{\ell(t)} = \widehat{\boldsymbol{w}}' + (\boldsymbol{I} - \eta' \boldsymbol{H}')^{\ell(t)}(\boldsymbol{w}_{k(t)} - \widehat{\boldsymbol{w}}')\,.$$

Consequently, under $\|\cdot\|_{\mathtt{M}}$ (using either $\|\cdot\|_{\boldsymbol{H}'}$ or $\|\cdot\|_{\widehat{\boldsymbol{\Sigma}}'}$), we have

$$f_{\mathtt{M}}(t) = \|(\boldsymbol{I} - (\boldsymbol{I} - \eta' \boldsymbol{H}')^{\ell(t)})(\boldsymbol{I} - \eta \boldsymbol{H})^{k(t)}\widehat{\boldsymbol{w}} - (\boldsymbol{I} - \eta' \boldsymbol{H}')^{\ell(t)}\boldsymbol{\Delta}\|_{\mathtt{M}}^2.$$

Note that, due to the step-sizes choice, the matrices $\boldsymbol{I} - \eta \boldsymbol{H}$ and $\boldsymbol{I} - \eta' \boldsymbol{H}'$ are positive definite.

Remark 3. Let us consider the extreme cases: $t = 0$ represents no pretraining, so $\boldsymbol{\Delta}_0 = \boldsymbol{w}'_{\ell(0)}$ and $f_{\mathtt{M}}(0) = \|\widehat{\boldsymbol{w}} - (\boldsymbol{I} - \eta' \boldsymbol{H}')^{B/c_{\mathrm{f}}}\widehat{\boldsymbol{w}}'\|_{\mathtt{M}}^2$; $t = T$ corresponds to no fine-tuning, so $\boldsymbol{\Delta}_T = 0$ and $f_{\mathtt{M}}(T) = \|\boldsymbol{\Delta}\|_{\mathtt{M}}^2$. While these two quantities are not immediately comparable, we prove in Theorem 7 that, at the global minima identified in Theorem 6, the value of $f_{\boldsymbol{H}'}$ represents an exponential improvement over very little pretraining (small k) or fine-tuning (small ℓ).

3.1 Warm-up with special cases of $t^* = 0$ or $t^* = T$

As a starting point, in this section we discuss three degenerate cases that arise due to the interplay between $\widehat{\boldsymbol{w}}$, $\boldsymbol{\Delta}$ and the spectral structure of the positive definite Hessians $\boldsymbol{H}$ and $\boldsymbol{H}'$.

Let $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_r$ be the *distinct* eigenvalues of $\boldsymbol{H}$, with corresponding orthogonal eigenspace projectors $\{\boldsymbol{P}_i\}_{i=1}^r$. Thus, if $\boldsymbol{U}_i$ contains as columns an orthonormal basis for the eigenspace associated with λ_i, then $\boldsymbol{P}_i = \boldsymbol{U}_i\boldsymbol{U}_i^\top$. Define $\{\lambda'_j, \boldsymbol{P}'_j\}_{j=1}^{r'}$ analogously for $\boldsymbol{H}'$, so that

$$\boldsymbol{H} = \sum_{i=1}^{r} \lambda_i \boldsymbol{P}_i, \quad \boldsymbol{H}' = \sum_{j=1}^{r'} \lambda'_j \boldsymbol{P}'_j, \quad \sum_{i=1}^{r} \boldsymbol{P}_i = \sum_{j=1}^{r'} \boldsymbol{P}'_j = \boldsymbol{I}\,.$$

Consider the *active* sets for pretraining and fine-tuning through the geometry induced by $\|\cdot\|_{\mathtt{M}}$:

$$\mathcal{I} := \{i : \|\boldsymbol{P}_i\widehat{\boldsymbol{w}}\|_{\mathtt{M}} > 0\}, \quad \mathcal{I}' := \{j : \|\boldsymbol{P}'_j\boldsymbol{\Delta}\|_{\mathtt{M}} > 0\}\,.$$

For instance, under the downstream prediction seminorm $\|\cdot\|_{\widehat{\boldsymbol{\Sigma}}'}$, $\mathcal{I}$ contains the pretraining eigenspaces whose residual signal is visible to fine-tuning prediction, while $\mathcal{I}'$ contains the fine-tuning shift eigenspaces that are visible in downstream data geometry.

We deliberately work with projectors onto eigenspaces of distinct eigenvalues, rather than with an arbitrary eigenvector basis. This distinction matters when eigenvalues have multiplicity: individual eigenvectors are not canonical, but the projector $\boldsymbol{P}_i$ onto the full eigenspace is basis-invariant. The active sets are therefore defined through eigenspace-level quantities.

We now outline three degenerate cases, when either one or both of these active sets are empty.

Theorem 4 (Special cases for t^*)**.** *Let t^* be a global minimiser of $f_{\mathtt{M}} : [T] \to [0, \infty)$ defined as $f_{\mathtt{M}}(t) := \|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|_{\mathtt{M}}^2$. (i) If $\mathcal{I} \neq \emptyset$, $\mathcal{I}' = \emptyset$, then $T \in \mathcal{T}^*$. (ii) If $\mathcal{I} = \emptyset$, $\mathcal{I}' \neq \emptyset$, then $0 \in \mathcal{T}^*$. (iii) If $\mathcal{I} = \mathcal{I}' = \emptyset$, then $\mathcal{T}^* = [T]$.*

Remark 5. Case (i) indicates that the pretraining solution $\widehat{\boldsymbol{w}}$ has components that matter for the fine-tuning task, but the true task shift $\boldsymbol{\Delta}$ is invisible under $\|\cdot\|_{\mathtt{M}}$. In other words, the optimal allocation is full pretraining and no fine-tuning. Case (ii) is the opposite of Case (i): the pretraining solution $\widehat{\boldsymbol{w}}$ is completely irrelevant for the fine-tuning task, but the fine-tuning shift does matter. So, we take a no pretraining but full fine-tuning strategy. Case (iii) is the trivial case.

Theorem 4 characterises degenerate regimes where either pretraining or fine-tuning is completely useless, and a boundary allocation $t = 0$ or $t = T$ is optimal. From now on, we will assume $\mathcal{I}, \mathcal{I}' \neq \emptyset$, in which case t^* can be an interior point, as we show next.

3.2 General cases: $\mathcal{I}, \mathcal{I}' \neq \emptyset$ and fine-tuning ridge Hessian evaluation

The results for this general case are more complex. For ease of description, we first state our results for $\mathtt{M} = \boldsymbol{H}'$, which are clearer, and then provide in Section 3.3 the corresponding results for $\mathtt{M} = \widehat{\boldsymbol{\Sigma}}'$.

Theorem 6 (General case under the $\boldsymbol{H}'$ norm)**.** *Assume $\mathcal{I}, \mathcal{I}' \neq \emptyset$ under the $\boldsymbol{H}'$ norm and consider $i^* = \min \mathcal{I}$, $j^* = \min \mathcal{I}'$, and θ^* as defined in (2). Then, there exist an offset $s^* \in \mathbb{R}$, a rate $\delta > 0$, and $C, T_0 > 0$, all functions of the problem parameters $L, c_{\mathrm{p}}, c_{\mathrm{f}}, \gamma, \gamma', \widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\Sigma}}', \widehat{\boldsymbol{w}}, \boldsymbol{\Delta}, \eta, \eta'$, such that the following holds: for all $T \geq T_0$, the evaluation metric $f_{\boldsymbol{H}'} : [T] \to [0, \infty)$, $f_{\boldsymbol{H}'}(t) = \|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|^2_{\boldsymbol{H}'}$ has at most two global minimisers, each $t^* \in \mathcal{T}^*$ satisfying*

$$|t^* - (s^* + \theta^* T)| \leq 1 + Ce^{-\delta T}.$$

As the proof of Theorem 6 reveals, for sufficiently large T, the relaxation of $f_{\boldsymbol{H}'}$ to $[0, T]$ has a unique global minimiser, and every discrete global minimiser is its floor or ceiling. Thus, up to one feasible allocation step, t^* is approximated by two terms with an exponentially decaying error. The term linear in T is determined by the *contraction rates* $-\ln(1 - \eta\lambda_{i^*})$ and $-\ln(1 - \eta'\lambda'_{j^*})$ of smallest active eigenvalues λ_{i^*} and λ'_{j^*} of $\boldsymbol{H}$ and $\boldsymbol{H}'$, respectively, and the compute normalisation $1/c_{\mathrm{p}}$ and $1/c_{\mathrm{f}}$. The offset s^* is independent of T, hence, as $T \to \infty$, t^*/T converges to the *leading fraction* θ^*. The corresponding results for the optimal iteration counts $k(t^*)$ and $\ell(t^*)$, accompanied by a geometric interpretation, can be found in Section 3.2.1.

The offset s^* is the correction to the leading allocation $\theta^* T$, and its explicit expression can be found in Appendix A. Geometrically, it is quantified by (i) the misalignment in $\boldsymbol{H}'$-geometry between the two slow 'bottleneck' directions via $\langle \boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}, \boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\rangle_{\boldsymbol{H}'}$, induced by the non-commutativity of $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}'$; (ii) the strength of the dominant pretraining $\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\|^2_{\boldsymbol{H}'}$ and fine-tuning $\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\|^2_{\boldsymbol{H}'}$ components.

The rate δ, whose explicit expression can also be found in Appendix A, depends on the separation between the bottleneck contraction rates and the remaining active rates. Small spectral separations can therefore delay the budget regime in which the two-term approximation becomes accurate.

Next, Theorem 7 shows that the value of $f_{\boldsymbol{H}'}$ at interior points represents an exponential improvement (in $T \propto B$) over extreme allocations with a fixed number M of units assigned to either phase.

Theorem 7 (Exponential improvement over extreme allocations)**.** *Recall the setting of Theorem 6. Fix any $\widehat{\theta} \in (0, 1)$ and define the feasible allocation $\widehat{t} \in [T]$ and rate $m(\widehat{\theta}) > 0$ as*

$$\widehat{t} \in \arg\min_{t \in [T]} \left| t - \widehat{\theta} T \right|, \qquad m(\widehat{\theta}) := \min\left\{ -\frac{L\widehat{\theta}}{c_{\mathrm{p}}} \ln(1 - \eta\lambda_{i^*}), -\frac{L(1-\widehat{\theta})}{c_{\mathrm{f}}} \ln(1 - \eta'\lambda'_{j^*}) \right\}.$$

Fix any integer $M \geq 1$. Then, there exists $C_1 > 0$ and $T_0 > 2M$, depending on M and the fixed problem parameters, but independent of $\widehat{\theta}$, such that for any $T \geq T_0$,

$$f_{\boldsymbol{H}'}(\,\widehat{t}\,) \leq C_1 \exp(-2m(\widehat{\theta})T) \cdot \min\{f_{\boldsymbol{H}'}(t) : t \in \{0, \ldots, M\} \cup \{T - M, \ldots, T\}\}.$$

The decay rate $m(\widehat{\theta})$ is maximised for $\widehat{\theta} = \theta^*$ defined in (2).

Remark 8. Although the offset s^* in Theorem 6 shifts the optimal t^* away from $\theta^* T$, this offset depends on solution-level information, including $\widehat{\boldsymbol{w}}$ and $\widehat{\boldsymbol{w}}'$. Thus, evaluating t^* can be as computationally intensive as solving each ridge problem separately. Theorem 7 guarantees an exponential improvement over extreme allocations *without computing the offset s^**. Since an estimate $\widehat{\theta} \approx \theta^*$ retains the improvement with $m(\widehat{\theta})$ close to $m(\theta^*)$, one can instead use the split induced by an integer nearest to $\widehat{\theta} T$. Section 3.4 describes how to perform such an estimation of θ^* in $O((n + n')d)$ time using matrix-free Lanczos iterations.

3.2.1 Optimal iteration counts

We now return to the original iteration counts k and ℓ. Using the representation (1), Theorem 6 gives the following corollary, in which the $O(1)$ terms absorb the offset and integral allocation error:

Corollary 9. *Under the assumptions of Theorem 6, every optimal schedule $(k(t^*), \ell(t^*))$ satisfies*

$$k(t^*) = \theta^* \frac{B}{c_{\mathrm{p}}} + O(1), \quad \ell(t^*) = (1 - \theta^*)\frac{B}{c_{\mathrm{f}}} + O(1).$$

In particular, as $B = LT \to \infty$,

$$\frac{k(t^*)}{\ell(t^*)} \to \frac{\ln(1 - \eta'\lambda'_{j^*})}{\ln(1 - \eta\lambda_{i^*})}.$$

Remark 10. Our two-phase algorithm optimises a strongly convex quadratic in each phase. GD progresses faster in high-curvature directions, and slower in low-curvature directions. Thus, the relevant bottlenecks are the flattest active directions, indexed by i^* and j^*. Each GD iteration contracts the hardest active component by $1-\eta\lambda_{i^*}$ or $1-\eta'\lambda'_{j^*}$, respectively. The inverse logarithmic contraction rates therefore measure the iteration scales for constant progress:

$$k(t^*) \propto \frac{1}{-\ln(1-\eta\lambda_{i^*})} \quad \text{and} \qquad \ell(t^*) \propto \frac{1}{-\ln(1-\eta'\lambda'_{j^*})},$$

with a common proportionality factor fixed by the compute constraint $c_{\mathrm{p}}k(t^*) + c_{\mathrm{f}}\ell(t^*) = B$. Up to the offset and integer allocation, this insight recovers the iteration counts of Corollary 9.

Remark 11 (Sample sizes and compute allocation). For dense full-batch GD, the iteration costs scale as $c_{\mathrm{p}} = \Theta(nd)$ and $c_{\mathrm{f}} = \Theta(n'd)$, so that $c_{\mathrm{p}}/c_{\mathrm{f}} \approx n/n'$. According to Corollary 9, at leading order, the optimal pretraining compute is:

$$c_{\mathrm{p}}k(t^*) \approx \frac{n}{n + n'\frac{\ln(1-\eta\lambda_{i^*})}{\ln(1-\eta'\lambda'_{j^*})}}\, B. \tag{3}$$

Thus, holding the contraction rates fixed, increasing n/n' shifts the leading allocation towards pretraining, because each upstream pass becomes relatively more expensive. The ratio of the contraction rates adjusts this sample-size weighting for the relative difficulty of the two phases. In Fig. 1, the ratio of contraction rates is approximately $2.626/2.626/1.003$ in the left/middle/right panels, respectively. In particular, when $n = n' = 80$, a fraction of $1/(1 + 2.626) \approx 27.58\%$ of compute is allocated by (3) to pretraining, close to the numerical optimum of 27.21%.

3.3 General cases: $\mathcal{I}, \mathcal{I}' \neq \emptyset$ under the fine-tuning data geometry

We now provide our results on the optimal set $\mathcal{T}^*$ under the $\widehat{\boldsymbol{\Sigma}}'$-seminorm. Compared to the $\boldsymbol{H}'$ norm evaluation metric case, these results require an additional assumption that we discuss in detail.

3.3.1 Localisation of t^*

Theorem 12 (General case under the $\|\cdot\|_{\widehat{\boldsymbol{\Sigma}}'}$ seminorm)**.** *Assume $\mathcal{I}, \mathcal{I}' \neq \emptyset$ under the $\|\cdot\|_{\widehat{\boldsymbol{\Sigma}}'}$ seminorm and consider $i^* = \min \mathcal{I}$ and $j^* = \min \mathcal{I}'$ and θ^* as defined in (2). Then, there exist an offset $s^* \in \mathbb{R}$, a rate $\delta > 0$, an integer burn-in $K \geq 0$, and $C > 0$, $T_0 \geq K$, all functions of the problem parameters L, c_{p}, c_{f}, γ, γ', $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\boldsymbol{\Sigma}}'$, $\widehat{\boldsymbol{w}}$, $\boldsymbol{\Delta}$, η, η', such that the following holds: Assume the* no-kernel-hitting *condition:*

$$\textit{for all } t \in [K]: \quad \left\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_{k(t)}\right\|_{\widehat{\boldsymbol{\Sigma}}'} \neq 0. \tag{NKH}$$

Then, for all $T \geq T_0$, the evaluation metric $f_{\widehat{\boldsymbol{\Sigma}}'} : [T] \to [0, \infty)$, $f_{\widehat{\boldsymbol{\Sigma}}'}(t) = \|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|^2_{\widehat{\boldsymbol{\Sigma}}'}$ has at most two global minimisers, each $t^* \in \mathcal{T}^*$ satisfying

$$|t^* - (s^* + \theta^* T)| \leq 1 + Ce^{-\delta T}.$$

These results under the $\|\cdot\|_{\widehat{\boldsymbol{\Sigma}}'}$ seminorm are very similar to the ones of Theorem 6 under the $\boldsymbol{H}'$ norm, and the expression for K, s^*, and δ can be explicitly found in Appendix A.

Different from the $\|\cdot\|_{\boldsymbol{H}'}$ evaluation metric, $\widehat{\boldsymbol{\Sigma}}'$ could in principle have 0 eigenvalues, causing $\|\cdot\|_{\widehat{\boldsymbol{\Sigma}}'}$ to become a seminorm that vanishes on $\ker(\widehat{\boldsymbol{\Sigma}}')$. Hence, we make the 'no-kernel-hitting' assumption (NKH): for a burn-in period K, the pretraining GD trajectory $t \mapsto \widehat{\boldsymbol{w}} - \boldsymbol{w}_{k(t)}$ does not enter $\ker(\widehat{\boldsymbol{\Sigma}}')$, remaining visible to downstream prediction. As we show next, there are cases where (NKH) does not hold and, unlike in the conclusion of Theorem 12, t^* may not scale linearly with T.

Proposition 13. *There exist $(\boldsymbol{X}, \boldsymbol{y}), (\boldsymbol{X}', \boldsymbol{y}') \in \mathbb{R}^{2\times 2} \times \mathbb{R}^2$ and $c_{\mathrm{p}}, c_{\mathrm{f}}, \eta, \eta', \gamma, \gamma' > 0$, such that, if $T \geq 3$, then the (NKH) condition does not hold for any $K \geq 3$, i.e. $\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_{k(t)}\|_{\widehat{\boldsymbol{\Sigma}}'} = 0$ for some $t \in [K]$. Additionally, $f_{\widehat{\boldsymbol{\Sigma}}'}(t)$ has a unique global minimiser at $t^* = 1$, so that $t^*/T \to 0$ as $T \to \infty$.*

However, under a standard random-design sampling mechanism, the condition (NKH) holds almost surely, and it is weaker than a commutativity assumption on $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}'$, as we show next.

Proposition 14. *Assume that $n' \geq 1$ and the rows of $\boldsymbol{X}' \in \mathbb{R}^{n' \times d}$ are drawn from a distribution supported on $\mathbb{R}^d$ that has a density with respect to the Lebesgue measure, and are independent of the (potentially deterministic) pretraining data. Moreover, assume $\widehat{\boldsymbol{w}} \neq 0$ (otherwise $\mathcal{I} = \emptyset$, see Theorem 4). Then, almost surely, $\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_{k(t)}\|_{\widehat{\boldsymbol{\Sigma}}'} > 0$ for every integer $t \geq 0$.*

Proposition 15. *Assume that $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}'$ commute. Then (deterministically):*

(i) If $\widehat{\boldsymbol{w}} \in \ker(\widehat{\boldsymbol{\Sigma}}')$, then $\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_{k(t)}\|_{\widehat{\boldsymbol{\Sigma}}'} = 0$ for all integers $t \geq 0$, but $\mathcal{I} = \emptyset$ (c.f. Theorem 4).

(ii) If $\widehat{\boldsymbol{w}} \notin \ker(\widehat{\boldsymbol{\Sigma}}')$, then (NKH) holds.

3.4 Estimating the leading fraction θ^*

Evaluating the offset s^* in Theorems 6 and 12, whose explicit expression can be found in Appendix A, requires information about projections of $\widehat{\boldsymbol{w}}$ and $\boldsymbol{\Delta}$ onto the active eigenspaces. However, estimating the leading fraction θ^* can overcome a potential lack of solution-level information. Indeed,

$$\frac{t^*}{T} = \theta^* + O(T^{-1}).$$

The offset and integer-allocation error contribute only $O(T^{-1})$ to the compute fraction. This suggests estimating $\widehat{\theta} \approx \theta^*$ through the two bottleneck rates in the definition (2) of θ^*. We next describe one possible approach for the $\boldsymbol{H}'$-norm criterion, deployed in Fig. 1.

Estimating the upstream bottleneck λ_{i^*}: Let $\boldsymbol{g} := \boldsymbol{X}^\top \boldsymbol{y}/n = -\nabla R_{\mathrm{pre}}(\boldsymbol{0}) = \boldsymbol{H}\widehat{\boldsymbol{w}}$, which is already needed for the first pretraining iteration. Since $\boldsymbol{P}_i \boldsymbol{g} = \lambda_i \boldsymbol{P}_i \widehat{\boldsymbol{w}}$ and $\lambda_i > 0$, the upstream bottleneck satisfies $\lambda_{i^*} = \min\{\lambda_i : \boldsymbol{P}_i \boldsymbol{g} \neq \boldsymbol{0}\}$. Consequently, λ_{i^*} can be estimated using a matrix-free Lanczos iteration (see, for instance, [22]) applied to $\boldsymbol{H}$ and initialised with $\boldsymbol{g}$, without first computing $\widehat{\boldsymbol{w}}$. Indeed, $\boldsymbol{H}^m \boldsymbol{g} = \sum_{i \in \mathcal{I}} \lambda_i^m \boldsymbol{P}_i \boldsymbol{g}$, so no inactive eigenspaces enter the subspace explored by the Lanczos method. The smallest eigenvalue of $\boldsymbol{H}$ on the full subspace generated by these vectors $\boldsymbol{H}^m \boldsymbol{g}$ is exactly λ_{i^*}.

Estimating the downstream bottleneck λ'_{j^*}: If $\boldsymbol{P}'_1 \boldsymbol{\Delta} \neq \boldsymbol{0}$, as one expects, for instance, in a random-design setting, then $j^* = 1$. Consequently, the relevant downstream eigenvalue is simply the smallest eigenvalue λ'_1 of $\boldsymbol{H}'$. When $d > n'$ we can simply use $\lambda'_1 = \gamma'$, and otherwise it can be estimated, once again, by Lanczos, using a starting vector with a non-zero $\boldsymbol{P}'_1$ projection.

Computational requirements: Since they deploy matrix-free Lanczos iterations, these procedures use Hessian–vector products rather than explicit Hessian matrices. For example, $\boldsymbol{H}\boldsymbol{v} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{v})/n + \gamma \boldsymbol{v}$, which costs $O(nd)$ for a dense design matrix, with the analogous downstream product costing $O(n'd)$. For a fixed number of Lanczos iterations, estimating the allocation θ^* costs $O((n+n')d)$. For dense full-batch GD, the per-iteration costs are $c_{\mathrm{p}} = \Theta(nd)$, $c_{\mathrm{f}} = \Theta(n'd)$. As $B = LT$ and $L \geq \max\{c_{\mathrm{p}}, c_{\mathrm{f}}\}$, we have $(n+n')d = O(B/T)$.

In comparison, an exhaustive GD sweep over all $T+1$ allocations costs $\Theta(BT)$, even when the pretraining trajectory is shared. Alternatively, solving each ridge problem directly requires computing $\widehat{\boldsymbol{w}}$ and $\widehat{\boldsymbol{w}}'$, at a cost of $O((n+n')d^2)$ when $d \leq n, n'$ or $O((n^2 + (n')^2)d)$ when $d \geq n, n'$. In either case, matrix-free Lanczos estimation is a more computationally efficient alternative.

By Theorem 7, for any fixed estimate $\widehat{\theta} \in (0,1)$, the feasible split nearest to $\widehat{\theta}T$ achieves improvement of $\exp(-2m(\widehat{\theta})T)$ over extreme allocations. Lanczos estimation error therefore affects the guaranteed exponential decay rate, rather than eliminating the guarantee, and this rate approaches its maximum as $\widehat{\theta} \to \theta^*$. This supports estimating the leading fraction θ^* without computing the ridge solutions or the offset s^*. The accuracy of such an estimation requires a separate analysis that we leave as future work.

4 Conclusion

In this paper, we study compute allocation in a two-stage procedure for ridge regression: running k GD iterations for pretraining, followed by fine-tuning with another ℓ GD iterations, under the budget constraint $c_{\mathrm{p}}k + c_{\mathrm{f}}\ell = B$. We characterise the optimal allocation $(k(t^*), \ell(t^*))$, showing that it is governed by the slowest upstream and downstream-visible bottleneck curvatures, with a correction

term capturing their alignment. This characterisation is structural, requiring access to solution-level and spectral quantities, which in general can be as costly as solving the underlying optimisation problem itself (to compute the transfer update $\mathbf{\Delta}$). Algorithmically, we propose estimating the optimal allocation via matrix-free Lanczos iterations. This estimated interior allocation can be computed without access to solution-level information, and is more efficient than a checkpoint sweep or full spectral decompositions. To complement these results, we show that interior allocations are exponentially better (in B) than extreme splits.

Extending the theory to more general settings calls for new tools and perspectives. Our analysis relies on full-batch gradient descent, where the dynamics are linear and noise-free. Moving to stochastic optimisation would introduce gradient noise effects that interact non-trivially with the underlying geometry. More fundamentally, the ridge setting yields fixed Hessians and a static spectral decomposition, whereas in nonlinear models the representation, and thus the effective geometry, evolves during training. Understanding whether the same bottleneck-driven allocation principle persists in such feature-learning regimes, and how it can be approximated in an adaptive and computationally efficient way, remains an important direction for future work.

Acknowledgments and Disclosure of Funding

Alex Buna was supported by EPSRC through the StatML CDT programme [project reference EP/S023151/1]. Fanghui Liu was supported by the National Natural Science Foundation of China [grant number 12671631]. Patrick Rebeschini was supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number EP/Y028333/1]

References

[1] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd international conference on artificial intelligence and statistics*, pages 1370–1378. PMLR, 2019.

[2] Francis Bach. *Learning theory from first principles*. MIT press, 2024.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, pages 1877–1901, 2020.

[4] Xingyi Cheng, Bo Chen, Pan Li, Jing Gong, Jie Tang, and Le Song. Training compute-optimal protein language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 69386–69418. Curran Associates, Inc., 2024.

[5] Arif Kerem Dayı and Sitan Chen. Low-rank fine-tuning lies between lazy training and feature learning. *Proceedings of Machine Learning Research*, 291:1415–1471, 2025.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[7] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*, 2022.

[8] Jiawei Ge, Shange Tang, Jianqing Fan, and Chi Jin. On the provable advantage of unsupervised pretraining. *arXiv preprint arXiv:2303.01566*, 2023.

[9] L Lo Gerfo, Lorenzo Rosasco, Francesca Odone, E De Vito, and Alessandro Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

[10] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2 edition, 2009.

[11] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

[12] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[13] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[14] 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. In *International Conference on Learning Representations*, 2022.

[15] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

[16] Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. Scaling laws for downstream task performance of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

[17] Dhruva Karkada, Joseph Turnbull, Yuxi Liu, and James Simon. Predicting kernel regression learning curves from only raw data statistics. In *International Conference on Learning Representations*, pages 118355–118425, 2026.

[18] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

[19] Binghui Li, Fengling Chen, Zixun Huang, Lean Wang, and Lei Wu. Functional scaling laws in kernel regression: Loss dynamics and learning rate schedules. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 101211–101269. Curran Associates, Inc., 2025.

[20] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+ 3 phases of compute-optimal neural scaling laws. In *Advances in Neural Information Processing Systems*, volume 37, pages 16459–16537, 2024.

[21] Lorenzo Rosasco, Ernesto De Vito, and Alessandro Verri. Spectral methods for regularization in learning theory. *DISI, Universita degli Studi di Genova, Italy, Technical Report DISI-TR-05-18*, 2005.

[22] Yousef Saad. *Numerical methods for large eigenvalue problems: revised edition*. SIAM, 2011.

[23] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[24] Gal Shachaf, Alon Brutzkus, and Amir Globerson. A theoretical analysis of fine-tuning with linear teachers. *Advances in Neural Information Processing Systems*, 34:15382–15394, 2021.

[25] Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. *Advances in Neural Information Processing Systems*, 31, 2018.

[26] Kaiser Sun and Mark Dredze. Amuro & char: Analyzing the relationship between pre-training and fine-tuning of large language models. In *Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*, pages 131–151, 2025.

[27] Salma Tarmoun, Guilherme Franca, Benjamin D Haeffele, and Rene Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In *International Conference on Machine Learning*, pages 10153–10161. PMLR, 2021.

[28] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[29] Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. The power and limitation of pretraining-finetuning for linear regression under covariate shift. In *Advances in Neural Information Processing Systems*, volume 35, pages 33041–33053, 2022.

[30] Jingfeng Wu, Peter L Bartlett, Jason D Lee, Sham M Kakade, and Bin Yu. Risk comparisons in linear regression: Implicit regularization dominates explicit regularization. *arXiv preprint arXiv:2509.17251*, 2025.

[31] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*, 2024.

[32] Yunzhang Zhu and Renxiong Liu. An algorithmic view of ℓ_2 regularization and some path-following algorithms. *Journal of Machine Learning Research*, 22(138):1–62, 2021.

A Additional notation

We introduce additional notation to aid with the proofs. When the eigendecomposition of a matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is given by $(\lambda_i, \boldsymbol{v}_i)_{i=1}^d$ and f is a scalar function, we write $f(\boldsymbol{A}) = \sum_{i=1}^d f(\lambda_i)\boldsymbol{v}_i\boldsymbol{v}_i^\top$.

Primarily, our analysis relies on the relaxation $f_{\mathtt{M}} : [0,T] \to [0,\infty)$, $f_{\mathtt{M}}(t) = \|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}\|_{\mathtt{M}}^2$, which allows for differentiation. This can be expressed in a more convenient form using the two generating matrices

$$\boldsymbol{G} := -\frac{L}{c_{\mathrm{p}}}\ln(\boldsymbol{I} - \eta\boldsymbol{H}), \quad \boldsymbol{G}' := -\frac{L}{c_{\mathrm{f}}}\ln(\boldsymbol{I} - \eta'\boldsymbol{H}').$$

Note that these are well-defined, symmetric, and positive definite, due to the step-size choice $0 < \eta < \|\boldsymbol{H}\|_2^{-1}$, $0 < \eta' < \|\boldsymbol{H}'\|_2^{-1}$. We work with these generating matrices, rather than the contractions $\boldsymbol{I} - \eta\boldsymbol{H} = \exp(-c_{\mathrm{p}}\boldsymbol{G}/L)$, $\boldsymbol{I} - \eta'\boldsymbol{H}' = \exp(-c_{\mathrm{f}}\boldsymbol{G}'/L)$, in order to allow for differentiation of exponentials with respect to t in our subsequent analysis.

With this notation, note that the conclusions of Lemma 2 are

$$\begin{aligned}
\boldsymbol{w}_{k(t)} &= (\boldsymbol{I} - e^{-t\boldsymbol{G}})\widehat{\boldsymbol{w}}\,,\\
\boldsymbol{w}'_{\ell(t)} &= \widehat{\boldsymbol{w}}' + e^{-(T-t)\boldsymbol{G}'}(\boldsymbol{w}_{k(t)} - \widehat{\boldsymbol{w}}'),\\
f_{\mathtt{M}}(t) &= \|(\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}} - e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\|_{\mathtt{M}}^2\,.
\end{aligned}$$

Additionally, recalling the basis-invariant eigendecompositions of $\boldsymbol{H}$ and $\boldsymbol{H}'$ as $(\lambda_i, \boldsymbol{P}_i)_{i=1}^r$ and $(\lambda'_j, \boldsymbol{P}'_j)_{j=1}^{r'}$, respectively, we denote the eigenvalues of $\boldsymbol{G}$ and $\boldsymbol{G}'$ as

$$\mu_i := -\frac{L}{c_{\mathrm{p}}}\ln(1 - \eta\lambda_i), \quad \mu'_j := -\frac{L}{c_{\mathrm{f}}}\ln(1 - \eta'\lambda'_j),$$

where $1 \le i \le r$ and $1 \le j \le r'$. Note that the corresponding orthogonal eigenspace projectors of $\boldsymbol{G}$ and $\boldsymbol{G}'$ remain $\{\boldsymbol{P}_i\}_{i=1}^r$ and $\{\boldsymbol{P}'_j\}_{j=1}^{r'}$, respectively, and recall that we denote $i^* = \min\mathcal{I}$, $j^* = \min\mathcal{I}'$, potentially taking different values when $\mathtt{M} = \boldsymbol{H}'$ or $\mathtt{M} = \widehat{\boldsymbol{\Sigma}}'$. With this notation in place, the leading fraction becomes

$$\theta^* = \frac{\mu'_{j^*}}{\mu_{i^*} + \mu'_{j^*}}.$$

The expression of the offset term in both Theorems 6 and 12 is

$$s^* := \frac{1}{\mu_{i^*} + \mu'_{j^*}}\ln\left(\frac{(\mu'_{j^*} - \mu_{i^*})d^* + \sqrt{(\mu'_{j^*} - \mu_{i^*})^2(d^*)^2 + 4\mu_{i^*}\mu'_{j^*}a^*b^*}}{2\mu'_{j^*}b^*}\right),$$

where $a^* := \|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\|_{\mathtt{M}}^2 > 0$, $b^* := \|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\|_{\mathtt{M}}^2 > 0$, and $d^* := \langle \boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}, \boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\rangle_{\mathtt{M}}$. Moreover, the convergence rate $\delta > 0$ is given by

$$\delta := \min\left(\delta_1, \delta_2, \frac{\mu_{i^*}\mu'_1}{\mu_{i^*} + \mu'_{j^*}}\right),$$

where $\delta_1 := \begin{cases} +\infty & \text{if } \mathcal{I} = \{i^*\},\\ \min\limits_{i\in\mathcal{I}\setminus\{i^*\}} \frac{(\mu_i - \mu_{i^*})\mu'_{j^*}}{\mu_{i^*} + \mu'_{j^*}}, & \text{otherwise,}\end{cases}$ $\delta_2 := \begin{cases} +\infty & \text{if } \mathcal{I}' = \{j^*\},\\ \min\limits_{j\in\mathcal{I}'\setminus\{j^*\}} \frac{\mu_{i^*}(\mu'_j - \mu'_{j^*})}{\mu_{i^*} + \mu'_{j^*}}, & \text{otherwise.}\end{cases}$

Additionally, the expression of the burn-in period in the (NKH) condition of Theorem 12 (specialised for $\mathtt{M} = \widehat{\boldsymbol{\Sigma}}'$) is given by:

$$K := \begin{cases} 0, & \text{if } \mathcal{I} = \{i^*\},\\ \max\left(0, \left\lceil \frac{1}{\min\limits_{i\in\mathcal{I}\setminus\{i^*\}}(\mu_i - \mu_{i^*})}\ln\left(\frac{4\sum_{i\in\mathcal{I}\setminus\{i^*\}}\|\boldsymbol{P}_i\widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'}}{\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'}}\right)\right\rceil\right) & \text{otherwise.}\end{cases}$$

Using the generator $\boldsymbol{G}$, the (NKH) condition can be equivalently expressed as

$$\min_{t\in[K]} \left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_{\widehat{\boldsymbol{\Sigma}}'} > 0\,.$$

B Proof sketch of Theorem 12

In this section we give the proof sketch of Theorem 12 under the $\|\cdot\|_{\widehat{\boldsymbol{\Sigma}}'}$ seminorm. The proof of Theorem 6 under the $\boldsymbol{H}'$-induced norm is similar with the exception of Step 5 below, so we omit the sketch of its proof in the main text. The main difference is that, by the invertibility of $e^{-t\boldsymbol{G}}$ and as $\widehat{\boldsymbol{w}} \neq \mathbf{0}$, $\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\|_2 > 0$ always and we do not need to impose the (NKH) assumption. The relevant notation is presented in Appendix A, and full details and derivations can be found in Appendix C. The proof is split into seven steps and it relies on analysing $f'_{\widehat{\boldsymbol{\Sigma}}'}$, the derivative of the interpolator $f_{\widehat{\boldsymbol{\Sigma}}'} : [0, T] \to [0, \infty)$. It requires identifying an approximate zero of $f'_{\widehat{\boldsymbol{\Sigma}}'}$, and developing a perturbative control of non-commuting dynamics at exponentially small scales.

Step 1: Decomposing f': First of all, we recenter f around the limiting

$$t_0 := \frac{\mu'_{j^*}}{\mu_{i^*} + \mu'_{j^*}} T\,,$$

so that we mostly work with $f_{\widehat{\boldsymbol{\Sigma}}'}(t_0 + s)$ with $s \in [-t_0, T - t_0]$. Under this reparametrisation, a bit of algebra shows that

$$\frac{f'_{\widehat{\boldsymbol{\Sigma}}'}(t_0 + s)}{2e^{-2m^*T}} = \Gamma(s) + \Xi(s),$$

where we recall that $m^* = \mu_{i^*}\mu'_{j^*}/(\mu_{i^*} + \mu'_{j^*})$, and the two newly introduced functions of s have the following roles: $\Gamma(s)$ gathers the three terms a^*, b^* and d^* (defined in Appendix A) from the expansion terms of $f'_{\widehat{\boldsymbol{\Sigma}}'}$ that depend on i^* and j^* only, and hence are the relevant asymptotic bottlenecks, i.e.

$$\Gamma(s) = \mu'_{j^*} \left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'} e^{2\mu'_{j^*}s} - \mu_{i^*} \left\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'} e^{-2\mu_{i^*}s} - (\mu'_{j^*} - \mu_{i^*})\langle \boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}, \boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\rangle_{\widehat{\boldsymbol{\Sigma}}'} e^{(\mu'_{j^*} - \mu_{i^*})s};$$

while $\Xi(s)$ encompasses the remaining terms in f' that behave like 'small' perturbations.

Step 2: Analysing $\Gamma(s)$: It turns out that $\Gamma(s)$ has as its unique root the offset s^*. Thus, if $\Xi(s)$ has a small additive contribution, we would expect $f'_{\widehat{\boldsymbol{\Sigma}}'}(t_0 + s^*) \approx 0$.

Moreover, $\Gamma(s)$ is strictly increasing in a neighborhood $[s^* - S_0, s^* + S_0]$ centred at s^*, for some S_0 that does not depend on T, and we denote by

$$m_{S_0} := \min_{|s^* - s| \leq S_0} \Gamma'(s) > 0.$$

Moreover, as $\Gamma(s^*) = 0$, we have that $\Gamma(s^* - S_0) < 0$ and $\Gamma(s^* + S_0) > 0$.

Step 3: Uniform bound for $\Xi(s)$ in any compact interval centred at s^*: We show that for any $S > 0$, the perturbation term $\Xi(s)$ and its derivative are 'small' in the sense that

$$\sup_{|s - s^*| \leq S} (|\Xi(s)| + |\Xi'(s)|) \leq C_S e^{-\delta T}, \tag{4}$$

where C_S does not depend on T and δ was introduced in the statement of Theorem 12. Thus, we can expect Ξ to not shift the zero of f' far away from $t_0 + s^*$.

Step 4: Existence, uniqueness and localisation of the zero of $\Gamma(s) + \Xi(s)$ inside $[s^* - S_0, s^* + S_0]$: Since S_0 is independent of T, using Eq. (4) for T large enough we can make $|\Xi'(s)|$ appropriately small on $[s^* - S_0, s^* + S_0]$ such that

$$\Gamma'(s) + \Xi'(s) \geq m_{S_0} - |\Xi'(s)| \geq m_{S_0}/2 > 0$$

So, if $\Gamma + \Xi$ has a zero in this interval, it is unique.

And $\Gamma + \Xi$ indeed has a zero in $[s^* - S_0, s^* + S_0]$: As we argued in Step 2, $\Gamma(s^* - S_0) < 0$ and by Eq. (4) we can make sure that the small additive contribution of $\Xi(s^* - S_0)$ does not change the sign of Γ. Similarly $(\Gamma + \Xi)(s^* + S_0) > 0$, and by the Intermediate Value Theorem we get the existence of a zero, call it z_T.

To localise z_T, we can use the Mean Value Theorem on the interval $I \subseteq [s^* - S_0, s^* + S_0]$ with endpoints z_T and s^* and the fact that $\Gamma(s^*) = 0$:

$$\begin{aligned} C_{S_0} e^{-\delta T} &\geq |\Xi(s^*)| = |(\Xi + \Gamma)(s^*) - (\Xi + \Gamma)(z_T)| \\ &\geq |s^* - z_T| \cdot \inf_{\xi \in \mathcal{I}} |\Gamma'(\xi) + \Xi'(\xi)| \geq |s^* - z_T| m_{S_0}/2, \end{aligned}$$

which, after defining $\bar{t}_T := t_0 + z_T$ (to translate it back to t-coordinates), rearranges as

$$|\bar{t}_T - (t_0 + s^*)| \leq \frac{2C_{S_0}}{m_{S_0}} e^{-\delta T}.$$

This localises the continuous critical point $\bar{t}_T$. In the next steps we show that every discrete global minimiser is one of its adjacent integers.

Step 5: Proving the discrete global minimisers lie inside $t_0 + [S_-, S_+]$ for some $S_- < 0$, $S_- < s^* < S_+$: To do so, we first upper-bound $f_{\widehat{\boldsymbol{\Sigma}}'}(\hat{t}_T) \leq C_0 e^{-2m^* T}$, where $\hat{t}_T$ is the nearest integer to $t_0 + s^*$. This can be easily done using the technical tools developed in Step 3. Next, we show that $f_{\widehat{\boldsymbol{\Sigma}}'}(t_0 + s) \geq C_{S_-} e^{-2m^* T}$ for all $s \leq S_-$ and $f_{\widehat{\boldsymbol{\Sigma}}'}(t_0 + s) \geq C_{S_+} e^{-2m^* T}$ for all $s \geq S_+$, in both cases working only with s for which $t_0 + s \in [T]$. Once we show how to pick $S_- < s^* < S_+$ such that $C_{S_\pm} > C_0$, we can conclude that the discrete global minimisers lie inside $t_0 + [S_-, S_+]$. While the details are rather technical, we focus on the left-tail $s \leq S_-$ lower bound that uses assumption (NKH):

One can show that the following holds for all $t \leq S_- + t_0$ for a carefully chosen $S_- < 0$:

$$\sqrt{f_{\widehat{\boldsymbol{\Sigma}}'}(t)} \geq \frac{3}{4} \left\| e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} - e^{-\mu'_{j^*}(T-t)} \left\| \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'}.$$

When $t \geq K$ introduced in Appendix A, it can be shown that $\|e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'} \geq 3e^{-\mu_{i^*} t} \|\boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'}/4$ and adjusting S_- appropriately (dependent on the data, but independent of T), we can achieve the desired lower-bound on $f_{\widehat{\boldsymbol{\Sigma}}'}(t_0 + s)$ for all $s \leq S_-$. However, such a uniformly positive lower-bound on $\|e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'}$ cannot be guaranteed uniformly near 0 (for example, at $t = 0$, it could be the case that $\widehat{\boldsymbol{w}} \in \ker(\widehat{\boldsymbol{\Sigma}}')$ and the respective quantity is 0), hence why at the very least we need $\|e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'}$ to be bounded away from 0 for $t \in [K]$.

Thus, all discrete global minimisers must be in the central interval $[S_- + t_0, S_+ + t_0]$.

Step 6: Sign of the derivative in the central interval: In this step, we combine the results from Steps 4 and 5, together with a perturbative sign argument based on Step 3, to conclude that on the central interval the derivative changes sign at $\bar{t}_T$.

Step 7: Localisation of the discrete global minimisers: Finally, we argue that each discrete minimiser is at one of the two integers adjacent to $\bar{t}_T$.

C Missing proofs from Section 3

C.1 Proof of Lemma 1

Denote by $\boldsymbol{u} = \boldsymbol{\Delta}_t - \boldsymbol{\Delta}$, so that $\widehat{\boldsymbol{w}} + \boldsymbol{\Delta}_t = \widehat{\boldsymbol{w}}' + \boldsymbol{u}$ and so

$$\begin{aligned} R_{\text{fine}}(\widehat{\boldsymbol{w}} + \boldsymbol{\Delta}_t) - R_{\text{fine}}(\widehat{\boldsymbol{w}}') &= R_{\text{fine}}(\widehat{\boldsymbol{w}}' + \boldsymbol{u}) - R_{\text{fine}}(\widehat{\boldsymbol{w}}') \\ &= \frac{1}{2n'} \left(\|\boldsymbol{X}'(\widehat{\boldsymbol{w}}' + \boldsymbol{u}) - \boldsymbol{y}'\|_2^2 - \|\boldsymbol{X}'\widehat{\boldsymbol{w}}' - \boldsymbol{y}'\|_2^2 \right) \\ &+ \frac{\gamma'}{2} \left(\|\widehat{\boldsymbol{w}}' + \boldsymbol{u}\|_2^2 - \|\widehat{\boldsymbol{w}}'\|_2^2 \right) \\ &= \frac{1}{n'} (\boldsymbol{X}'\widehat{\boldsymbol{w}}' - \boldsymbol{y}')^\top \boldsymbol{X}'\boldsymbol{u} + \frac{1}{2n'} \|\boldsymbol{X}'\boldsymbol{u}\|_2^2 + \gamma' (\widehat{\boldsymbol{w}}')^\top \boldsymbol{u} + \frac{\gamma'}{2} \|\boldsymbol{u}\|_2^2 \\ &= \left(\frac{1}{n'} (\boldsymbol{X}')^\top (\boldsymbol{X}'\widehat{\boldsymbol{w}}' - \boldsymbol{y}') + \gamma' \widehat{\boldsymbol{w}}' \right)^\top \boldsymbol{u} + \frac{1}{2n'} \|\boldsymbol{X}'\boldsymbol{u}\|_2^2 + \frac{\gamma'}{2} \|\boldsymbol{u}\|_2^2. \end{aligned}$$

Using the first order optimality condition for $\widehat{\boldsymbol{w}}'$ with respect to $R_{\text{fine}}(\cdot)$,

$$\frac{1}{n'}(\boldsymbol{X}')^\top(\boldsymbol{X}'\widehat{\boldsymbol{w}}'-\boldsymbol{y}')+\gamma'\widehat{\boldsymbol{w}}'=0,$$

and thus

$$\begin{aligned}R_{\text{fine}}(\widehat{\boldsymbol{w}}+\boldsymbol{\Delta}_t)-R_{\text{fine}}(\widehat{\boldsymbol{w}}')&=\frac{1}{2n'}\left\|\boldsymbol{X}'\boldsymbol{u}\right\|_2^2+\frac{\gamma'}{2}\left\|\boldsymbol{u}\right\|_2^2=\frac{1}{2n'}\left\|\boldsymbol{X}'(\boldsymbol{\Delta}_t-\boldsymbol{\Delta})\right\|_2^2+\frac{\gamma'}{2}\left\|\boldsymbol{\Delta}_t-\boldsymbol{\Delta}\right\|_2^2\\&=\frac{1}{2}\left\|\boldsymbol{\Delta}_t-\boldsymbol{\Delta}\right\|_{\widehat{\boldsymbol{\Sigma}}'}^2+\frac{\gamma'}{2}\left\|\boldsymbol{\Delta}_t-\boldsymbol{\Delta}\right\|_2^2=\frac{1}{2}\left\|\boldsymbol{\Delta}_t-\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'}^2,\end{aligned}$$

where the second to last equality follows as $\widehat{\boldsymbol{\Sigma}}'=((\boldsymbol{X}')^\top\boldsymbol{X}')/n'$ and $\boldsymbol{H}'=\widehat{\boldsymbol{\Sigma}}'+\gamma'\boldsymbol{I}$.

C.2 Proof of Lemma 2

Define $\boldsymbol{C}:=\boldsymbol{I}-\eta\boldsymbol{H}$ and $\boldsymbol{C}':=\boldsymbol{I}-\eta'\boldsymbol{H}'$. For the pretraining GD, we have for all $0\le s\le k(t)-1$,

$$\boldsymbol{w}_{s+1}-\widehat{\boldsymbol{w}}=\boldsymbol{C}(\boldsymbol{w}_s-\widehat{\boldsymbol{w}}),$$

and unfolding the recursion together with $\boldsymbol{w}_0=0$ we have

$$\boldsymbol{w}_{k(t)}=(\boldsymbol{I}-\boldsymbol{C}^{k(t)})\widehat{\boldsymbol{w}}.$$

Similarly, for the fine-tuning GD, for all $0\le s\le \ell(t)-1$,

$$\boldsymbol{w}'_{s+1}-\widehat{\boldsymbol{w}}'=\boldsymbol{C}'(\boldsymbol{w}'_s-\widehat{\boldsymbol{w}}'),$$

and unfolding the recursion together with $\boldsymbol{w}'_0=\boldsymbol{w}_{k(t)}$ we arrive at

$$\boldsymbol{w}'_{\ell(t)}=\widehat{\boldsymbol{w}}'+(\boldsymbol{C}')^{\ell(t)}(\boldsymbol{w}_{k(t)}-\widehat{\boldsymbol{w}}').$$

For the expression of $f_{\mathtt{M}}(t)=\|\boldsymbol{\Delta}_t-\boldsymbol{\Delta}\|_{\mathtt{M}}^2$, some algebra and the representation (1) give:

$$\begin{aligned}\boldsymbol{\Delta}_t-\boldsymbol{\Delta}&=(\boldsymbol{w}'_{\ell(t)}-\boldsymbol{w}_{k(t)})-\boldsymbol{\Delta}\\&=\widehat{\boldsymbol{w}}'+(\boldsymbol{C}')^{\ell(t)}(\boldsymbol{w}_{k(t)}-\widehat{\boldsymbol{w}}')-\boldsymbol{w}_{k(t)}-\boldsymbol{\Delta}\\&=((\boldsymbol{C}')^{\ell(t)}-\boldsymbol{I})\boldsymbol{w}_{k(t)}+(\boldsymbol{I}-(\boldsymbol{C}')^{\ell(t)})\widehat{\boldsymbol{w}}'-\boldsymbol{\Delta}\\&=((\boldsymbol{C}')^{\ell(t)}-\boldsymbol{I})(\boldsymbol{I}-\boldsymbol{C}^{k(t)})\widehat{\boldsymbol{w}}+(\boldsymbol{I}-(\boldsymbol{C}')^{\ell(t)})\widehat{\boldsymbol{w}}'-\boldsymbol{\Delta}\\&=(\boldsymbol{I}-(\boldsymbol{C}')^{\ell(t)})\boldsymbol{C}^{k(t)}\widehat{\boldsymbol{w}}+(\boldsymbol{I}-(\boldsymbol{C}')^{\ell(t)})(\widehat{\boldsymbol{w}}'-\widehat{\boldsymbol{w}})-\boldsymbol{\Delta}\\&=(\boldsymbol{I}-(\boldsymbol{C}')^{\ell(t)})\boldsymbol{C}^{k(t)}\widehat{\boldsymbol{w}}+(\boldsymbol{I}-(\boldsymbol{C}')^{\ell(t)})\boldsymbol{\Delta}-\boldsymbol{\Delta}\\&=(\boldsymbol{I}-(\boldsymbol{C}')^{\ell(t)})\boldsymbol{C}^{k(t)}\widehat{\boldsymbol{w}}-(\boldsymbol{C}')^{\ell(t)}\boldsymbol{\Delta},\end{aligned}$$

which concludes the proof. ■

C.3 Proof of Theorem 4

We first prove Case (i) for the $\|\cdot\|_{\widehat{\boldsymbol{\Sigma}}'}$ seminorm, then Cases (ii) and (iii).

We start with $\mathcal{I}'=\emptyset$, which implies $\|\boldsymbol{P}'_j\boldsymbol{\Delta}\|_{\widehat{\boldsymbol{\Sigma}}'}^2=0$ for all $1\le j\le r'$. Using $\boldsymbol{H}'=\widehat{\boldsymbol{\Sigma}}'+\gamma'\boldsymbol{I}=\sum_{j=1}^{r'}\lambda'_j\boldsymbol{P}'_j$, $\widehat{\boldsymbol{\Sigma}}'=\sum_{j=1}^{r'}(\lambda'_j-\gamma')\boldsymbol{P}'_j$, we then obtain $\widehat{\boldsymbol{\Sigma}}'\boldsymbol{P}'_j=(\lambda'_j-\gamma')\boldsymbol{P}'_j$ via the orthogonality of $\{\boldsymbol{P}'_j\}_{j=1}^{r'}$. Hence, for all $1\le j\le r'$, we have

$$0=\left\|\boldsymbol{P}'_j\boldsymbol{\Delta}\right\|_{\widehat{\boldsymbol{\Sigma}}'}^2=\boldsymbol{\Delta}^\top(\boldsymbol{P}'_j)^\top\widehat{\boldsymbol{\Sigma}}'\boldsymbol{P}'_j\boldsymbol{\Delta}=(\lambda'_j-\gamma')\boldsymbol{\Delta}^\top(\boldsymbol{P}'_j)^\top\boldsymbol{P}'_j\boldsymbol{\Delta}=(\lambda'_j-\gamma')\left\|\boldsymbol{P}'_j\boldsymbol{\Delta}\right\|_2^2,$$

so $\boldsymbol{P}'_j\boldsymbol{\Delta}=0$ for all $1\le j\le r'$ when $\lambda'_j>\gamma'$. Then,

$$\left\|\widehat{\boldsymbol{\Sigma}}'\boldsymbol{\Delta}\right\|_2^2=\boldsymbol{\Delta}^\top(\widehat{\boldsymbol{\Sigma}}')^2\boldsymbol{\Delta}=\sum_{j:\lambda'_j=\gamma'}(\lambda'_j-\gamma')^2\left\|\boldsymbol{P}'_j\boldsymbol{\Delta}\right\|_2^2+\sum_{j:\lambda'_j>\gamma'}(\lambda'_j-\gamma')^2\left\|\boldsymbol{P}'_j\boldsymbol{\Delta}\right\|_2^2=0,$$

so that $\widehat{\boldsymbol{\Sigma}}'\boldsymbol{\Delta} = 0$. Then, using the fact that $\widehat{\boldsymbol{\Sigma}}'$ and $\boldsymbol{H}'$, hence $\widehat{\boldsymbol{\Sigma}}'$ and $e^{-(T-t)\boldsymbol{G}'}$ commute

$$\begin{aligned} f_{\mathtt{M}}(t) &= \left\|(\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}} - e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'} \\ &= \left\|(\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'} \\ &\quad - 2((\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}})^\top e^{-(T-t)\boldsymbol{G}'}\widehat{\boldsymbol{\Sigma}}'\boldsymbol{\Delta} + (e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta})^\top e^{-(T-t)\boldsymbol{G}'}\widehat{\boldsymbol{\Sigma}}'\boldsymbol{\Delta} \\ &= \left\|(\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'}. \end{aligned}$$

As $f_{\mathtt{M}}(T) = 0$ and $f_{\mathtt{M}} \geq 0$, it is the case that $T \in \mathcal{T}^*$. This concludes (i).

On the other hand, if $\mathcal{I} = \emptyset$, then $\|\boldsymbol{P}_i\widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'} = 0$ for all $1 \leq i \leq r$, so that $(\widehat{\boldsymbol{\Sigma}}')^{1/2}\boldsymbol{P}_i\widehat{\boldsymbol{w}} = \boldsymbol{0}$. Then

$$\widehat{\boldsymbol{\Sigma}}'e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}} = (\widehat{\boldsymbol{\Sigma}}')^{1/2}\sum_{i=1}^{r} e^{-\mu_i t}(\widehat{\boldsymbol{\Sigma}}')^{1/2}\boldsymbol{P}_i\widehat{\boldsymbol{w}} = \boldsymbol{0}\,.$$

Again, using the fact that $\widehat{\boldsymbol{\Sigma}}'$ and $\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'}$ commute, we have $f_{\mathtt{M}}(t) = \|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\|^2_{\widehat{\boldsymbol{\Sigma}}'}$ which is minimised at $t^* = 0$, proving (ii).

These two cases combine to give (iii).

For the $\boldsymbol{H}'$ norm case, note that $\mathcal{I}' = \emptyset$ already implies that $e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta} = \boldsymbol{0}$, which proves (i), and $\mathcal{I} = \emptyset$ implies that $e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}} = \boldsymbol{0}$, proving (ii). These implications are due to the fact that $\boldsymbol{H}'$ is positive definite (unlike $\widehat{\boldsymbol{\Sigma}}'$), so $\|\boldsymbol{v}\|_{\boldsymbol{H}'} = 0$ if and only if $\boldsymbol{v} = \boldsymbol{0}$. ■

C.4 Proof of Theorem 6

We split the proof into several steps. Throughout the Steps 1-6 of the proof, we will work with the interpolator of $f_{\boldsymbol{H}'}$, which we will define by overloading

$$f_{\boldsymbol{H}'} : [0, T] \to [0, \infty), \quad f_{\boldsymbol{H}'}(t) = \left\|(\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}} - e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|^2_{\boldsymbol{H}'}.$$

We will revert back to the original discrete $f_{\boldsymbol{H}'} : [T] \to [0, \infty)$ at the end, in Step 7.

Step 1: Decomposing $f'_{\boldsymbol{H}'}$: Define

$$\boldsymbol{g}(t) := (\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}} - e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\,.$$

Recall that, as $\boldsymbol{H}'$ is positive definite, $i \notin \mathcal{I}$ or $j \notin \mathcal{I}'$ implies that $\boldsymbol{P}_i\widehat{\boldsymbol{w}} = 0$ or $\boldsymbol{P}'_j\boldsymbol{\Delta} = 0$, respectively. Using the spectral transformations of $\boldsymbol{G}$ and $\boldsymbol{G}'$, we have

$$\begin{aligned} \boldsymbol{g}(t) &= \sum_{i\in\mathcal{I}} e^{-\mu_i t}\boldsymbol{P}_i\widehat{\boldsymbol{w}} - \sum_{i\in\mathcal{I}}\sum_{j=1}^{r'} e^{-\mu_i t - \mu'_j(T-t)}\boldsymbol{P}'_j\boldsymbol{P}_i\widehat{\boldsymbol{w}} - \sum_{j\in\mathcal{I}'} e^{-\mu'_j(T-t)}\boldsymbol{P}'_j\boldsymbol{\Delta}, \\ \boldsymbol{g}'(t) &= -\sum_{i\in\mathcal{I}} \mu_i e^{-\mu_i t}\boldsymbol{P}_i\widehat{\boldsymbol{w}} + \sum_{i\in\mathcal{I}}\sum_{j=1}^{r'}(\mu_i - \mu'_j)e^{-\mu_i t - \mu'_j(T-t)}\boldsymbol{P}'_j\boldsymbol{P}_i\widehat{\boldsymbol{w}} - \sum_{j\in\mathcal{I}'}\mu'_j e^{-\mu'_j(T-t)}\boldsymbol{P}'_j\boldsymbol{\Delta}. \end{aligned}$$

Further, note that $f'_{\boldsymbol{H}'}(t) = 2g(t)^\top\boldsymbol{H}'g'(t)$.

For ease of presentation, we use the following notations:

$$c^* := \frac{\mu'_{j^*}}{\mu_{i^*} + \mu'_{j^*}}, \quad t_0 := c^*T, \quad m^* := \frac{\mu_{i^*}\mu'_{j^*}}{\mu_{i^*} + \mu'_{j^*}}, \quad \kappa^* := e^{-2m^*T}\,.$$

Using the transformation $t = t_0 + s$ with $s \in [-t_0, T - t_0]$, we can decompose

$$\begin{aligned}
\boldsymbol{g}(t_0+s) &= e^{-\mu_{i^*}(t_0+s)}\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}} - e^{-\mu'_{j^*}(T-(t_0+s))}\boldsymbol{P}'_{j^*}\boldsymbol{\Delta} + \sum_{i\in\mathcal{I}\setminus\{i^*\}} e^{-\mu_i(t_0+s)}\boldsymbol{P}_i\widehat{\boldsymbol{w}} \\
&\quad - \sum_{i\in\mathcal{I}}\sum_{j=1}^{r'} e^{-\mu_i(t_0+s)-\mu'_j(T-(t_0+s))}\boldsymbol{P}'_j\boldsymbol{P}_i\widehat{\boldsymbol{w}} - \sum_{j\in\mathcal{I}'\setminus\{j^*\}} e^{-\mu'_j(T-(t_0+s))}\boldsymbol{P}'_j\boldsymbol{\Delta} \\
&:= \underbrace{e^{-\mu_{i^*}(t_0+s)}\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}} - e^{-\mu'_{j^*}(T-(t_0+s))}\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}}_{\boldsymbol{g}_{\text{lead}}} + \boldsymbol{g}_{\text{rest}}(t_0+s), \\
\boldsymbol{g}'(t_0+s) &= \underbrace{-\mu_{i^*}e^{-\mu_{i^*}(t_0+s)}\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}} - \mu'_{j^*}e^{-\mu'_{j^*}(T-(t_0+s))}\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}}_{\boldsymbol{g}'_{\text{lead}}} + \boldsymbol{g}'_{\text{rest}}(t_0+s)
\end{aligned}$$

and further, taking the $\boldsymbol{H}'$-inner product between these two, we have

$$\begin{aligned}
\frac{f'_{\boldsymbol{H}'}(t_0+s)}{2} &= \boldsymbol{g}_{\text{lead}}(t_0+s)^\top\boldsymbol{H}'\boldsymbol{g}'_{\text{lead}}(t_0+s) \\
&\quad + \underbrace{(\boldsymbol{g}_{\text{rest}}^\top\boldsymbol{H}'\boldsymbol{g}'_{\text{lead}} + \boldsymbol{g}_{\text{lead}}^\top\boldsymbol{H}'\boldsymbol{g}'_{\text{rest}} + \boldsymbol{g}_{\text{rest}}^\top\boldsymbol{H}'\boldsymbol{g}'_{\text{rest}})(t_0}_{:=\kappa^*\cdot\Xi} + s),
\end{aligned}$$

where we have introduced, for notation simplicity, the function $\Xi : [-t_0, T - t_0] \to \mathbb{R}$.

Now, note that $\sqrt{\kappa^*} = e^{-m^*T} = e^{-\mu_{i^*}t_0} = e^{-\mu'_{j^*}(T-t_0)}$, so that

$$\begin{aligned}
\Gamma(s) &:= \frac{\boldsymbol{g}_{\text{lead}}(t_0+s)^\top\boldsymbol{H}'\boldsymbol{g}'_{\text{lead}}(t_0+s)}{\kappa^*} \\
&= \mu'_{j^*}\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|^2_{\boldsymbol{H}'} e^{2\mu'_{j^*}s} - \mu_{i^*}\left\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\right\|^2_{\boldsymbol{H}'} e^{-2\mu_{i^*}s} - (\mu'_{j^*}-\mu_{i^*})\widehat{\boldsymbol{w}}^\top\boldsymbol{P}_{i^*}\boldsymbol{H}'\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}e^{(\mu'_{j^*}-\mu_{i^*})s},
\end{aligned}$$

and further we have decomposed f' into (as we show below) a quadratic Γ in the exponential space plus a perturbation term Ξ:

$$\frac{f'_{\boldsymbol{H}'}(t_0+s)}{2\kappa^*} = \Gamma(s) + \Xi(s). \tag{5}$$

We next compute the single zero of Γ and show that around this zero the function Γ is increasing:

Step 2: Analysing $\Gamma(s)$: For brevity, we will use the notation

$$a^* = \left\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\right\|^2_{\boldsymbol{H}'}, \quad b^* := \left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|^2_{\boldsymbol{H}'}, \quad d^* := \widehat{\boldsymbol{w}}^\top\boldsymbol{P}_{i^*}\boldsymbol{H}'\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}. \tag{6}$$

With $q := e^{(\mu_{i^*}+\mu'_{j^*})s} > 0$, we have

$$e^{2\mu_{i^*}s}\Gamma(s) = \mu'_{j^*}b^*q^2 - (\mu'_{j^*} - \mu_{i^*})d^*q - \mu_{i^*}a^*.$$

This is a quadratic in $q > 0$. The product of the roots is negative, so the quadratic has exactly one positive root. Taking its $\log$ and dividing by $\mu_{i^*} + \mu'_{j^*}$ we arrive at the unique zero of Γ:

$$s^* = \frac{1}{\mu_{i^*}+\mu'_{j^*}}\ln\left(\frac{(\mu'_{j^*}-\mu_{i^*})d^* + \sqrt{(\mu'_{j^*}-\mu_{i^*})^2(d^*)^2 + 4\mu_{i^*}\mu'_{j^*}a^*b^*}}{2\mu'_{j^*}b^*}\right).$$

Next, using the fact that $\Gamma(s^*) = 0 \iff (\mu'_{j^*} - \mu_{i^*})d^*e^{(\mu'_{j^*}-\mu_{i^*})s^*} = \mu'_{j^*}b^*e^{2\mu'_{j^*}s^*} - \mu_{i^*}a^*e^{-2\mu_{i^*}s^*}$, we have

$$\begin{aligned}
\Gamma'(s^*) &= 2(\mu'_{j^*})^2b^*e^{2\mu'_{j^*}s^*} + 2(\mu_{i^*})^2a^*e^{-2\mu_{i^*}s^*} - (\mu'_{j^*}-\mu_{i^*})^2d^*e^{(\mu'_{j^*}-\mu_{i^*})s^*} \\
&= 2(\mu'_{j^*})^2b^*e^{2\mu'_{j^*}s^*} + 2(\mu_{i^*})^2a^*e^{-2\mu_{i^*}s^*} - (\mu'_{j^*}-\mu_{i^*})(\mu'_{j^*}b^*e^{2\mu'_{j^*}s^*} - \mu_{i^*}a^*e^{-2\mu_{i^*}s^*}) \\
&= (\mu_{i^*}+\mu'_{j^*})(\mu'_{j^*}b^*e^{2\mu'_{j^*}s^*} + \mu_{i^*}a^*e^{-2\mu_{i^*}s^*}) > 0.
\end{aligned}$$

So, there exists $S_0 > 0$ (independent of T) such that $m_{S_0} := \min_{|s-s^*|\leq S_0} \Gamma'(s) > 0$. In this interval $[s^* - S_0, s^* + S_0]$ the function Γ is thus increasing. It is negative on $[s^* - S_0, s^*)$ and positive on $(s^*, s^* + S_0]$. Moreover, one can pick T large enough to make sure $[t_0 + s^* - S_0, t_0 + s^* + S_0] \subseteq [0, T]$.

Next, we show that the additive perturbation term Ξ is small on any compact interval around s^*:

Step 3: Uniform bound for the perturbation term $\Xi(s)$ in any compact $[s^* - S, s^* + S]$: We show that for any $S > 0$ (so, in particular, for the previous choice of S_0), there exist $C_S > 0$ that may change from line to line, but stay independent of T and just absorb other constants, and $\delta > 0$ (also independent of T) such that for T large enough

$$\sup_{|s-s^*|\leq S} (|\Xi(s)| + |\Xi'(s)|) \leq C_S e^{-\delta T}.$$

Recalling the definition of Ξ from Eq. (5) and using Cauchy-Schwarz,

$$\begin{aligned} |\Xi| &= \frac{1}{\kappa^*} |\boldsymbol{g}_{\text{rest}}^\top \boldsymbol{H}' \boldsymbol{g}'_{\text{lead}} + \boldsymbol{g}_{\text{lead}}^\top \boldsymbol{H}' \boldsymbol{g}'_{\text{rest}} + \boldsymbol{g}_{\text{rest}}^\top \boldsymbol{H}' \boldsymbol{g}'_{\text{rest}}| \\ &\leq \frac{1}{\kappa^*} \left(\|\boldsymbol{g}_{\text{rest}}\|_{\boldsymbol{H}'} \|\boldsymbol{g}'_{\text{lead}}\|_{\boldsymbol{H}'} + \|\boldsymbol{g}_{\text{lead}}\|_{\boldsymbol{H}'} \|\boldsymbol{g}'_{\text{rest}}\|_{\boldsymbol{H}'} + \|\boldsymbol{g}_{\text{rest}}\|_{\boldsymbol{H}'} \|\boldsymbol{g}'_{\text{rest}}\|_{\boldsymbol{H}'} \right). \end{aligned} \tag{7}$$

We next bound each norm. For s with $|s - s^*| \leq S$, as $|s| \leq |s^*| + S := M_S$, we have

$$\left|e^{\pm \mu_i s}\right| = e^{\pm \mu_i s} \leq e^{\mu_i |s|} \leq e^{\mu_i M_S} \leq e^{\mu_r M_S} \leq e^{(\mu_r + \mu'_{r'}) M_S} =: C_S,$$

and similarly $|e^{\pm \mu'_j s}| \leq C_S$, $|e^{\pm(\mu'_j - \mu_i)s}| \leq C_S$.

Recall that $\boldsymbol{g}_{\text{lead}}(t_0 + s) = e^{-\mu_{i^*}(t_0+s)} \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} - e^{-\mu'_{j^*}(T-(t_0+s))} \boldsymbol{P}'_{j^*} \boldsymbol{\Delta}$, $t_0 = c^* T = \mu'_{j^*} T / (\mu_{i^*} + \mu'_{j^*})$ and the notations introduced in Eq. (6). Then, by a triangle inequality,

$$\begin{aligned} \|\boldsymbol{g}_{\text{lead}}(t_0 + s)\|_{\boldsymbol{H}'} &\leq e^{-\mu_{i^*} t_0} e^{-\mu_{i^*} s} \sqrt{a^*} + e^{-\mu'_{j^*}(T - t_0)} e^{\mu'_{j^*} s} \sqrt{b^*} \\ &= e^{-m^* T} e^{-\mu_{i^*} s} \sqrt{a^*} + e^{-m^* T} e^{\mu'_{j^*} s} \sqrt{b^*} \leq C_S e^{-m^* T}. \end{aligned} \tag{8}$$

A similar computation (by letting C_S absorb μ_{i^*} and μ'_{j^*}) gives $\|\boldsymbol{g}'_{\text{lead}}(t_0 + s)\|_{\boldsymbol{H}'} \leq C_S e^{-m^* T}$.

By the definition of μ_{i^*} and μ'_{j^*} , $\delta_1 := \min_{i \in \mathcal{I} \setminus \{i^*\}} (c^*(\mu_i - \mu_{i^*})) > 0$ (taken to be $+\infty$ if $\mathcal{I} = \{i^*\}$), $\delta_2 := \min_{j \in \mathcal{I}' \setminus \{j^*\}} ((1 - c^*)(\mu'_j - \mu'_{j^*})) > 0$ (taken to be $+\infty$ if $\mathcal{I}' = \{j^*\}$). Additionally, as $\boldsymbol{H}'$ is positive definite, $(1 - c^*)\mu'_1 > 0$. We take $\delta := \min(\delta_1, \delta_2, (1 - c^*)\mu'_1) > 0$.

Next, we deal with bounding $\boldsymbol{g}_{\text{rest}}$ and its derivative. However, for the derivative, we can simply absorb the eigenvalues into the constants, the analysis is otherwise identical. There are three types of contribution in it:

- From the first summation, terms with $i \neq i^*$: using the fact that $\mu_{i^*} c^* = m^*$, these terms can be bounded as

$$\begin{aligned} \left\| e^{-\mu_i (t_0+s)} \boldsymbol{P}_i \widehat{\boldsymbol{w}} \right\|_{\boldsymbol{H}'} &= e^{-\mu_i t_0} e^{-\mu_i s} \|\boldsymbol{P}_i \widehat{\boldsymbol{w}}\|_{\boldsymbol{H}'} \leq e^{-\mu_i c^* T} C_S \|\boldsymbol{P}_i \widehat{\boldsymbol{w}}\|_{\boldsymbol{H}'} \\ &= e^{-(\mu_i - \mu_{i^*}) c^* T} e^{-\mu_{i^*} c^* T} C_S \|\boldsymbol{P}_i \widehat{\boldsymbol{w}}\|_{\boldsymbol{H}'} \\ &\leq C_S e^{-\delta T} e^{-m^* T} \|\boldsymbol{P}_i \widehat{\boldsymbol{w}}\|_{\boldsymbol{H}'} = C_S e^{-(m^* + \delta) T}. \end{aligned}$$

  where the constant (independent of T) C_S absorbed $\|\boldsymbol{P}_i \widehat{\boldsymbol{w}}\|_{\boldsymbol{H}'}$.
- Similarly, from the third summation, the terms with $j \neq j^*$ can be bounded as

$$\left\| e^{-\mu'_j (T - (t_0+s))} \boldsymbol{P}'_j \boldsymbol{\Delta} \right\|_{\boldsymbol{H}'} \leq C_S e^{-(m^* + \delta) T}.$$

- For the second summation over both $i \in I$ and $j \in \{1, \ldots, r'\}$, the contribution of each pair (i, j) is:

$$\left\| e^{-\mu_i (t_0+s) - \mu'_j (T - (t_0+s))} \boldsymbol{P}'_j \boldsymbol{P}_i \widehat{\boldsymbol{w}} \right\|_{\boldsymbol{H}'} \leq C_S e^{-(\mu_i c^* + \mu'_j (1 - c^*)) T}.$$

  Now, note that

$$\mu_i c^* + \mu'_j (1 - c^*) \geq \mu_{i^*} c^* + \mu'_1 (1 - c^*) = m^* + \mu'_1 (1 - c^*) \geq m^* + \delta,$$

  so that the contribution for each pair (i, j) is again at most $C_S e^{-(\delta + m^*) T}$.

Putting everything together and absorbing the sum of all constants into a single constant C_S (which depends on r' and $|\mathcal{I}|$), we get that for any $s \in [s^* - S, s^* + S]$,

$$\|\boldsymbol{g}_{\text{rest}}(t_0 + s)\|_{\boldsymbol{H}'} \leq C_S e^{-(m^* + \delta)T} \quad \text{and} \quad \|\boldsymbol{g}'_{\text{rest}}(t_0 + s)\|_{\boldsymbol{H}'} \leq C_S e^{-(m^* + \delta)T}, \tag{9}$$

for some appropriate (independent of T, but containing $|\mathcal{I}|$ and r') constant C_S.

Using Eq. (7), we get that

$$|\Xi(s)| \leq \frac{C_S}{\kappa^*}(e^{-(2m^* + \delta)T} + e^{-(2m^* + \delta)T} + e^{-2(m^* + \delta)T}) \leq C_S e^{-\delta T}.$$

The bound for $\Xi'(s)$ is almost identical, with the only difference being that

$$\kappa^* \Xi' = 2(\boldsymbol{g}'_{\text{rest}})^\top \boldsymbol{H}' \boldsymbol{g}'_{\text{lead}} + \boldsymbol{g}_{\text{rest}}^\top \boldsymbol{H}' \boldsymbol{g}''_{\text{lead}} + (\boldsymbol{g}''_{\text{rest}})^\top \boldsymbol{H}' \boldsymbol{g}_{\text{lead}} + \|\boldsymbol{g}'_{\text{rest}}\|^2_{\boldsymbol{H}'} + \boldsymbol{g}_{\text{rest}}^\top \boldsymbol{H}' \boldsymbol{g}''_{\text{rest}},$$

and now the bounds for the second derivatives are (up to constants independent of T) identical.

Given this bound on the perturbation term Ξ on any compact interval around s^*, we show that it cannot perturb the zero of $\Gamma(s)$ by too much:

Step 4: Existence, uniqueness and localisation of the zero of $\Gamma(s) + \Xi(s)$ inside $[s^* - S_0, s^* + S_0]$: First, we show that $\Gamma(s) + \Xi(s)$ has a unique zero in $(s^* - S_0, s^* + S_0)$ for T large enough. Start by defining

$$\eta_{S_0} := \min\{-\Gamma(s^* - S_0), \Gamma(s^* + S_0)\}$$

and note that $\eta_{S_0} > 0$, as the derivative in $[s^* - S_0, s^* + S_0]$ is positive by the choice of S_0 and $\Gamma(s^*) = 0$. Moreover, pick T_0 large enough such that, for all $T \geq T_0$,

$$C_{S_0} e^{-\delta T} \leq \min\{\eta_{S_0}/2, m_{S_0}/2\},$$

where we recall that $m_{S_0} := \min_{|s - s^*| \leq S_0} \Gamma'(s) > 0$ and C_{S_0} is the constant from the previous Step 3 when instantiating $S := S_0$. This is possible as S_0 is independent of T. Then, for all $s \in [s^* - S_0, s^* + S_0]$, $|\Xi(s)| \leq \eta_{S_0}/2$ and $|\Xi'(s)| \leq m_{S_0}/2$. This implies that on that interval

$$\Gamma'(s) + \Xi'(s) \geq m_{S_0} - |\Xi'(s)| \geq m_{S_0}/2 > 0, \tag{10}$$

so if there is a zero in this interval, it is unique. And there is indeed a zero: at the endpoints

$$\begin{aligned}
&\Gamma(s^* - S_0) + \Xi(s^* - S_0) \leq -\eta_{S_0} + \eta_{S_0}/2 = -\eta_{S_0}/2 < 0 \quad \text{and} \\
&\Gamma(s^* + S_0) + \Xi(s^* + S_0) \geq \eta_{S_0} - \eta_{S_0}/2 = \eta_{S_0}/2 > 0.
\end{aligned}$$

and the existence follows by the Intermediate Value Theorem. Let z_T be this unique zero. Since $f'(s + t_0) = 2\kappa^*(\Gamma(s) + \Xi(s))$ is negative at the left end of the interval, 0 at z_T, positive at the right end, it follows that $z_T + t_0$ is the minimum of f in this interval.

Further, note that z_T satisfies $(\Gamma + \Xi)(z_T) = 0$ and recall that $\Gamma(s^*) = 0$. Then, by the Mean Value Theorem on the interval $I \subseteq [s^* - S_0, s^* + S_0]$ with endpoints z_T and s^*,

$$\begin{aligned}
|\Xi(s^*)| &= |(\Xi + \Gamma)(s^*) - (\Xi + \Gamma)(z_T)| \geq |s^* - z_T| \inf_{\xi \in \mathcal{I}} |\Gamma'(\xi) + \Xi'(\xi)| \\
&\geq |s^* - z_T| \inf_{\xi \in [s^* - S_0, s^* + S_0]} |\Gamma'(\xi) + \Xi'(\xi)| \geq |s^* - z_T| m_{S_0}/2,
\end{aligned}$$

where the last inequality follows from Eq. (10). Next, recall that $|\Xi(s^*)| \leq C_{S_0} e^{-\delta T}$, so that

$$|z_T - s^*| \leq \frac{2C_{S_0}}{m_{S_0}} e^{-\delta T}.$$

Defining $\bar{t}_T := t_0 + z_T$, we arrive at

$$|\bar{t}_T - (t_0 + s^*)| \leq \frac{2C_{S_0}}{m_{S_0}} e^{-\delta T},$$

and it remains to show that this is indeed the global minimum of the interpolator $f_{\boldsymbol{H}'}$, having previously argued it is a minimum in $[t_0 + s^* - S_0, t_0 + s^* + S_0]$.

We do this by showing that:

- (Step 5) the global minimum must be in some (other) compact interval $t_0 + [S_-, S_+]$ around $t_0 + s^*$;
- (Step 6) The function $s \mapsto f'_{\boldsymbol{H}'}(t_0+s) = 2\kappa^*(\Gamma(s)+\Xi(s))$ changes sign exactly once on $[S_-, S_+]$.

Step 5: Proving the global minimum lies inside $t_0 + [S_-, S_+]$ for some $S_- < 0$, $S_- < s^* < S_+$: We show that the minimum of f has to lie in an interval containing $t_0 + s^*$ and not close to the endpoints 0 or T of the domain of f. To do so, note that $\min_{t\in[0,T]} f_{\boldsymbol{H}'}(t) \leq f_{\boldsymbol{H}'}(t_0 + s^*)$, which we estimate, and then argue that around 0 and T the function f takes a higher value.

Upper-bounding $f_{\boldsymbol{H}'}(t_0 + s^*)$: Recall that $f_{\boldsymbol{H}'}(t_0 + s^*) = \|\boldsymbol{g}(t_0 + s^*)\|^2_{\boldsymbol{H}'}$. For later use in Step 7, define

$$\hat{t}_T \in \arg\min_{q\in[T]} |q - (t_0 + s^*)|, \quad \hat{s}_T := \hat{t}_T - t_0.$$

For a sufficiently large T, we have $|\hat{s}_T - s^*| \leq 1/2$.

Now, we can apply the bounds derived in Step 3. To be more precise, consider $S = 1$, then, according to the triangle inequality, Eq. (8) and Eq. (9):

$$\begin{aligned}
\sup_{|s-s^*|\leq 1} f_{\boldsymbol{H}'}(t_0 + s) &= \sup_{|s-s^*|\leq 1} \|\boldsymbol{g}(t_0 + s)\|^2_{\boldsymbol{H}'} \\
&\leq \sup_{|s-s^*|\leq 1} (\|\boldsymbol{g}_{\text{lead}}(t_0 + s)\|_{\boldsymbol{H}'} + \|\boldsymbol{g}_{\text{rest}}(t_0 + s)\|_{\boldsymbol{H}'})^2 \\
&\leq C_1^2(e^{-m^*T} + e^{-(m^*+\delta)T})^2 \leq 2C_1^2 e^{-2m^*T} = 2C_1^2\kappa^*.
\end{aligned}$$

Importantly, the constant does not depend on T. Since both $s := s^*$ and $s := \hat{s}_T$ satisfy $|s - s^*| \leq 1$, this inequality gives both

$$f_{\boldsymbol{H}'}(t_0 + s^*) \leq 2C_1^2\kappa^*, \tag{11}$$

and

$$f_{\boldsymbol{H}'}(\hat{t}_T) \leq 2C_1^2\kappa^*. \tag{12}$$

Left-tail lower bound: We argue that for a certain S_- (independent of T) and all T large enough, i.e. $T \geq T_-(S_-)$ depending on S_-, we have for all $s \leq S_-$

$$f_{\boldsymbol{H}'}(t_0 + s) \geq \frac{\gamma' \|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\|_2^2 e^{-2\mu_{i^*}S_-}}{4}\kappa^*.$$

Since $\boldsymbol{H}' = \widehat{\boldsymbol{\Sigma}}' + \gamma'\boldsymbol{I} \succeq \gamma' I$ and $f_{\boldsymbol{H}'}(t_0 + s) = \|\boldsymbol{g}(t)\|^2_{\boldsymbol{H}'} \geq \gamma' \|\boldsymbol{g}(t)\|_2^2$, it's in fact enough to show that

$$\|\boldsymbol{g}(t_0 + s)\|_2^2 \geq \frac{\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\|_2^2 e^{-2\mu_{i^*}S_-}}{4}\kappa^*. \tag{13}$$

By the triangle inequality applied twice and the definition of the operator norm,

$$\begin{aligned}
\|\boldsymbol{g}(t)\|_2 &\geq \left\|(\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2 - \left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2 \\
&\geq \left(\left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2 - \left\|e^{-(T-t)\boldsymbol{G}'}e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2\right) - \left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2 \\
&\geq \left(\left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2 - \left\|e^{-(T-t)\boldsymbol{G}'}\right\|_2 \left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2\right) - \left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2 \\
&= \left(1 - \left\|e^{-(T-t)\boldsymbol{G}'}\right\|_2\right) \left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2 - \left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2.
\end{aligned}$$

Now, consider $S_- < s^*$ to be fixed later and recall that we translated $t = t_0 + s$. For $s \leq S_-$, $T - (t_0 + s) \geq (1 - c^*)T - S_- \to \infty$ as $T \to \infty$, so we can pick T large enough (and dependent on S_-) such that

$$\left\|e^{-(T-t)\boldsymbol{G}'}\right\|_2 \leq \frac{1}{4} \quad \text{for all } s \leq S_-,$$

and then we get that for any $s \leq S_-$,

$$\|\boldsymbol{g}(t_0+s)\|_2 \geq \frac{3}{4}\left\|e^{-(t_0+s)\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2 - \left\|e^{-(T-(t_0+s))\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2. \tag{14}$$

We now bound the two norms:

- Because $\boldsymbol{P}_i$ are orthogonal projectors, the vectors $\boldsymbol{P}_i\widehat{\boldsymbol{w}}$ are orthogonal and hence

$$\begin{aligned}\left\|e^{-(t_0+s)\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2^2 &= \left\|\sum_{i\in\mathcal{I}} e^{-\mu_i(t_0+s)}\boldsymbol{P}_i\widehat{\boldsymbol{w}}\right\|_2^2 = \sum_{i\in\mathcal{I}} e^{-2\mu_i(t_0+s)}\left\|\boldsymbol{P}_i\widehat{\boldsymbol{w}}\right\|_2^2 \\ &\geq e^{-2\mu_{i^*}(t_0+s)}\left\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\right\|_2^2,\end{aligned}$$

  so that

$$\left\|e^{-(t_0+s)\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2 \geq e^{-\mu_{i^*}(t_0+s)}\left\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\right\|_2 = e^{-m^*T}e^{-\mu_{i^*}s}\left\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\right\|_2 \tag{15}$$

- Since $\boldsymbol{P}'_j\boldsymbol{\Delta} = 0$ for $j \notin \mathcal{I}'$ and μ'_{j^*} is the smallest eigenvalue over $\mathcal{I}'$,

$$\left\|e^{-(T-(t_0+s))\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2 \leq e^{-\mu'_{j^*}(T-t)}\left\|\boldsymbol{\Delta}\right\|_2$$

Then, for all $s \leq S_-$,

$$\frac{\left\|e^{-(T-(t_0+s))\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2}{\left\|e^{-(t_0+s)\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2} \leq e^{s(\mu'_{j^*}+\mu_{i^*})}\frac{\left\|\boldsymbol{\Delta}\right\|_2}{\left\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\right\|_2} \leq e^{S_-(\mu'_{j^*}+\mu_{i^*})}\frac{\left\|\boldsymbol{\Delta}\right\|_2}{\left\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\right\|_2}. \tag{16}$$

Hence, we can choose $S_- < s^*$ negative enough so that the right-hand side of Eq. (16) is $\leq 1/4$, from which

$$\left\|e^{-(T-(t_0+s))\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2 \leq \frac{1}{4}\left\|e^{-(t_0+s)\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2 \quad \text{for all } s \leq S_-.$$

Combining this with Eq. (14) and Eq. (15), we get

$$\|\boldsymbol{g}(t_0+s)\|_2 \geq \frac{1}{2}\left\|e^{-(t_0+s)\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2 \geq \frac{1}{2}e^{-m^*T}e^{-\mu_{i^*}s}\left\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\right\|_2,$$

and, finally, as $\kappa^* = e^{-2m^*T}$ and $f_{\boldsymbol{H}'}(t_0+s) = \|\boldsymbol{g}(t_0+s)\|^2_{\boldsymbol{H}'}$, we arrive at Eq. (13) by squaring the above.

Right-tail lower-bound: We argue that for a certain S_+ independent of T, we have for all $s \geq S_+$

$$f_{\boldsymbol{H}'}(t_0+s) \geq \frac{9}{16}\gamma'\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_2^2 e^{2\mu'_{j^*}S_+}\kappa^*.$$

Similarly to the left-tail lower bound, it's in fact enough to show

$$\|\boldsymbol{g}(t_0+s)\|_2^2 \geq \frac{9}{16}\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_2^2 e^{2\mu'_{j^*}S_+}\kappa^*. \tag{17}$$

For this, by the triangle inequality and the definition of the spectral norm,

$$\begin{aligned}\|\boldsymbol{g}(t)\|_2 &\geq \left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2 - \left\|(\boldsymbol{I}-e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2 \\ &\geq \left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2 - \left\|\boldsymbol{I}-e^{-(T-t)\boldsymbol{G}'}\right\|_2\left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2 \\ &\geq \left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2 - \left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2.\end{aligned}$$

We now bound the two norms:

- Since $e^{-\mu'_{j^*}(T-t)}\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}$ is an orthogonal component of $e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}$:

$$\left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2 \geq e^{-\mu'_{j^*}(T-t)}\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_2. \tag{18}$$

- Since $\boldsymbol{P}_i\widehat{\boldsymbol{w}} = 0$ for $i \notin \mathcal{I}$ and μ_{i^*} is the smallest eigenvalue over $\mathcal{I}$,

$$\left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2 \leq e^{-t\mu_{i^*}}\left\|\widehat{\boldsymbol{w}}\right\|_2.$$

Putting these together and recalling that $t = t_0 + s$ with $\mu_{i^*}t_0 = \mu'_{j^*}(T - t_0)$, we obtain that for all $s \geq S_+$,

$$\frac{\left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2}{\left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2} \leq \frac{e^{-\mu_{i^*}t}\left\|\widehat{\boldsymbol{w}}\right\|_2}{e^{-\mu'_{j^*}(T-t)}\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_2} = \frac{e^{-(\mu_{i^*}+\mu'_{j^*})s}\left\|\widehat{\boldsymbol{w}}\right\|_2}{\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_2} \leq \frac{e^{-(\mu_{i^*}+\mu'_{j^*})S_+}\left\|\widehat{\boldsymbol{w}}\right\|_2}{\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_2}. \tag{19}$$

Consequently, we can pick $S_+ > s^*$ large enough so that the right-hand side of Eq. (19) is $\leq 1/4$. Then, $\left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2 \leq \left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2/4$ so that, together with Eq. (18),

$$\|\boldsymbol{g}(t)\|_2 \geq \frac{3}{4}\left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_2 \geq \frac{3}{4}e^{-\mu'_{j^*}(T-t)}\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_2.$$

Squaring this and using $f_{\boldsymbol{H}'}(t) = \|\boldsymbol{g}(t)\|^2_{\boldsymbol{H}'} \geq \gamma'\|\boldsymbol{g}(t)\|_2^2$ and $e^{-2\mu'_{j^*}(T-t)} = \kappa^* e^{2\mu'_{j^*}s}$, we arrive at Eq. (17).

Putting everything together: By decreasing S_- and increasing S_+, if necessary, we have proved the following: there are $S_- < s^* - 1 < s^* + 1 < S_+$ independent of T and T large enough such that

- $f_{\boldsymbol{H}'}(t_0 + s^*) \leq 2C_1^2\kappa^*$;
- $f_{\boldsymbol{H}'}(t_0 + s) \geq (\gamma'\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\|_2^2\, e^{-2\mu_{i^*}S_-}/4)\kappa^*$ for all $s \leq S_-$;
- $f_{\boldsymbol{H}'}(t_0 + s) \geq (9\gamma'\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_2^2\, e^{2\mu'_{j^*}S_+}/16)\kappa^*$ for all $s \geq S_+$.

Recall that C_1 is a constant independent of $S_\pm$ or T. We can accordingly increase S_+ and decrease (and make negative) S_- such that

$$\gamma'\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\|_2^2\, e^{-2\mu_{i^*}S_-}/4 \geq 3C_1^2 \quad \text{and} \quad 9\gamma'\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_2^2 e^{2\mu'_{j^*}S_+}/16 \geq 3C_1^2.$$

So, for these choices of S_-, S_+ and sufficiently large T depending on the chosen S_-, S_+, we have

$$\begin{aligned}
&\min_{S_- \leq s \leq S_+} f_{\boldsymbol{H}'}(t_0 + s) \leq f_{\boldsymbol{H}'}(t_0 + s^*) \leq 2C_1^2\kappa^* \leq f_{\boldsymbol{H}'}(t_0 + s) && \text{for all } s \leq S_-,\\
&\min_{S_- \leq s \leq S_+} f_{\boldsymbol{H}'}(t_0 + s) \leq f_{\boldsymbol{H}'}(t_0 + s) && \text{for all } s \geq S_+.
\end{aligned}$$

Accordingly, any global minimiser of $f_{\boldsymbol{H}'}$ must be of the form $t_0 + s$ with $s \in [S_-, S_+]$. Note that, as $t_0 = \mu'_{j^*}T/(\mu_{i^*} + \mu'_{j^*})$, we can enlarge T such that this minimiser is indeed in $[0, T]$, i.e. $t_0 + S_- \geq 0$, $t_0 + S_+ \leq T$.

Moreover, note that the discrete $\hat{t}_T \in [S_- + t_0, S_+ + t_0]$, and thus, using Eq. (12),

$$\min_{\substack{q \in [T] \\ t_0 + S_- \leq q \leq t_0 + S_+}} f_{\boldsymbol{H}'}(q) \leq f_{\boldsymbol{H}'}(\hat{t}_T) \leq 2C_1^2\kappa^*,$$

whereas from the tail bounds we have that for every $q \in [T]$ outside $[t_0 + S_-, t_0 + S_+]$,

$$f_{\boldsymbol{H}'}(q) \geq 3C_1^2\kappa^*.$$

Therefore, the set of global minimisers of the discrete $f_{\boldsymbol{H}'}$ satisfies

$$\mathcal{T}^* \subseteq [T] \cap [t_0 + S_-, t_0 + S_+]. \tag{20}$$

Step 6: Uniqueness of the global minimiser: Recall that we are still working with the interpolator $f_{\boldsymbol{H}'} : [0, T] \to [0, \infty)$. Also, recall the definition of S_0 from Step 2. Now, decrease S_- and increase

S_+, if needed, such that $S_- < s^* - S_0 - 1 < s^* + S_0 + 1 < S_+$, and furthermore consider $R > 0$ for which $[S_-, S_+] \subseteq [-R, R]$. Additionally, we can increase T until $0 < t_0 - R < t_0 + R < T$, ensuring thus that any global minimiser in $[S_- + t_0, S_+ + t_0]$ is an interior point of the domain $[0, T]$.

By Step 3, we have

$$\sup_{-R \le s \le R} |\Xi(s)| \le C_R e^{-\delta T}.$$

Note that, as $\Gamma(s) = 0$ has a unique solution at s^*, on $[-R, R] \setminus [s^* - S_0, s^* + S_0]$ the function Γ does not have a zero and moreover

$$\min\{|\Gamma(s)| : s \in [-R, s^* - S_0] \cup [s^* + S_0, R]\} > 0,$$

which is independent of T. So, we can take T large enough so that

$$\sup_{-R \le s \le R} |\Xi(s)| \le \frac{1}{2} \min\{|\Gamma(s)| : s \in [-R, s^* - S_0] \cup [s^* + S_0, R]\}.$$

Accordingly, for every $s \in [-R, R] \setminus [s^* - S_0, s^* + S_0]$,

$$\text{sign}(\Xi(s) + \Gamma(s)) = \text{sign}(\Gamma(s)),$$

as $|\Xi| \le |\Gamma|/2$ implies that addition of Ξ cannot change the sign of Γ. But, as we showed in Step 2, s^* is the unique root of Γ and $\Gamma(s) < 0$ for $s < s^*$, $\Gamma(s) > 0$ for $s > s^*$, so there is no root of $\Xi + \Gamma$ in $[-R, R] \setminus [s^* - S_0, s^* + S_0]$.

By Step 4, z_T is the unique root of $\Gamma + \Xi$ in $[s^* - S_0, s^* + S_0]$, while the above argument shows that there are no roots in $[S_-, S_+] \setminus [s^* - S_0, s^* + S_0]$. Hence, $t_0 + z_T$ is the unique stationary point of $f_{\boldsymbol{H}'}$ in $[S_- + t_0, S_+ + t_0]$. But, by Step 5, every global minimiser lies within this interval and, since the interval is a subset of $(0, T)$, every such minimiser must be stationary. Thus, $\bar{t}_T = t_0 + z_T$ is the unique global minimiser of the interpolator $f_{\boldsymbol{H}'}$.

Step 7: Localisation of the discrete global minimisers. By Step 5 Eq. (20), every $t^* \in \mathcal{T}^*$ lies in $[t_0 + S_-, t_0 + S_+]$. From Steps 4 and 6, the interpolator $f_{\boldsymbol{H}'}$ is strictly decreasing on $[t_0 + S_-, \bar{t}_T]$, and strictly increasing on $[\bar{t}_T, t_0 + S_+]$. Thus, its minimum over integer points in this interval can occur only at the two integers adjacent to $\bar{t}_T$, i.e.

$$\mathcal{T}^* \subseteq \{\lfloor \bar{t}_T \rfloor, \lceil \bar{t}_T \rceil\}.$$

If $\bar{t}_T$ is an integer, these two points coincide. Hence, $\mathcal{T}^*$ contains at most two elements. Moreover, for every $t^* \in \mathcal{T}^*$, by the triangle inequality

$$|t^* - (t_0 + s^*)| \le |t^* - \bar{t}_T| + |\bar{t}_T - (t_0 + s^*)| \le 1 + \frac{2C_{S_0}}{m_{S_0}} e^{-\delta T}.$$

Absorbing $2C_{S_0}/m_{S_0}$ into C and recalling $t_0 = \mu'_{j^*} T/(\mu_{i^*} + \mu'_{j^*})$ gives the stated conclusion. ■

C.5 Proof of Theorem 7

Recall that we write $f_{\boldsymbol{H}'}(t) = \|\boldsymbol{g}(t)\|^2_{\boldsymbol{H}'}$, where

$$\boldsymbol{g}(t) = (\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'}) e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} - e^{-(T-t)\boldsymbol{G}'} \boldsymbol{\Delta}.$$

Unlike in the Proof of Theorem 6, here $f_{\boldsymbol{H}'} : [T] \to [0, \infty)$, that is, we do not work again with a differentiable interpolator.

For the **lower-tail** $t \in [M]$, define

$$m_M = \min_{t \in [M]} \left\| e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\boldsymbol{H}'} ,$$

and note that, as $\boldsymbol{G}, \boldsymbol{H}'$ are both positive definite and $\widehat{\boldsymbol{w}} \neq 0$ as $\mathcal{I} \neq \emptyset$, it must be the case that $m_M > 0$. Now, using the eigendecomposition of $\boldsymbol{H}'$ and the fact that $\boldsymbol{H}' \succeq \gamma' \boldsymbol{I}$, for any $\boldsymbol{u} \in \mathbb{R}^d$:

$$\left\| e^{-(T-t)\boldsymbol{G}'} \boldsymbol{u} \right\|^2_{\boldsymbol{H}'} = \sum_{j=1}^{r'} e^{-2(T-t)\mu'_j} \left\| \boldsymbol{P}'_j \boldsymbol{u} \right\|^2_{\boldsymbol{H}'} \le e^{-2(T-t)\mu'_1} \|\boldsymbol{u}\|^2_{\boldsymbol{H}'} ,$$

and similarly

$$\left\|(\boldsymbol{I}-e^{-(T-t)\boldsymbol{G}'})\boldsymbol{u}\right\|_{\boldsymbol{H}'}^2 \geq (1-e^{-(T-t)\mu_1'})^2 \left\|\boldsymbol{u}\right\|_{\boldsymbol{H}'}^2.$$

Applying these two inequalities for $\boldsymbol{u}=e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}$ and $\boldsymbol{u}=\boldsymbol{\Delta}$, and using the triangle inequality, the definition of m_M and $T-t\geq T-M$,

$$\begin{aligned}\left\|\boldsymbol{g}(t)\right\|_{\boldsymbol{H}'} &\geq \left\|(\boldsymbol{I}-e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_{\boldsymbol{H}'} - \left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'} \\ &\geq (1-e^{-\mu_1'(T-t)})\left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_{\boldsymbol{H}'} - e^{-\mu_1'(T-t)}\left\|\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'} \\ &\geq (1-e^{-\mu_1'(T-M)})m_M - e^{-\mu_1'(T-M)}\left\|\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'}.\end{aligned}$$

We can pick T large enough so that

$$e^{-\mu_1'(T-M)} \leq \min\left(\frac{1}{8}, \frac{m_M}{4\left\|\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'}}\right),$$

for which we will have, for every $t\in[M]$

$$\left\|\boldsymbol{g}(t)\right\|_{\boldsymbol{H}'} \geq \left(1-\frac{1}{8}\right)m_M - \frac{1}{4}m_M = \frac{5}{8}m_M.$$

Thus, as $f_{\boldsymbol{H}'}(t)=\left\|\boldsymbol{g}(t)\right\|_{\boldsymbol{H}'}^2$, we arrive at

$$\min_{t\in[M]} f_{\boldsymbol{H}'}(t) \geq \frac{25}{64}m_M^2. \tag{21}$$

For the **upper-tail** $t\in\{T-M,\ldots,T\}$, we first start with the triangle inequality:

$$\left\|\boldsymbol{g}(t)\right\|_{\boldsymbol{H}'} \geq \left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'} - \left\|(\boldsymbol{I}-e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_{\boldsymbol{H}'}. \tag{22}$$

We bound each term on the right hand side individually. For the first term,

$$\begin{aligned}\left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'}^2 &= \sum_{j\in\mathcal{I}'} e^{-2(T-t)\mu_j'}\left\|\boldsymbol{P}_j'\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'}^2 \geq e^{-2(T-t)\mu_{j^*}'}\left\|\boldsymbol{P}_{j^*}'\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'}^2 \\ &\geq e^{-2\mu_{j^*}'M}\left\|\boldsymbol{P}_{j^*}'\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'}^2,\end{aligned}$$

so, after taking the square root,

$$\left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'} \geq e^{-\mu_{j^*}'M}\left\|\boldsymbol{P}_{j^*}'\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'}.$$

For the second term, denoting by $\left\|\boldsymbol{H}'\right\|_2$ the spectral norm of $\boldsymbol{H}'$, we proceed as

$$\begin{aligned}\left\|(\boldsymbol{I}-e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_{\boldsymbol{H}'} &\leq \left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_{\boldsymbol{H}'} \leq \left\|\boldsymbol{H}'\right\|_2^{1/2}\left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2 \\ &\leq \left\|\boldsymbol{H}'\right\|_2^{1/2} e^{-\mu_1(T-M)}\left\|\widehat{\boldsymbol{w}}\right\|_2.\end{aligned}$$

Substituting back into Eq. (22) we arrive at

$$\left\|\boldsymbol{g}(t)\right\|_{\boldsymbol{H}'} \geq e^{-\mu_{j^*}'M}\left\|\boldsymbol{P}_{j^*}'\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'} - \left\|\boldsymbol{H}'\right\|_2^{1/2} e^{-\mu_1(T-M)}\left\|\widehat{\boldsymbol{w}}\right\|_2.$$

Now, picking T large enough so that

$$\left\|\boldsymbol{H}'\right\|_2^{1/2} e^{-\mu_1(T-M)}\left\|\widehat{\boldsymbol{w}}\right\|_2 \leq \frac{1}{2}e^{-\mu_{j^*}'M}\left\|\boldsymbol{P}_{j^*}'\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'},$$

we get that

$$\left\|\boldsymbol{g}(t)\right\|_{\boldsymbol{H}'} \geq \frac{1}{2}e^{-\mu_{j^*}'M}\left\|\boldsymbol{P}_{j^*}'\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'},$$

and so

$$\min_{t\in\{T-M,\dots,T\}} f_{\boldsymbol{H}'}(t) \geq \frac{1}{4} e^{-2\mu'_{j^*}M} \left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|^2_{\boldsymbol{H}'}. \tag{23}$$

Combining the two tails from Eqs. (21) and (23), we have

$$\min\{f_{\boldsymbol{H}'}(t) : t \in \{0,\dots,M\}\cup\{T-M,\dots,T\}\} \geq \underbrace{\min\left(\frac{25}{64}m_M^2, \frac{1}{4}e^{-2\mu'_{j^*}M}\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|^2_{\boldsymbol{H}'}\right)}_{:=C_M}. \tag{24}$$

We next bound $f_{\boldsymbol{H}'}$ at an arbitrary feasible allocation $t \in [T]$:

Since the eigenvalues of $\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'}$ lie in $[0,1]$ and G' and H' have the same eigenspaces, for every $\boldsymbol{v} \in \mathbb{R}^d$,

$$\left\|(\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'})\boldsymbol{v}\right\|_{\boldsymbol{H}'} \leq \|\boldsymbol{v}\|_{\boldsymbol{H}'}.$$

Moreover, since $\boldsymbol{P}_i\widehat{\boldsymbol{w}} = 0$ for $i \notin \mathcal{I}$,

$$\left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_2^2 = \sum_{i\in\mathcal{I}} e^{-2\mu_i t}\left\|\boldsymbol{P}_i\widehat{\boldsymbol{w}}\right\|_2^2 \leq e^{-2\mu_{i^*}t}\left\|\widehat{\boldsymbol{w}}\right\|_2^2,$$

and similarly

$$\left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|^2_{\boldsymbol{H}'} = \sum_{j\in\mathcal{I}'} e^{-2\mu'_j(T-t)}\left\|\boldsymbol{P}'_j\boldsymbol{\Delta}\right\|^2_{\boldsymbol{H}'} \leq e^{-2\mu'_{j^*}(T-t)}\left\|\boldsymbol{\Delta}\right\|^2_{\boldsymbol{H}'}.$$

These three bounds, together with the triangle inequality in the expression of $g(t)$, give

$$\begin{aligned}\sqrt{f_{\boldsymbol{H}'}(t)} &\leq \left\|(\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_{\boldsymbol{H}'} + \left\|e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'} \\ &\leq e^{-\mu_{i^*}t}\left\|\boldsymbol{H}'\right\|_2^{1/2}\left\|\widehat{\boldsymbol{w}}\right\|_2 + e^{-\mu'_{j^*}(T-t)}\left\|\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'}.\end{aligned} \tag{25}$$

Now, we can equivalently rewrite $\widehat{t}$ in the theorem statement as $\widehat{t} = \widehat{\theta}T + \varepsilon$ with $|\varepsilon| \leq 1/2$. Substituting this into Eq. (25) and recalling that $m(\widehat{\theta}) = \min\{\mu_{i^*}\widehat{\theta}, \mu'_{j^*}(1-\widehat{\theta})\}$, we have

$$\begin{aligned}\sqrt{f_{\boldsymbol{H}'}(\widehat{t})} &\leq e^{-\mu_{i^*}\widehat{\theta}T - \mu_{i^*}\varepsilon}\left\|\boldsymbol{H}'\right\|_2^{1/2}\left\|\widehat{\boldsymbol{w}}\right\|_2 + e^{-\mu'_{j^*}(1-\widehat{\theta})T + \mu'_{j^*}\varepsilon}\left\|\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'} \\ &\leq e^{-m(\widehat{\theta})T}\left(\underbrace{\left\|\boldsymbol{H}'\right\|_2^{1/2}\left\|\widehat{\boldsymbol{w}}\right\|_2 e^{\mu_{i^*}/2} + \left\|\boldsymbol{\Delta}\right\|_{\boldsymbol{H}'} e^{\mu'_{j^*}/2}}_{:=C'}\right).\end{aligned}$$

Note that C' is independent of $\widehat{\theta}$ and T. Combining this bound with Eq. (24), we arrive at

$$f_{\boldsymbol{H}'}(\widehat{t}) \leq \frac{(C')^2}{C_M} e^{-2m(\widehat{\theta})T}\min\{f_{\boldsymbol{H}'}(t) : t \in \{0,\dots,M\}\cup\{T-M,\dots,T\}\},$$

and the inequality in the theorem statement follows by renaming $C_1 := (C')^2/C_M$.

Finally, the $\widehat{\theta}$ that maximises the rate $m(\widehat{\theta})$ equates the two terms in the minimum defining $m(\widehat{\theta})$. This is precisely $\widehat{\theta} = \theta^*$ defined in (2). ■

C.6 Proof of Theorem 12

Although this proof is similar to the Proof of Theorem 6, with the exception of Step 5, we provide it in full below.

We split the proof into several steps. Throughout the Steps 1-6 of the proof, we will work with the interpolator of $f_{\widehat{\boldsymbol{\Sigma}}'}$, which we will define by overloading

$$f_{\widehat{\boldsymbol{\Sigma}}'} : [0,T] \to [0,\infty), \quad f_{\widehat{\boldsymbol{\Sigma}}'}(t) = \left\| (\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'}) e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} - e^{-(T-t)\boldsymbol{G}'} \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'}^2 .$$

We will revert back to the original discrete $f_{\widehat{\boldsymbol{\Sigma}}'} : [T] \to [0,\infty)$ at the end, in Step 7.

Step 1: Decomposing $f'_{\widehat{\boldsymbol{\Sigma}}'}$: Define

$$\boldsymbol{g}(t) := (\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'}) e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} - e^{-(T-t)\boldsymbol{G}'} \boldsymbol{\Delta} .$$

Using the spectral transformations of $\boldsymbol{G}$ and $\boldsymbol{G}'$, we have

$$\begin{aligned}
\boldsymbol{g}(t) &= \sum_{i=1}^{r} e^{-\mu_i t} \boldsymbol{P}_i \widehat{\boldsymbol{w}} - \sum_{i=1}^{r} \sum_{j=1}^{r'} e^{-\mu_i t - \mu'_j (T-t)} \boldsymbol{P}'_j \boldsymbol{P}_i \widehat{\boldsymbol{w}} - \sum_{j=1}^{r'} e^{-\mu'_j (T-t)} \boldsymbol{P}'_j \boldsymbol{\Delta} , \\
\boldsymbol{g}'(t) &= - \sum_{i=1}^{r} \mu_i e^{-\mu_i t} \boldsymbol{P}_i \widehat{\boldsymbol{w}} + \sum_{i=1}^{r} \sum_{j=1}^{r'} (\mu_i - \mu'_j) e^{-\mu_i t - \mu'_j (T-t)} \boldsymbol{P}'_j \boldsymbol{P}_i \widehat{\boldsymbol{w}} - \sum_{j=1}^{r'} \mu'_j e^{-\mu'_j (T-t)} \boldsymbol{P}'_j \boldsymbol{\Delta} .
\end{aligned}$$

Further, we have $f'_{\widehat{\boldsymbol{\Sigma}}'}(t) = 2\boldsymbol{g}(t)^\top \widehat{\boldsymbol{\Sigma}}' \boldsymbol{g}'(t)$.

For ease of presentation, we use the following notations:

$$c^* := \frac{\mu'_{j^*}}{\mu_{i^*} + \mu'_{j^*}}, \quad t_0 := c^* T, \quad m^* := \frac{\mu_{i^*} \mu'_{j^*}}{\mu_{i^*} + \mu'_{j^*}}, \quad \kappa^* := e^{-2m^* T} .$$

Using the transformation $t = t_0 + s$ with $s \in [-t_0, T - t_0]$, we can decompose

$$\begin{aligned}
\boldsymbol{g}(t_0 + s) &= e^{-\mu_{i^*}(t_0+s)} \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} - e^{-\mu'_{j^*}(T-(t_0+s))} \boldsymbol{P}'_{j^*} \boldsymbol{\Delta} + \sum_{i \neq i^*, i=1}^{r} e^{-\mu_i (t_0+s)} \boldsymbol{P}_i \widehat{\boldsymbol{w}} \\
&\quad - \sum_{i=1}^{r} \sum_{j=1}^{r'} e^{-\mu_i (t_0+s) - \mu'_j (T-(t_0+s))} \boldsymbol{P}'_j \boldsymbol{P}_i \widehat{\boldsymbol{w}} - \sum_{j \neq j^*, j=1}^{r'} e^{-\mu'_j (T-(t_0+s))} \boldsymbol{P}'_j \boldsymbol{\Delta} \\
&:= \underbrace{e^{-\mu_{i^*}(t_0+s)} \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} - e^{-\mu'_{j^*}(T-(t_0+s))} \boldsymbol{P}'_{j^*} \boldsymbol{\Delta}}_{\boldsymbol{g}_{\text{lead}}} + \boldsymbol{g}_{\text{rest}}(t_0 + s), \\
\boldsymbol{g}'(t_0 + s) &= \underbrace{-\mu_{i^*} e^{-\mu_{i^*}(t_0+s)} \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} - \mu'_{j^*} e^{-\mu'_{j^*}(T-(t_0+s))} \boldsymbol{P}'_{j^*} \boldsymbol{\Delta}}_{\boldsymbol{g}'_{\text{lead}}} + \boldsymbol{g}'_{\text{rest}}(t_0 + s) ,
\end{aligned}$$

and further, taking the inner product between these two, we have

$$\begin{aligned}
\frac{f'_{\widehat{\boldsymbol{\Sigma}}'}(t_0 + s)}{2} &= \boldsymbol{g}_{\text{lead}}(t_0 + s)^\top \widehat{\boldsymbol{\Sigma}}' \boldsymbol{g}'_{\text{lead}}(t_0 + s) \\
&\quad + \underbrace{(\boldsymbol{g}_{\text{rest}}^\top \widehat{\boldsymbol{\Sigma}}' \boldsymbol{g}'_{\text{lead}} + \boldsymbol{g}_{\text{lead}}^\top \widehat{\boldsymbol{\Sigma}}' \boldsymbol{g}'_{\text{rest}} + \boldsymbol{g}_{\text{rest}}^\top \widehat{\boldsymbol{\Sigma}}' \boldsymbol{g}'_{\text{rest}})(t_0}_{:= \kappa^* \cdot \Xi} + s) ,
\end{aligned}$$

where we have introduced, for notation simplicity, the function $\Xi : [-t_0, T - t_0] \to \mathbb{R}$.

Now, note that $\sqrt{\kappa^*} = e^{-m^* T} = e^{-\mu_{i^*} t_0} = e^{-\mu'_{j^*}(T - t_0)}$, so that

$$\begin{aligned}
\Gamma(s) &:= \frac{\boldsymbol{g}_{\text{lead}}(t_0 + s)^\top \widehat{\boldsymbol{\Sigma}}' \boldsymbol{g}'_{\text{lead}}(t_0 + s)}{\kappa^*} \\
&= \mu'_{j^*} \left\| \boldsymbol{P}'_{j^*} \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'}^2 e^{2\mu'_{j^*} s} - \mu_{i^*} \left\| \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'}^2 e^{-2\mu_{i^*} s} - (\mu'_{j^*} - \mu_{i^*}) \widehat{\boldsymbol{w}}^\top \boldsymbol{P}_{i^*} \widehat{\boldsymbol{\Sigma}}' \boldsymbol{P}'_{j^*} \boldsymbol{\Delta} e^{(\mu'_{j^*} - \mu_{i^*}) s},
\end{aligned}$$

and further we have decomposed f' into (as we show below) a quadratic Γ in the exponential space plus a perturbation term Ξ:

$$\frac{f'_{\widehat{\boldsymbol{\Sigma}}'}(t_0 + s)}{2\kappa^*} = \Gamma(s) + \Xi(s). \tag{26}$$

We next compute the single zero of Γ and show that around this zero the function Γ is increasing:

Step 2: Analysing $\Gamma(s)$: Recall the notations

$$a^* = \|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'}^2, \quad b^* := \left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_{\widehat{\boldsymbol{\Sigma}}'}^2, \quad d^* := \widehat{\boldsymbol{w}}^\top \boldsymbol{P}_{i^*}\widehat{\boldsymbol{\Sigma}}'\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}. \tag{27}$$

With $q := e^{(\mu_{i^*}+\mu'_{j^*})s} > 0$, we have

$$e^{2\mu_{i^*}s}\Gamma(s) = \mu'_{j^*}b^*q^2 - (\mu'_{j^*} - \mu_{i^*})d^*q - \mu_{i^*}a^*.$$

This is a quadratic in $q > 0$. The product of the roots is negative, so the quadratic has exactly one positive root. Taking its log and dividing by $\mu_{i^*} + \mu'_{j^*}$ we arrive at the unique zero of Γ:

$$s^* = \frac{1}{\mu_{i^*} + \mu'_{j^*}} \ln\left(\frac{(\mu'_{j^*} - \mu_{i^*})d^* + \sqrt{(\mu'_{j^*} - \mu_{i^*})^2(d^*)^2 + 4\mu_{i^*}\mu'_{j^*}a^*b^*}}{2\mu'_{j^*}b^*}\right).$$

Next, using the fact that $\Gamma(s^*) = 0 \iff (\mu'_{j^*} - \mu_{i^*})d^*e^{(\mu'_{j^*}-\mu_{i^*})s^*} = \mu'_{j^*}b^*e^{2\mu'_{j^*}s^*} - \mu_{i^*}a^*e^{-2\mu_{i^*}s^*}$, we have

$$\begin{aligned}
\Gamma'(s^*) &= 2(\mu'_{j^*})^2b^*e^{2\mu'_{j^*}s^*} + 2(\mu_{i^*})^2a^*e^{-2\mu_{i^*}s^*} - (\mu'_{j^*} - \mu_{i^*})^2d^*e^{(\mu'_{j^*}-\mu_{i^*})s^*} \\
&= 2(\mu'_{j^*})^2b^*e^{2\mu'_{j^*}s^*} + 2(\mu_{i^*})^2a^*e^{-2\mu_{i^*}s^*} - (\mu'_{j^*} - \mu_{i^*})(\mu'_{j^*}b^*e^{2\mu'_{j^*}s^*} - \mu_{i^*}a^*e^{-2\mu_{i^*}s^*}) \\
&= (\mu_{i^*} + \mu'_{j^*})(\mu'_{j^*}b^*e^{2\mu'_{j^*}s^*} + \mu_{i^*}a^*e^{-2\mu_{i^*}s^*}) > 0\,.
\end{aligned}$$

So, there exists $S_0 > 0$ (independent of T) such that $m_{S_0} := \min_{|s-s^*|\le S_0}\Gamma'(s) > 0$. In this interval $[s^* - S_0, s^* + S_0]$ the function Γ is thus increasing. It is negative on $[s^* - S_0, s^*)$ and positive on $(s^*, s^*+S_0]$. Moreover, one can pick T large enough to make sure $[t_0+s^*-S_0, t_0+s^*+S_0] \subseteq [0, T]$.

Next, we show that the additive perturbation term Ξ is small on any compact interval around s^*:

Step 3: Uniform bound for the perturbation term $\Xi(s)$ in any compact $[s^* - S, s^* + S]$: We show that for any $S > 0$ (so, in particular, for the previous choice of S_0), there exist $C_S > 0$ that may change from line to line – but stay independent of T and just absorb other constants – and $\delta > 0$ (also independent of T) such that for T large enough

$$\sup_{|s-s^*|\le S}(|\Xi(s)| + |\Xi'(s)|) \le C_S e^{-\delta T}.$$

Recalling the definition of Ξ from Eq. (26) and using Cauchy-Schwarz inequality,

$$|\Xi| \le \frac{1}{\kappa^*}\left(\|\boldsymbol{g}_{\text{rest}}\|_{\widehat{\boldsymbol{\Sigma}}'}\|\boldsymbol{g}'_{\text{lead}}\|_{\widehat{\boldsymbol{\Sigma}}'} + \|\boldsymbol{g}_{\text{lead}}\|_{\widehat{\boldsymbol{\Sigma}}'}\|\boldsymbol{g}'_{\text{rest}}\|_{\widehat{\boldsymbol{\Sigma}}'} + \|\boldsymbol{g}_{\text{rest}}\|_{\widehat{\boldsymbol{\Sigma}}'}\|\boldsymbol{g}'_{\text{rest}}\|_{\widehat{\boldsymbol{\Sigma}}'}\right). \tag{28}$$

We next bound each $\widehat{\boldsymbol{\Sigma}}'$-seminorm. For s with $|s - s^*| \le S$, as $|s| \le |s^*| + S := M_S$, we have

$$\left|e^{\pm\mu_i s}\right| = e^{\pm\mu_i s} \le e^{\mu_i|s|} \le e^{\mu_i M_S} \le e^{\mu_r M_S} \le e^{(\mu_r+\mu'_{r'})M_S} =: C_S,$$

and similarly $|e^{\pm\mu'_j s}| \le C_S$, $|e^{\pm(\mu'_j-\mu_i)s}| \le C_S$.

Recall the notation $\boldsymbol{g}_{\text{lead}}(t_0 + s) = e^{-\mu_{i^*}(t_0+s)}\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}} - e^{-\mu'_{j^*}(T-(t_0+s))}\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}$ and $t_0 = c^*T = \mu'_{j^*}T/(\mu_{i^*} + \mu'_{j^*})$ and the notations introduced in Eq. (27). Then, by a triangle inequality,

$$\begin{aligned}
\|\boldsymbol{g}_{\text{lead}}(t_0 + s)\|_{\widehat{\boldsymbol{\Sigma}}'} &\le e^{-\mu_{i^*}t_0}e^{-\mu_{i^*}s}\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'} + e^{-\mu'_{j^*}(T-t_0)}e^{\mu'_{j^*}s}\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_{\widehat{\boldsymbol{\Sigma}}'} \\
&= e^{-m^*T}e^{-\mu_{i^*}s}\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'} + e^{-m^*T}e^{\mu'_{j^*}s}\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_{\widehat{\boldsymbol{\Sigma}}'} \le C_S e^{-m^*T}.
\end{aligned} \tag{29}$$

A similar computation (by letting C_S absorb μ_{i^*} and μ'_{j^*}) gives $\|\boldsymbol{g}'_{\text{lead}}(t_0 + s)\|_{\widehat{\boldsymbol{\Sigma}}'} \le C_S e^{-m^*T}$.

By the definition of μ_{i^*} and μ'_{j^*}, $\delta_1 := \min_{i\in\mathcal{I}\setminus\{i^*\}}(c^*(\mu_i - \mu_{i^*})) > 0$ (taken to be $+\infty$ if $\mathcal{I} = \{i^*\}$), $\delta_2 := \min_{j\in\mathcal{I}'\setminus\{j^*\}}((1 - c^*)(\mu'_j - \mu'_{j^*})) > 0$ (taken to be $+\infty$ if $\mathcal{I}' = \{j^*\}$). Additionally, as $\boldsymbol{H}'$ is positive definite, $(1 - c^*)\mu'_1 > 0$. We take $\delta := \min(\delta_1, \delta_2, (1 - c^*)\mu'_1) > 0$.

Next, we deal with bounding $\boldsymbol{g}_{\text{rest}}$ and its derivative. However, for the derivative, we can simply absorb the eigenvalues into the constants, the analysis is otherwise identical. There are three types of contribution in it:

- From the first summation, terms with $i \notin \mathcal{I}$ have zero seminorm. For $i \neq i^*$, $i \in \mathcal{I}$, using the fact that $\mu_{i^*} c^* = m^*$, these terms can be bounded as

$$\begin{aligned}\left\| e^{-\mu_i (t_0+s)} \boldsymbol{P}_i \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} &= e^{-\mu_i t_0} e^{-\mu_i s} \left\| \boldsymbol{P}_i \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \leq e^{-\mu_i c^* T} C_S \left\| P_i \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \\ &= e^{-(\mu_i - \mu_{i^*}) c^* T} e^{-\mu_{i^*} c^* T} C_S \left\| \boldsymbol{P}_i \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \\ &\leq C_S e^{-\delta T} e^{-m^* T} \left\| \boldsymbol{P}_i \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} = C_S e^{-(m^* + \delta) T}.\end{aligned}$$

  where the constant (independent of T) C_S absorbed $\left\| \boldsymbol{P}_i \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'}$.
- Similarly, from the third summation, the terms with $j \neq j^*$ can be bounded as

$$\left\| e^{-\mu'_j (T - (t_0+s))} \boldsymbol{P}'_j \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \leq C_S e^{-(m^* + \delta) T}.$$

- For the second summation over both $i \in \{1, \ldots, r\}$ and $j \in \{1, \ldots, r'\}$, the contribution of each pair (i, j) is:

$$\left\| e^{-\mu_i (t_0+s) - \mu'_j (T - (t_0+s))} \boldsymbol{P}'_j \boldsymbol{P}_i \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \leq C_S e^{-(\mu_i c^* + \mu'_j (1 - c^*)) T}.$$

  Now, note that if $i \notin \mathcal{I}$, the contribution of (i, j) is 0: as $\boldsymbol{P}'_j$ commutes with $\widehat{\boldsymbol{\Sigma}}'$, we have $\widehat{\boldsymbol{\Sigma}}' \boldsymbol{P}'_j \boldsymbol{P}_i \widehat{\boldsymbol{w}} = \boldsymbol{P}'_j \widehat{\boldsymbol{\Sigma}}' \boldsymbol{P}_i \widehat{\boldsymbol{w}} = \boldsymbol{0}$. Moreover, for $i \in \mathcal{I}$ we have

$$\mu_i c^* + \mu'_j (1 - c^*) \geq \mu_{i^*} c^* + \mu'_1 (1 - c^*) = m^* + \mu'_1 (1 - c^*) \geq m^* + \delta \,,$$

  so that the contribution for each pair (i, j) is again at most $C_S e^{-(\delta + m^*) T}$.

Putting everything together and absorbing the sum of all constants into a single constant C_S (which depends on r, r'), we get that for any $s \in [s^* - S, s^* + S]$,

$$\left\| \boldsymbol{g}_{\text{rest}}(t_0 + s) \right\|_{\widehat{\boldsymbol{\Sigma}}'} \leq C_S e^{-(m^* + \delta) T} \quad \text{and} \quad \left\| \boldsymbol{g}'_{\text{rest}}(t_0 + s) \right\|_{\widehat{\boldsymbol{\Sigma}}'} \leq C_S e^{-(m^* + \delta) T}, \tag{30}$$

for some appropriate (independent of T, but containing r and r') constant C_S.

Using Eq. (28), we get that

$$|\Xi(s)| \leq \frac{C_S}{\kappa^*} (e^{-(2m^* + \delta) T} + e^{-(2m^* + \delta) T} + e^{-2(m^* + \delta) T}) \leq C_S e^{-\delta T}.$$

The bound for $\Xi'(s)$ is almost identical, with the only difference being that

$$\kappa^* \Xi' = 2 (\boldsymbol{g}'_{\text{rest}})^\top \widehat{\boldsymbol{\Sigma}}' \boldsymbol{g}'_{\text{lead}} + \boldsymbol{g}_{\text{rest}}^\top \widehat{\boldsymbol{\Sigma}}' \boldsymbol{g}''_{\text{lead}} + (g''_{\text{rest}})^\top \widehat{\boldsymbol{\Sigma}}' g_{\text{lead}} + \left\| \boldsymbol{g}'_{\text{rest}} \right\|^2_{\widehat{\boldsymbol{\Sigma}}'} + \boldsymbol{g}_{\text{rest}}^\top \widehat{\boldsymbol{\Sigma}}' g''_{\text{rest}},$$

and now the bounds for the second derivatives are (up to constants independent of T) identical.

Given this bound on the perturbation term Ξ on any compact interval around s^*, we show that it cannot perturb the zero of $\Gamma(s)$ by too much:

Step 4: Existence, uniqueness and localisation of the zero of $\Gamma(s) + \Xi(s)$ inside $[s^* - S_0, s^* + S_0]$: First, we show that $\Gamma(s) + \Xi(s)$ has a unique zero in $(s^* - S_0, s^* + S_0)$ for T large enough. Start by defining

$$\eta_{S_0} := \min\{-\Gamma(s^* - S_0), \Gamma(s^* + S_0)\}$$

and note that $\eta_{S_0} > 0$, as the derivative in $[s^* - S_0, s^* + S_0]$ is positive by the choice of S_0 and $\Gamma(s^*) = 0$. Moreover, pick T_0 large enough such that, for all $T \geq T_0$,

$$C_{S_0} e^{-\delta T} \leq \min\{\eta_{S_0}/2, m_{S_0}/2\},$$

where we recall that $m_{S_0} := \min_{|s - s^*| \leq S_0} \Gamma'(s) > 0$ and C_{S_0} is the constant from the previous Step 3 when instantiating $S := S_0$. This is possible as S_0 is independent of T. Then, for all $s \in [s^* - S_0, s^* + S_0]$, $|\Xi(s)| \leq \eta_{S_0}/2$ and $|\Xi'(s)| \leq m_{S_0}/2$. This implies that on that interval

$$\Gamma'(s) + \Xi'(s) \geq m_{S_0} - |\Xi'(s)| \geq m_{S_0}/2 > 0, \tag{31}$$

so if there is a zero in this interval, it is unique. And there is indeed a zero: at the endpoints

$$\Gamma(s^* - S_0) + \Xi(s^* - S_0) \le -\eta_{S_0} + \eta_{S_0}/2 = -\eta_{S_0}/2 < 0 \quad \text{and}$$
$$\Gamma(s^* + S_0) + \Xi(s^* + S_0) \ge \eta_{S_0} - \eta_{S_0}/2 = \eta_{S_0}/2 > 0.$$

and the existence follows by the Intermediate Value Theorem. Let z_T be this unique zero. Since $f'(s + t_0) = 2\kappa^*(\Gamma(s) + \Xi(s))$ is negative at the left end of the interval, 0 at z_T, positive at the right end, it follows that $z_T + t_0$ is the minimum of f in this interval.

Further, note that z_T satisfies $(\Gamma + \Xi)(z_T) = 0$ and recall that $\Gamma(s^*) = 0$. Then, by the Mean Value Theorem on the interval $I \subseteq [s^* - S_0, s^* + S_0]$ with endpoints z_T and s^*,

$$|\Xi(s^*)| = |(\Xi + \Gamma)(s^*) - (\Xi + \Gamma)(z_T)| \ge |s^* - z_T| \inf_{\xi \in \mathcal{I}} |\Gamma'(\xi) + \Xi'(\xi)|$$
$$\ge |s^* - z_T| \inf_{\xi \in [s^* - S_0, s^* + S_0]} |\Gamma'(\xi) + \Xi'(\xi)| \ge |s^* - z_T| m_{S_0}/2,$$

where the last inequality follows from Eq. (31). Next, recall that $|\Xi(s^*)| \le C_{S_0} e^{-\delta T}$, so that

$$|z_T - s^*| \le \frac{2C_{S_0}}{m_{S_0}} e^{-\delta T}.$$

Defining $\bar{t}_T := t_0 + z_T$, we arrive at

$$|\bar{t}_T - (t_0 + s^*)| \le \frac{2C_{S_0}}{m_{S_0}} e^{-\delta T}.$$

It remains to locate the discrete global minimisers of $f_{\widehat{\mathbf{\Sigma}}'}$. We do this as follows: Step 5 proves that these are within $[t_0 + S_-, t_0 + S_+]$, Step 6 establishes the sign of $f'_{\widehat{\mathbf{\Sigma}}'}$ on this central interval, and finally Step 7 shows that these global minimisers lie at the integers adjacent to $\bar{t}_T$.

Step 5: Proving the discrete global minimum lies inside $t_0 + [S_-, S_+]$ for some $S_- < 0$, $S_- < s^* < S_+$:

Upper-bounding $f_{\widehat{\mathbf{\Sigma}}'}(t_0 + s)$ around s^*: Recall that $f_{\widehat{\mathbf{\Sigma}}'}(t_0 + s) = \|\boldsymbol{g}(t_0 + s)\|^2_{\widehat{\mathbf{\Sigma}}'}$. Define

$$\hat{t}_T \in \arg\min_{q \in [T]} |q - (t_0 + s^*)|, \quad \hat{s}_T := \hat{t}_T - t_0.$$

For a sufficiently large T, we have $|\hat{s}_T - s^*| \le 1/2$.

Now, we can apply the bounds derived in Step 3. To be more precise, consider $S = 1$, then, according to the triangle inequality, Eq. (29) and Eq. (30):

$$\begin{aligned}\sup_{|s - s^*| \le 1} f_{\widehat{\mathbf{\Sigma}}'}(t_0 + s) &= \sup_{|s - s^*| \le 1} \|\boldsymbol{g}(t_0 + s)\|^2_{\widehat{\mathbf{\Sigma}}'} \\ &\le \sup_{|s - s^*| \le 1} (\|\boldsymbol{g}_{\text{lead}}(t_0 + s)\|_{\widehat{\mathbf{\Sigma}}'} + \|\boldsymbol{g}_{\text{rest}}(t_0 + s)\|_{\widehat{\mathbf{\Sigma}}'})^2 \\ &\le C_1^2 (e^{-m^* T} + e^{-(m^* + \delta)T})^2 \le 2C_1^2 e^{-2m^* T} = 2C_1^2 \kappa^*.\end{aligned}$$

Importantly, the constant does not depend on T. Since $s := \hat{s}_T$ satisfies $|s - s^*| \le 1$, this inequality gives

$$f_{\widehat{\mathbf{\Sigma}}'}(\hat{t}_T) \le 2C_1^2 \kappa^*. \tag{32}$$

Left-tail lower bound: Note that if $t = t_0 + s \ge 0$, then $s \ge -t_0$. We argue that for a certain $S_- < s^*$ (independent of T) and all T large enough (depending on S_-) we have

$$f_{\widehat{\mathbf{\Sigma}}'}(t_0 + s) \ge \frac{81}{1024} e^{-2\mu_{i^*} S_-} \|\boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}}\|^2_{\widehat{\mathbf{\Sigma}}'} \kappa^*$$

for every integer $t_0 + s \in [T]$ with $s \le S_-$.

By the triangle inequality,

$$\|\boldsymbol{g}(t)\|_{\widehat{\mathbf{\Sigma}}'} \ge \left\| (\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'}) e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\widehat{\mathbf{\Sigma}}'} - \left\| e^{-(T-t)\boldsymbol{G}'} \boldsymbol{\Delta} \right\|_{\widehat{\mathbf{\Sigma}}'}. \tag{33}$$

We now bound each of the norms.

Since $\boldsymbol{H}' = \widehat{\boldsymbol{\Sigma}}' + \gamma' \boldsymbol{I}$ has eigenvalue/projector pairs $(\lambda'_j, \boldsymbol{P}'_j)_{j=1}^{r'}$, the eigenvalue/projector pairs for $\widehat{\boldsymbol{\Sigma}}'$ are $(\lambda'_j - \gamma', \boldsymbol{P}'_j)_{j=1}^{r'}$. Now, for an arbitrary vector $\boldsymbol{v}$,

$$\begin{aligned}
\left\| e^{-(T-t)\boldsymbol{G}'} \boldsymbol{v} \right\|_{\widehat{\boldsymbol{\Sigma}}'}^2 &= (e^{-(T-t)\boldsymbol{G}'} \boldsymbol{v})^\top \widehat{\boldsymbol{\Sigma}}' (e^{-(T-t)\boldsymbol{G}'} \boldsymbol{v}) \\
&= \left(\sum_{j=1}^{r'} e^{-\mu'_j (T-t)} \boldsymbol{P}'_j \boldsymbol{v} \right)^\top \left(\sum_{j=1}^{r'} (\lambda'_j - \gamma') \boldsymbol{P}'_j \right) \left(\sum_{j=1}^{r'} e^{-\mu'_j (T-t)} \boldsymbol{P}'_j \boldsymbol{v} \right) \\
&= \sum_{j=1}^{r'} (\lambda'_j - \gamma') e^{-2\mu'_j (T-t)} \boldsymbol{v}^\top \boldsymbol{P}'_j \boldsymbol{v} \qquad (34) \\
&\leq e^{-2\mu'_1 (T-t)} \sum_{j=1}^{r'} (\lambda'_j - \gamma') \boldsymbol{v}^\top \boldsymbol{P}'_j \boldsymbol{v} = e^{-2\mu'_1 (T-t)} \left\| \boldsymbol{v} \right\|_{\widehat{\boldsymbol{\Sigma}}'}^2 .
\end{aligned}$$

Now, applying this to $\boldsymbol{v} = e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}}$ and using the triangle inequality,

$$\begin{aligned}
\left\| (\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'}) e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} &\geq \left\| e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} - \left\| e^{-(T-t)\boldsymbol{G}'} e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \\
&\geq (1 - e^{-(T-t)\mu'_1}) \left\| e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} .
\end{aligned}$$

Next, consider $S_- < s^*$ to be fixed later and recall that we translated $t = t_0 + s$. For $s \leq S_-$, $T - (t_0 + s) \geq (1 - c^*)T - S_- \to \infty$ as $T \to \infty$. So, we can pick T large enough (and dependent on S_-) such that

$$\left\| (I - e^{-(T-t)\boldsymbol{G}'}) e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \geq \frac{3}{4} \left\| e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} . \qquad (35)$$

For the other norm in Eq. (33), continuing from Eq. (34) for $v = \boldsymbol{\Delta}$,

$$\begin{aligned}
\left\| e^{-(T-t)\boldsymbol{G}'} \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'}^2 &= \sum_{j=1}^{r'} (\lambda'_j - \gamma') e^{-2\mu'_j (T-t)} \boldsymbol{\Delta}^\top \boldsymbol{P}'_j \boldsymbol{\Delta} = \sum_{j=1}^{r'} e^{-2\mu'_j (T-t)} \left\| \boldsymbol{P}'_j \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'}^2 \\
&= \sum_{j \in \mathcal{I}'} e^{-2\mu'_j (T-t)} \left\| \boldsymbol{P}'_j \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'}^2 \qquad (36) \\
&\leq e^{-2\mu'_{j^*} (T-t)} \sum_{j \in \mathcal{I}'} \left\| \boldsymbol{P}'_j \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'}^2 = e^{-2\mu'_{j^*} (T-t)} \left\| \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'}^2 ,
\end{aligned}$$

from which $\| e^{-(T-t)\boldsymbol{G}'} \boldsymbol{\Delta} \|_{\widehat{\boldsymbol{\Sigma}}'} \leq e^{-\mu'_{j^*} (T-t)} \| \boldsymbol{\Delta} \|_{\widehat{\boldsymbol{\Sigma}}'}$. Substituting this and taking Eq. (35) back into Eq. (33), and using Eq. (36), we arrive at

$$\begin{aligned}
\| \boldsymbol{g}(t) \|_{\widehat{\boldsymbol{\Sigma}}'} &\geq \frac{3}{4} \left\| e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} - \left\| e^{-(T-t)\boldsymbol{G}'} \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \\
&\geq \frac{3}{4} \left\| e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} - e^{-\mu'_{j^*} (T-t)} \| \boldsymbol{\Delta} \|_{\widehat{\boldsymbol{\Sigma}}'} \qquad (37)
\end{aligned}$$

Now, we claim that when K is defined as in Appendix A (depending on $\boldsymbol{H}$, $\widehat{\boldsymbol{w}}$, $\widehat{\boldsymbol{\Sigma}}'$ but not T), for all $t \in \{K, \ldots, T\}$,

$$\left\| e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \geq \frac{3}{4} e^{-\mu_{i^*} t} \left\| \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} . \qquad (38)$$

If $\mathcal{I} = \{i^*\}$, this clearly holds with $K = 0$. Otherwise, defining $\xi := \min_{i \in \mathcal{I} \setminus \{i^*\}} (\mu_i - \mu_{i^*}) > 0$, by the triangle inequality,

$$\begin{aligned}
\left\| e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} &= \left\| \sum_{i=1}^{r} e^{-\mu_i t} \boldsymbol{P}_i \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \geq e^{-\mu_{i^*} t} \left\| \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} - \sum_{i \neq i^*, i=1}^{r} e^{-\mu_i t} \left\| P_i \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \\
&= e^{-\mu_{i^*} t} \left\| \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} - \sum_{i \in \mathcal{I} \setminus \{i^*\}}^{r} e^{-\mu_i t} \left\| P_i \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \\
&= e^{-\mu_{i^*} t} \left(\left\| \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} - \sum_{i \in \mathcal{I} \setminus \{i^*\}} e^{-(\mu_i - \mu_{i^*}) t} \left\| P_i \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \right) \\
&\geq e^{-\mu_{i^*} t} \left(\left\| \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} - e^{-\xi t} \sum_{i \in \mathcal{I} \setminus \{i^*\}} \left\| P_i \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \right).
\end{aligned}$$

By the definition of K in Appendix A, for any $t \in \{K, \ldots, T\}$, we have $e^{-\xi t} \sum_{i \in \mathcal{I} \setminus \{i^*\}} \left\| \boldsymbol{P}_i \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \leq \left\| \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} / 4$, and then Eq. (38) holds. We next consider two different regimes:

- $t \in [K]$: We use the assumption that

  $$\min_{t \in [K]} \left\| e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} := m_K > 0.$$

  We consider T large enough so that $e^{-\mu'_{j^*}(T-K)} \left\| \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \leq m_K / 8$. This holds, for example, for

  $$T \geq K + \frac{1}{\mu'_{j^*}} \ln \left(\frac{8 \left\| \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'}}{m_K} \right).$$

  Plugging this back into Eq. (37), we have

  $$\left\| \boldsymbol{g}(t) \right\|_{\widehat{\boldsymbol{\Sigma}}'} \geq \frac{3}{4} m_K - \frac{1}{8} m_K = \frac{5}{8} m_K.$$

  Recalling that $f_{\widehat{\boldsymbol{\Sigma}}'}(t) = \left\| \boldsymbol{g}(t) \right\|^2_{\widehat{\boldsymbol{\Sigma}}'}$, we arrive at

  $$f_{\widehat{\boldsymbol{\Sigma}}'}(t) \geq \frac{25}{64} m_K^2 \quad \text{for all } t \in [K], \tag{39}$$

  or, equivalently for integers $t = t_0 + s$ with $s \in [-t_0, K - t_0]$. As $\kappa^* = e^{-2m^* T} \to 0$ as $T \to \infty$, by choosing T large enough we can make

  $$f_{\widehat{\boldsymbol{\Sigma}}'}(t_0 + s) \geq \frac{25 m_K^2}{64} \geq \frac{9}{32} e^{-2\mu_{i^*} S_-} \left\| \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} \right\|^2_{\widehat{\boldsymbol{\Sigma}}'} \kappa^* \text{ for all } s \in [K] - t_0$$

  for any T-independent S_-.

- t is an integer with $K \leq t \leq t_0 + S_-$: We can now apply Eq. (38) to Eq. (37):

  $$\left\| \boldsymbol{g}(t) \right\|_{\widehat{\boldsymbol{\Sigma}}'} \geq \frac{9}{16} e^{-\mu_{i^*} t} \left\| \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} - e^{-\mu'_{j^*}(T-t)} \left\| \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'}.$$

  Setting $t = t_0 + s$, we have

  $$\begin{aligned}
  \left\| \boldsymbol{g}(t_0 + s) \right\|_{\widehat{\boldsymbol{\Sigma}}'} &\geq e^{-\mu_{i^*} t_0} \left(\frac{9}{16} e^{-\mu_{i^*} s} \left\| \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} - e^{-\mu'_{j^*}(T - t_0 - s) + \mu_{i^*} t_0} \left\| \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \right) \\
  &= e^{-m^* T} \left(\frac{9}{16} e^{-\mu_{i^*} s} \left\| \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'} - e^{\mu'_{j^*} s} \left\| \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \right).
  \end{aligned}$$

  We can adjust $S_- < 0$ and $S_- < s^*$ independent of T such that for all $s \leq S_- < 0$,

  $$e^{\mu'_{j^*} s} \left\| \boldsymbol{\Delta} \right\|_{\widehat{\boldsymbol{\Sigma}}'} \leq \frac{9}{32} e^{-\mu_{i^*} s} \left\| \boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}} \right\|_{\widehat{\boldsymbol{\Sigma}}'},$$

which gives

$$\|\boldsymbol{g}(t_0+s)\|_{\widehat{\boldsymbol{\Sigma}}'} \geq \frac{9}{32} e^{-m^* T} e^{-\mu_{i^*} s} \|\boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'} \geq \frac{9}{32} e^{-m^* T} e^{-\mu_{i^*} S_-} \|\boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'} .$$

Squaring and using $\kappa^* = e^{-2m^* T}$, we arrive again at

$$f_{\widehat{\boldsymbol{\Sigma}}'}(t_0+s) \geq \frac{81}{1024} e^{-2\mu_{i^*} S_-} \|\boldsymbol{P}_{i^*} \widehat{\boldsymbol{w}}\|^2_{\widehat{\boldsymbol{\Sigma}}'} \kappa^*$$

for every integer $t_0 + s \in [T]$ with $t_0 + s \geq K$ and $s \leq S_-$.

If needed, we can increase T to ensure that $K \leq t_0 + S_-$, so that the two cases analysed cover the full discrete left tail.

Right-tail lower-bound: Note that if $t = t_0 + s \leq T$, then $s \leq T - t_0$. We argue that for a certain $S_+ > s^*$ independent of T (and for T large enough depending on S_+), we have

$$f_{\widehat{\boldsymbol{\Sigma}}'}(t_0+s) \geq \frac{9}{16} \left\|\boldsymbol{P}'_{j^*} \boldsymbol{\Delta}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'} e^{2\mu'_{j^*} S_+} \kappa^*$$

for every integer $t_0 + s \in [T]$ and $s \geq S_+$.

For this, by the triangle inequality and the definition of the spectral norm,

$$\|\boldsymbol{g}(t)\|_{\widehat{\boldsymbol{\Sigma}}'} \geq \left\|e^{-(T-t)\boldsymbol{G}'} \boldsymbol{\Delta}\right\|_{\widehat{\boldsymbol{\Sigma}}'} - \left\|(\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'}) e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}}\right\|_{\widehat{\boldsymbol{\Sigma}}'} . \tag{40}$$

We bound each of the two norms.

For the first norm on the right-hand side, continuing from Eq. (36),

$$\left\|e^{-(T-t)\boldsymbol{G}'} \boldsymbol{\Delta}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'} = \sum_{j \in \mathcal{I}'} e^{-2\mu'_j (T-t)} \left\|\boldsymbol{P}'_j \boldsymbol{\Delta}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'} \geq e^{-2\mu'_{j^*}(T-t)} \left\|\boldsymbol{P}'_{j^*} \boldsymbol{\Delta}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'} ,$$

so that

$$\|e^{-(T-t)\boldsymbol{G}'} \boldsymbol{\Delta}\|_{\widehat{\boldsymbol{\Sigma}}'} \geq e^{-\mu'_{j^*}(T-t)} \|\boldsymbol{P}'_{j^*} \boldsymbol{\Delta}\|_{\widehat{\boldsymbol{\Sigma}}'} . \tag{41}$$

For the second norm, as $\boldsymbol{I} = \sum_{j=1}^{r'} \boldsymbol{P}'_j$, proceeding analogously as we have arrived at Eq. (34),

$$\begin{aligned}
\left\|(\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'}) e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'} &= \sum_{j=1}^{r'} (\lambda'_j - \gamma')(1 - e^{-\mu'_j (T-t)})^2 (e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}})^\top \boldsymbol{P}'_j e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \\
&\leq \sum_{j=1}^{r'} (\lambda'_j - \gamma')(e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}})^\top \boldsymbol{P}'_j e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}} \\
&= \left\|e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'} .
\end{aligned} \tag{42}$$

Further, taking the square root, using the eigendecomposition of $\boldsymbol{G}$ and the triangle inequality,

$$\begin{aligned}
\left\|(\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'}) e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}}\right\|_{\widehat{\boldsymbol{\Sigma}}'} \leq \left\|e^{-t\boldsymbol{G}} \widehat{\boldsymbol{w}}\right\|_{\widehat{\boldsymbol{\Sigma}}'} &= \left\|\sum_{i \in \mathcal{I}} e^{-t\mu_i} \boldsymbol{P}_i \widehat{\boldsymbol{w}}\right\|_{\widehat{\boldsymbol{\Sigma}}'} \\
&\leq \sum_{i \in \mathcal{I}} e^{-t\mu_i} \|\boldsymbol{P}_i \widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'} \leq e^{-\mu_{i^*} t} \sum_{i \in \mathcal{I}} \|P_i \widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'} .
\end{aligned}$$

Substituting this together with Eq. (41) back into Eq. (40), we arrive at

$$\|\boldsymbol{g}(t)\|_{\widehat{\boldsymbol{\Sigma}}'} \geq e^{-\mu'_{j^*}(T-t)} \|\boldsymbol{P}'_{j^*} \boldsymbol{\Delta}\|_{\widehat{\boldsymbol{\Sigma}}'} - e^{-\mu_{i^*} t} \sum_{i \in \mathcal{I}} \|P_i \widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'} .$$

With the substitution $t = t_0 + s$, we get

$$\|\boldsymbol{g}(t)\|_{\widehat{\boldsymbol{\Sigma}}'} \geq e^{-\mu'_{j^*}(T-t_0-s)} \left(\left\|\boldsymbol{P}'_{j^*} \boldsymbol{\Delta}\right\|_{\widehat{\boldsymbol{\Sigma}}'} - e^{-(\mu_{i^*} + \mu'_{j^*}) s} \sum_{i \in \mathcal{I}} \|\boldsymbol{P}_i \widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'} \right) .$$

Now, we can choose $S_+ > \max(0, s^*)$ large enough so that for all $s \geq S_+$,

$$\frac{1}{4}\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_{\widehat{\boldsymbol{\Sigma}}'} \geq e^{-(\mu_{i^*}+\mu'_{j^*})s}\sum_{i\in\mathcal{I}}\|\boldsymbol{P}_i\widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'},$$

resulting in

$$\begin{aligned}\|\boldsymbol{g}(t)\|_{\widehat{\boldsymbol{\Sigma}}'} &\geq \frac{3}{4}e^{-\mu'_{j^*}(T-t_0-s)}\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_{\widehat{\boldsymbol{\Sigma}}'} = \frac{3}{4}e^{-m^*T}e^{\mu'_{j^*}s}\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_{\widehat{\boldsymbol{\Sigma}}'}\\ &\geq \frac{3}{4}e^{-m^*T}e^{\mu'_{j^*}S_+}\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|_{\widehat{\boldsymbol{\Sigma}}'}.\end{aligned}$$

After squaring using $\kappa^* = e^{-2m^*T}$, this gives

$$f_{\widehat{\boldsymbol{\Sigma}}'}(t_0+s) \geq \frac{9}{16}\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'}e^{2\mu'_{j^*}S_+}\kappa^*$$

for every integer $t_0 + s \in [T]$ with $s \geq S_+$.

Putting everything together: After decreasing S_- and increasing S_+ if necessary, we have proved the following: there are $S_- < s^* - S_0 - 1$, $S_+ > s^* + S_0 + 1$, both $S_\pm$ independent of T and T large enough (depending on both S_- and S_+) such that

- according to Eq. (32), $f_{\widehat{\boldsymbol{\Sigma}}'}(\hat{t}_T) \leq 2C_1^2\kappa^*$;
- $f_{\widehat{\boldsymbol{\Sigma}}'}(t_0+s) \geq (81e^{-2\mu_{i^*}S_-}\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\|^2_{\widehat{\boldsymbol{\Sigma}}'}/1024)\kappa^*$ for all $s \leq S_-$ with $t_0+s \in [T]$;
- $f_{\widehat{\boldsymbol{\Sigma}}'}(t_0+s) \geq (9\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'}e^{2\mu'_{j^*}S_+}/16)\kappa^*$ for all $s \geq S_+$ with $t_0+s \in [T]$.

Recall that C_1 is a constant independent of $S_\pm$ or T. We can accordingly increase S_+ and decrease (and make negative) S_- such that

$$81e^{-2\mu_{i^*}S_-}\|\boldsymbol{P}_{i^*}\widehat{\boldsymbol{w}}\|^2_{\widehat{\boldsymbol{\Sigma}}'}/1024 \geq 3C_1^2 \quad \text{and} \quad 9\left\|\boldsymbol{P}'_{j^*}\boldsymbol{\Delta}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'}e^{2\mu'_{j^*}S_+}/16 \geq 3C_1^2.$$

So, for these choices of S_-, S_+ and T, we have that

$$\mathcal{T}^* \subseteq [T] \cap [S_- + t_0, S_+ + t_0]. \tag{43}$$

Step 6: Derivative sign in the central interval: Recall that we are still working with the interpolator $f_{\widehat{\boldsymbol{\Sigma}}'} : [0,T] \to [0,\infty)$. Also, recall the definition of S_0 from Step 2. Now, decrease S_- and increase S_+, if needed, such that $S_- < s^* - S_0 - 1 < s^* + S_0 + 1 < S_+$, and furthermore consider $R > 0$ for which $[S_-, S_+] \subseteq [-R, R]$. Additionally, we can increase T until $0 < t_0 - R < t_0 + R < T$.

By Step 3, we have

$$\sup_{-R\leq s\leq R}|\Xi(s)| \leq C_R e^{-\delta T}.$$

Note that, as $\Gamma(s) = 0$ has a unique solution at s^*, on $[-R,R] \setminus [s^* - S_0, s^* + S_0]$ the function Γ does not have a zero and moreover

$$\min\{|\Gamma(s)| : s \in [-R, s^* - S_0] \cup [s^* + S_0, R]\} > 0,$$

which is independent of T. So, we can take T large enough so that

$$\sup_{-R\leq s\leq R}|\Xi(s)| \leq \frac{1}{2}\min\{|\Gamma(s)| : s \in [-R, s^* - S_0] \cup [s^* + S_0, R]\}.$$

Accordingly, for every $s \in [-R,R] \setminus [s^* - S_0, s^* + S_0]$,

$$\text{sign}(\Xi(s) + \Gamma(s)) = \text{sign}(\Gamma(s)),$$

as $|\Xi| \leq |\Gamma|/2$ implies that addition of Ξ cannot change the sign of Γ. But, as we showed in Step 2, s^* is the unique root of Γ and $\Gamma(s) < 0$ for $s < s^*$, $\Gamma(s) > 0$ for $s > s^*$, so there is no root of $\Xi + \Gamma$ in $[-R,R] \setminus [s^* - S_0, s^* + S_0]$.

By Step 4, z_T is the unique root of $\Gamma + \Xi$ in $[s^* - S_0, s^* + S_0]$, on which this function is strictly increasing. Together with the sign argument above, and recalling that $f'_{\widehat{\boldsymbol{\Sigma}}'}(t_0+s) = 2\kappa^*(\Gamma(s)+\Xi(s))$ with $\kappa^* > 0$, we obtain that for $\bar{t}_T := t_0 + z_T$,

$$f'_{\widehat{\boldsymbol{\Sigma}}'}(t) < 0 \quad \text{for } t \in [t_0 + S_-, \bar{t}_T),$$

and

$$f'_{\widehat{\boldsymbol{\Sigma}}'}(t) > 0 \quad \text{for } t \in (\bar{t}_T, t_0 + S_+].$$

Thus, on the central interval $[t_0 + S_-, t_0 + S_+]$ the interpolator $f_{\widehat{\boldsymbol{\Sigma}}'}$ is strictly decreasing before $\bar{t}_T$ and strictly increasing afterwards.

Step 7: Localisation of the discrete global minimisers. By Step 5 Eq. (43), every $t^* \in \mathcal{T}^*$ lies in $[t_0+S_-, t_0+S_+]$. From Step 6, the interpolator $f_{\widehat{\boldsymbol{\Sigma}}'}$ is strictly decreasing on $[t_0+S_-, \bar{t}_T]$, and strictly increasing on $[\bar{t}_T, t_0 + S_+]$. By the choices of $S_\pm$ in Step 6 and the fact that $z_T \in [s^* - S_0, s^* + S_0]$, both $\lfloor \bar{t}_T \rfloor$ and $\lceil \bar{t}_T \rceil$ lie in the central interval $[t_0 + S_-, t_0 + S_+]$. Thus, the minimum of the discrete $f_{\widehat{\boldsymbol{\Sigma}}'} : [T] \to [0, \infty)$ over integer points in this interval can occur only at the two integers adjacent to $\bar{t}_T$, i.e.

$$\mathcal{T}^* \subseteq \{\lfloor \bar{t}_T \rfloor, \lceil \bar{t}_T \rceil\}.$$

If $\bar{t}_T$ is an integer, these two points coincide. Hence, $\mathcal{T}^*$ contains at most two elements. Moreover, for every $t^* \in \mathcal{T}^*$, by the triangle inequality

$$|t^* - (t_0 + s^*)| \le |t^* - \bar{t}_T| + |\bar{t}_T - (t_0 + s^*)| \le 1 + \frac{2C_{S_0}}{m_{S_0}} e^{-\delta T}.$$

Absorbing $2C_{S_0}/m_{S_0}$ into C and recalling $t_0 = \mu'_{j^*} T/(\mu_{i^*} + \mu'_{j^*})$ gives the stated conclusion. ■

C.7 Proof of Proposition 13

We will work over $d = 2$ dimensions. Define the following quantities:

- Pretraining: take $n = 2$, ridge parameter $\gamma = 1/2$ and consider

  $$\boldsymbol{X} = \begin{pmatrix} \sqrt{3} & 2/\sqrt{3} \\ 0 & \sqrt{5/3} \end{pmatrix}, \quad \widehat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{n} = \begin{pmatrix} 3/2 & 1 \\ 1 & 3/2 \end{pmatrix}, \quad \boldsymbol{H} = \widehat{\boldsymbol{\Sigma}} + \gamma \boldsymbol{I} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

  Additionally, as $\boldsymbol{X}$ is invertible, we can select $\boldsymbol{y}$ appropriately such that

  $$\widehat{\boldsymbol{w}} = \frac{1}{n} \boldsymbol{H}^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

  (For example, for $\boldsymbol{y} = (8/\sqrt{3}, 14/\sqrt{15})^\top$).
- Fine-tuning: take $n' = 2$, ridge parameter $\gamma' = 1$ and consider

  $$\boldsymbol{X}' = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad \widehat{\boldsymbol{\Sigma}}' = \frac{1}{n'} (\boldsymbol{X}')^\top \boldsymbol{X}' = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \boldsymbol{H}' = \widehat{\boldsymbol{\Sigma}}' + \gamma' \boldsymbol{I} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

  Additionally, for $\boldsymbol{y}' = (4, 4)^\top$, we have

  $$\widehat{\boldsymbol{w}}' = \frac{1}{n'} (\boldsymbol{H}')^{-1} (\boldsymbol{X}')^\top \boldsymbol{y}' = \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

- Thus, the empirical update is

  $$\boldsymbol{\Delta} = \widehat{\boldsymbol{w}}' - \widehat{\boldsymbol{w}} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

- We will take $c_{\mathrm{p}} = c_{\mathrm{f}} = 1$ and $\eta = \eta' = 1/4$.

Then,

$$e^{-\boldsymbol{G}} = \boldsymbol{I} - \frac{1}{4}\boldsymbol{H} = \begin{pmatrix} 1/2 & -1/4 \\ -1/4 & 1/2 \end{pmatrix}, \quad e^{-\boldsymbol{G}}\widehat{\boldsymbol{w}} = \begin{pmatrix} 0 \\ 3/4 \end{pmatrix} \in \mathrm{Ker}(\widehat{\boldsymbol{\Sigma}}').$$

Thus, (NKH) fails at $t = 1$. The burn-in defined in Appendix A is $K = 3$, since:

$$\mu_1 = \ln(4/3), \quad \mu_2 = \ln(4), \quad \frac{\|\boldsymbol{P}_2\widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'}}{\|\boldsymbol{P}_1\widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'}} = 3, \quad \text{therefore } K = \left\lceil \frac{\ln(12)}{\ln(3)} \right\rceil = 3.$$

Now, let's analyse $f_{\widehat{\boldsymbol{\Sigma}}'}(t)$. Recall that, due to the $\widehat{\boldsymbol{\Sigma}}'$-seminorm, $f_{\widehat{\boldsymbol{\Sigma}}'}(t) = g_1(t)^2$, where $g_1(t)$ is the first coordinate of

$$\boldsymbol{g}(t) = (\boldsymbol{I} - e^{-(T-t)\boldsymbol{G}'})e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}} - e^{-(T-t)\boldsymbol{G}'}\boldsymbol{\Delta}\,.$$

For an integer t, let

$$\beta_t := (e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}})_1 = \frac{3 - 3^t}{2 \cdot 4^t},$$

so that

$$g_1(t) = (1 - 2^{-(T-t)})\beta_t - 2^{-(T-t)}.$$

At $t := 1$, $\beta_1 = 0$ and $|g_1(1)| = 2^{-(T-1)}$. For every $t \in \{2, \ldots, T\}$, $\beta_t < 0$, so $|g_1(t)| \geq 2^{-(T-t)} > 2^{-(T-1)}$, while at $t := 0$, $g_1(0) = 1 - 2^{1-T}$, with magnitude greater than $2^{-(T-1)}$ for $T \geq 3$.

Hence, $t^* = 1$ is the unique discrete minimiser for every $T \geq 3$, and $t^*/T \to 0$. ■

C.8 Proof of Proposition 14

For fixed pretraining optimisation parameters, condition, if random, on the pretraining data $(\boldsymbol{X}, \boldsymbol{y})$, so that $\boldsymbol{G}$ and $\widehat{\boldsymbol{w}}$ are fixed. All subsequent probabilities are implicitly conditioned. Since $e^{-t\boldsymbol{G}}$ is invertible and $\widehat{\boldsymbol{w}} \neq 0$, we have $e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}} \neq 0$. Now, note that, writing the ith row of $\boldsymbol{X}'$ as $(\boldsymbol{x}'_i)^\top$,

$$\left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'} = \frac{1}{n'}\sum_{i=1}^{n'}((\boldsymbol{x}'_i)^\top e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}})^2,$$

so that

$$\{\left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_{\widehat{\boldsymbol{\Sigma}}'} = 0\} \subseteq \{(\boldsymbol{x}'_1)^\top e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}} = 0\}.$$

For each fixed integer $t \geq 0$, the set $\{\boldsymbol{x}' \in \mathbb{R}^d : (\boldsymbol{x}')^\top e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}} = 0\}$ is a hyperplane with Lebesgue measure 0. Therefore,

$$\mathbb{P}\left(\left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_{\widehat{\boldsymbol{\Sigma}}'} = 0\right) \leq \mathbb{P}\left((\boldsymbol{x}'_1)^\top e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}} = 0\right) = 0.$$

Taking a countable union over all non-negative integers t, and averaging over the pretraining data if needed, gives the unconditional almost-sure conclusion. ■

C.9 Proof of Proposition 15

Since $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}'$ commute, $\widehat{\boldsymbol{\Sigma}}'$ and $\boldsymbol{G}$ commute, so they can be simultaneously diagonalised by the same orthonormal basis $(\boldsymbol{u}_k)_{k=1}^d$, with corresponding eigenvalues $(g_k)_{k=1}^d$ for $\boldsymbol{G}$ and $(\sigma_k)_{k=1}^d$ for $\widehat{\boldsymbol{\Sigma}}'$ (here, we consider all eigenvalues). Let's write

$$\widehat{\boldsymbol{w}} = \sum_{k=1}^d a_k \boldsymbol{u}_k,$$

so that

$$e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}} = \sum_{k=1}^d a_k e^{-tg_k}\boldsymbol{u}_k, \quad \left\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|^2_{\widehat{\boldsymbol{\Sigma}}'} = \sum_{k=1}^d \sigma_k a_k^2 e^{-2tg_k}. \tag{44}$$

(i) If $\widehat{\boldsymbol{w}} \in \ker(\widehat{\boldsymbol{\Sigma}}')$, then $a_k = 0$ for all $1 \le k \le d$ with $\sigma_k > 0$, hence $\|e^{-t\boldsymbol{G}}\widehat{\boldsymbol{w}}\|_{\widehat{\boldsymbol{\Sigma}}'} = 0$.

To prove that $\mathcal{I} = \emptyset$, note that each projector $\boldsymbol{P}_i$ (for $1 \le i \le r$) has the form $\boldsymbol{P}_i = \boldsymbol{U}\boldsymbol{D}_i\boldsymbol{U}^\top$, where $\boldsymbol{U}$ is the $d \times d$ matrix containing all eigenvectors $(\boldsymbol{u}_k)_{k=1}^d$ as columns, and $\boldsymbol{D}_i$ is a $d \times d$ diagonal matrix with $(\boldsymbol{D}_i)_{kk} = 1$ whenever $\mu_i = g_k$, and 0 otherwise. Note that $\widehat{\boldsymbol{w}} \in \ker(\widehat{\boldsymbol{\Sigma}}')$ implies that

$$0 = \boldsymbol{U}^\top(\widehat{\boldsymbol{\Sigma}}'\widehat{\boldsymbol{w}}) = (\boldsymbol{U}^\top\widehat{\boldsymbol{\Sigma}}'\boldsymbol{U})(\boldsymbol{U}^\top\widehat{\boldsymbol{w}}) = \mathbf{diag}(\sigma_1, \dots, \sigma_d)(\boldsymbol{U}^\top\widehat{\boldsymbol{w}}),$$

so that $\sigma_k(\boldsymbol{U}^\top\widehat{\boldsymbol{w}})_k = 0$ for all $1 \le k \le d$. Now, fix any $1 \le i \le r$. Then

$$\boldsymbol{U}^\top(\widehat{\boldsymbol{\Sigma}}'\boldsymbol{P}_i\widehat{\boldsymbol{w}}) = (\boldsymbol{U}^\top\widehat{\boldsymbol{\Sigma}}'\boldsymbol{U})(\boldsymbol{U}^\top\boldsymbol{P}_i\boldsymbol{U})(\boldsymbol{U}^\top\widehat{\boldsymbol{w}}) = \mathbf{diag}(\sigma_1, \dots, \sigma_d)\boldsymbol{D}_i(\boldsymbol{U}^\top\widehat{\boldsymbol{w}}).$$

But, for each $1 \le k \le d$, the k'th entry of $\mathbf{diag}(\sigma_1, \dots, \sigma_d)\boldsymbol{D}_i(\boldsymbol{U}^\top\widehat{\boldsymbol{w}})$ is $\sigma_k(\boldsymbol{D}_i)_{kk}(\boldsymbol{U}^\top\widehat{\boldsymbol{w}})_k = 0$, so $\boldsymbol{U}^\top(\widehat{\boldsymbol{\Sigma}}'\boldsymbol{P}_i\widehat{\boldsymbol{w}}) = \mathbf{0}$. Because $\boldsymbol{U}$ is invertible we get $\widehat{\boldsymbol{\Sigma}}'\boldsymbol{P}_i\widehat{\boldsymbol{w}} = 0$, hence $\boldsymbol{P}_i\widehat{\boldsymbol{w}} \in \ker(\widehat{\boldsymbol{\Sigma}}')$ and finally $\mathcal{I} = \emptyset$.

(ii) Assume $\widehat{\boldsymbol{w}} \notin \ker(\widehat{\boldsymbol{\Sigma}}')$. Then $\widehat{\boldsymbol{\Sigma}}'\widehat{\boldsymbol{w}} = \sum_{k=1}^d a_k\sigma_k\boldsymbol{u}_k \neq 0$ and hence there exists $1 \le k \le d$ with $\sigma_k > 0$ and $a_k \neq 0$. So, the corresponding term in the summation in Eq. (44) is non-zero, and as $\sigma_k \ge 0$, $a_k^2 \ge 0$, the whole sum strictly decreases with t. Hence, its minimum over $[K]$ is achieved at $t = K$. Finally, as μ_r is the largest eigenvalue of $\boldsymbol{G}$,

$$\left\|e^{-K\boldsymbol{G}}\widehat{\boldsymbol{w}}\right\|_{\widehat{\boldsymbol{\Sigma}}'}^2 = \sum_{k=1}^d \sigma_k a_k^2 e^{-2Kg_k} \ge e^{-2K\mu_r}\sum_{k=1}^d \sigma_k a_k^2 = e^{-2K\mu_r}\left\|\widehat{\boldsymbol{w}}\right\|_{\widehat{\boldsymbol{\Sigma}}'}^2,$$

which concludes the proof. ■

D Experimental details

Data generation. We generate 320 observations for each phase in $d = 400$ dimensions. The rows of the random designs $\boldsymbol{X}, \boldsymbol{X}'$ are independently sampled from centred Gaussian distributions whose covariance matrices have the same eigenvalues, log-spaced from 1 to 10^{-3}, and independent eigenbases obtained by QR decompositions of matrices with entries sampled from standard normals. We draw independent $\boldsymbol{w}_\star, \boldsymbol{\delta}_\star \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d/d)$ and generate the responses $\boldsymbol{y}, \boldsymbol{y}'$ as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}_\star + \boldsymbol{\varepsilon}, \quad \boldsymbol{y}' = \boldsymbol{X}'(\boldsymbol{w}_\star + \boldsymbol{\delta}_\star) + \boldsymbol{\varepsilon}'$$

with independent Gaussian noise entries in $\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'$ of mean 0 and variance 0.01. Panels with 80 observations use the first 80 rows and corresponding labels from the corresponding dataset.

Training parameters. All panels use ridge penalties $\gamma = \gamma' = 0.05$, step-sizes $\eta = \eta' = 0.05$, zero initialisation for pretraining, and the resulting checkpoint as the fine-tuning initialisation. The step-size conditions are checked numerically. We use costs $c_\mathrm{p} = nd$, $c_\mathrm{f} = n'd$ and the same training budget $B = 65{,}536 \times 80d$, corresponding, under representation (1), to $T = B/\operatorname{lcm}(c_\mathrm{p}, c_\mathrm{f})$. The training budget excludes estimation or any other computational overhead.

Predictions and numerical evaluation. We use thin SVD to compute the ridge solutions and spectral quantities for the two-term prediction $\theta^* T + s^*$ (red dotted line). For the Lanczos estimation (grey dashed line) in Section 3.4, the upstream bottleneck λ_{i^*} is estimated using 40 matrix-free Lanczos iterations, initialised at $\boldsymbol{X}^\top\boldsymbol{y}/n$, with two-pass full reorthogonalisation. We estimate λ_{i^*} as the smallest eigenvalue of the obtained tridiagonal matrix. The downstream bottleneck is $\lambda'_{j^*} = \gamma'$ in these instances, hence no downstream Lanczos iterations are needed. Every feasible integer allocation is evaluated to obtain the numerical optimum (black dot). The plots use logarithmic vertical axes with separate ranges.

The experiment uses NumPy's `default_rng` with seed 2026 and NumPy and SciPy in double-precision floating-point arithmetic. We use exponential rescaling to compute the closed-form GD residuals directly, and retain objective values in logarithmic form to avoid cancellation or underflow.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Contributions are summarised in Section 1.1. The claims are formally proved in Theorems 6, 7 and 12, to name a few.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 4.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All proofs can be found in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The full information for the experiment in Fig. 1 can be found in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: Our paper is theoretical and the role of the experiments is to support our theoretical findings.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The full information for the experiment in Fig. 1 can be found in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: Our paper is theoretical and the role of the experiments is to support our theoretical findings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The full information for the experiment in Fig. 1 can be found in its caption.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This is a theory paper.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is a theory paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This is a theory paper.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve either.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve either.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.