THEORY OF CERTIFIED UNLEARNING FOR A FAMILY OF REGULARIZATION-BASED CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Integrating certified unlearning into continual learning, an emerging paradigm termed continual learning-unlearning (CLU), enables models to learn sequentially while retaining past knowledge and reliably removing requested information to protect privacy. Despite its importance, CLU remains largely unexplored. We provide the first unified (ε, δ)-certified CLU framework for a broad family of regularization-based CL algorithms, with the objective of jointly minimizing CL excess risk and unlearning loss. We derive an algorithm-dependent CL excess-risk bound for nonconvex losses, showing that regularizers that better accommodate task drift, such as Hessian-based regularization, improve learning performance. We then develop two complementary certified-unlearning mechanisms. Our first-order mechanism exploits natural forgetting induced by CL updates with minimal additional cost. Our second-order method enables precise request-specific unlearning for arbitrary requests, with unlearning loss strongly depending on request order through prior-correction propagation and regularizer mismatch. Both methods show that CL algorithms with stronger regularization, such as MAS, achieve lower unlearning loss, revealing a fundamental trade-off between retaining undeleted-task knowledge and removing deleted-task influence. Empirical results reveal distinct Pareto frontiers between certified-unlearning loss and CL excess risk, showing that no single method minimizes both. Thus, method selection should prioritize lower unlearning loss for stringent privacy requirements and lower CL excess risk for better predictive performance.

1 INTRODUCTION

Machine learning systems are increasingly deployed as long-lived services that interact with evolving data streams. As these systems absorb data for personalization, adaptation, and continual improvement, they must also support deletion requests that remove the influence of specified data after training. This capability has become a critical requirement under privacy regulations such as the General Data Protection Regulation (GDPR) (Voigt & von dem Bussche, 2017), which protects individuals' "right to be forgotten". Machine unlearning addresses this requirement by removing the influence of requested data without computationally prohibitive retraining from scratch (Nguyen et al., 2025). Among existing approaches, (ε, δ)-certified unlearning is a key framework for provable machine unlearning. Building on the indistinguishability notion of differential privacy, it provides a rigorous distributional guarantee: the distribution of the model produced by unlearning is close to that of a model retrained from scratch on the retained data (Guo et al., 2019). Existing certified unlearning approaches can be broadly divided into two categories. First-order methods use gradient-based fine-tuning on the retained data (Neel et al., 2021; Chien et al., 2024a;b; Koloskova et al., 2025; Mu & Klabjan, 2026), while second-order methods use Newton-style approximations to estimate the retrained model (Sekhari et al., 2021; Suriyakumar & Wilson, 2022; Liu et al., 2023; Qiao et al., 2024; Basaran et al., 2025).

Most of the certified-unlearning literature considers static learning settings, in which a model is trained once on a fixed dataset and subsequently processes deletion requests (Mu & Klabjan, 2026; Koloskova et al., 2025). However, modern learning systems increasingly operate within continual learning (CL) frameworks rather than conventional offline learning settings, with tasks arriving sequentially and models being updated throughout their lifetimes. A CL model must retain knowledge from earlier tasks while adapting to new ones, yet storing all historical data is infeasible because of

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

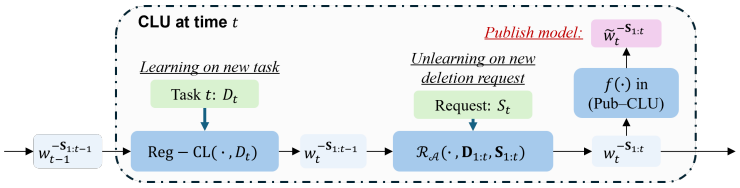


Figure 1: CLU update at time t. Starting from $w_{t-1}^{-\mathbf{S}_{1:t-1}}$, the learner first trains on the new dataset D_t to obtain $w_t^{-\mathbf{S}_{1:t-1}}$ and then processes the unlearning request set S_t using $\mathcal{R}_{\mathcal{A}}$ in (1), producing $w_t^{-\mathbf{S}_{1:t}}$. The randomized mechanism f in (Pub-CLU) then releases $\tilde{w}_t^{-\mathbf{S}_{1:t}}$, whose distribution satisfies the certified-unlearning guarantee in Definition 2.1.

privacy, memory, or system constraints (Wang et al., 2024). Given that modern learning systems, such as continually updated large language models (LLMs) and CL-enabled AI agents, increasingly operate in rapidly evolving environments and learn from continuously changing data streams, achieving certified unlearning in CL is of critical importance.

Several empirical studies have recently explored joint continual learning and unlearning, but they focus primarily on empirical forgetting and do not provide certified distributional guarantees relative to retraining (Liu et al., 2022; Chatterjee et al., 2024; Cha et al., 2024; Huang et al., 2025). The importance of such guarantees is underscored by Hu et al. (2025), who show that non-certified methods may provide only the illusion of successful unlearning. Against this backdrop, Hu et al. (2026) initiated the theoretical study of certified unlearning in CL under a simple ℓ_2-regularized update. Nevertheless, ℓ_2 regularization is an overly restrictive abstraction that cannot capture the task-adaptive regularization mechanisms widely used in modern CL systems. Consequently, a substantial gap remains between certified-unlearning theory and practical CL.

In this paper, we develop a unified certified unlearning method for the continual learning-unlearning (CLU) problem, situated within the broad and widely used family of regularization-based CL (Reg-CL) methods, which encompasses many prominent algorithms, such as those proposed by Kirkpatrick et al. (2017); Aljundi et al. (2018); Yin et al. (2020); Zenke et al. (2017). Our goal is to minimize the post-unlearning excess risk subject to certified unlearning, while reducing the computational and storage costs. A fundamental challenge in CLU is balancing knowledge retention on undeleted tasks against influence removal for deleted tasks. We show that, within regularization-based CL, this trade-off is governed by the choice of CL algorithm by analyzing how algorithm-specific regularizers jointly affect CL excess risk and unlearning loss. This provides a basis for selecting CL algorithms for different unlearning scenarios. Our contributions are summarized as follows.

- We are the first to develop certified unlearning methods for a general family of Reg-CL algorithms in the sequential CLU setting, formulated in Section 2. In Section 3, we derive a new algorithm-dependent upper bound on the CL excess risk that lays the foundation for CLU. We also compare two representative CL algorithms, ℓ_2 and Hessian regularization, to illustrate how more structured regularization can reduce CL excess risk.
- In Section 4, we propose a simple first-order certified-unlearning method for the family of Reg-CL by exploiting the natural-forgetting effect of subsequent CL updates. By deriving an algorithm-specific upper bound on the unlearning loss, we show that general Reg-CL methods, such as MAS, can achieve low unlearning loss under strong regularization and when deleting later tasks.
- In Section 5, we develop a second-order certified-unlearning method for Reg-CL that enables precise request-specific unlearning. Our analysis reveals how unlearning loss and operational costs depend on the request order, providing practical insights into unlearning-request scheduling. We further show that this method can outperform natural forgetting when later tasks are deleted, particularly under strong regularization.
- In Section 6, we conduct comprehensive experiments on real-world datasets to compare the CL excess risk and unlearning loss of different Reg-CL algorithms under the proposed certified-unlearning methods, identifying their suitable operating regimes across different CLU settings. The results corroborate our theoretical findings and suggest prioritizing lower unlearning loss such as MAS for stringent privacy requirements and lower CL excess risk for better predictive performance such as EWC.

2 PROBLEM FORMULATION FOR CLU

This section formalizes the sequential CLU process with learning updates and arbitrary unlearning requests over time (Fig. 1), then defines the (ε, δ)-certified continual-unlearning requirement and post-unlearning excess risk.

2.1 TWO-STAGE CLU FRAMEWORK

At each time $t = 1, \dots, T$, the learner receives a new task with dataset $D_t = (z_{t,i})_{i=1}^{n_t} \in \mathcal{Z}^{n_t}$, where n_t denotes the number of samples in task t. We denote the sequence of task datasets observed through time t by $\mathbf{D}_{1:t} = (D_1, \dots, D_t)$.

After learning task t, the system may receive an unlearning request S_t, a subset of previous task indices identifying the datasets $\{D_i\}_{i \in S_t}$ whose influence must be removed from $w_t^{-\mathbf{S}_{1:t-1}}$ (Fig. 1). If $S_t = \emptyset$, no request arrives. We define the cumulative requested tasks as $S_{\le t} = \bigcup_{i=1}^{t} S_i$ and require each admissible request to satisfy $S_t \subseteq [t] \setminus S_{\le t-1}$. Denote the sequence of unlearning requests as $\mathbf{S}_{1:t} = (S_1, \dots, S_t)$. For $t' \le t$, $w_t^{-\mathbf{S}_{1:t'}}$ denotes the internal model after learning through time t and processing requests through t'. In particular, $w_t^{-\mathbf{S}_{1:t-1}}$ and $w_t^{-\mathbf{S}_{1:t}}$ are the pre- and post-unlearning models at time t, respectively, as shown in Fig. 1.

Continual learning from D_t. Let $\ell(w, z)$ denote the loss of a model parameter $w \in \mathcal{W} \subseteq \mathbb{R}^d$ on an example $z \in \mathcal{Z}$. As in Fig. 1, starting from $w_{t-1}^{-\mathbf{S}_{1:t-1}}$, the learner trains on D_t according to

$$w_t^{-\mathbf{S}_{1:t-1}} \in \arg\min_{w \in \mathcal{W}} \left\{ L(w, D_t) + \Omega_{\mathbf{D}_{1:t-1}^{-\mathbf{S}_{1:t-1}}}\left(w_{t-1}^{-\mathbf{S}_{1:t-1}}, w\right) \right\}. \tag{Reg-CL}$$

Here, $L(w, D_t) = n_t^{-1} \sum_{i=1}^{n_t} \ell(w, z_{t,i})$ denotes the empirical loss on the current task. In standard CL, $\Omega_{\mathbf{D}_{1:t-1}}$ denotes an algorithm-specific regularizer that penalizes deviations from the previous model w_{t-1} to mitigate forgetting. For example, EWC (Kirkpatrick et al., 2017) uses $\Omega_{\mathbf{D}_{1:t-1}}(w_{t-1}, w_t) = \|w_t - w_{t-1}\|_{P_{t-1}}^2$, where P_{t-1} is the accumulated Fisher information matrix through task $t-1$. A key challenge in integrating unlearning into CL is redesigning $\Omega_{\mathbf{D}_{1:t-1}}$ as $\Omega_{\mathbf{D}_{1:t-1}^{-\mathbf{S}_{1:t-1}}}$ in (Reg-CL) to remove the information of deleted tasks $\mathbf{S}_{1:t-1}$, specified later.

Continual Unlearning of S_t. After Stage I produces $w_t^{-\mathbf{S}_{1:t-1}}$, the unlearning algorithm $\mathcal{R}_{\mathcal{A}}$ processes S_t and produces the updated internal model $w_t^{-\mathbf{S}_{1:t}}$ according to (1) as in Fig. 1. The randomized publishing mechanism f then produces the published model $\tilde{w}_t^{-\mathbf{S}_{1:t}} = f(w_t^{-\mathbf{S}_{1:t}})$. The internal model $w_t^{-\mathbf{S}_{1:t}}$ is carried forward to time $t+1$ for subsequent CLU.

2.2 OBJECTIVES AND GUARANTEES IN CLU

Given a request history $\mathbf{S}_{1:t}$, certified unlearning aims to efficiently approximate the model obtained by perfect retraining on the retained task sequence. The update at time t is

$$w_t^{-\mathbf{S}_{1:t}} = \mathcal{R}_{\mathcal{A}}\left(w_t^{-\mathbf{S}_{1:t-1}}, \mathbf{D}_{1:t}, \mathbf{S}_{1:t}\right), \tag{1}$$

which aims to approximate the perfect retraining model

$$w_t^{-S_{\le t}} = \mathcal{A}\left((D_i)_{i \in [t] \setminus S_{\le t}}\right), \tag{2}$$

where $\mathcal{A}$ denotes the sequential application of the CL procedure in (Reg-CL) to the retained task datasets in their original chronological order, without unlearning $\mathbf{S}_{1:t}$.

For a certified guarantee, the published model $\tilde{w}_t^{-\mathbf{S}_{1:t}} = f(w_t^{-\mathbf{S}_{1:t}})$ must be distributionally indistinguishable from the randomized retraining output $\tilde{w}_t^{-S_{\le t}} = f(w_t^{-S_{\le t}})$, formalized as follows.

Definition 2.1 ((ε, δ)-certified continual unlearning (Sekhari et al., 2021; Hu et al., 2026))**.** *Let $\varepsilon > 0$ and $\delta \ge 0$. An unlearning algorithm $f \circ \mathcal{R}_{\mathcal{A}}$ satisfies* (ε, δ)-certified continual unlearning *if the following inequalities hold for every time t, dataset sequence $\mathbf{D}_{1:t}$, admissible unlearning request history $\mathbf{S}_{1:t}$, and measurable subset $\mathcal{O} \subseteq \mathcal{W}$.*

$$\Pr\left(\tilde{w}_t^{-\mathbf{S}_{1:t}} \in \mathcal{O}\right) \le e^{\varepsilon} \Pr\left(\tilde{w}_t^{-S_{\le t}} \in \mathcal{O}\right) + \delta, \ \Pr\left(\tilde{w}_t^{-S_{\le t}} \in \mathcal{O}\right) \le e^{\varepsilon} \Pr\left(\tilde{w}_t^{-\mathbf{S}_{1:t}} \in \mathcal{O}\right) + \delta. \tag{3}$$

We define the post-unlearning excess risk over all retained tasks as follows.

Definition 2.2 (Post-unlearning excess risk (Hu et al., 2026)). *Let $\mathcal{D}_\tau$ denote the data-generating distribution of task τ, from which D_τ is drawn. The population loss for task τ is $R_\tau(w) = \mathbb{E}_{z\sim\mathcal{D}_\tau}[\ell(w,z)]$. At time t, the post-unlearning excess risk of a CLU algorithm is defined as $\mathcal{E}_{\mathrm{LU}} = \mathcal{E}_{\mathrm{U}} + \mathcal{E}_{\mathrm{L}}$, where*

$$\mathcal{E}_{\mathrm{U}} = \underbrace{\mathbb{E}\left[\frac{1}{N_t}\sum_{\substack{\tau=1\\ \tau\notin S_{\le t}}}^{t}\left(R_\tau(\tilde{w}_t^{-\mathbf{S}_{1:t}}) - R_\tau(w_t^{-S_{\le t}})\right)\right]}_{\textit{unlearning loss}}, \tag{4}$$

$$\mathcal{E}_{\mathrm{L}} = \underbrace{\mathbb{E}\left[\frac{1}{N_t}\sum_{\substack{\tau=1\\ \tau\notin S_{\le t}}}^{t} R_\tau(w_t^{-S_{\le t}})\right] - \min_{w\in\mathcal{W}}\frac{1}{N_t}\sum_{\substack{\tau=1\\ \tau\notin S_{\le t}}}^{t} R_\tau(w)}_{\textit{continual learning excess risk}}. \tag{5}$$

Here $N_t := t - |S_{\le t}|$ denotes the number of retained tasks, and we consider times t for which $N_t \ge 1$. The expectation is taken over the randomness of the task datasets $\{D_\tau\}_{\tau=1}^t$ and the noise introduced by f when producing $\tilde{w}_t^{-\mathbf{S}_{1:t}}$.

In CLU, we minimize the post-unlearning excess risk $\mathcal{E}_{\mathrm{LU}}$ subject to the (ε,δ)-certified continual-unlearning constraint. Its two components isolate unlearning and learning: $\mathcal{E}_{\mathrm{U}}$ in (4) measures the gap between the published CLU model $\tilde{w}_t^{-\mathbf{S}_{1:t}}$ and the retrained model $w_t^{-S_{\le t}}$, whereas $\mathcal{E}_{\mathrm{L}}$ measures the gap between $w_t^{-S_{\le t}}$ and the optimal model on retained tasks. Thus, $\mathcal{E}_{\mathrm{U}}$ depends on the unlearning algorithm $\mathcal{R}_{\mathcal{A}}$ in (1), while $\mathcal{E}_{\mathrm{L}}$ depends on the CL algorithm $\mathcal{A}$ in (2).

This decomposition reveals a fundamental tension between unlearning and knowledge retention (Zhao et al., 2024): a CL algorithm that strongly mitigates forgetting may preserve deleted tasks more persistently, making their influence harder to remove. Reg-CL algorithms in (Reg-CL) therefore provide a rich design space for balancing $\mathcal{E}_{\mathrm{L}}$ and $\mathcal{E}_{\mathrm{U}}$, which we investigate throughout the paper. We also aim to reduce CLU computation and storage.

We introduce two mild assumptions that are compatible with modern neural networks.

Assumption 2.3. *There exists a radius $r \ge 0$ such that, for every $z \in \mathcal{Z}$, the loss $\ell(\cdot,z)$ is L-Lipschitz continuous and twice continuously differentiable almost everywhere on $\overline{\mathcal{B}}(w_0,r)$. At every point where $\ell(\cdot,z)$ is twice continuously differentiable, its Hessian satisfies $\mu I \preceq \nabla^2\ell(w,z) \preceq MI$, with any $z \in \mathrm{supp}(\mathcal{D}_t)$ for every $t = 1,\dots,T$ and finite constants $\mu \le M$.*

This assumption applies only to the bounded region traversed by the practical CLU trajectory, imposing no strong global restriction on the parameter space. The Hessian naturally remains finite in this region. In particular, Assumption 2.3 allows $\mu < 0$, accommodating nonconvex losses.

Assumption 2.4. *Let the parameter space be $\mathcal{W} = \mathcal{B}(w_0,r)$. For each task t, define the set of empirical risk minimizers as $W_t = \{w \in \mathcal{W} \mid L(w,D_t) = \min_{w'\in\mathcal{W}} L(w',D_t)\}$. We assume that an empirical minimizer $\hat{w}_t \in W_t$ exists for every task t. Assume that the deterministic internal states generated by the CLU and retraining procedures exist and remain in $\mathcal{W}$.*

Assumption 2.4 is supported by results showing that sufficiently wide neural networks can attain empirical minimizers while remaining close to their initialization (Jacot et al., 2018; Chizat et al., 2019; Li & Liang, 2018). Thus, a localized parameter space suffices for good practical results.

The complete statements and proofs of all subsequent results are deferred to the Appendix.

3 Preliminary Results for Reg-CL

We first present preliminary results on the CL excess risk $\mathcal{E}_{\mathrm{L}}$ for the general Reg-CL framework in (Reg-CL), laying the foundation for CLU. We compare general (Reg-CL) with ℓ_2-regularized CL and establish a new CL excess-risk bound for Reg-CL with nonconvex models, a setting rarely analyzed in the literature. We then recall an existing result on the unlearning loss $\mathcal{E}_{\mathrm{U}}$.

3.1 Reg-CL Excess Risk $\mathcal{E}_{\mathrm{L}}$

Certified CLU remains largely unexplored, with only Hu et al. (2026) studying it under ℓ_2 regularization. However, this basic CL algorithm can perform poorly in terms of the CL excess risk $\mathcal{E}_{\mathrm{L}}$, which

substantially limits the practical usefulness of certified-unlearning methods built exclusively on it. To illustrate this limitation, we show that scalar ℓ_2 regularization cannot guarantee convergence even in a simple linear CL setting in which all tasks share the same ground-truth model. Suppose each task has a dataset $D_t = (X_t, y_t)$ with $X_t \in \mathbb{R}^{n\times p}$ and $y_t \in \mathbb{R}^n$ generated according to

$$y_t = X_t w^\star + \eta_t,\ (X_t)_{i,j} \sim \mathcal{N}(0,1),\ \eta_t \sim \mathcal{N}(0, \sigma^2 I_n),$$

where $n > p$. The loss is $L(w, D_t) = \|X_t w - y_t\|^2 / n$. The following lemma characterizes the CL excess risk $\mathcal{E}_{\mathrm{L}}$ under ℓ_2 regularization. Define $[t] \setminus S_{\le t} := \{\tau_1, \dots, \tau_{N_t}\}$.

Lemma 3.1 (ℓ_2 regularization Hu et al. (2026))**.** *Let w_T denote the CL model produced by* (Reg-CL) *on N_T retained tasks with $\Omega_{(D_{\tau_i})_{i=1}^{N_T}}(x, y) = \lambda\|x - y\|_2^2/n$. Then*

$$\mathcal{E}_{\mathrm{L}} \approx a_\lambda^{2N_T} \|w^\star\|_2^2 + b_\lambda \frac{1 - a_\lambda^{2N_T}}{1 - a_\lambda^2}, \tag{6}$$

where $b_\lambda > 0$ and $a_\lambda \in (0, 1)$ are constants.

Even in this benign linear setting with a shared ground-truth model and informative tasks, the second excess-risk term under ℓ_2 regularization remains bounded away from zero as $N_T \to \infty$. In contrast, the more structured Hessian-regularized CL method has excess risk converging to zero under the same setting, as shown below.

Lemma 3.2 (Accumulated Hessian regularization Zhao et al. (2024))**.** *Let w_T denote the CL model produced by* (Reg-CL) *on T tasks with the accumulated-Hessian regularizer $\Omega_{\mathbf{D}_{1:t-1}}(x, y) = \|x - y\|_{H_t}^2$, where $H_t = \lambda I + \sum_{i=1}^{t-1} \frac{1}{n} X_i^\top X_i$. Then we have $\lim_{T\to\infty} \mathcal{E}_{\mathrm{L}} = 0$.*

General Reg-CL therefore offers broader algorithmic choices for reducing the post-unlearning excess risk $\mathcal{E}_{\mathrm{LU}}$.

The preceding lemmas cover two Reg-CL algorithms in a linear setting. We now extend the analysis to general Reg-CL and bound $\mathcal{E}_{\mathrm{L}}$ in (5) for nonconvex models.

Theorem 3.3. *Suppose Assumptions 2.3 and 2.4 hold, and let $B = \sup_{w\in\mathcal{W},\, z\in\mathcal{Z}} |\ell(w, z)|$. Let $\hat{w}_{\tau_i}$ be the empirical minimzer of each retained dataset D_{τ_i}. For the perfect retraining model $w_t^{-S_{\le t}} = \mathcal{A}\big((D_i)_{i\in[t]\setminus S_{\le t}}\big)$, the CL excess risk defined in* (5) *satisfies $\mathcal{E}_{\mathrm{L}} \le \mathcal{E}^{-S_{\le t}}$, where $\mathcal{E}^{-S_{\le t}}$ equals*

$$\frac{L}{N_t}\left[\sum_{k=1}^{N_t}\sum_{i=2}^{N_t}\left(\prod_{j=i+1}^{N_t} q_j\right)\bar{q}_i \mathbb{E}[\|\hat{w}_{\tau_i} - \hat{w}_{\tau_k}\|] + \sum_{k=2}^{N_t}\left(\prod_{j=2}^{N_t} q_j\right)\mathbb{E}[\|\hat{w}_{\tau_k} - \hat{w}_{\tau_1}\|]\right] + \sum_{k=1}^{N_t} \frac{O\left(n_{\tau_k}^{-\frac{1}{4}}\right)}{N_t} \tag{7}$$

where $q_j = \|P_j\|$ and $\bar{q}_j = \|I - P_j\|$ for $j = 2, \dots, N_t$, with P_j denoting a CL-algorithm-specific matrix.

In (7), the first term captures model-drift propagation through retained tasks via q_i and $\bar{q}_i$. Since they cannot be simultaneously small, methods that better balance this propagation, such as EWC and Hessian regularization, yield smaller CL excess risk bounds. The second term is the algorithm-independent finite sample generalization error.

3.2 Preliminaries on Unlearning Loss $\mathcal{E}_{\mathrm{U}}$

A key certified-unlearning result is that, for the CLU process in Fig. 1, if $\|w_t^{-\mathbf{S}_{1:t}} - w_t^{-S_{\le t}}\| \le \gamma_t(\mathbf{S}_{1:t})$, then adding Gaussian noise via

$$f(w_t^{-\mathbf{S}_{1:t}}) = w_t^{-\mathbf{S}_{1:t}} + \xi_t,\ \text{where}\ \ \xi_t \sim \mathcal{N}\big(\mathbf{0}, \gamma_t(\mathbf{S}_{1:t})^2 2\ln(1.25/\delta)/\varepsilon^2 I\big) \tag{Pub-CLU}$$

guarantees (ε, δ)-certified continual unlearning in Definition 2.1, with unlearning loss (4) given by

$$\mathcal{E}_{\mathrm{U}} \le L\gamma_t(\mathbf{S}_{1:t})\left(\sqrt{2d\ln(1.25/\delta)}/\varepsilon + 1\right), \tag{8}$$

where L is the Lipschitz constant assumed in Assumption 2.3 Hu et al. (2026). Accordingly, in the following sections, we focus on deriving the approximation error bound $\gamma_t(\mathbf{S}_{1:t})$ under different unlearning algorithms.

The two main lines of certified-unlearning research in offline learning are first-order gradient-based fine-tuning and second-order Hessian-based correction. Motivated by these approaches, we propose two corresponding unlearning algorithms in Sections 4 and 5, respectively.

4 NATURAL FORGETTING UNLEARNING FOR CLU

Continual learning gradually attenuates the influence of earlier tasks as later updates are applied. Similar dynamics have been used for gradient-based certified unlearning in offline settings (Neel et al., 2021; Chien et al., 2024a; Mu & Klabjan, 2026). We first use this natural-forgetting effect as a simple unlearning mechanism that returns the Stage-I CL model:

$$w_t^{-\mathbf{S}_{1:t}} = \mathcal{R}_{\mathcal{A}}\Big(w_t^{-\mathbf{S}_{1:t-1}}, \mathbf{D}_{1:t}, \mathbf{S}_{1:t}\Big) = w_t^{-\mathbf{S}_{1:t-1}}. \tag{9}$$

In subsequent stages, we retain the original CL regularizer over all observed tasks without removing deleted-task contributions, that is, $\Omega_{\mathbf{D}_{1:t}^{-\mathbf{S}_{1:t}}} = \Omega_{\mathbf{D}_{1:t}}$. This exploits future CL updates' natural forgetting, requiring no additional storage and only $O(1)$ computation beyond the underlying CL algorithm. Nevertheless, certified unlearning requires careful analysis of $\gamma(\mathbf{S}_{1:t})$ to calibrate the noise in (Pub-CLU).

Theorem 4.1. *Assume $S_{\le t} \neq \emptyset$, and let $\underline{s}^t := \min S_{\le t}$. If using* (9) *for each time step, we have*

$$\gamma(\mathbf{S}_{1:t}) = \underbrace{\sum_{s\in S_{\le t}} \left(\prod_{\substack{j=s+1\\ j\notin S_{\le t}}}^{t} \rho_j^t\right)\Gamma_s}_{\textit{deletion-induced loss}} + \underbrace{\sum_{\substack{i=\underline{s}^t+1\\ i\notin S_{\le t}}}^{t} \left(\prod_{\substack{j=i+1\\ j\notin S_{\le t}}}^{t} \rho_j^t\right) g_i^t}_{\textit{regularizer mismatch loss}}, \tag{10}$$

where ρ_j^t, g_i^t, and Γ_s are algorithm-specific parameters of (Reg-CL)*.*

The first summation in (10) captures the direct **approximation error from each deleted task** $s \in S_{\le t}$, propagated through later retained updates by ρ_j^t. The second captures the **regularizer-mismatch error**: exact retraining constructs $\Omega_{(D_i)_{i\in[n]\setminus S_{\le t}}}$ along a trajectory excluding deleted tasks, whereas natural-forgetting CLU constructs $\Omega_{\mathbf{D}_{1:n}}$ along the full trajectory.

4.1 REG-CL ALGORITHM CHOICE UNDER NATURAL FORGETTING

We instantiate Theorem 4.1 for task-additive regularizers to analyze how the CL algorithm affects natural forgetting. This class includes popular CL algorithms such as EWC (Kirkpatrick et al., 2017), MAS (Aljundi et al., 2018), and Hessian-based regularization (Yin et al., 2020). Task-additive regularizers $\Omega_{\mathbf{D}_{1:t-1}}(w_{t-1}^{-\mathbf{S}_{1:t-1}}, w_t^{-\mathbf{S}_{1:t-1}})$ take the quadratic form

$$\frac{1}{2}\left\| w_t^{-\mathbf{S}_{1:t-1}} - w_{t-1}^{-\mathbf{S}_{1:t-1}} \right\|_{\mathbf{F}_t}^2, \tag{11}$$

where $\|v\|_{\mathbf{F}_t}^2 := v^\top \mathbf{F}_t v$, $\mathbf{F}_t := \lambda I + \sum_{i=1}^{t-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i)$. For example, in EWC, $F(w_i, D_i)$ denotes the Fisher information matrix. Analogous to Assumption 2.3, we impose a standard boundedness condition that holds naturally along the algorithm trajectory Li & Liang (2018).

Assumption 4.2. *For any $w \in \overline{\mathcal{B}}(w_0, r)$ and any dataset D, $\mu_\Omega I \preceq F(w, D) \preceq M_\Omega I$.*

Proposition 4.3. *Define $S_{\le t}(j) := |S_{\le t} \cap \{1, \dots, j\}|$, $N_{S_{\le t}}(j) := j - S_{\le t}(j)$, and $N_{S_{\le t}}([a, b]) := b - a + 1 - |S_{\le t} \cap \{a, \dots, b\}|$. Let $\lambda > \mu$. For the CL algorithm defined in* (11)*, under Assumptions 2.3, 2.4, and 4.2, natural forgetting yields the following approximation error upper bound:*

$$\gamma(\mathbf{S}_{1:t}) = \sum_{s\in S_{\le t}} \alpha(s) + \sum_{\substack{i=\underline{s}^t+1\\ i\notin S_{\le t}}}^{t} \alpha(i) \frac{\Big[S_{\le t}(i-1)M_\Omega + N_{S_{\le t}}([\underline{s}^t, i-1])(M_\Omega - \mu_\Omega)\Big]}{\mu + \lambda + N_{S_{\le t}}(i-1)\mu_\Omega}, \tag{12}$$

where $\alpha(i) := \left(\prod_{j=N_{S_{\le t}}(i)}^{N_{S_{\le t}}(t)-1} \frac{\lambda + jM_\Omega}{\mu+\lambda+j\mu_\Omega}\right) \frac{L}{\lambda+(i-1)\mu_\Omega}$.

By contrast, the approximation-error bound under the ℓ_2 regularizer is obtained from (12) by setting $\mu_\Omega = M_\Omega = 0$.

Compared with ℓ_2, the second summation in (12) captures the regularizer mismatch in (10). It is absent under ℓ_2 because $\lambda \|w_t - w_{t-1}\|_2^2$ is history-independent, generally yielding lower unlearning loss. Moreover, the anisotropic matrix regularizer in (11) introduces the larger contraction factor $\frac{\lambda + jM_\Omega}{\mu+\lambda+j\mu_\Omega}$, compared with $\frac{\lambda}{\mu+\lambda}$ under ℓ_2 regularization, in both terms of (12). Only with strong

regularization, a small spectral gap $M_\Omega - \mu_\Omega$, and large μ_Ω can the effects of regularizer mismatch and anisotropy be sufficiently controlled, allowing a task-additive regularizer to outperform ℓ_2 regularization. This indicates an inevitable cost of unlearning with a richer family of CL algorithms that achieve lower CL excess risk.

5 Hessian-Based Unlearning for CLU

Although efficient, natural forgetting essentially forgets all tasks together. A CL algorithm that unlearns well through natural forgetting may therefore struggle to retain non-deleted tasks, creating an imbalance between unlearning and knowledge retention. This motivates a Hessian-based method that more precisely removes target-task influence.

To accommodate the family of Reg-CL algorithms, we use an oracle estimator for $\Omega_{\mathbf{D}_{1:n}^{-\mathbf{S}_{1:t}}}$ in (Reg-CL) to approximate the retraining-induced regularizer $\Omega_{(D_i)_{i\in[n]\setminus S_{\le t}}}$, and use $\widehat{\nabla^2 L}$ instead of the exact loss Hessian to avoid costly computation. Suppressing the shared regularizer arguments $(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}})$, define

$$\hat{P}_n^t := -\left[\widehat{\nabla^2 L}\Big(w_n^{-\mathbf{S}_{1:n-1}}, D_n\Big) + \nabla_{22}^2 \Omega_{\mathbf{D}_{1:n-1}^{-\mathbf{S}_{1:t}}}\right]^{-1} \nabla_{21}^2 \Omega_{\mathbf{D}_{1:n-1}^{-\mathbf{S}_{1:t}}},$$

$$\hat{G}_n^t := \left[\widehat{\nabla^2 L}\Big(w_n^{-\mathbf{S}_{1:n-1}}, D_n\Big) + \nabla_{22}^2 \Omega_{\mathbf{D}_{1:n-1}^{-\mathbf{S}_{1:t}}}\right]^{-1} \left[\nabla_2 \Omega_{D_{1:n-1}^{-\mathbf{S}_{1:n-1}}} - \nabla_2 \Omega_{\mathbf{D}_{1:n-1}^{-\mathbf{S}_{1:t}}}\right]. \tag{13}$$

The correction-propagation operator $\hat{P}_n^t$ uses local second-order curvature to propagate corrections through subsequent CL updates, while $\hat{G}_n^t$ corrects the mismatch between the original and retraining-induced regularizers.

The Hessian-based algorithm operates as follows. If $S_t = \emptyset$, it leaves the internal model unchanged and returns $\mathcal{R}_{\mathcal{A}}\big(w_t^{-\mathbf{S}_{1:t-1}}, \mathbf{D}_{1:t}, \mathbf{S}_{1:t}\big) = w_t^{-\mathbf{S}_{1:t-1}}$. Upon the k-th nonempty request S_{t_k} at time t_k, with previous unlearning times $t_1 < \cdots < t_{k-1}$, the update is $\mathcal{R}_{\mathcal{A}}(w_{t_k}^{-\mathbf{S}_{1:t_k-1}}, \mathbf{D}_{1:t_k}, \mathbf{S}_{1:t_k}) = w_{t_k}^{-\mathbf{S}_{1:t_k-1}} + \bar{\Delta}_{t_k}$, where

$$\bar{\Delta}_{t_k} = \sum_{n\in S_{\le t_k}} \Bigg(\prod_{\substack{j=n+1\\ j\notin S_{\le t_k}}}^{t_k} \hat{P}_j^{t_k}\Bigg)\Delta_n + \sum_{\substack{n=\underline{s}^{t_k}+1\\ n\notin S_{\le t_k}}}^{t_k} \Bigg(\prod_{\substack{j=n+1\\ j\notin S_{\le t_k}}}^{t_k} \hat{P}_j^{t_k}\Bigg)\hat{G}_n^{t_k} - \sum_{i=1}^{k-1} \Bigg(\prod_{\substack{j=t_i+1\\ j\notin S_{\le t_k}}}^{t_k} \hat{P}_j^{t_k}\Bigg)\bar{\Delta}_{t_i}, \tag{14}$$

with $\Delta_i = w_{i-1}^{-\mathbf{S}_{1:i-1}} - w_i^{-\mathbf{S}_{1:i-1}}$ and $\bar{\Delta}_i = w_i^{-\mathbf{S}_{1:i}} - w_i^{-\mathbf{S}_{1:i-1}}$. At each subsequent learning stage $n > t_k$ in (Reg-CL), the model is trained using the oracle estimator $\Omega_{\mathbf{D}_{1:n}^{-\mathbf{S}_{1:t_k}}}$ of the retraining-induced regularizer.

5.1 Forward-Synchronous Unlearning

The update in (14) removes newly deleted tasks' influence and corrects the resulting regularizer mismatch and disruptions to previous corrections. Crucially, these corrections depend strongly on request order. We define a **forward-synchronous** sequence as satisfying

$$s > t_i, \qquad \forall s \in S_{t_{i+1}}, \qquad i = 1, \ldots, k-1. \tag{15}$$

A sequence violating this condition is asynchronous. Under forward synchronicity, (14) simplifies: the first summation reduces to S_{t_k}, the second to $n = t_{k-1}+1, \ldots, t_k$, and the third vanishes. Thus, each request targets only recent tasks arriving after the previous unlearning time, avoiding disruptions to prior operations and simplifying the corrections.

Lemma 5.1. *Under an arbitrary unlearning order,* (14) *requires $O(t(d_1+d_2))$ storage and computation, where d_1 and d_2 are the dimensions of the approximate Hessian and regularizer, respectively. In particular, the forward-synchronous order in* (15) *reduces both costs to $O((d_1+d_2)\max_i(t_i - t_{i-1}))$, where t_i is the i-th unlearning time.*

Beyond costs, approximation error depends on whether the sequence is synchronous. Since asynchrony adds error, we first bound the error in (8) under a forward-synchronous sequence.

Theorem 5.2. *Let $t_0 := \min S_{t_1}$, and suppose $\{t_1, \ldots, t_k\}$ are the unlearning times. For an algorithm of the form* (Reg-CL) *under the forward-synchronous sequence in* (15)*, the Hessian-*

unlearning error $\|w_{t_k}^{-\mathbf{S}_{1:t_k}} - w_{t_k}^{-S_{\le t_k}}\|$ *is upper bounded by* $\gamma_{\mathrm{FS}}(\mathbf{S}_{1:t_k})$*, where* $\gamma_{\mathrm{FS}}(\mathbf{S}_{1:t_k})$ *equals*

$$\underbrace{\sum_{j=1}^{k}\sum_{s\in S_{t_j}}\left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k}\rho_a^{t_k}\right)\sum_{\substack{q=s+1\\ q\notin S_{\le t_k}}}^{t_j}\frac{\bar{\rho}_q^{t_k}}{\rho_q^{t_k}}\Gamma_s}_{\text{deletion induced loss}}+\underbrace{\sum_{j=1}^{k}\sum_{\substack{s=t_{j-1}\\ s\notin S_{t_j}}}^{t_j}\left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k}\rho_a^{t_k}\right)\left[\left(\sum_{\substack{q=s+1\\ q\notin S_{\le t_k}}}^{t_j}\frac{\bar{\rho}_q^{t_k}}{\rho_q^{t_k}}g_s^{t_k}\right)+\bar{g}_s^{t_k}\right]}_{\text{regularizer mismatch loss}},\quad(16)$$

where Γ_s*,* $\rho_a^{t_k}$*,* $\bar{\rho}_q^{t_k}$*,* $g_s^{t_k}$*, and* $\bar{g}_s^{t_k}$ *are algorithm-specific parameters of* (Reg-CL)*.*

In (16), the first term captures the Hessian-unlearning error from each deleted task, and the second captures the error from approximating the retraining regularizer $\Omega_{(D_i)_{i\in[n]\setminus S_{\le t}}}$ with the oracle $\Omega_{\mathbf{D}_{1:n}^{-\mathbf{S}_{1:t}}}$ in (13) and (14).

5.2 Asynchronous Unlearning Sequence

Under an order violating (15), each asynchronous task adds an error term. A key analytical challenge is that a retained task contributing regularizer-mismatch error may later be deleted, causing mismatch and deletion errors to interact. By sequentially tracking each error source across tasks and unlearning times, we establish a general bound for arbitrary request orders. Its full form is deferred to Appendix D, and an intuitive illustration is as follows.

Consider $t_1 < t_2$, with $S_{t_1} = \{s_1\}$, $S_{t_2} = \{s_2, s_3\}$, and $s_1 < s_2 < s_3 < t_1 < t_2$. At t_1, the correction for s_1 propagates through retained tasks $n \in [s_1, t_1]$. Later deleting s_2 and s_3 creates two forms of asynchrony. A later-deleted task between n and t_1 disrupts correction propagation from n to t_1, shown in blue. One before n changes the retraining history at n, increasing regularizer mismatch, shown in purple. Both effects occur when later-deleted tasks lie on both sides of n. Consequently, using $\gamma_{\mathrm{FS}}(\mathbf{S}_{1:t_2})$ from (16), $\left\|w_{t_2}^{-\mathbf{S}_{1:t_2}} - w_{t_2}^{-S_{\le t_2}}\right\|$ is bounded by

$$\sum_{i=s_1}^{s_2-1} c_1(i) + \sum_{i=s_2+1}^{s_3-1}\big(c_1(i)+c_2(i)\big) + \sum_{i=s_3+1}^{t_1-1} c_2(i) + \gamma_{\mathrm{FS}}(\mathbf{S}_{1:t_2}),\quad(17)$$

where $c_1(i)$ denotes the error for task i induced by correction-interruption asynchrony, and $c_2(i)$ denotes the error induced by regularizer-mismatch asynchrony.

5.3 Reg-CL Algorithm Choice under Hessian Unlearning

For the task-additive regularizers in (11), we use the following natural estimator $\Omega_{\mathbf{D}_{1:n-1}^{-\mathbf{S}_{1:t}}}(x,y)$ for the retraining regularizer: $\frac{1}{2}\|y-x\|^2_{\lambda I+\sum_{\substack{i=1\\ i\notin S_{\le t}}}^{n-1} F\left(w_i^{-\mathbf{S}_{1:i-1}}, D_i\right)}$. For clarity, Appendix D.3 instantiates Theorem 5.2 for task-additive regularization under a single request, and we summarize its implications below. Choosing between task-additive and ℓ_2 regularization reveals a curvature-mismatch trade-off. Task-additive regularization provides richer curvature information for Hessian correction but introduces mismatch after deletion. It outperforms the history-independent ℓ_2 regularizer when the curvature gain dominates this mismatch, particularly under strong, well-conditioned regularization. Similarly, natural forgetting favors early deletions with more subsequent updates, whereas Hessian unlearning directly corrects requested tasks and favors recent deletions with s close to t.

6 Experiments

We simulate CLU on CIFAR-100 (Krizhevsky, 2009) using a pretrained ResNet-18 whose final classification layer is replaced by a trainable three-layer neural-network head, following Koloskova et al. (2025); Hu et al. (2026).

We divide the stream into 10 tasks and issue one request at $t = 10$, varying the deleted task from 1 to 10 to study deletion timing. We evaluate natural forgetting and Hessian-based unlearning across EWC, MAS, diagonal-Hessian, scalar ℓ_2, and accumulated ℓ_2 regularization. For each combination, we measure the final approximation error $\left\|w_{10}^{-S_{\le 10}} - w_{10}^{-\mathbf{S}_{1:10}}\right\|$ and retrained-model accuracy, corresponding to $\mathcal{E}_{\mathrm{U}}$ in (4) and $\mathcal{E}_{\mathrm{L}}$ in (5). Table 1 reports results for which neither metric can improve without worsening the other.

Table 1 identifies the preferred regimes of different CL regularizers. Diag-Hessian achieves the highest accuracy across all three deletion settings, consistent with Section 3, where structured reg-

Table 1: Pareto-optimal results for jointly minimizing unlearning loss and CL excess risk in single-request CLU.

Del. Task	Reg-CL Alg.	Approx. err. ↓	Acc. ↑
2	MAS	**0.0045** (H)	0.4979
	Accum-ℓ_2	0.0094 (H)	0.4987
	Diag-Hessian	0.2070 (N)	**0.5949**
5	Accum-ℓ_2	**0.000951** (H)	0.4987
	EWC	0.0772 (N)	0.5903
	Diag-Hessian	0.0825 (N)	**0.5949**
8	MAS	**0.0000545** (H)	0.4979
	Accum-ℓ_2	0.000185 (H)	0.4987
	Hessian	0.0550 (H)	**0.5949**

H and **N** denote Hessian unlearning and natural forgetting, respectively.

Table 2: Runtime, storage, and final test accuracy under multi-request CLU.

(a) Runtime and storage

Reg. Alg.	UL Method	Time (s)	Storage (MiB)
Diag-Hessian	Hessian	231.6	161.6
	Natural	233.8	71.1
	Enhanced	235.1	114.6
	Retrain	455.5	165.2
MAS	Hessian	258.7	759.4
	Natural	218.4	595.7
	Enhanced	258.4	683.4
	Retrain	436.0	691.2

(b) Final test accuracy

Reg. Alg.	Enhanced + noise (%)	Retrain, no noise (%)
ℓ_2	52.80 ± 0.17	52.28
Diag-Hessian	55.65 ± 0.38	56.21
MAS	49.42 ± 0.03	49.41

Storage is the sum of peak GPU allocation and peak CPU data storage. Enhanced denotes forgetting-enhanced Hessian unlearning.

ularization better balances model drift. MAS gives the lowest unlearning loss for tasks 2 and 8, while accumulated ℓ_2 performs best for task 5, supporting Sections 4-5: stronger regularization can improve unlearning. The preferred unlearning method also depends on deletion position. Under Diag-Hessian, natural forgetting performs better for early deletions, which benefit from more subsequent updates, whereas Hessian unlearning is more effective for recent deletions, consistent with Sections 4-5. Overall, these results guide method selection by deletion regime and the relative priority of predictive accuracy and privacy.

We next consider a 20-task stream with multiple requests, deleting $\{4, 8\}$, $\{5, 10, 13\}$, and $\{14\}$ at times 12, 16, and 18, respectively. We compare natural forgetting, Hessian unlearning, forgetting-enhanced Hessian, and exact retraining under ℓ_2, Diag-Hessian, and MAS. Forgetting-enhanced Hessian selects between natural forgetting and Hessian correction for each deleted task, as detailed in Appendix E. Figure 2 shows that each request increases approximation error, which subsequent learning gradually reduces. Under ℓ_2, natural forgetting performs better for the first request and Hessian unlearning for later ones. Combining these regimes gives forgetting-enhanced Hessian the lowest final error. Diag-Hessian shows a similar transition but incurs additional error from its anisotropic, history-dependent regularizer. MAS yields substantially smaller errors across all requests, with Hessian unlearning performing best throughout under stronger regularization. These trends extend the single-request results to multiple deletions and agree with Theorems 4.1 and 5.2.

Table 2(a) shows that exact retraining takes approximately twice as long because it repeats CL after each request. We omit ℓ_2, whose costs are similar to Diag-Hessian. It retains all historical raw datasets and thus requires more storage than Hessian unlearning under Diag-Hessian and ℓ_2, despite the latter storing historical Hessians. Natural forgetting requires the least additional state, while the enhanced method saves storage by applying Hessian corrections selectively. Under MAS, however, Hessian unlearning exceeds retraining's storage because of its additional regularizer state. Table 2(b) reports final accuracy as an empirical measure of excess risk. Retrained-model accuracy reflects the CL excess risk $\mathcal{E}_{\mathrm{L}}$ in (5), while forgetting-enhanced Hessian accuracy reflects the post-unlearning excess risk in Definition 2.2. Across all settings, our CLU model remains comparable to retraining, and Diag-Hessian achieves the highest accuracy, consistent with Section 3. Overall, forgetting-enhanced Hessian balances post-unlearning excess risk and costs.

7 Conclusion

We studied certified unlearning for regularization-based continual learning and proposed complementary natural-forgetting and Hessian-based methods. Analysis and experiments reveal how CL regularization and request order govern the trade-off between knowledge retention, unlearning loss, and cost, providing guidance for CLU design.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript and assist with running experiments. We did not use generative AI tools to develop the ideas or prove the theoretical results. The remaining disclosure categories are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the datasets, data-stream construction, model architectures, training procedures, hyperparameters, and evaluation protocols in Appendix G.2. All assumptions and complete proofs of our theoretical results are included in the corresponding appendices. We will provide the complete experimental code after acceptance.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 139–154, 2018.

Umit Yigit Basaran, Sk Miraj Ahmed, Amit Roy-Chowdhury, and Basak Guler. A certified unlearning approach without access to source data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=8lt5776GLB.

Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 11186–11194, 2024.

Romit Chatterjee, Vikram Chundawat, Ayush Tarun, Ankur Mali, and Murari Mandal. A unified framework for continual learning and machine unlearning. *arXiv e-prints*, pp. arXiv–2408, 2024.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Certified machine unlearning via noisy stochastic gradient descent. *Advances in Neural Information Processing Systems*, 37:38852–38887, 2024a.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Langevin unlearning: A new perspective of noisy gradient descent for machine unlearning. *Advances in neural information processing systems*, 37: 79666–79703, 2024b.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yiting Hu, Lingjie Duan, and Qian Zhang. The forgetting-retention dilemma: Certified unlearning theory in continual learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=zlIyJEp9nM.

Zhehao Huang, Xinwen Cheng, Jie Zhang, Jinghao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. A unified gradient-based framework for task-agnostic continual learning-unlearning. *arXiv preprint arXiv:2505.15178*, 2025.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. Certified unlearning for neural networks. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=3rWQlV3s1I.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/54fe976ba170c19ebae453679b362263-Paper.pdf.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pp. 243–254. PMLR, 2022.

Jiaqi Liu, Jian Lou, Zhan Qin, and Kui Ren. Certified minimax unlearning with generalization rates and deletion capacity. *Advances in Neural Information Processing Systems*, 36:62821–62852, 2023.

Siqiao Mu and Diego Klabjan. Rewind-to-delete: Certified machine unlearning for nonconvex functions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=FgjcLXIUjr.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46, 2025.

Xinbao Qiao, Meng Zhang, Ming Tang, and Ermin Wei. Hessian-free online certified unlearning. *arXiv preprint arXiv:2404.01712*, 2024.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Vinith Suriyakumar and Ashia C Wilson. Algorithms that approximate data removal: New results and limitations. *Advances in Neural Information Processing Systems*, 35:18892–18903, 2022.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer International Publishing, Cham, 2017. doi: 10.1007/978-3-319-57959-7.

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

Dong Yin, Mehrdad Farajtabar, Ang Li, Nir Levine, and Alex Mott. Optimization and generalization of regularization-based continual learning: a loss approximation viewpoint. *arXiv preprint arXiv:2006.10974*, 2020.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. Pmlr, 2017.

Xuyang Zhao, Huiyuan Wang, Weiran Huang, and Wei Lin. A statistical theory of regularization-based continual learning. *arXiv preprint arXiv:2406.06213*, 2024.

A Notation and Common Operators

We first collect the common trajectory and operator notation used throughout the appendix. Request-order-specific notation is introduced later in Appendix D.1.

Operator Convention. For a matrix-valued sequence, products are time ordered as $\prod_{j=a}^{b} A_j := A_b A_{b-1} \cdots A_a$, and an empty product equals the identity matrix. Unless stated otherwise, all suprema below are over admissible data streams, request histories, and iterates satisfying the standing assumptions.

Basic Quantities and Averaged Derivatives. For $n \le t$, let $D_{1:n}^{-S_{\le t}} = (D_i)_{i\in[n]\setminus S_{\le t}}$ denote the datasets through time n retained at time t, in their original chronological order, and let $\Omega_{D_{1:n}^{-S_{\le t}}}$ be the regularizer induced by retraining on this sequence. We further define

$$\Delta_s := w_{s-1}^{-\mathbf{S}_{1:s-1}} - w_s^{-\mathbf{S}_{1:s-1}},$$

$$H_n := \nabla^2 L(w_n^{-\mathbf{S}_{1:n-1}}, D_n), \tag{18}$$

$$\hat{H}_n := \widehat{\nabla^2 L}(w_n^{-\mathbf{S}_{1:n-1}}, D_n), \tag{19}$$

$$\widetilde{H}_n(x,y) := \int_0^1 \nabla^2 L\big(x + u(y-x), D_n\big)\, du, \tag{20}$$

For $(a,b) \in \{1,2\}^2$, define the line-averaged derivatives

$$\nabla^2_{ab}\widetilde{\Omega}_{D_{1:n}^{-S_{\le t}}}(x_1,y_1,x_2,y_2) = \int_0^1 \nabla^2_{ab}\Omega_{D_{1:n}^{-S_{\le t}}}\big(x_1 + u(y_1-x_1), x_2 + u(y_2-x_2)\big)\, du, \tag{21}$$

Exact Local Operators. The following operators describe exact local correction propagation and regularizer mismatch:

$$P_n^t := -\left[H_n + \nabla^2_{22}\Omega_{D_{1:n-1}^{-S_{\le t}}}\left(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}\right)\right]^{-1} \nabla^2_{21}\Omega_{D_{1:n-1}^{-S_{\le t}}}\left(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}\right), \tag{22}$$

$$G_n^t := \left[H_n + \nabla^2_{22}\Omega_{D_{1:n-1}^{-S_{\le t}}}\left(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}\right)\right]^{-1} \left[\nabla_2\Omega_{D_{1:n-1}^{-\mathbf{S}_{1:n-1}}}\left(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}\right) - \nabla_2\Omega_{D_{1:n-1}^{-\mathbf{S}_{1:t}}}\left(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}\right)\right]. \tag{23}$$

Approximate Local Operators. Replacing the exact loss Hessian H_n by its computational approximation $\hat{H}_n$ gives the operators used by Hessian unlearning:

$$\hat{P}_n^t := -\left[\hat{H}_n + \nabla^2_{22}\Omega_{D_{1:n-1}^{-S_{\le t}}}\left(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}\right)\right]^{-1} \nabla^2_{21}\Omega_{D_{1:n-1}^{-S_{\le t}}}\left(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}\right), \tag{24}$$

$$\hat{G}_n^t := \left[\hat{H}_n + \nabla^2_{22}\Omega_{D_{1:n-1}^{-S_{\le t}}}\left(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}\right)\right]^{-1} \left[\nabla_2\Omega_{D_{1:n-1}^{-\mathbf{S}_{1:n-1}}}\left(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}\right) - \nabla_2\Omega_{D_{1:n-1}^{-\mathbf{S}_{1:t}}}\left(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}\right)\right]. \tag{25}$$

Path-Averaged Operators. For the approximation-error analysis, the corresponding path-averaged operators connect the retraining and CLU trajectories:

$$\widetilde{P}_n^t := -\left[\widetilde{H}_n\left(w_n^{-S_{\le t}}, w_n^{-\mathbf{S}_{1:n-1}}\right) + \nabla_{22}^2\widetilde{\Omega}_{D_{1:n-1}^{-S_{\le t}}}\begin{pmatrix} w_{n-1}^{-S_{\le t}}, w_{n-1}^{-\mathbf{S}_{1:n-1}}, \\ w_n^{-S_{\le t}}, w_n^{-\mathbf{S}_{1:n-1}}\end{pmatrix}\right]^{-1} \tag{26}$$

$$\nabla_{21}^2\widetilde{\Omega}_{D_{1:n-1}^{-S_{\le t}}}\begin{pmatrix} w_{n-1}^{-S_{\le t}}, w_{n-1}^{-\mathbf{S}_{1:n-1}}, \\ w_n^{-S_{\le t}}, w_n^{-\mathbf{S}_{1:n-1}}\end{pmatrix}, \tag{27}$$

$$\widetilde{G}_n^t := \left[\widetilde{H}_n\left(w_n^{-S_{\le t}}, w_n^{-\mathbf{S}_{1:n-1}}\right) + \nabla_{22}^2\widetilde{\Omega}_{D_{1:n-1}^{-S_{\le t}}}\begin{pmatrix} w_{n-1}^{-S_{\le t}}, w_{n-1}^{-\mathbf{S}_{1:n-1}}, \\ w_n^{-S_{\le t}}, w_n^{-\mathbf{S}_{1:n-1}}\end{pmatrix}\right]^{-1}$$
$$\times\left[\nabla_2\Omega_{D_{1:n-1}^{-\mathbf{S}_{1:n-1}}}\left(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}\right) - \nabla_2\Omega_{D_{1:n-1}^{-S_{\le t}}}\left(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}\right)\right]. \tag{28}$$

B Proofs for the Reg-CL Excess-Risk Analysis

Proof of Lemma 3.1. Under ℓ_2 regularization, the objective in (Reg-CL) becomes

$$\frac{1}{n}\|X_t w - y_t\|_2^2 + \frac{\lambda}{n}\|w - w_{t-1}\|_2^2.$$

The first-order condition for this objective, multiplied by n, gives

$$\left(X_t^\top X_t + \lambda I\right)w_t = X_t^\top y_t + \lambda w_{t-1}.$$

Substituting $y_t = X_t w^\star + \eta_t$ and subtracting $\left(X_t^\top X_t + \lambda I\right)w^\star$ from both sides yield

$$w_t - w^\star = \lambda\left(X_t^\top X_t + \lambda I\right)^{-1}\left(w_{t-1} - w^\star\right) + \left(X_t^\top X_t + \lambda I\right)^{-1}X_t^\top\eta_t.$$

It follows that

$$\mathbb{E}\|w_t - w^\star\|_2^2 = a_\lambda^2 E\|w_{t-1} - w^\star\|_2^2 + b_\lambda^2,$$

where $a_\lambda^2 = \frac{\lambda^2}{p}\mathbb{E}\left[\mathrm{tr}(X_t^\top X_t + \lambda I_p)^{-2}\right]$, and $b_\lambda^2 = \frac{\sigma^2}{n}\mathbb{E}\left[\mathrm{tr}\left(X_t^\top X_t(X_t^\top X_t + \lambda I_p)^{-1}\right)\right]$. Repeatedly applying this recursion and using $w_0 = 0$ give

$$\mathbb{E}\|w_T - w^\star\|_2^2 \approx a_\lambda^{2T}\|w^\star\|_2^2 + b_\lambda^2\sum_{j=0}^{T-1}a_\lambda^{2j} = a_\lambda^{2T}\|w^\star\|_2^2 + b_\lambda^2\frac{1 - a_\lambda^{2T}}{1 - a_\lambda^2}.$$

Since $X_t^\top X_t$ is positive semidefinite, its eigenvalues $s_1, \ldots, s_p$ are nonnegative. Hence,

$$\frac{\lambda^2}{p}\,\mathrm{tr}\left(X_t^\top X_t + \lambda I_p\right)^{-2} = \frac{1}{p}\sum_{i=1}^{p}\frac{\lambda^2}{(s_i + \lambda)^2} = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{\lambda}{s_i + \lambda}\right)^2.$$

Because $s_i \ge 0$ and $\lambda > 0$, for every i,

$$0 < \left(\frac{\lambda}{s_i + \lambda}\right)^2 \le 1.$$

Therefore, pointwise,

$$0 < \frac{\lambda^2}{p}\,\mathrm{tr}\left(X_t^\top X_t + \lambda I_p\right)^{-2} \le 1.$$

Then, since $\mathbb{P}(X_t \ne 0) > 0$ we have

$$0 < \frac{\lambda^2}{p}\mathbb{E}\left[\mathrm{tr}\left(X_t^\top X_t + \lambda I_p\right)^{-2}\right] < 1.$$

This proves the result. □

Proof of Lemma 3.2. Under accumulated-Hessian regularization, the objective in (Reg-CL) becomes

$$\frac{1}{n}\|X_t w - y_t\|_2^2 + \|w - w_{t-1}\|_{H_t}^2.$$

The first-order condition for this objective gives

$$\left(\frac{1}{n}X_t^\top X_t + H_t\right)w_t = \frac{1}{n}X_t^\top y_t + H_t w_{t-1}.$$

Since $H_{t+1} = H_t + X_t^\top X_t/n$, substituting $y_t = X_t w^\star + \eta_t$ and subtracting $H_{t+1}w^\star$ yield

$$H_{t+1}\big(w_t - w^\star\big) = H_t\big(w_{t-1} - w^\star\big) + \frac{1}{n}X_t^\top \eta_t.$$

Iterating this identity and using $H_1 = \lambda I$ and $w_0 = 0$ give

$$H_{T+1}\big(w_T - w^\star\big) = -\lambda w^\star + \frac{1}{n}\sum_{t=1}^{T} X_t^\top \eta_t.$$

Thus,

$$w_T - w^\star = -\lambda H_{T+1}^{-1} w^\star + H_{T+1}^{-1}\frac{1}{n}\sum_{t=1}^{T} X_t^\top \eta_t.$$

Conditioned on $X_{1:T}$, the second term has mean zero and covariance

$$\mathbb{E}_\eta\left[H_{T+1}^{-1}\left(\frac{1}{n}\sum_{t=1}^{T} X_t^\top \eta_t\right)\left(\frac{1}{n}\sum_{t=1}^{T} X_t^\top \eta_t\right)^\top H_{T+1}^{-1}\,\middle|\, X_{1:T}\right] = \frac{\sigma^2}{n}H_{T+1}^{-1}\big(H_{T+1} - \lambda I\big)H_{T+1}^{-1}.$$

The conditional bias–variance decomposition therefore yields

$$\begin{aligned}&\mathbb{E}_\eta\left[\|w_T - w^\star\|_2^2 \,\middle|\, X_{1:T}\right]\\ &= \lambda^2 (w^\star)^\top H_{T+1}^{-2} w^\star + \frac{\sigma^2}{n}\operatorname{tr}\left(\big(H_{T+1} - \lambda I\big)H_{T+1}^{-2}\right).\end{aligned} \tag{29}$$

Let

$$\bar{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_T \end{bmatrix} \in \mathbb{R}^{nT\times p}.$$

Then $H_{T+1} = \lambda I + \bar{X}^\top \bar{X}/n$. For every $0 < u < \sqrt{nT} - \sqrt{p}$, the standard smallest-singular-value bound for a Gaussian matrix (Vershynin, 2018) gives

$$\lambda_{\min}\left(\frac{1}{n}\bar{X}^\top \bar{X}\right) \geq m_u = \frac{\big(\sqrt{nT} - \sqrt{p} - u\big)^2}{n}$$

with probability at least $1 - \exp(-u^2/2)$. On this event, if $d_1, \ldots, d_p$ are the eigenvalues of $\bar{X}^\top \bar{X}/n$, then

$$\lambda^2 (w^\star)^\top H_{T+1}^{-2} w^\star \leq \left(\frac{\lambda}{\lambda + m_u}\right)^2 \|w^\star\|_2^2,$$

$$\operatorname{tr}\left(\big(H_{T+1} - \lambda I\big)H_{T+1}^{-2}\right) = \sum_{j=1}^{p}\frac{d_j}{(\lambda + d_j)^2} \leq \frac{p}{\lambda + m_u}.$$

Choosing $u = (\sqrt{nT} - \sqrt{p})/2$ gives

$$m_u = \frac{(\sqrt{nT} - \sqrt{p})^2}{4n} = \frac{T}{4}\left(1 - \sqrt{\frac{p}{nT}}\right)^2 \geq \frac{T}{4}\left(1 - \frac{1}{\sqrt{2}}\right)^2,$$

where the final inequality follows from $n > p$ and $T \geq 2$. Thus $m_u \geq c_0 T$ for an absolute constant $c_0 > 0$. Substituting these bounds into (29) gives

$$\mathbb{E}_\eta\big[\|w_T - w^\star\|_2^2 \mid X_{1:T}\big] \lesssim \left(\frac{\lambda}{\lambda + T}\right)^2 \|w^\star\|_2^2 + \frac{\sigma^2 p}{n(\lambda + T)}$$

with probability at least $1 - \exp\big(-(\sqrt{nT} - \sqrt{p})^2/8\big)$ over $X_{1:T}$. Moreover, since $H_{T+1} \succeq \lambda I$ and $d/(\lambda + d)^2 \leq 1/(4\lambda)$, (29) gives the uniform bound

$$\mathbb{E}_\eta\big[\|w_T - w^\star\|_2^2 \mid X_{1:T}\big] \leq \|w^\star\|_2^2 + \frac{\sigma^2 p}{4n\lambda}.$$

Taking expectation over $X_{1:T}$ and separating the preceding high-probability event from its complement yield

$$\begin{aligned}\mathbb{E}_{X,\eta}\|w_T - w^\star\|_2^2 \lesssim\ &\left(\frac{\lambda}{\lambda + T}\right)^2 \|w^\star\|_2^2 + \frac{\sigma^2 p}{n(\lambda + T)}\\ &+ \left(\|w^\star\|_2^2 + \frac{\sigma^2 p}{4n\lambda}\right)\exp\left(-\frac{(\sqrt{nT} - \sqrt{p})^2}{8}\right).\end{aligned}$$

Every term on the right-hand side converges to zero as $T \to \infty$. This proves the result. □

Theorem B.1 (Complete version of Theorem 3.3)**.** *Suppose Assumptions 2.3 and 2.4 hold, and let* $B = \sup_{w\in\mathcal{W},\, z\in\mathcal{Z}} |\ell(w,z)|$. *Let* $1 \le \tau_1 < \cdots < \tau_{N_t} \le t$ *denote the retained task indices at time* t, *so that* $\{\tau_1, \ldots, \tau_{N_t}\} = [t] \setminus S_{\le t}$. *For the perfect retraining model* $w_t^{-S_{\le t}} = \mathcal{A}\big((D_i)_{i\in[t]\setminus S_{\le t}}\big)$, *the CL excess risk defined in* (5) *satisfies* $\mathcal{E}_{\mathrm{L}} \le \mathcal{E}^{-S_{\le t}}(\lambda)$, *where*

$$\mathcal{E}^{-S_{\le t}}(\lambda) = \frac{L}{N_t}\left[\sum_{k=1}^{N_t}\sum_{i=2}^{N_t}\left(\prod_{j=i+1}^{N_t} q_j\right)\bar{q}_i \left\|\hat{w}_{\tau_i} - \hat{w}_{\tau_k}\right\| + \sum_{k=2}^{N_t}\left(\prod_{j=2}^{N_t} q_j\right)\left\|\hat{w}_{\tau_k} - \hat{w}_{\tau_1}\right\|\right] + \Gamma\big((n_{\tau_k})_{k=1}^{N_t}\big) \tag{30}$$

and, for $i = 2, \ldots, N_t$,

$$q_i = \sup\left\|\left[\widetilde{H}_{\tau_i}(w_{\tau_i}^{-S_{\le t}}, \hat{w}_{\tau_i}) + \nabla_{22}^2\widetilde{\Omega}_{D_{1:\tau_{i-1}}^{-S_{\le t}}}\left(w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_i}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}\right)\right]^{-1} \nabla_{22}^2\widetilde{\Omega}_{D_{1:\tau_{i-1}}^{-S_{\le t}}}\left(w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_i}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}\right)\right\|, \tag{31}$$

$$\bar{q}_i = \sup\left\|\left[\widetilde{H}_{\tau_i}(w_{\tau_i}^{-S_{\le t}}, \hat{w}_{\tau_i}) + \nabla_{22}^2\widetilde{\Omega}_{D_{1:\tau_{i-1}}^{-S_{\le t}}}\left(w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_i}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}\right)\right]^{-1} \widetilde{H}_{\tau_i}(w_{\tau_i}^{-S_{\le t}}, \hat{w}_{\tau_i})\right\|. \tag{32}$$

Here, the generalization term $\Gamma\big((n_{\tau_k})_{k=1}^{N_t}\big)$ *is*

$$\Gamma\big((n_{\tau_k})_{k=1}^{N_t}\big) := \sum_{k=1}^{N_t}\left[\frac{4B}{N_t}\sqrt{\frac{2d\ln\big(2Lr\sqrt{n_{\tau_k}}/B + 1\big)}{n_{\tau_k}}} + \frac{2B}{N_t\sqrt{n_{\tau_k}}}\right]. \tag{33}$$

Proof of Theorem B.1. We work directly with the retained sequence $\tau_1 < \cdots < \tau_{N_t}$. Since $\Omega_\emptyset = 0$, the first retained update minimizes only $L(w, D_{\tau_1})$. We therefore have

$$w_{\tau_1}^{-S_{\le t}} = \hat{w}_{\tau_1}. \tag{34}$$

For every $i = 2, \ldots, N_t$, the first-order conditions for the retraining model $w_{\tau_i}^{-S_{\le t}}$ and the empirical minimizer $\hat{w}_{\tau_i}$ are

$$\nabla L(w_{\tau_i}^{-S_{\le t}}, D_{\tau_i}) + \nabla_2\Omega_{D_{1:\tau_{i-1}}^{-S_{\le t}}}(w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_i}^{-S_{\le t}}) = 0, \tag{35}$$

$$\nabla L(\hat{w}_{\tau_i}, D_{\tau_i}) = 0. \tag{36}$$

The integral Taylor identities in (20) and (21) give

$$\nabla L(w_{\tau_i}^{-S_{\le t}}, D_{\tau_i}) - \nabla L(\hat{w}_{\tau_i}, D_{\tau_i}) = \widetilde{H}_{\tau_i}(w_{\tau_i}^{-S_{\le t}}, \hat{w}_{\tau_i})(w_{\tau_i}^{-S_{\le t}} - \hat{w}_{\tau_i}), \tag{37}$$

$$\nabla_2\Omega_{D_{1:\tau_{i-1}}^{-S_{\le t}}}(w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_i}^{-S_{\le t}}) - \nabla_2\Omega_{D_{1:\tau_{i-1}}^{-S_{\le t}}}(w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}) \tag{38}$$

$$= \nabla_{22}^2\widetilde{\Omega}_{D_{1:\tau_{i-1}}^{-S_{\le t}}}\left(w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_i}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}\right)(w_{\tau_i}^{-S_{\le t}} - w_{\tau_{i-1}}^{-S_{\le t}}). \tag{39}$$

Because the regularizer is minimized at equal arguments, the second gradient on the left-hand side of (39) is zero. Combining (35)– (39) gives

$$\widetilde{H}_{\tau_i}(w_{\tau_i}^{-S_{\le t}}, \hat{w}_{\tau_i})(w_{\tau_i}^{-S_{\le t}} - \hat{w}_{\tau_i}) + \nabla_{22}^2\widetilde{\Omega}_{D_{1:\tau_{i-1}}^{-S_{\le t}}}\left(w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_i}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}\right)(w_{\tau_i}^{-S_{\le t}} - w_{\tau_{i-1}}^{-S_{\le t}}) = 0. \tag{40}$$

Define

$$P_i := \left[\widetilde{H}_{\tau_i}(w_{\tau_i}^{-S_{\le t}}, \hat{w}_{\tau_i}) + \nabla_{22}^2 \widetilde{\Omega}_{D_{1:\tau_{i-1}}^{-S_{\le t}}}\left(w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_i}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}\right)\right]^{-1} \times \nabla_{22}^2 \widetilde{\Omega}_{D_{1:\tau_{i-1}}^{-S_{\le t}}}\left(w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_i}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}\right), \tag{41}$$

$$Q_i := \left[\widetilde{H}_{\tau_i}(w_{\tau_i}^{-S_{\le t}}, \hat{w}_{\tau_i}) + \nabla_{22}^2 \widetilde{\Omega}_{D_{1:\tau_{i-1}}^{-S_{\le t}}}\left(w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}, w_{\tau_i}^{-S_{\le t}}, w_{\tau_{i-1}}^{-S_{\le t}}\right)\right]^{-1} \times \widetilde{H}_{\tau_i}(w_{\tau_i}^{-S_{\le t}}, \hat{w}_{\tau_i}). \tag{42}$$

Equation (40) then becomes

$$w_{\tau_i}^{-S_{\le t}} = Q_i \hat{w}_{\tau_i} + P_i w_{\tau_{i-1}}^{-S_{\le t}}. \tag{43}$$

Moreover,

$$P_i + Q_i = I, \qquad \|P_i\| \le q_i, \qquad \|Q_i\| \le \bar{q}_i. \tag{44}$$

For matrix products, we use $\prod_{j=a}^{b} P_j = P_b P_{b-1} \cdots P_a$, with every empty product equal to I. Iterating (43) from the base case in (34) yields

$$w_t^{-S_{\le t}} = \sum_{i=2}^{N_t} \left(\prod_{j=i+1}^{N_t} P_j\right) Q_i \hat{w}_{\tau_i} + \left(\prod_{j=2}^{N_t} P_j\right) \hat{w}_{\tau_1}. \tag{45}$$

Using $P_i + Q_i = I$ recursively also gives

$$\sum_{i=2}^{N_t} \left(\prod_{j=i+1}^{N_t} P_j\right) Q_i + \prod_{j=2}^{N_t} P_j = I. \tag{46}$$

Subtracting $\hat{w}_{\tau_k}$ from (45) and using (46), for every $k = 1, \ldots, N_t$ we obtain

$$w_t^{-S_{\le t}} - \hat{w}_{\tau_k} = \sum_{i=2}^{N_t} \left(\prod_{j=i+1}^{N_t} P_j\right) Q_i (\hat{w}_{\tau_i} - \hat{w}_{\tau_k}) + \left(\prod_{j=2}^{N_t} P_j\right) (\hat{w}_{\tau_1} - \hat{w}_{\tau_k}). \tag{47}$$

The triangle inequality, submultiplicativity, and (44) imply

$$\|w_t^{-S_{\le t}} - \hat{w}_{\tau_k}\| \le \sum_{i=2}^{N_t} \left(\prod_{j=i+1}^{N_t} q_j\right) \bar{q}_i \|\hat{w}_{\tau_i} - \hat{w}_{\tau_k}\| + \left(\prod_{j=2}^{N_t} q_j\right) \|\hat{w}_{\tau_1} - \hat{w}_{\tau_k}\|. \tag{48}$$

It remains to convert this model-distance bound into a population-risk bound. The minimum of a sum is no smaller than the sum of the individual minima, so

$$\min_{w \in \mathcal{W}} \frac{1}{N_t} \sum_{k=1}^{N_t} R_{\tau_k}(w) \ge \frac{1}{N_t} \sum_{k=1}^{N_t} \min_{w \in \mathcal{W}} R_{\tau_k}(w). \tag{49}$$

Consequently, adding and subtracting $R_{\tau_k}(\hat{w}_{\tau_k})$ for every retained task gives

$$\mathcal{E}_{\mathrm{L}} \le \frac{1}{N_t} \sum_{k=1}^{N_t} \mathbb{E}\Big[R_{\tau_k}(w_t^{-S_{\le t}}) - R_{\tau_k}(\hat{w}_{\tau_k})\Big] + \frac{1}{N_t} \sum_{k=1}^{N_t} \left[\mathbb{E} R_{\tau_k}(\hat{w}_{\tau_k}) - \min_{w \in \mathcal{W}} R_{\tau_k}(w)\right]. \tag{50}$$

Since $\ell(\cdot, z)$ is L-Lipschitz, so is every population loss R_{τ_k}. Therefore, the first term in (50) is at most

$$\frac{L}{N_t}\sum_{k=1}^{N_t}\mathbb{E}\|w_t^{-S_{\le t}} - \hat{w}_{\tau_k}\|. \tag{51}$$

For the second term, Lemma C.1 of Hu et al. (2026) gives

$$\mathbb{E}R_{\tau_k}(\hat{w}_{\tau_k}) - \min_{w\in\mathcal{W}} R_{\tau_k}(w) \le \Gamma_{\tau_k}, \tag{52}$$

where Γ_{τ_k} is defined in (33). For fixed B, L, r, and d, Lemma C.1 in fact gives $\Gamma_{\tau_k} = O(\sqrt{\log(n_{\tau_k})/n_{\tau_k}})$, which is also $O(n_{\tau_k}^{-1/4})$.

Finally, substitute (48), (51), and (52) into (50). The second norm in (48) vanishes when $k = 1$, so its summation may start at $k = 2$. After summing over all retained tasks, the resulting expression is exactly (7). This completes the proof. □

C Natural-Forgetting Analysis

C.1 Complete Bound and Proof of Theorem 4.1

Theorem C.1 (Complete version of Theorem 4.1)**.** *Assume $S_{\le t} \neq \emptyset$, and let $\underline{s}^t := \min S_{\le t}$. If $w_t^{-\mathbf{S}_{1:t}} = \mathcal{R}_{\mathcal{A}}\Big(w_t^{-\mathbf{S}_{1:t-1}}, \mathbf{D}_{1:t}, \mathbf{S}_{1:t}\Big) = w_t^{-\mathbf{S}_{1:t-1}}$ is used in Fig. 1, then*

$$\gamma(\mathbf{S}_{1:t}) = \sum_{s\in S_{\le t}}\left(\prod_{\substack{j=s+1\\ j\notin S_{\le t}}}^{t}\rho_j^t\right)\Gamma_s + \sum_{\substack{i=\underline{s}^t+1\\ i\notin S_{\le t}}}^{t}\left(\prod_{\substack{j=i+1\\ j\notin S_{\le t}}}^{t}\rho_j^t\right)g_i^t, \tag{53}$$

where ρ_j^t, g_j^t, and Γ_j are CL algorithm specific parameters defined as follows:

$$\rho_j^t := \sup\left\|\widetilde{P}_j^t\right\|, \quad g_j^t := \sup\left\|\widetilde{G}_j^t\right\|, \Gamma_j \quad := \sup\left\|w_j^{-\mathbf{S}_{1:j-1}} - w_{j-1}^{-\mathbf{S}_{1:j-1}}\right\|. \tag{54}$$

Proof. For every retained task $n \notin S_{\le t}$, the first-order conditions for the retraining and natural-forgetting iterates are

$$\nabla L(w_n^{-S_{\le t}}, D_n) + \nabla_2\Omega_{D_{1:n-1}^{-S_{\le t}}}(w_{n-1}^{-S_{\le t}}, w_n^{-S_{\le t}}) = 0, \tag{55}$$

$$\nabla L(w_n^{-\mathbf{S}_{1:n-1}}, D_n) + \nabla_2\Omega_{D_{1:n-1}^{-\mathbf{S}_{1:n-1}}}(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}) = 0. \tag{56}$$

Subtracting (56) from (55), adding and subtracting

$$\nabla_2\Omega_{D_{1:n-1}^{-S_{\le t}}}(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}),$$

and applying the integral Taylor identities in (20) and (21) yield

$$\begin{aligned}
&\left[\widetilde{H}_n(w_n^{-S_{\le t}}, w_n^{-\mathbf{S}_{1:n-1}}) + \nabla_{22}^2\widetilde{\Omega}_{D_{1:n-1}^{-S_{\le t}}}\begin{pmatrix} w_{n-1}^{-S_{\le t}}, w_{n-1}^{-\mathbf{S}_{1:n-1}},\\ w_n^{-S_{\le t}}, w_n^{-\mathbf{S}_{1:n-1}}\end{pmatrix}\right]\left(w_n^{-S_{\le t}} - w_n^{-\mathbf{S}_{1:n-1}}\right)\\
&+ \nabla_{21}^2\widetilde{\Omega}_{D_{1:n-1}^{-S_{\le t}}}\begin{pmatrix} w_{n-1}^{-S_{\le t}}, w_{n-1}^{-\mathbf{S}_{1:n-1}},\\ w_n^{-S_{\le t}}, w_n^{-\mathbf{S}_{1:n-1}}\end{pmatrix}\left(w_{n-1}^{-S_{\le t}} - w_{n-1}^{-\mathbf{S}_{1:n-1}}\right)\\
&- \left[\nabla_2\Omega_{D_{1:n-1}^{-\mathbf{S}_{1:n-1}}}(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}}) - \nabla_2\Omega_{D_{1:n-1}^{-S_{\le t}}}(w_{n-1}^{-\mathbf{S}_{1:n-1}}, w_n^{-\mathbf{S}_{1:n-1}})\right] = 0.
\end{aligned} \tag{57}$$

By (27) and (28), (57) is equivalent to

$$w_n^{-S_{\le t}} - w_n^{-\mathbf{S}_{1:n-1}} = \widetilde{P}_n^t\left(w_{n-1}^{-S_{\le t}} - w_{n-1}^{-\mathbf{S}_{1:n-1}}\right) + \widetilde{G}_n^t. \tag{58}$$

Therefore,

$$\left\|w_n^{-S_{\le t}} - w_n^{-\mathbf{S}_{1:n-1}}\right\| \le \rho_n^t\left\|w_{n-1}^{-S_{\le t}} - w_{n-1}^{-\mathbf{S}_{1:n-1}}\right\| + g_n^t. \tag{59}$$

For a deleted task $s \in S_{\le t}$, exact retraining skips D_s, so $w_s^{-S_{\le t}} = w_{s-1}^{-S_{\le t}}$. Hence,

$$\begin{aligned}
w_s^{-S_{\le t}} - w_s^{-\mathbf{S}_{1:s-1}} &= w_{s-1}^{-S_{\le t}} - w_{s-1}^{-\mathbf{S}_{1:s-1}}\\
&\quad - \left(w_s^{-\mathbf{S}_{1:s-1}} - w_{s-1}^{-\mathbf{S}_{1:s-1}}\right),
\end{aligned} \tag{60}$$

which, by (54), implies

$$\left\|w_s^{-S_{\le t}} - w_s^{-\mathbf{S}_{1:s-1}}\right\| \le \left\|w_{s-1}^{-S_{\le t}} - w_{s-1}^{-\mathbf{S}_{1:s-1}}\right\| + \Gamma_s. \tag{61}$$

Before the first deleted task $\underline{s}^t$, the retraining and natural-forgetting trajectories coincide. Starting from their zero difference at $\underline{s}^t - 1$ and iterating (59) over retained tasks and (61) over deleted tasks gives

$$\begin{aligned}\left\|w_t^{-S_{\le t}} - w_t^{-\mathbf{S}_{1:t}}\right\| &\le \sum_{s\in S_{\le t}} \left(\prod_{\substack{j=s+1\\ j\notin S_{\le t}}}^{t} \rho_j^t\right)\Gamma_s \\ &+ \sum_{\substack{i=\underline{s}^t+1\\ i\notin S_{\le t}}}^{t} \left(\prod_{\substack{j=i+1\\ j\notin S_{\le t}}}^{t} \rho_j^t\right) g_i^t = \gamma(\mathbf{S}_{1:t}),\end{aligned} \tag{62}$$

which proves (53). □

The result of Proposition 4.3 is directly obtained by substituting the bounds on Γ_s, ρ_j^t, g_j^t, from Lemmas F.1, F.2, F.5 into Theorem 4.1.

D Hessian-Unlearning Analysis

D.1 Auxiliary Notation for General Request Orders

We next introduce the error envelopes, product-asymmetry quantities, and time partitions needed to analyze arbitrary unlearning-request orders.

Operator Envelopes. For the path-averaged dynamics defined in Appendix A, let

$$\begin{aligned}\rho_j^t &:= \sup\left\|\widetilde{P}_j^t\right\|, & g_j^t &:= \sup\left\|\widetilde{G}_j^t\right\|, \\ \Gamma_j &:= \sup\left\|w_j^{-\mathbf{S}_{1:j-1}} - w_{j-1}^{-\mathbf{S}_{1:j-1}}\right\|.\end{aligned} \tag{63}$$

For the approximate operators and their discrepancies from the path-averaged operators, define

$$\begin{aligned}\hat\rho_j^t &:= \sup\left\|\hat P_j^t\right\|, & \hat g_j^t &:= \sup\left\|\hat G_j^t\right\|, & \eta_j^\tau &:= \sup\left\|I - \hat P_j^\tau\right\|, \\ \bar\rho_j^t &:= \sup\left\|\widetilde P_j^t - \hat P_j^t\right\|, & \bar g_j^t &:= \sup\left\|\widetilde G_j^t - \hat G_j^t\right\|, & & \\ \kappa_j^\tau &:= \sup\left\|\Delta_j - \widetilde G_j^\tau\right\|, & \hat\kappa_j^\tau &:= \sup\left\|\Delta_j - \hat G_j^\tau\right\|. & &\end{aligned} \tag{64}$$

Product-Asymmetry Notation. For an interval $[a,b]$ and two request times $t_i < t_j$, let

$$D_{j\setminus i} := ([a,b]\cap S_{\le t_j}) \setminus ([a,b]\cap S_{\le t_i}) = \{a_1,\dots,a_n\}, \qquad a_1 < \cdots < a_n.$$

Define the changes in the approximate operators between these request times by

$$\chi_c^{t_i\to t_j} := \sup\left\|\hat P_c^{t_j} - \hat P_c^{t_i}\right\|, \qquad \zeta_c^{t_i\to t_j} := \sup\left\|\hat G_c^{t_j} - \hat G_c^{t_i}\right\|,$$

and let

$$X_{[a,b]}^{j\to i} := \sum_{r=1}^{n}\left(\prod_{\ell=r+1}^{n}\hat\rho_{a_\ell}^{t_i}\right)\eta_{a_r}^{t_i}.$$

The resulting product-asymmetry envelope is

$$Z_{[a,b]}^{j\to i} := \left(\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{b}(\hat\rho_c^{t_i}/\hat\rho_c^{t_j})\right)\left[\sum_{\substack{c=a\\ c\notin S_{\le t_j}}}^{b}\frac{\chi_c^{t_i\to t_j}}{\hat\rho_c^{t_i}}\left(\prod_{\substack{r=c+1\\ r\notin S_{\le t_j}}}^{b}\frac{\hat\rho_r^{t_j}}{\hat\rho_r^{t_i}}\right) + X_{[a,b]}^{j\to i}\right]. \tag{65}$$

Unlearning-Time Partitions. Let $\underline{s}_m := \min S_{\le t_m}$ for $m = 1, \dots, k$. At the k-th unlearning time t_k, partition the relevant task indices into $I^k_{i,j}$, $0 \le j \le i \le k-1$, as follows:

$$I^k_{i,j} = \big(\big[\underline{s}_{k-i}+1, \underline{s}_{k-i-1}-1\big] \cup [t_{k-i-1}+1, t_{k-i}]\big) \cap S_{t_{k-i+j+1}}, \quad i = 1, \dots, k-2, \quad j \le i-1, \tag{66}$$

$$I^k_{k-1,j} = \big[\underline{s}_1+1, t_1\big] \cap S_{t_{j+2}}, \qquad j = 0, \dots, k-2, \tag{67}$$

$$I^k_{i,i} = \big(\big[\underline{s}_{k-i}+1, \underline{s}_{k-i-1}-1\big] \cup [t_{k-i-1}+1, t_{k-i}]\big) \setminus S_{\le t_k}, \qquad i = 0, \dots, k-2, \tag{68}$$

$$I^k_{k-1,k-1} = \big[\underline{s}_1+1, t_1\big] \setminus S_{\le t_k}. \tag{69}$$

Coefficient and Term Notation. For $j = 0, \dots, k-1$ and $s \in S_{t_{k-j}}$, let $d_j(t_k, s)$ denote the coefficient of Δ_s in $\bar{\Delta}_{t_k}$, and let $D_j(t_k, s)$ denote its coefficient in $w_{t_k}^{-S_{\le t_k}} - w_{t_k}^{-\mathbf{S}_{1:t_k}}$.

For an unlearning time t_h and $s \in I^h_{i,j}$, let $b_{i,j}(t_h, s)$ denote the term in $\bar{\Delta}_{t_h}$ associated with the $\hat{G}_s$-factor, regardless of its time superscript. Similarly, let $B_{i,j}(t_h, s)$ denote the corresponding G_s-type term in $w_{t_h}^{-S_{\le t_h}} - w_{t_h}^{-\mathbf{S}_{1:t_h}}$, regardless of its time superscript or whether it is hatted or tilded.

We now bound the approximation error $\left\| w_t^{-S_{\le t}} - w_t^{-\mathbf{S}_{1:t}} \right\|$ of the Hessian-based unlearning method (14).

D.2 Approximation-Error Bounds

D.2.1 Arbitrary Request Orders

Theorem D.1. *At time t, let $U_t = \{t_1, \dots, t_k\}$ be the ordered set of unlearning times up to t. Then*

$$\left\| w_t^{-S_{\le t}} - w_t^{-\mathbf{S}_{1:t}} \right\| \le \left(\prod_{j=t_k+1}^{t} \rho_j^{t_k} \right) \gamma(\mathbf{S}_{1:t_k}) + \sum_{i=t_k+1}^{t} \left(\prod_{a=i+1}^{t} \rho_a^{t_k} \right) g_i^{t_k}, \tag{70}$$

where

$$\begin{aligned} \gamma(\mathbf{S}_{1:t_k}) = &\sum_{i=2}^{k} \sum_{a=k+1-i}^{k-1} \sum_{s \in I^k_{a,a+i-k-1}} \Big(\mathbf{D}\mathbf{B}_{k-i}(t_k, s) + \mathbf{B}_{a,a}(t_k, s) \Big) \\ &+ \sum_{a=0}^{k-1} \sum_{s \in I^k_{a,a}} \mathbf{B}_{a,a}(t_k, s) \\ &+ \sum_{i=2}^{k} \sum_{s \in S_{t_i} \setminus \left([\underline{s}_{i-1}+1, t_{i-1}] \setminus S_{\le t_{i-1}} \right)} \mathbf{D}_{k-i}(t_k, s)\, \Gamma_s \\ &+ \sum_{s \in S_{t_1}} \mathbf{D}_{k-1}(t_k, s)\, \Gamma_s, \end{aligned} \tag{71}$$

where $\mathbf{D}_j(t_k, s)$, $\mathbf{B}_{j,j}(t_k, s)$, and $\mathbf{D}\mathbf{B}_j(t_k, s)$ are defined in Lemma D.10 and Corollary D.11, and all other notations are defined in Appendix D.1.

Proof of Theorem D.1. Firstly, note that $\gamma(\mathbf{S}_{1:t_k})$ is defined as the upper bound of $\left\| w_{t_k}^{-S_{\le t_k}} - w_{t_k}^{-\mathbf{S}_{1:t_k}} \right\|$. Applying the triangle inequality and submultiplicativity to the first identity in Lemma D.7, and then using the definitions of $\rho_j^{t_k}$ and $g_i^{t_k}$, gives (70).

It remains to prove for $\gamma(\mathbf{S}_{1:t_k})$. By the second identity in Lemma D.7 and the coefficient notation introduced in Appendix D.1, we have

$$w_{t_k}^{-S_{\le t_k}} - w_{t_k}^{-\mathbf{S}_{1:t_k}} = \sum_{i=1}^{k} \sum_{s \in S_{t_i}} D_{k-i}(t_k, s) \Delta_s + \sum_{a=0}^{k-1} \sum_{b=0}^{a} \sum_{s \in I^k_{a,b}} B_{a,b}(t_k, s). \tag{72}$$

For every $i = 2, \dots, k$, we have

$$\begin{aligned} S_{t_i} =& S_{t_i} \cap (([\underline{s}_i, \underline{s}_{i-1}] \cup [t_{i-1}+1, t_i]) \cup [\underline{s}_{i-1}+1, t_{i-1}]) \\ =& (S_{t_i} \cap ([\underline{s}_i, \underline{s}_{i-1}] \cup [t_{i-1}+1, t_i])) \cup (S_{t_i} \cap [\underline{s}_{i-1}+1, t_{i-1}]). \end{aligned}$$

Then, by the definition in Appendix D.1, it is easy to verify

$$\left([\underline{s}_{i-1}+1, t_{i-1}] \setminus S_{\le t_{i-1}}\right) \cap S_{t_i} = [\underline{s}_{i-1}+1, t_{i-1}] \cap S_{t_i} = \bigcup_{a=k+1-i}^{k-1} I^k_{a,a+i-k-1}. \tag{73}$$

Therefore, the off-diagonal B-terms can be paired with the corresponding Δ_s-terms, while deletion indices outside the set on the left-hand side of (73) remain unpaired. Reordering (72) gives

$$\begin{aligned} w_{t_k}^{-S_{\le t_k}} - w_{t_k}^{-\mathbf{s}_{1:t_k}} =& \sum_{i=2}^{k} \sum_{a=k+1-i}^{k-1} \sum_{s \in I^k_{a,a+i-k-1}} (D_{k-i}(t_k, s)\Delta_s + B_{a,a+i-k-1}(t_k, s)) \\ &+ \sum_{a=0}^{k-1} \sum_{s \in I^k_{a,a}} B_{a,a}(t_k, s) \\ &+ \sum_{i=2}^{k} \sum_{s \in S_{t_i} \setminus \left([\underline{s}_{i-1}+1, t_{i-1}] \setminus S_{\le t_{i-1}}\right)} D_{k-i}(t_k, s)\Delta_s \\ &+ \sum_{s \in S_{t_1}} D_{k-1}(t_k, s)\Delta_s. \end{aligned} \tag{74}$$

Taking norms in (74), applying the triangle inequality, and using $\|\Delta_s\| \le \Gamma_s$, we obtain

$$\begin{aligned} &\left\| w_{t_k}^{-S_{\le t_k}} - w_{t_k}^{-\mathbf{s}_{1:t_k}} \right\| \\ \le& \sum_{i=2}^{k} \sum_{a=k+1-i}^{k-1} \sum_{s \in I^k_{a,a+i-k-1}} \|D_{k-i}(t_k, s)\Delta_s + B_{a,a+i-k-1}(t_k, s)\| \\ &+ \sum_{a=0}^{k-1} \sum_{s \in I^k_{a,a}} \|B_{a,a}(t_k, s)\| \\ &+ \sum_{i=2}^{k} \sum_{s \in S_{t_i} \setminus \left([\underline{s}_{i-1}+1, t_{i-1}] \setminus S_{\le t_{i-1}}\right)} \|D_{k-i}(t_k, s)\| \, \Gamma_s \\ &+ \sum_{s \in S_{t_1}} \|D_{k-1}(t_k, s)\| \, \Gamma_s. \end{aligned} \tag{75}$$

The bounds for the second through fourth lines are given in Lemma D.10. We now simplify each paired term in the first line. Fix $i = 2, \dots, k$, $a = k+1-i, \dots, k-1$, and $s \in I^k_{a,a+i-k-1}$. Lemma D.13 gives the first equality below.

$$\begin{aligned} & D_{k-i}(t_k, s)\Delta_s + B_{a,a+i-k-1}(t_k, s) \\ =& D_{k-i}(t_k, s)\Delta_s - \sum_{n=0}^{a-1-k+i} D_{k-i+n}(t_k, t_{i-n-1}) b_{0,0}(t_{i-n-1}, s) \\ =& D_{k-i}(t_k, s)\Delta_s - B_{k-i,k-i}(t_k, s) \\ &+ \left(B_{k-i,k-i}(t_k, s) - \sum_{n=0}^{a-1-k+i} D_{k-i+n}(t_k, t_{i-n-1}) b_{0,0}(t_{i-n-1}, s) \right) \\ =& D_{k-i}(t_k, s)\Delta_s - B_{k-i,k-i}(t_k, s) + B_{a,a}(t_k, s), \end{aligned}$$

where the last equality follows by iterating the diagonal recursion in Lemma D.8. Consequently,

$$\begin{aligned} & \|D_{k-i}(t_k, s)\Delta_s + B_{a,a+i-k-1}(t_k, s)\| \\ \le& \|D_{k-i}(t_k, s)\Delta_s - B_{k-i,k-i}(t_k, s)\| + \|B_{a,a}(t_k, s)\|. \end{aligned}$$

Substituting this inequality into (75) yields (71). The individual terms are then controlled by Lemma D.10, Corollary D.11, and their zero-index base cases, as stated in the theorem. □

D.2.2 FORWARD-SYNCHRONOUS REQUEST ORDERS

Proof of Theorem 5.2. By the definition of a forward-synchronous unlearning sequence in (15), we have $I^k_{i,a} = \emptyset$ for $a = 0, \dots, i-1$ and $i = 0, \dots, k-1$, and $S_{t_i} \subseteq [t_{i-1}+1, t_i]$. Thus, Theorem D.1 reduces to

$$\left\| w_t^{-S_{\le t}} - w_t^{-\mathbf{S}_{1:t}} \right\| \le \left(\prod_{j=t_k+1}^{t} \rho_j^{t_k} \right) \gamma(\mathbf{S}_{1:t_k}) + \sum_{i=t_k+1}^{t} \left(\prod_{a=i+1}^{t} \rho_a^{t_k} \right) g_i^{t_k}, \tag{76}$$

where

$$\gamma(\mathbf{S}_{1:t_k}) = \sum_{a=0}^{k-1} \sum_{s \in I^k_{a,a}} \mathbf{B}_{a,a}(t_k, s) + \sum_{i=1}^{k} \sum_{s \in S_{t_i}} \mathbf{D}_{k-i}(t_k, s)\, \Gamma_s. \tag{77}$$

Then, according to the Definition of $\mathbf{B}_{a,a}(t_k, s)$ and $\mathbf{D}_{k-i}(t_k, s)$ in Lemma D.8, all the $C(\cdots)$ and $\hat{C}(\cdots)$ terms are zero. Then, taking all the $\hat{\rho} = \rho$ and $\hat{g} = g$ in Lemma D.8 since we use exact Hessian in Theorem 5.2, we obtain the results. □

D.3 SINGLE-REQUEST BOUNDS FOR TASK-ADDITIVE REGULARIZERS

Corollary D.2. *Suppose a single request $S_t = \{s\}$ arrives at time t without previous unlearning request, and Assumptions 2.3, 2.4, and 4.2 hold. Hessian unlearning in* (14)*, using the exact Hessian and a task-additive regularizer, achieves the following approximation-error bound:*

$$\left\| w_t^{-S_{\le t}} - w_t^{-\mathbf{S}_{1:t}} \right\| \tag{78}$$

$$\le \alpha(s) \sum_{n=s+1}^{t} \beta(n) + \sum_{i=s+1}^{t} \alpha(i) \left[\frac{(i-s-1)(M_\Omega - \mu_\Omega)}{\mu + \lambda + (i-2)\mu_\Omega} + \frac{M_\Omega}{\mu + \lambda + (i-2)\mu_\Omega} \sum_{n=i}^{t} \beta(n) \right]. \tag{79}$$

where $\alpha(i)$ is defined in Proposition 4.3, and $\beta(i)$ is

$$\beta(n) := \frac{(n-s-1)(M_\Omega - \mu_\Omega)}{\lambda + (n-2)M_\Omega} + \frac{M - \mu + (n-s-1)(M_\Omega - \mu_\Omega)}{\mu + \lambda + (n-2)\mu_\Omega}. \tag{80}$$

The bound for ℓ_2 is obtained by setting $M_\Omega = \mu_\Omega = 0$ in (79)*.*

Proof. The result follows by substituting the bounds on Γ_s, ρ_j^t, g_j^t, $\bar{\rho}_j^t$, and $\bar{g}_j^t$ from Lemmas F.1, F.2, F.5, F.3, and F.7 into Theorem 5.2 for the single-request sequence $S_t = \{s\}$.

□

D.4 AUXILIARY LEMMAS

Lemma D.3 (Product perturbation)**.** *For any $1 \le a \le b \le t$ and any τ,*

$$\left\| \prod_{\substack{j=a \\ j \notin S_{\le t}}}^{b} \widetilde{P}_j^\tau - \prod_{\substack{j=a \\ j \notin S_{\le t}}}^{b} \hat{P}_j^\tau \right\| \le \prod_{\substack{i=a \\ i \notin S_{\le t}}}^{b} \hat{\rho}_i^\tau \sum_{\substack{j=a \\ j \notin S_{\le t}}}^{b} \frac{\bar{\rho}_j^\tau}{\hat{\rho}_j^\tau} \left(\prod_{\substack{i=j+1 \\ i \notin S_{\le t}}}^{b} \frac{\rho_i^\tau}{\hat{\rho}_i^\tau} \right). \tag{81}$$

Proof. Using the telescoping identity, we have

$$\left\| \prod_{\substack{j=a \\ j \notin S_{\le t}}}^{b} \widetilde{P}_j^\tau - \prod_{\substack{j=a \\ j \notin S_{\le t}}}^{b} \hat{P}_j^\tau \right\| = \left\| \sum_{\substack{j=a \\ j \notin S_{\le t}}}^{b} \left(\prod_{\substack{i=j+1 \\ i \notin S_{\le t}}}^{b} \widetilde{P}_i^\tau \right) (\widetilde{P}_j^\tau - \hat{P}_j^\tau) \left(\prod_{\substack{i=a \\ i \notin S_{\le t}}}^{j-1} \hat{P}_i^\tau \right) \right\|.$$

Using the definitions in (63) and (64), we have

$$\left\| \sum_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \left(\prod_{\substack{i=j+1 \\ i\notin S_{\le t}}}^{b} \widetilde{P}_i^\tau \right) (\widetilde{P}_j^\tau - \hat{P}_j^\tau) \left(\prod_{\substack{i=a \\ i\notin S_{\le t}}}^{j-1} \hat{P}_i^\tau \right) \right\| \le \sum_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \left(\prod_{\substack{i=j+1 \\ i\notin S_{\le t}}}^{b} \left\|\widetilde{P}_i^\tau\right\| \right) \left\|\widetilde{P}_j^\tau - \hat{P}_j^\tau\right\| \left(\prod_{\substack{i=a \\ i\notin S_{\le t}}}^{j-1} \left\|\hat{P}_i^\tau\right\| \right)$$

$$\le \sum_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \left(\prod_{\substack{i=j+1 \\ i\notin S_{\le t}}}^{b} \rho_i^\tau \right) \bar{\rho}_j^\tau \left(\prod_{\substack{i=a \\ i\notin S_{\le t}}}^{j-1} \hat{\rho}_i^\tau \right)$$

$$= \prod_{\substack{i=a \\ i\notin S_{\le t}}}^{b} \hat{\rho}_i^\tau \sum_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \frac{\bar{\rho}_j^\tau}{\hat{\rho}_j^\tau} \left(\prod_{\substack{i=j+1 \\ i\notin S_{\le t}}}^{b} \frac{\rho_i^\tau}{\hat{\rho}_i^\tau} \right),$$

which completes the proof. □

Lemma D.4. *For any $1 \le a \le b \le t$ and any τ,*

$$\left\| \left(\prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \widetilde{P}_j^\tau \right) \widetilde{G}_{a-1}^\tau - \left(\prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \hat{P}_j^\tau \right) \hat{G}_{a-1}^\tau \right\| \tag{82}$$

$$\le \left(\prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \rho_j^\tau \right) \bar{g}_{a-1}^\tau + \prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \hat{\rho}_j^\tau \sum_{\substack{i=a \\ i\notin S_{\le t}}}^{b} \frac{\bar{\rho}_i^\tau}{\hat{\rho}_i^\tau} \left(\prod_{\substack{r=i+1 \\ r\notin S_{\le t}}}^{b} \frac{\rho_r^\tau}{\hat{\rho}_r^\tau} \right) \hat{g}_{a-1}^\tau. \tag{83}$$

Proof. Add and subtract $\left(\prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \widetilde{P}_j^\tau \right) \hat{G}_{a-1}^\tau$. Then, by the triangle inequality and submultiplicativity,

$$\left\| \left(\prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \widetilde{P}_j^\tau \right) \widetilde{G}_{a-1}^\tau - \left(\prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \hat{P}_j^\tau \right) \hat{G}_{a-1}^\tau \right\|$$

$$\le \left\| \left(\prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \widetilde{P}_j^\tau \right) (\widetilde{G}_{a-1}^\tau - \hat{G}_{a-1}^\tau) \right\| + \left\| \left(\prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \widetilde{P}_j^\tau - \prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \hat{P}_j^\tau \right) \hat{G}_{a-1}^\tau \right\|$$

$$\le \left(\prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \left\|\widetilde{P}_j^\tau\right\| \right) \left\|\widetilde{G}_{a-1}^\tau - \hat{G}_{a-1}^\tau\right\| + \left\| \prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \widetilde{P}_j^\tau - \prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \hat{P}_j^\tau \right\| \left\|\hat{G}_{a-1}^\tau\right\|$$

$$\le \left(\prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \rho_j^\tau \right) \bar{g}_{a-1}^\tau + \left\| \prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \widetilde{P}_j^\tau - \prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \hat{P}_j^\tau \right\| \hat{g}_{a-1}^\tau.$$

Applying Lemma D.3 gives

$$\left\| \prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \widetilde{P}_j^\tau - \prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \hat{P}_j^\tau \right\| \le \prod_{\substack{j=a \\ j\notin S_{\le t}}}^{b} \hat{\rho}_j^\tau \sum_{\substack{i=a \\ i\notin S_{\le t}}}^{b} \frac{\bar{\rho}_i^\tau}{\hat{\rho}_i^\tau} \left(\prod_{\substack{r=i+1 \\ r\notin S_{\le t}}}^{b} \frac{\rho_r^\tau}{\hat{\rho}_r^\tau} \right).$$

Combining the two displays completes the proof. □

Lemma D.5. *Fix integers $1 \le a \le b \le t$ and two request times $t_i < t_j$. Then*

$$\left\| \prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_j} - \prod_{\substack{c=a \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \right\| \le \left(\prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{\rho}_c^{t_j} \right) Z_{[a,b]}^{j\to i}. \tag{84}$$

Proof. Firstly, we have

$$\left\| \prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_j} - \prod_{\substack{c=a \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \right\| \le \left\| \prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_j} - \prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_i} \right\| + \left\| \prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_i} - \prod_{\substack{c=a \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \right\|.$$

For the first term, the telescoping identity gives

$$\prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_j} - \prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_i} = \sum_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \left(\prod_{\substack{r=c+1 \\ r\notin S_{\le t_j}}}^{b} \hat{P}_r^{t_j} \right) \left(\hat{P}_c^{t_j} - \hat{P}_c^{t_i} \right) \left(\prod_{\substack{r=a \\ r\notin S_{\le t_j}}}^{c-1} \hat{P}_r^{t_i} \right).$$

Hence, by the triangle inequality and submultiplicativity,

$$\begin{aligned}
\left\| \prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_j} - \prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_i} \right\| &\le \sum_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \left(\prod_{\substack{r=c+1 \\ r\notin S_{\le t_j}}}^{b} \left\| \hat{P}_r^{t_j} \right\| \right) \left\| \hat{P}_c^{t_j} - \hat{P}_c^{t_i} \right\| \left(\prod_{\substack{r=a \\ r\notin S_{\le t_j}}}^{c-1} \left\| \hat{P}_r^{t_i} \right\| \right) \\
&\le \sum_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \left(\prod_{\substack{r=c+1 \\ r\notin S_{\le t_j}}}^{b} \hat{\rho}_r^{t_j} \right) \chi_c^{t_i\to t_j} \left(\prod_{\substack{r=a \\ r\notin S_{\le t_j}}}^{c-1} \hat{\rho}_r^{t_i} \right) \\
&= \left(\prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{\rho}_c^{t_i} \right) \sum_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \frac{\chi_c^{t_i\to t_j}}{\hat{\rho}_c^{t_i}} \left(\prod_{\substack{r=c+1 \\ r\notin S_{\le t_j}}}^{b} \frac{\hat{\rho}_r^{t_j}}{\hat{\rho}_r^{t_i}} \right).
\end{aligned}$$

For the second term, the only difference between $[a,b]\cap S_{\le t_i}$ and $[a,b]\cap S_{\le t_j}$ is the set $D_{j\setminus i} = \{a_1,\dots,a_n\}$, whose factors are present in the $[a,b]\cap S_{\le t_i}$-product and absent from the $[a,b]\cap S_{\le t_j}$-product. Thus,

$$\begin{aligned}
&\prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_i} - \prod_{\substack{c=a \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \\
&= \prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_i} + \sum_{r=1}^{n} - \left(\prod_{\substack{c=a_r \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \right) \left(\prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{a_r-1} \hat{P}_c^{t_i} \right) + \left(\prod_{\substack{c=a_r \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \right) \left(\prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{a_r-1} \hat{P}_c^{t_i} \right) - \prod_{\substack{c=a \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \\
&= \sum_{r=1}^{n} \left(\prod_{\substack{c=a_r+1 \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \right) \left(I - \hat{P}_{a_r}^{t_i} \right) \left(\prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{a_r-1} \hat{P}_c^{t_i} \right),
\end{aligned}$$

where the first term is zero. For the intermediate terms, we use

$$\begin{aligned}
&\left(\prod_{\substack{c=a_r\\ c\notin S_{\le t_i}}}^{b}\hat{P}_c^{t_i}\right)\left(\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{a_r-1}\hat{P}_c^{t_i}\right)-\left(\prod_{\substack{c=a_{r-1}\\ c\notin S_{\le t_i}}}^{b}\hat{P}_c^{t_i}\right)\left(\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{a_{r-1}-1}\hat{P}_c^{t_i}\right)\\
&=\left(\prod_{\substack{c=a_r\\ c\notin S_{\le t_i}}}^{b}\hat{P}_c^{t_i}\right)\left(\prod_{\substack{c=a_{r-1}\\ c\notin S_{\le t_j}}}^{a_r-1}\hat{P}_c^{t_i}-\prod_{\substack{c=a_{r-1}\\ c\notin S_{\le t_i}}}^{a_r-1}\hat{P}_c^{t_i}\right)\left(\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{a_{r-1}-1}\hat{P}_c^{t_i}\right)\\
&=\left(\prod_{\substack{c=a_r\\ c\notin S_{\le t_i}}}^{b}\hat{P}_c^{t_i}\right)\left(\prod_{\substack{c=a_{r-1}+1\\ c\notin S_{\le t_j}}}^{a_r-1}\hat{P}_c^{t_i}\right)\left(I-\hat{P}_{a_{r-1}}^{t_i}\right)\left(\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{a_{r-1}-1}\hat{P}_c^{t_i}\right)\\
&=\left(\prod_{\substack{c=a_{r-1}+1\\ c\notin S_{\le t_i}}}^{b}\hat{P}_c^{t_i}\right)\left(I-\hat{P}_{a_{r-1}}^{t_i}\right)\left(\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{a_{r-1}-1}\hat{P}_c^{t_i}\right),
\end{aligned}$$

For the last term, we use

$$\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{b}\hat{P}_c^{t_i}-\left(\prod_{\substack{c=a_n\\ c\notin S_{\le t_i}}}^{b}\hat{P}_c^{t_i}\right)\left(\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{a_n-1}\hat{P}_c^{t_i}\right)=\left(\prod_{\substack{c=a_n+1\\ c\notin S_{\le t_i}}}^{b}\hat{P}_c^{t_i}\right)\left(I-\hat{P}_{a_n}^{t_i}\right)\left(\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{a_n-1}\hat{P}_c^{t_i}\right).$$

Therefore, taking norms and using the notation introduced above together with submultiplicativity gives

$$\begin{aligned}
\left\|\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{b}\hat{P}_c^{t_i}-\prod_{\substack{c=a\\ c\notin S_{\le t_i}}}^{b}\hat{P}_c^{t_i}\right\| &\le\sum_{r=1}^{n}\left(\prod_{\substack{c=a_r+1\\ c\notin S_{\le t_i}}}^{b}\left\|\hat{P}_c^{t_i}\right\|\right)\left\|I-\hat{P}_{a_r}^{t_i}\right\|\left(\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{a_r-1}\left\|\hat{P}_c^{t_i}\right\|\right)\\
&\le\sum_{r=1}^{n}\left(\prod_{\substack{c=a_r+1\\ c\notin S_{\le t_i}}}^{b}\hat{\rho}_c^{t_i}\right)\eta_{a_r}^{t_i}\left(\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{a_r-1}\hat{\rho}_c^{t_i}\right)\\
&=\left(\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{b}\hat{\rho}_c^{t_i}\right)\sum_{r=1}^{n}\prod_{c=r+1}^{n}\hat{\rho}_{a_c}^{t_i}\eta_{a_r}^{t_i}=\left(\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{b}\hat{\rho}_c^{t_i}\right)X_{[a,b]}^{j\to i}.
\end{aligned}$$

Combining the two upper bounds proves (84). □

Lemma D.6. *Fix integers $2\le a\le b\le t$ and two request times $t_i<t_j$. Then we have*

$$\left\|\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{b}\hat{P}_c^{t_j}\hat{G}_{a-1}^{t_j}-\prod_{\substack{c=a\\ c\notin S_{\le t_i}}}^{b}\hat{P}_c^{t_i}\hat{G}_{a-1}^{t_i}\right\|\le\left(\prod_{\substack{c=a\\ c\notin S_{\le t_j}}}^{b}\hat{\rho}_c^{t_j}\right)Z_{[a,b]}^{j\to i}\hat{g}_{a-1}^{t_j}+\prod_{\substack{c=a\\ c\notin S_{\le t_i}}}^{b}\hat{\rho}_c^{t_i}\zeta_{a-1}^{t_i\to t_j},$$

where $\chi_c^{t_i\to t_j}$ and $X_{[a,b]}^{j\to i}$ are defined as in Lemma D.5.

Proof.

$$\begin{aligned}
&\left\| \prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_j} \hat{G}_{a-1}^{t_j} - \prod_{\substack{c=a \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \hat{G}_{a-1}^{t_i} \right\| \\
&= \left\| \prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_j} \hat{G}_{a-1}^{t_j} - \prod_{\substack{c=a \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \hat{G}_{a-1}^{t_j} + \prod_{\substack{c=a \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \hat{G}_{a-1}^{t_j} - \prod_{\substack{c=a \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \hat{G}_{a-1}^{t_i} \right\| \\
&\le \left\| \prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{P}_c^{t_j} \hat{G}_{a-1}^{t_j} - \prod_{\substack{c=a \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \hat{G}_{a-1}^{t_j} \right\| + \left\| \prod_{\substack{c=a \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \hat{G}_{a-1}^{t_j} - \prod_{\substack{c=a \\ c\notin S_{\le t_i}}}^{b} \hat{P}_c^{t_i} \hat{G}_{a-1}^{t_i} \right\| \\
&\le \left(\prod_{\substack{c=a \\ c\notin S_{\le t_j}}}^{b} \hat{\rho}_c^{t_j} \right) Z_{[a,b]}^{j\to i} \hat{g}_{a-1}^{t_j} + \prod_{\substack{c=a \\ c\notin S_{\le t_i}}}^{b} \hat{\rho}_c^{t_i} \zeta_{a-1}^{t_i\to t_j},
\end{aligned}$$

where the last inequality uses Lemma D.5 and the definitions of $\hat{g}_{a-1}^{t_j}$ and $\zeta_{a-1}^{t_i\to t_j}$. □

Lemma D.7. *At time t, suppose $U_t = \{t_1, \dots, t_k\}$ is the ordered set of historical unlearning times up to t. The following identities hold:*

$$w_t^{-S_{\le t}} - w_t^{-\mathbf{S}_{1:t}} = \prod_{j=t_k+1}^{t} \tilde{P}_j^{t_k} (w_{t_k}^{-S_{\le t_k}} - w_{t_k}^{-\mathbf{S}_{1:t_k}}) + \sum_{i=t_k+1}^{t} \prod_{a=i+1}^{t} \tilde{P}_a^{t_k} \tilde{G}_i^{t_k}, \tag{85}$$

and

$$\begin{aligned}
&w_{t_k}^{-S_{\le t_k}} - w_{t_k}^{-\mathbf{S}_{1:t_k}} \qquad (86)\\
&= \sum_{s\in S_{\le t_k}} \left(\prod_{\substack{j=s+1 \\ j\notin S_{\le t_k}}}^{t_k} \tilde{P}_j^{t_k} - \prod_{\substack{j=s+1 \\ j\notin S_{\le t_k}}}^{t_k} \hat{P}_j^{t_k} \right) \Delta_s + \sum_{\substack{n=s^{t_k}+1 \\ n\notin S_{\le t_k}}}^{t_k} \left(\prod_{\substack{j=n+1 \\ j\notin S_{\le t_k}}}^{t_k} \tilde{P}_j^{t_k} \tilde{G}_n^{t_k} - \prod_{\substack{j=n+1 \\ j\notin S_{\le t_k}}}^{t_k} \hat{P}_j^{t_k} \hat{G}_n^{t_k} \right) \\
&\quad - \sum_{\tau\in U_{t_k-1}} \left(\prod_{\substack{j=\tau+1 \\ j\notin S_{\le t_k}}}^{t_k} \tilde{P}_j^{t_k} - \prod_{\substack{j=\tau+1 \\ j\notin S_{\le t_k}}}^{t_k} \hat{P}_j^{t_k} \right) \bar{\Delta}_\tau. \qquad (87)
\end{aligned}$$

Proof. Because $t_k = \max U_t$, there is no unlearning request after t_k. Thus $S_{\le i} = S_{\le t_k} = S_{\le t}$ and $w_i^{-\mathbf{S}_{1:i}} = w_i^{-\mathbf{S}_{1:i-1}}$ for every $i = t_k+1, \dots, t$. If $t = t_k$, the first identity is immediate. Otherwise, at each $i = t_k+1, \dots, t$, the first-order conditions of the exact retraining step and the CLU learning step are, respectively,

$$\begin{aligned}
0 &= \nabla L(w_i^{-S_{\le t}}, D_i) + \nabla_2 \Omega_{D_{1:i}^{-S_{\le t}}}\left(w_{i-1}^{-S_{\le t}}, w_i^{-S_{\le t}}\right),\\
0 &= \nabla L(w_i^{-\mathbf{S}_{1:i-1}}, D_i) + \nabla_2 \Omega_{D_{1:i}^{-\mathbf{S}_{1:i-1}}}\left(w_{i-1}^{-\mathbf{S}_{1:i-1}}, w_i^{-\mathbf{S}_{1:i-1}}\right).
\end{aligned}$$

After taking the mean-value expansion of the first equality around

$$\left(w_{i-1}^{-\mathbf{S}_{1:i-1}}, w_i^{-\mathbf{S}_{1:i-1}}\right)$$

and subtracting the second equality, we use the definitions of $\tilde{P}_i^{t_k}$ and $\tilde{G}_i^{t_k}$ to obtain

$$\begin{aligned}
&w_i^{-S_{\le t}} - w_i^{-\mathbf{S}_{1:i-1}}\\
&= \tilde{P}_i^{t_k} \left(w_{i-1}^{-S_{\le t}} - w_{i-1}^{-\mathbf{S}_{1:i-1}} \right) + \tilde{G}_i^{t_k}. \qquad (88)
\end{aligned}$$

Repeatedly substituting (88), with an empty product interpreted as the identity operator, yields

$$w_t^{-S_{\le t}} - w_t^{-\mathbf{S}_{1:t-1}} = \left(\prod_{j=t_k+1}^{t} \tilde{P}_j^{t_k}\right)\left(w_{t_k}^{-S_{\le t_k}} - w_{t_k}^{-\mathbf{S}_{1:t_k}}\right) + \sum_{i=t_k+1}^{t}\left(\prod_{a=i+1}^{t} \tilde{P}_a^{t_k}\right)\tilde{G}_i^{t_k},$$

and $w_t^{-\mathbf{S}_{1:t-1}} = w_t^{-\mathbf{S}_{1:t}}$ because $S_t = \emptyset$. This proves the first identity.

We next prove the expansion at t_k by the original index-wise induction. For any $i = \underline{s}^{t_k}, \dots, t_k$, we claim that

$$\begin{aligned} & w_i^{-S_{\le t_k}} - w_i^{-\mathbf{S}_{1:i-1}} \\ &= \sum_{s \in S_{\le t_k} \cap [\underline{s}^{t_k}, i]} \left(\prod_{\substack{j=s+1 \\ j \notin S_{\le t_k}}}^{i} \tilde{P}_j^{t_k}\right)\Delta_s + \sum_{\substack{n=\underline{s}^{t_k}+1 \\ n \notin S_{\le t_k}}}^{i}\left(\prod_{\substack{j=n+1 \\ j \notin S_{\le t_k}}}^{i} \tilde{P}_j^{t_k}\right)\tilde{G}_n^{t_k} \\ &\quad - \sum_{\tau \in U_{i-1}}\left(\prod_{\substack{j=\tau+1 \\ j \notin S_{\le t_k}}}^{i} \tilde{P}_j^{t_k}\right)\bar{\Delta}_\tau. \end{aligned} \tag{89}$$

When $i = \underline{s}^{t_k}$, the first deleted task contributes exactly $\Delta_{\underline{s}^{t_k}}$, while there is no retained index after it and no earlier unlearning correction to propagate. Hence (89) holds in the base case.

Suppose (89) holds at $i-1$. If $i \in S_{\le t_k}$, the exact retraining path skips the learning update on task i. By the definitions of Δ_i and $\bar{\Delta}_{i-1}$,

$$\begin{aligned} & w_i^{-S_{\le t_k}} - w_i^{-\mathbf{S}_{1:i-1}} \\ &= \left(w_{i-1}^{-S_{\le t_k}} - w_{i-1}^{-\mathbf{S}_{1:i-2}}\right) - \bar{\Delta}_{i-1} + \Delta_i. \end{aligned}$$

Substituting the induction hypothesis into the right-hand side gives (89) at i, because the products ending at i are unchanged when $i \in S_{\le t_k}$.

If $i \notin S_{\le t_k}$, both sequences perform the regularized learning step on task i. Their complete first-order conditions are

$$\begin{aligned} 0 &= \nabla L(w_i^{-S_{\le t_k}}, D_i) + \nabla_2 \Omega_{D_{1:i}^{-S_{\le t_k}}}\left(w_{i-1}^{-S_{\le t_k}}, w_i^{-S_{\le t_k}}\right), \\ 0 &= \nabla L(w_i^{-\mathbf{S}_{1:i-1}}, D_i) + \nabla_2 \Omega_{D_{1:i}^{-\mathbf{S}_{1:i-1}}}\left(w_{i-1}^{-\mathbf{S}_{1:i-1}}, w_i^{-\mathbf{S}_{1:i-1}}\right). \end{aligned}$$

After taking the mean-value expansion of the first equality around

$$\left(w_{i-1}^{-\mathbf{S}_{1:i-1}}, w_i^{-\mathbf{S}_{1:i-1}}\right)$$

and subtracting the second equality, we obtain

$$\begin{aligned} & w_i^{-S_{\le t_k}} - w_i^{-\mathbf{S}_{1:i-1}} \\ &= \tilde{P}_i^{t_k}\left(w_{i-1}^{-S_{\le t_k}} - w_{i-1}^{-\mathbf{S}_{1:i-2}} - \bar{\Delta}_{i-1}\right) + \tilde{G}_i^{t_k}. \end{aligned}$$

Substituting the induction hypothesis appends the factor $\tilde{P}_i^{t_k}$ to every previously propagated term, adds the local term $\tilde{G}_i^{t_k}$, and adds $-\tilde{P}_i^{t_k}\bar{\Delta}_{i-1}$. This is exactly (89) at i. Therefore, the claim holds

for every $i = \underline{s}^{t_k}, \dots, t_k$. Taking $i = t_k$ gives

$$\begin{aligned} & w_{t_k}^{-S_{\le t_k}} - w_{t_k}^{-\mathbf{S}_{1:t_k-1}} \\ &= \sum_{s \in S_{\le t_k}} \left(\prod_{\substack{j=s+1 \\ j \notin S_{\le t_k}}}^{t_k} \tilde{P}_j^{t_k} \right) \Delta_s + \sum_{\substack{n=\underline{s}^{t_k}+1 \\ n \notin S_{\le t_k}}}^{t_k} \left(\prod_{\substack{j=n+1 \\ j \notin S_{\le t_k}}}^{t_k} \tilde{P}_j^{t_k} \right) \tilde{G}_n^{t_k} \\ &\quad - \sum_{\tau \in U_{t_k-1}} \left(\prod_{\substack{j=\tau+1 \\ j \notin S_{\le t_k}}}^{t_k} \tilde{P}_j^{t_k} \right) \bar{\Delta}_\tau. \end{aligned} \tag{90}$$

The Hessian-based expansion follows by the same induction. If $i \notin S_{\le t_k}$, the Hessian approximation in (13) gives

$$\begin{aligned} & w_i^{-\mathbf{S}_{1:t_k}} - w_i^{-\mathbf{S}_{1:i-1}} \\ &= \hat{P}_i^{t_k} \left(w_{i-1}^{-\mathbf{S}_{1:t_k}} - w_{i-1}^{-\mathbf{S}_{1:i-2}} - \bar{\Delta}_{i-1} \right) + \hat{G}_i^{t_k}. \end{aligned}$$

If $i \in S_{\le t_k}$, the skipped-task update gives

$$\begin{aligned} & w_i^{-\mathbf{S}_{1:t_k}} - w_i^{-\mathbf{S}_{1:i-1}} \\ &= \left(w_{i-1}^{-\mathbf{S}_{1:t_k}} - w_{i-1}^{-\mathbf{S}_{1:i-2}} \right) - \bar{\Delta}_{i-1} + \Delta_i. \end{aligned}$$

Thus, starting from the same base case and repeating the preceding induction with $\hat{P}_i^{t_k}$ and $\hat{G}_i^{t_k}$ in place of $\tilde{P}_i^{t_k}$ and $\tilde{G}_i^{t_k}$, respectively, gives

$$\begin{aligned} & w_{t_k}^{-\mathbf{S}_{1:t_k}} - w_{t_k}^{-\mathbf{S}_{1:t_k-1}} \\ &= \sum_{s \in S_{\le t_k}} \left(\prod_{\substack{j=s+1 \\ j \notin S_{\le t_k}}}^{t_k} \hat{P}_j^{t_k} \right) \Delta_s + \sum_{\substack{n=\underline{s}^{t_k}+1 \\ n \notin S_{\le t_k}}}^{t_k} \left(\prod_{\substack{j=n+1 \\ j \notin S_{\le t_k}}}^{t_k} \hat{P}_j^{t_k} \right) \hat{G}_n^{t_k} \\ &\quad - \sum_{\tau \in U_{t_k-1}} \left(\prod_{\substack{j=\tau+1 \\ j \notin S_{\le t_k}}}^{t_k} \hat{P}_j^{t_k} \right) \bar{\Delta}_\tau. \end{aligned} \tag{91}$$

Finally, subtracting (91) from (90) cancels the common Stage-I model $w_{t_k}^{-\mathbf{S}_{1:t_k-1}}$. Since t_k is the last unlearning time, $S_{\le t} = S_{\le t_k}$. The result is exactly the second identity. □

Lemma D.8. *For $j = 0, \dots, k-2$ and $s \in S_{t_{k-j-1}}$, the coefficients d_j and D_j satisfy*

$$d_0(t_k, s) = \prod_{\substack{i=s+1 \\ i \notin S_{\le t_k}}}^{t_k} \hat{P}_i^{t_k}, \qquad d_{j+1}(t_k, s) = d_j(t_k, s) - d_j(t_k, t_{k-j-1}) d_0(t_{k-j-1}, s), \tag{92}$$

and

$$D_0(t_k, s) = \prod_{\substack{i=s+1 \\ i \notin S_{\le t_k}}}^{t_k} \tilde{P}_i^{t_k} - \prod_{\substack{i=s+1 \\ i \notin S_{\le t_k}}}^{t_k} \hat{P}_i^{t_k}, \;\; D_{j+1}(t_k, s) = D_j(t_k, s) - D_j(t_k, t_{k-j-1}) d_0(t_{k-j-1}, s). \tag{93}$$

The corresponding G-type terms satisfy

$$b_{0,0}(t_k, s) = \prod_{\substack{i=s+1 \\ i \notin S_{\le t_k}}}^{t_k} \hat{P}_i^{t_k} \hat{G}_s^{t_k}, \;\; b_{j+1,j+1}(t_k, s) = b_{j,j}(t_k, s) - d_j(t_k, t_{k-j-1})\, b_{0,0}(t_{k-j-1}, s).$$

Moreover,

$$B_{0,0}(t_k, s) = \prod_{\substack{i=s+1 \\ i \notin S_{\le t_k}}}^{t_k} \tilde{P}_i^{t_k} \tilde{G}_s^{t_k} - \prod_{\substack{i=s+1 \\ i \notin S_{\le t_k}}}^{t_k} \hat{P}_i^{t_k} \hat{G}_s^{t_k},$$

$$B_{j+1,j+1}(t_k, s) = B_{j,j}(t_k, s) - D_j(t_k, t_{k-j-1})\, b_{0,0}(t_{k-j-1}, s).$$

Proof of Lemma D.8. We first prove the recursions for $d_\cdot(\cdot,\cdot)$ and $D_\cdot(\cdot,\cdot)$. The expression for d_0 follows directly from the unlearning update in (14). For any earlier unlearning time $t_a < t_k$, by definition,

$$\prod_{\substack{r=t_a+1 \\ r \notin S_{\le t_k}}}^{t_k} \hat{P}_r^{t_k} = d_0(t_k, t_a).$$

Since the term Δ_s with $s \in S_{t_{k-j-1}}$ first appears in $\bar{\Delta}_{t_{k-j-1}}$ and is then propagated recursively through each later $\bar{\Delta}_{t_m}$ with $k \ge m > k-j-1$, we have, for any $l = 0, \dots, j+1$ with an empty sum when $l = 0$,

$$d_l(t_k, s) = d_0(t_k, s) - \sum_{i=0}^{l-1} d_0(t_k, t_{k-l+i}) d_i(t_{k-l+i}, s). \tag{94}$$

Subtracting (94) with $l = j$ from the same identity with $l = j+1$ gives the recursion for d_{j+1}, as in Lemma E.8 of Hu et al. (2026).

Next, the expression for D_0 follows from Lemma D.7. For any earlier unlearning time $t_a < t_k$, by definition,

$$\prod_{\substack{r=t_a+1 \\ r \notin S_{\le t_k}}}^{t_k} \tilde{P}_r^{t_k} - \prod_{\substack{r=t_a+1 \\ r \notin S_{\le t_k}}}^{t_k} \hat{P}_r^{t_k} = D_0(t_k, t_a).$$

The same recursive decomposition gives, for any $l = 0, \dots, j+1$ with an empty sum when $l = 0$,

$$D_l(t_k, s) = D_0(t_k, s) - \sum_{i=0}^{l-1} D_0(t_k, t_{k-l+i}) d_i(t_{k-l+i}, s). \tag{95}$$

Subtracting (95) with $l = j$ from the same identity with $l = j+1$ gives the recursion for D_{j+1}.

We then prove the recursions for $b_\cdot(\cdot,\cdot)$ and $B_\cdot(\cdot,\cdot)$.

Consider first $b_{l,l}(t_k, s)$, where $0 \le l < k$. For $0 \le l \le k-2$, the relevant indices are in $([\underline{s}_{k-l}+1, \underline{s}_{k-l-1}-1] \cup [t_{k-l-1}+1, t_{k-l}]) \setminus S_{\le t_k}$; for $l = k-1$, they are in $[\underline{s}_1+1, t_1] \setminus S_{\le t_k}$. In both cases, these elements have not yet been forgotten by time t_k and first appear at time t_{k-l}. Therefore, by (14), for $l = 0, \dots, k-1$,

$$\begin{aligned} b_{l,l}(t_k, s) &= \prod_{\substack{n=s+1 \\ n \notin S_{\le t_k}}}^{t_k} \widehat{P}_n^{t_k} \widehat{G}_s^{t_k} - \sum_{m=1}^{l} \prod_{\substack{n=t_{k-m}+1 \\ n \notin S_{\le t_k}}}^{t_k} \widehat{P}_n^{t_k}\, b_{l-m,l-m}(t_{k-m}, s) \\ &= b_{0,0}(t_k, s) - \sum_{m=1}^{l} d_0(t_k, t_{k-m}) b_{l-m,l-m}(t_{k-m}, s). \end{aligned} \tag{96}$$

Hence

$$
\begin{aligned}
& b_{j,j}(t_k,s) - b_{j+1,j+1}(t_k,s) \\
=& b_{0,0}(t_k,s) - \sum_{i=0}^{j-1} d_0(t_k,t_{k-j+i}) b_{i,i}(t_{k-j+i},s) \\
& - \left(b_{0,0}(t_k,s) - \sum_{i=0}^{j} d_0(t_k,t_{k-j+i-1}) b_{i,i}(t_{k-j+i-1},s) \right) \\
=& - \sum_{i=0}^{j-1} d_0(t_k,t_{k-j+i}) b_{i,i}(t_{k-j+i},s) + \sum_{i=-1}^{j-1} d_0(t_k,t_{k-j+i}) b_{i+1,i+1}(t_{k-j+i},s) \\
=& - \sum_{i=0}^{j-1} d_0(t_k,t_{k-j+i}) \big(b_{i,i}(t_{k-j+i},s) - b_{i+1,i+1}(t_{k-j+i},s) \big) + d_0(t_k,t_{k-j-1}) b_{0,0}(t_{k-j-1},s) \\
=& - \sum_{i=0}^{j-1} d_0(t_k,t_{k-j+i}) d_i(t_{k-j+i},t_{k-j-1}) b_{0,0}(t_{k-j-1},s) + d_0(t_k,t_{k-j-1}) b_{0,0}(t_{k-j-1},s) \\
=& \left(d_0(t_k,t_{k-j-1}) - \sum_{i=0}^{j-1} d_0(t_k,t_{k-j+i}) d_i(t_{k-j+i},t_{k-j-1}) \right) b_{0,0}(t_{k-j-1},s) \\
=& d_j(t_k,t_{k-j-1}) b_{0,0}(t_{k-j-1},s),
\end{aligned}
$$

where the fourth equality follows from the induction hypothesis for $0 \le i \le j-1$, and the last equality uses (94) with $s = t_{k-j-1}$ and $l = j$. This completes the induction.

Next consider $B_{l,l}(t_k,s)$, where $0 \le l < k$. The same index decomposition as above shows that the corresponding elements have not yet been forgotten by time t_k and first appear at time t_{k-l}. Therefore, by (14), for $l = 0, \dots, k-1$,

$$
\begin{aligned}
& B_{l,l}(t_k,s) && (97) \\
&= \prod_{\substack{n=s+1 \\ n \notin S_{\le t_k}}}^{t_k} \tilde{P}_n^{t_k} \tilde{G}_s^{t_k} - \prod_{\substack{n=s+1 \\ n \notin S_{\le t_k}}}^{t_k} \widehat{P}_n^{t_k} \widehat{G}_s^{t_k} - \sum_{m=1}^{l} \left(\prod_{\substack{n=t_{k-m}+1 \\ n \notin S_{\le t_k}}}^{t_k} \tilde{P}_n^{t_k} - \prod_{\substack{n=t_{k-m}+1 \\ n \notin S_{\le t_k}}}^{t_k} \widehat{P}_n^{t_k} \right) b_{l-m,l-m}(t_{k-m},s) \\
&= B_{0,0}(t_k,s) - \sum_{n=1}^{l} D_0(t_k,t_{k-n})\, b_{l-n,l-n}(t_{k-n},s). && (98)
\end{aligned}
$$

Hence

$$
\begin{aligned}
& B_{j,j}(t_k,s) - B_{j+1,j+1}(t_k,s) \\
=& B_{0,0}(t_k,s) - \sum_{n=1}^{j} D_0(t_k,t_{k-n}) b_{j-n,j-n}(t_{k-n},s) \\
& - \left(B_{0,0}(t_k,s) - \sum_{n=1}^{j+1} D_0(t_k,t_{k-n}) b_{j+1-n,j+1-n}(t_{k-n},s) \right) \\
=& - \sum_{n=1}^{j} D_0(t_k,t_{k-n}) \big(b_{j-n,j-n}(t_{k-n},s) - b_{j+1-n,j+1-n}(t_{k-n},s) \big) \\
& + D_0(t_k,t_{k-j-1}) b_{0,0}(t_{k-j-1},s) \\
=& - \sum_{n=1}^{j} D_0(t_k,t_{k-n}) d_{j-n}(t_{k-n},t_{k-j-1}) b_{0,0}(t_{k-j-1},s) + D_0(t_k,t_{k-j-1}) b_{0,0}(t_{k-j-1},s) \\
=& \left(D_0(t_k,t_{k-j-1}) - \sum_{n=1}^{j} D_0(t_k,t_{k-n}) d_{j-n}(t_{k-n},t_{k-j-1}) \right) b_{0,0}(t_{k-j-1},s) \\
=& D_j(t_k,t_{k-j-1}) b_{0,0}(t_{k-j-1},s),
\end{aligned}
$$

where the third equality uses the recursion for b proved above, and the last equality uses (95) with $s = t_{k-j-1}$ and $l = j$. This completes the induction.

□

Lemma D.9. *Fix $j = 0, \dots, k-2$ and $s \in S_{t_{k-j-1}}$. For $D_{j+1}(t_k, s)$, we have*

$$D_{j+1}(t_k,s) = \sum_{i=0}^{j} D_{j-i}(t_k, t_{k-j-1}) \left(\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j+i}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+i}} - \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j+i-1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+i-1}} \right) \tag{99}$$

$$+ D_0(t_k,s) - D_0(t_k,t_{k-j-1}) \prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \hat{P}_a^{t_k}. \tag{100}$$

For $B_{j+1,j+1}(t_k, s)$, we have

$$\begin{aligned} &B_{j+1,j+1}(t_k,s) \\ &= \sum_{i=0}^{j} D_{j-i}(t_k, t_{k-j-1}) \left(\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j+i}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+i}} \hat{G}_s^{t_{k-j+i}} - \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j+i-1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+i-1}} \hat{G}_s^{t_{k-j+i-1}} \right) \\ &+ B_{0,0}(t_k,s) - D_0(t_k,t_{k-j-1}) \prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \hat{P}_a^{t_k} \hat{G}_s^{t_k}. \end{aligned} \tag{101}$$

Proof. Define

$$A_i(s) := \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j+i-1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+i-1}}, \qquad i = 0, \dots, j+1.$$

When $j = 0$, both identities follow directly from Lemma D.8 after adding and subtracting the corresponding $A_1(s)$-term. Hence, in the rest of the proof, assume $j \ge 1$. By Lemma D.8,

$$D_{j+1}(t_k,s) = D_j(t_k,s) - D_j(t_k,t_{k-j-1})A_0(s).$$

We claim that for $b = 0, \dots, j-1$,

$$\begin{aligned} D_{j+1}(t_k,s) = &\sum_{i=0}^{b} D_{j-i}(t_k,t_{k-j-1})\big(A_{i+1}(s) - A_i(s)\big) \\ &+ D_{j-b-1}(t_k,s) - D_{j-b-1}(t_k,t_{k-j-1})A_{b+1}(s). \end{aligned}$$

For $b = 0$ and $j \ge 1$, add and subtract $D_j(t_k, t_{k-j-1})A_1(s)$. Then

$$\begin{aligned} D_{j+1}(t_k,s) &= D_j(t_k,s) - D_j(t_k,t_{k-j-1})A_0(s) \\ &= D_j(t_k,t_{k-j-1})\big(A_1(s) - A_0(s)\big) + \big(D_j(t_k,s) - D_j(t_k,t_{k-j-1})A_1(s)\big). \end{aligned}$$

It remains to simplify the second term. By Lemma D.8, applied once with starting index s and once with starting index t_{k-j-1}, we have

$$D_j(t_k,s) = D_{j-1}(t_k,s) - D_{j-1}(t_k,t_{k-j}) \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j}} \hat{P}_a^{t_{k-j}},$$

$$D_j(t_k,t_{k-j-1}) = D_{j-1}(t_k,t_{k-j-1}) - D_{j-1}(t_k,t_{k-j}) \prod_{\substack{a=t_{k-j-1}+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j}} \hat{P}_a^{t_{k-j}}.$$

Moreover, by the product decomposition convention,

$$\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j}} \hat{P}_a^{t_{k-j}} = \left(\prod_{\substack{a=t_{k-j-1}+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j}} \hat{P}_a^{t_{k-j}} \right) A_1(s).$$

Therefore,

$$\begin{aligned}
&D_j(t_k,s) - D_j(t_k,t_{k-j-1})A_1(s)\\
&= D_{j-1}(t_k,s) - D_{j-1}(t_k,t_{k-j})\left(\prod_{\substack{a=t_{k-j-1}+1\\ a\notin S_{\le t_{k-j}}}}^{t_{k-j}} \hat{P}_a^{t_{k-j}}\right)A_1(s)\\
&\quad - D_{j-1}(t_k,t_{k-j-1})A_1(s) + D_{j-1}(t_k,t_{k-j})\left(\prod_{\substack{a=t_{k-j-1}+1\\ a\notin S_{\le t_{k-j}}}}^{t_{k-j}} \hat{P}_a^{t_{k-j}}\right)A_1(s)\\
&= D_{j-1}(t_k,s) - D_{j-1}(t_k,t_{k-j-1})A_1(s).
\end{aligned}$$

Thus,

$$D_{j+1}(t_k,s) = D_j(t_k,t_{k-j-1})\big(A_1(s)-A_0(s)\big) + D_{j-1}(t_k,s) - D_{j-1}(t_k,t_{k-j-1})A_1(s),$$

which proves the base case.

Assume the claim holds for $b = l < j-1$. Then we have

$$\begin{aligned}
D_{j+1}(t_k,s) = &\sum_{i=0}^{l} D_{j-i}(t_k,t_{k-j-1})\big(A_{i+1}(s)-A_i(s)\big)\\
&+ D_{j-l-1}(t_k,s)\\
&- D_{j-l-1}(t_k,t_{k-j-1})\prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l}}.
\end{aligned}$$

The second term can be written as

$$\begin{aligned}
&D_{j-l-1}(t_k,s) - D_{j-l-1}(t_k,t_{k-j-1})\prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l}}\\
=&D_{j-l-1}(t_k,t_{k-j-1})\left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l+1}} - \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l}}\right)\\
&+ D_{j-l-1}(t_k,s) - D_{j-l-1}(t_k,t_{k-j-1})\prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l+1}}.
\end{aligned}$$

It remains to simplify the second line. By Lemma D.8,

$$D_{j-l-1}(t_k,s) = D_{j-l-2}(t_k,s) - D_{j-l-2}(t_k,t_{k-j+l+1})\prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j+l+1}} \hat{P}_a^{t_{k-j+l+1}},$$

$$D_{j-l-1}(t_k,t_{k-j-1}) = D_{j-l-2}(t_k,t_{k-j-1}) - D_{j-l-2}(t_k,t_{k-j+l+1})\prod_{\substack{a=t_{k-j-1}+1\\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j+l+1}} \hat{P}_a^{t_{k-j+l+1}}.$$

Therefore,

$$
\begin{aligned}
&D_{j-l-1}(t_k,s) - D_{j-l-1}(t_k,t_{k-j-1}) \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l+1}} \\
&= D_{j-l-2}(t_k,s) - D_{j-l-2}(t_k,t_{k-j+l+1}) \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j+l+1}} \hat{P}_a^{t_{k-j+l+1}} \\
&\quad - D_{j-l-2}(t_k,t_{k-j-1}) \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l+1}} \\
&\quad + D_{j-l-2}(t_k,t_{k-j+l+1}) \prod_{\substack{a=t_{k-j-1}+1\\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j+l+1}} \hat{P}_a^{t_{k-j+l+1}} \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l+1}} \\
&= D_{j-l-2}(t_k,s) - D_{j-l-2}(t_k,t_{k-j-1}) \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l+1}}.
\end{aligned}
$$

This proves the induction step from $b = l$ to $b = l + 1$.

Taking $b = j - 1$, we obtain

$$
D_{j+1}(t_k,s) = \sum_{i=0}^{j-1} D_{j-i}(t_k,t_{k-j-1})(A_{i+1}(s) - A_i(s)) + D_0(t_k,s) - D_0(t_k,t_{k-j-1})A_j(s).
$$

Finally, add and subtract $D_0(t_k,t_{k-j-1})A_{j+1}(s)$. Since $A_{j+1}(s) = \prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \hat{P}_a^{t_k}$, this gives exactly (100).

To prove the second result, we claim that for $b = 0, \dots, j-1$,

$$
\begin{aligned}
B_{j+1,j+1}(t_k,s) = &\sum_{i=0}^{b} D_{j-i}(t_k,t_{k-j-1})\big(A_{i+1}(s)\hat{G}_s^{t_{k-j+i}} - A_i(s)\hat{G}_s^{t_{k-j+i-1}}\big) \\
&+ B_{j-b-1,j-b-1}(t_k,s) - D_{j-b-1}(t_k,t_{k-j-1})A_{b+1}(s)\hat{G}_s^{t_{k-j+b}}.
\end{aligned}
$$

For $b = 0$ and $j \ge 1$, add and subtract $D_j(t_k,t_{k-j-1})A_1(s)\hat{G}_s^{t_{k-j}}$ in Lemma D.8, then

$$
\begin{aligned}
&B_{j+1,j+1}(t_k,s) \\
&= B_{j,j}(t_k,s) - D_j(t_k,t_{k-j-1})A_0(s)\hat{G}_s^{t_{k-j-1}} \\
&= D_j(t_k,t_{k-j-1})\big(A_1(s)\hat{G}_s^{t_{k-j}} - A_0(s)\hat{G}_s^{t_{k-j-1}}\big) + \big(B_{j,j}(t_k,s) - D_j(t_k,t_{k-j-1})A_1(s)\hat{G}_s^{t_{k-j}}\big).
\end{aligned}
$$

It remains to simplify the second term. By Lemma D.8, applied once with starting index s and once with starting index t_{k-j-1}, we have

$$
B_{j,j}(t_k,s) = B_{j-1,j-1}(t_k,s) - D_{j-1}(t_k,t_{k-j}) \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j}}}}^{t_{k-j}} \hat{P}_a^{t_{k-j}}\hat{G}_s^{t_{k-j}},
$$

$$
D_j(t_k,t_{k-j-1}) = D_{j-1}(t_k,t_{k-j-1}) - D_{j-1}(t_k,t_{k-j}) \prod_{\substack{a=t_{k-j-1}+1\\ a\notin S_{\le t_{k-j}}}}^{t_{k-j}} \hat{P}_a^{t_{k-j}}.
$$

Moreover, by the product decomposition convention,

$$\prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j}}}}^{t_{k-j}} \hat{P}_a^{t_{k-j}} = \left(\prod_{\substack{a=t_{k-j-1}+1\\ a\notin S_{\le t_{k-j}}}}^{t_{k-j}} \hat{P}_a^{t_{k-j}} \right) A_1(s).$$

Therefore,

$$\begin{aligned}
&B_{j,j}(t_k,s) - D_j(t_k,t_{k-j-1})A_1(s)\hat{G}_s^{t_{k-j}} \\
&= B_{j-1,j-1}(t_k,s) - D_{j-1}(t_k,t_{k-j}) \left(\prod_{\substack{a=t_{k-j-1}+1\\ a\notin S_{\le t_{k-j}}}}^{t_{k-j}} \hat{P}_a^{t_{k-j}} \right) A_1(s)\hat{G}_s^{t_{k-j}} \\
&\quad - D_{j-1}(t_k,t_{k-j-1})A_1(s)\hat{G}_s^{t_{k-j}} + D_{j-1}(t_k,t_{k-j}) \left(\prod_{\substack{a=t_{k-j-1}+1\\ a\notin S_{\le t_{k-j}}}}^{t_{k-j}} \hat{P}_a^{t_{k-j}} \right) A_1(s)\hat{G}_s^{t_{k-j}} \\
&= B_{j-1,j-1}(t_k,s) - D_{j-1}(t_k,t_{k-j-1})A_1(s)\hat{G}_s^{t_{k-j}}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&B_{j+1,j+1}(t_k,s) \\
=&D_j(t_k,t_{k-j-1})\big(A_1(s)\hat{G}_s^{t_{k-j}} - A_0(s)\hat{G}_s^{t_{k-j-1}}\big) \\
&+ B_{j-1,j-1}(t_k,s) - D_{j-1}(t_k,t_{k-j-1})A_1(s)\hat{G}_s^{t_{k-j}},
\end{aligned}$$

which proves the base case.

Assume the claim holds for $b = l < j-1$. Then we have

$$\begin{aligned}
B_{j+1,j+1}(t_k,s) =& \sum_{i=0}^{l} D_{j-i}(t_k,t_{k-j-1})\big(A_{i+1}(s)\hat{G}_s^{t_{k-j+i}} - A_i(s)\hat{G}_s^{t_{k-j+i-1}}\big) \\
&+ B_{j-l-1,j-l-1}(t_k,s) - D_{j-l-1}(t_k,t_{k-j-1}) \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l}}\hat{G}_s^{t_{k-j+l}}.
\end{aligned}$$

The second term can be written as

$$\begin{aligned}
&B_{j-l-1,j-l-1}(t_k,s) - D_{j-l-1}(t_k,t_{k-j-1}) \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l}}\hat{G}_s^{t_{k-j+l}} \\
=&D_{j-l-1}(t_k,t_{k-j-1}) \left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l+1}}\hat{G}_s^{t_{k-j+l+1}} - \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l}}\hat{G}_s^{t_{k-j+l}} \right) \\
&+ B_{j-l-1,j-l-1}(t_k,s) - D_{j-l-1}(t_k,t_{k-j-1}) \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l+1}}\hat{G}_s^{t_{k-j+l+1}}.
\end{aligned}$$

It remains to simplify the second line. By Lemma D.8,

$$B_{j-l-1,j-l-1}(t_k,s) = B_{j-l-2,j-l-2}(t_k,s) - D_{j-l-2}(t_k,t_{k-j+l+1}) \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j+l+1}} \hat{P}_a^{t_{k-j+l+1}} \hat{G}_s^{t_{k-j+l+1}},$$

$$D_{j-l-1}(t_k,t_{k-j-1}) = D_{j-l-2}(t_k,t_{k-j-1}) - D_{j-l-2}(t_k,t_{k-j+l+1}) \prod_{\substack{a=t_{k-j-1}+1 \\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j+l+1}} \hat{P}_a^{t_{k-j+l+1}}.$$

Therefore,

$$\begin{aligned}
&B_{j-l-1,j-l-1}(t_k,s) - D_{j-l-1}(t_k,t_{k-j-1}) \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l+1}} \hat{G}_s^{t_{k-j+l+1}} \\
&= B_{j-l-2,j-l-2}(t_k,s) - D_{j-l-2}(t_k,t_{k-j+l+1}) \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j+l+1}} \hat{P}_a^{t_{k-j+l+1}} \hat{G}_s^{t_{k-j+l+1}} \\
&\quad - D_{j-l-2}(t_k,t_{k-j-1}) \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l+1}} \hat{G}_s^{t_{k-j+l+1}} \\
&\quad + D_{j-l-2}(t_k,t_{k-j+l+1}) \prod_{\substack{a=t_{k-j-1}+1 \\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j+l+1}} \hat{P}_a^{t_{k-j+l+1}} \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l+1}} \hat{G}_s^{t_{k-j+l+1}} \\
&= B_{j-l-2,j-l-2}(t_k,s) - D_{j-l-2}(t_k,t_{k-j-1}) \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j+l+1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+l+1}} \hat{G}_s^{t_{k-j+l+1}}.
\end{aligned}$$

This proves the induction step from $b = l$ to $b = l + 1$.

Taking $b = j - 1$, we obtain

$$\begin{aligned}
&B_{j+1,j+1}(t_k,s) \\
&= \sum_{i=0}^{j-1} D_{j-i}(t_k,t_{k-j-1})\big(A_{i+1}(s)\hat{G}_s^{t_{k-j+i}} - A_i(s)\hat{G}_s^{t_{k-j+i-1}}\big) \\
&\quad + B_{0,0}(t_k,s) - D_0(t_k,t_{k-j-1})A_j(s)\hat{G}_s^{t_{k-1}}.
\end{aligned}$$

Finally, add and subtract $D_0(t_k,t_{k-j-1})A_{j+1}(s)\hat{G}_s^{t_k}$. Since $A_{j+1}(s) = \prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \hat{P}_a^{t_k}$, this gives exactly (101). □

Lemma D.10. *Define*

$$C_1^1(t_k,t_{k-1},s) = Z_{[s+1,t_{k-1}]}^{k\to k-1}, \quad \hat{C}_1^1(t_k,t_{k-1},s) = Z_{[s+1,t_{k-1}]}^{k\to k-1}\hat{g}_s^{t_k} + \frac{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-1}}}}^{t_{k-1}} \widehat{\rho}_a^{t_{k-1}}}{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_{k-1}} \widehat{\rho}_a^{t_k}} \zeta_s^{t_{k-1}\to t_k}.$$

For $j \geq 1$, define C^i_{j+1} and $\hat{C}^i_{j+1}$ recursively by

$$C^i_{j+1}(t_k, t_{k-j-1}, s) = \begin{cases} C^i_j(t_k, t_{k-j}, s) & 1 \leq i \leq j, \\ \quad + C^i_j(t_k, t_{k-j}, t_{k-j-1}) Z^{k-j \to k-j-1}_{[s+1, t_{k-j-1}]} & \\ \qquad \times \dfrac{\prod^{t_{k-j-1}}_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j}}}} \widehat{\rho}^{t_{k-j}}_a}{\prod^{t_{k-j-1}}_{\substack{a=s+1 \\ a \notin S_{\leq t_k}}} \widehat{\rho}^{t_k}_a} & \\ C^j_j(t_k, t_{k-j}, t_{k-j-1}) & i = j+1. \\ \quad + Z^{k-j \to k-j-1}_{[s+1, t_{k-j-1}]} \dfrac{\prod^{t_{k-j-1}}_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j}}}} \widehat{\rho}^{t_{k-j}}_a}{\prod^{t_{k-j-1}}_{\substack{a=s+1 \\ a \notin S_{\leq t_k}}} \widehat{\rho}^{t_k}_a} & \end{cases}$$

$$\begin{aligned} &\hat{C}^i_{j+1}(t_k, t_{k-j-1}, s) \\ &= \begin{cases} \hat{C}^i_j(t_k, t_{k-j}, s) & 1 \leq i \leq j, \\ \quad + C^i_j(t_k, t_{k-j}, t_{k-j-1}) \Bigg(& \\ \quad \dfrac{\prod^{t_{k-j-1}}_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j}}}} \widehat{\rho}^{t_{k-j}}_a}{\prod^{t_{k-j-1}}_{\substack{a=s+1 \\ a \notin S_{\leq t_k}}} \widehat{\rho}^{t_k}_a} Z^{k-j \to k-j-1}_{[s+1, t_{k-j-1}]} \hat{g}^{t_{k-j}}_s & \\ \quad + \dfrac{\prod^{t_{k-j-1}}_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j-1}}}} \widehat{\rho}^{t_{k-j-1}}_a}{\prod^{t_{k-j-1}}_{\substack{a=s+1 \\ a \notin S_{\leq t_k}}} \widehat{\rho}^{t_k}_a} \zeta^{t_{k-j-1} \to t_{k-j}}_s \Bigg) & \\ \hat{C}^j_j(t_k, t_{k-j}, t_{k-j-1}) & i = j+1. \\ \quad + \Bigg(\dfrac{\prod^{t_{k-j-1}}_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j}}}} \widehat{\rho}^{t_{k-j}}_a}{\prod^{t_{k-j-1}}_{\substack{a=s+1 \\ a \notin S_{\leq t_k}}} \widehat{\rho}^{t_k}_a} Z^{k-j \to k-j-1}_{[s+1, t_{k-j-1}]} \hat{g}^{t_{k-j}}_s & \\ \quad + \dfrac{\prod^{t_{k-j-1}}_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j-1}}}} \widehat{\rho}^{t_{k-j-1}}_a}{\prod^{t_{k-j-1}}_{\substack{a=s+1 \\ a \notin S_{\leq t_k}}} \widehat{\rho}^{t_k}_a} \zeta^{t_{k-j-1} \to t_{k-j}}_s \Bigg) & \end{cases} \end{aligned}$$

Then, for every $j \geq 1$, $\|D_j(t_k, s)\| \leq \mathbf{D}_j(t_k, s)$ and $\|B_{j,j}(t_k, s)\| \leq \mathbf{B}_{j,j}(t_k, s)$ where

$$\begin{aligned} \mathbf{D}_j(t_k, s) := \left(\prod^{t_k}_{\substack{a=s+1 \\ a \notin S_{\leq t_k}}} \widehat{\rho}^{t_k}_a \right) \Bigg[& \sum^{t_{k-j}}_{\substack{q=s+1 \\ q \notin S_{\leq t_k}}} \frac{\bar{\rho}^{t_k}_q}{\widehat{\rho}^{t_k}_q} \prod^{t_k}_{\substack{b=q+1 \\ b \notin S_{\leq t_k}}} \frac{\rho^{t_k}_b}{\widehat{\rho}^{t_k}_b} \\ & + \sum^{j-1}_{i=0} \sum^{t_{k-i}}_{\substack{q=t_{k-i-1}+1 \\ q \notin S_{\leq t_k}}} \frac{\bar{\rho}^{t_k}_q}{\widehat{\rho}^{t_k}_q} \prod^{t_k}_{\substack{b=q+1 \\ b \notin S_{\leq t_k}}} \frac{\rho^{t_k}_b}{\widehat{\rho}^{t_k}_b} C^{i+1}_j(t_k, t_{k-j}, s) \Bigg]. \end{aligned}$$

$$\mathbf{B}_{j,j}(t_k,s) := \left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a^{t_k}\right)\Bigg[\sum_{\substack{q=s+1\\ q\notin S_{\le t_k}}}^{t_{k-j}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}\hat{g}_s^{t_k} + \sum_{i=0}^{j-1}\sum_{\substack{q=t_{k-i-1}+1\\ q\notin S_{\le t_k}}}^{t_{k-i}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}\hat{C}_j^{i+1}(t_k,t_{k-j},s)\Bigg] + \left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \rho_a^{t_k}\right)\bar{g}_s^{t_k}.$$

Corollary D.11. *Define*

$$\tilde{C}_1^1(t_k,t_{k-1},s) = Z^{k\to k-1}_{[s+1,t_{k-1}]}\hat{\kappa}_s^{t_k} + \frac{\prod_{a=s+1}^{t_{k-1}}{}_{a\notin S_{\le t_{k-1}}} \widehat{\rho}_a^{t_{k-1}}}{\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-1}} \widehat{\rho}_a^{t_k}}\zeta_s^{t_{k-1}\to t_k}.$$

For $j \ge 1$, define $\tilde{C}_{j+1}^i$ recursively by

$$\tilde{C}_{j+1}^i(t_k,t_{k-j-1},s) = \begin{cases} \tilde{C}_j^i(t_k,t_{k-j},s) & 1\le i\le j,\\ \quad + C_j^i(t_k,t_{k-j},t_{k-j-1})\Bigg(& \\ \quad \frac{\prod_{a=s+1}^{t_{k-j-1}}{}_{a\notin S_{\le t_{k-j}}} \widehat{\rho}_a^{t_{k-j}}}{\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_k}} Z^{k-j\to k-j-1}_{[s+1,t_{k-j-1}]}\hat{\kappa}_s^{t_{k-j}} & \\ \quad + \frac{\prod_{a=s+1}^{t_{k-j-1}}{}_{a\notin S_{\le t_{k-j-1}}} \widehat{\rho}_a^{t_{k-j-1}}}{\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_k}}\zeta_s^{t_{k-j-1}\to t_{k-j}}\Bigg) & \\ \tilde{C}_j^j(t_k,t_{k-j},t_{k-j-1}) & i=j+1.\\ \quad + \Bigg(\frac{\prod_{a=s+1}^{t_{k-j-1}}{}_{a\notin S_{\le t_{k-j}}} \widehat{\rho}_a^{t_{k-j}}}{\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_k}} Z^{k-j\to k-j-1}_{[s+1,t_{k-j-1}]}\hat{\kappa}_s^{t_{k-j}} & \\ \quad + \frac{\prod_{a=s+1}^{t_{k-j-1}}{}_{a\notin S_{\le t_{k-j-1}}} \widehat{\rho}_a^{t_{k-j-1}}}{\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_k}}\zeta_s^{t_{k-j-1}\to t_{k-j}}\Bigg) & \end{cases}$$

Then, for every $j \ge 1$, $\|D_j(t_k,s)\Delta_s - B_{j,j}(t_k,s)\| \le \mathbf{DB}_j(t_k,s)$, where

$$\mathbf{DB}_j(t_k,s) :$$
$$= \left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a^{t_k}\right)\Bigg[\sum_{\substack{q=s+1\\ q\notin S_{\le t_k}}}^{t_{k-j}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}\hat{\kappa}_s^{t_k} + \sum_{i=0}^{j-1}\sum_{\substack{q=t_{k-i-1}+1\\ q\notin S_{\le t_k}}}^{t_{k-i}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}\tilde{C}_j^{i+1}(t_k,t_{k-j},s)\Bigg]$$
$$+ \left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \rho_a^{t_k}\right)\bar{g}_s^{t_k}.$$

Proof of Lemma D.10 and Corollary D.11. We first prove Lemma D.10 by induction on j.

First, we prove the elementary bound for D_0. By definition,

$$D_0(t_k, s) = \prod_{\substack{a=s+1 \\ a \notin S_{\le t_k}}}^{t_k} \widetilde{P}_a^{t_k} - \prod_{\substack{a=s+1 \\ a \notin S_{\le t_k}}}^{t_k} \widehat{P}_a^{t_k}.$$

By Lemma D.3,

$$\|D_0(t_k, s)\| \le \left(\prod_{\substack{a=s+1 \\ a \notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a^{t_k} \right) \sum_{\substack{q=s+1 \\ q \notin S_{\le t_k}}}^{t_k} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b \notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}.$$

Now consider $j = 1$. By Lemma D.8, we have

$$\begin{aligned}
D_1(t_k, s) =& D_0(t_k, s) - D_0(t_k, t_{k-1}) d_0(t_{k-1}, s) && (102)\\
=& D_0(t_k, s) - D_0(t_k, t_{k-1}) \prod_{\substack{a=s+1 \\ a \notin S_{\le t_k}}}^{t_{k-1}} \widehat{P}_a^{t_k} \\
&+ D_0(t_k, t_{k-1}) \left(\prod_{\substack{a=s+1 \\ a \notin S_{\le t_k}}}^{t_{k-1}} \widehat{P}_a^{t_k} - \prod_{\substack{a=s+1 \\ a \notin S_{\le t_{k-1}}}}^{t_{k-1}} \widehat{P}_a^{t_{k-1}} \right) && (103)\\
=& \left(\prod_{\substack{a=t_{k-1}+1 \\ a \notin S_{\le t_k}}}^{t_k} \widetilde{P}_a^{t_k} \right) \left(\prod_{\substack{a=s+1 \\ a \notin S_{\le t_k}}}^{t_{k-1}} \widetilde{P}_a^{t_k} - \prod_{\substack{a=s+1 \\ a \notin S_{\le t_k}}}^{t_{k-1}} \widehat{P}_a^{t_k} \right) \\
&+ D_0(t_k, t_{k-1}) \left(\prod_{\substack{a=s+1 \\ a \notin S_{\le t_k}}}^{t_{k-1}} \widehat{P}_a^{t_k} - \prod_{\substack{a=s+1 \\ a \notin S_{\le t_{k-1}}}}^{t_{k-1}} \widehat{P}_a^{t_{k-1}} \right). && (104)
\end{aligned}$$

Taking norms gives

$$\begin{aligned}
\|D_1(t_k, s)\| \le & \left\| \prod_{\substack{a=t_{k-1}+1 \\ a \notin S_{\le t_k}}}^{t_k} \widetilde{P}_a^{t_k} \right\| \left\| \prod_{\substack{a=s+1 \\ a \notin S_{\le t_k}}}^{t_{k-1}} \widetilde{P}_a^{t_k} - \prod_{\substack{a=s+1 \\ a \notin S_{\le t_k}}}^{t_{k-1}} \widehat{P}_a^{t_k} \right\| \\
&+ \|D_0(t_k, t_{k-1})\| \left\| \prod_{\substack{a=s+1 \\ a \notin S_{\le t_k}}}^{t_{k-1}} \widehat{P}_a^{t_k} - \prod_{\substack{a=s+1 \\ a \notin S_{\le t_{k-1}}}}^{t_{k-1}} \widehat{P}_a^{t_{k-1}} \right\|.
\end{aligned}$$

By Lemma D.3, the first term is bounded by

$$\left(\prod_{\substack{a=t_{k-1}+1 \\ a \notin S_{\le t_k}}}^{t_k} \rho_a^{t_k} \right) \left(\prod_{\substack{a=s+1 \\ a \notin S_{\le t_k}}}^{t_{k-1}} \widehat{\rho}_a^{t_k} \right) \sum_{\substack{q=s+1 \\ q \notin S_{\le t_k}}}^{t_{k-1}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b \notin S_{\le t_k}}}^{t_{k-1}} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} = \left(\prod_{\substack{a=s+1 \\ a \notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a^{t_k} \right) \sum_{\substack{q=s+1 \\ q \notin S_{\le t_k}}}^{t_{k-1}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b \notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}.$$

For the second term, also by Lemma D.3 and Lemma D.8, we have

$$\|D_0(t_k,t_{k-1})\| \le \left(\prod_{\substack{a=t_{k-1}+1\\a\notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a^{t_k}\right) \sum_{\substack{q=t_{k-1}+1\\q\notin S_{\le t_k}}}^{t_k} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}. \tag{105}$$

Moreover, by Lemma D.5,

$$\left\| \prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_{k-1}} \widehat{P}_a^{t_k} - \prod_{\substack{a=s+1\\a\notin S_{\le t_{k-1}}}}^{t_{k-1}} \widehat{P}_a^{t_{k-1}} \right\| \le \left(\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_{k-1}} \hat{\rho}_a^{t_k}\right) Z_{[s+1,t_{k-1}]}^{k\to k-1}.$$

Hence the second term is bounded by

$$\left(\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a^{t_k}\right) \sum_{\substack{q=t_{k-1}+1\\q\notin S_{\le t_k}}}^{t_k} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} Z_{[s+1,t_{k-1}]}^{k\to k-1}.$$

Reorganizing the terms, we obtain

$$\|D_1(t_k,s)\| \le \left(\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a^{t_k}\right) \left[\sum_{\substack{q=s+1\\q\notin S_{\le t_k}}}^{t_{k-1}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} + \sum_{\substack{q=t_{k-1}+1\\q\notin S_{\le t_k}}}^{t_k} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} Z_{[s+1,t_{k-1}]}^{k\to k-1} \right].$$

Since

$$C_1^1(t_k,t_{k-1},s) = Z_{[s+1,t_{k-1}]}^{k\to k-1},$$

this is exactly the desired bound for $j=1$.

Now assume the estimate holds for some $j \ge 1$. We prove it for $j+1$. Lemma D.9 gives the decomposition

$$D_{j+1}(t_k,s) = D'_j(t_k,s) + D_j(t_k,t_{k-j-1}) \left(\prod_{\substack{a=s+1\\a\notin S_{\le t_{k-j}}}}^{t_{k-j-1}} \widehat{P}_a^{t_{k-j}} - \prod_{\substack{a=s+1\\a\notin S_{\le t_{k-j-1}}}}^{t_{k-j-1}} \widehat{P}_a^{t_{k-j-1}} \right),$$

where

$$\begin{aligned} D'_j(t_k,s) = &\sum_{x=1}^{j} D_{j-x}(t_k,t_{k-j-1}) \left(\prod_{\substack{a=s+1\\a\notin S_{\le t_{k-j+x}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+x}} - \prod_{\substack{a=s+1\\a\notin S_{\le t_{k-j+x-1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+x-1}} \right) \\ &+ D_0(t_k,s) - D_0(t_k,t_{k-j-1}) \prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_{k-j-1}} \hat{P}_a^{t_k}. \end{aligned}$$

For the first term, the induction hypothesis gives the following by replacing t_{k-j} with t_{k-j-1}:

$$\begin{aligned} \|D'_j(t_k,s)\| \le \left(\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a^{t_k}\right) \Bigg[&\sum_{\substack{q=s+1\\q\notin S_{\le t_k}}}^{t_{k-j-1}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} + \sum_{\substack{q=t_{k-j-1}+1\\q\notin S_{\le t_k}}}^{t_{k-j+1}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} C_j^j(t_k,t_{k-j-1},s) \\ &+ \sum_{i=0}^{j-2} \sum_{\substack{q=t_{k-i-1}+1\\q\notin S_{\le t_k}}}^{t_{k-i}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} C_j^{i+1}(t_k,t_{k-j-1},s) \Bigg]. \end{aligned}$$

For the second term, Lemma D.5 gives

$$\left\| \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j-1}} \widehat{P}_a^{t_{k-j}} - \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j-1}}}}^{t_{k-j-1}} \widehat{P}_a^{t_{k-j-1}} \right\| \le \left(\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j}} \right) Z_{[s+1,t_{k-j-1}]}^{k-j\to k-j-1}.$$

Then, applying the induction hypothesis to $D_j(t_k, t_{k-j-1})$ yields

$$\begin{aligned}
&\left\| D_j(t_k, t_{k-j-1}) \left(\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j-1}} \widehat{P}_a^{t_{k-j}} - \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j-1}}}}^{t_{k-j-1}} \widehat{P}_a^{t_{k-j-1}} \right) \right\| \\
&\le \left(\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j}} \right) Z_{[s+1,t_{k-j-1}]}^{k-j\to k-j-1} \left(\prod_{\substack{a=t_{k-j-1}+1 \\ a\notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a^{t_k} \right) \left[\sum_{\substack{q=t_{k-j-1}+1 \\ q\notin S_{\le t_k}}}^{t_{k-j}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} \right. \\
&\qquad \left. + \sum_{i=0}^{j-1} \sum_{\substack{q=t_{k-i-1}+1 \\ q\notin S_{\le t_k}}}^{t_{k-i}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} C_j^{i+1}(t_k, t_{k-j}, t_{k-j-1}) \right] \\
&= \left(\prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a^{t_k} \right) \\
&\qquad \times \left[\sum_{\substack{q=t_{k-j-1}+1 \\ q\notin S_{\le t_k}}}^{t_{k-j}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} + \sum_{i=0}^{j-1} \sum_{\substack{q=t_{k-i-1}+1 \\ q\notin S_{\le t_k}}}^{t_{k-i}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} C_j^{i+1}(t_k, t_{k-j}, t_{k-j-1}) \right] \\
&\qquad \frac{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j}}}{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_k}} Z_{[s+1,t_{k-j-1}]}^{k-j\to k-j-1}.
\end{aligned}$$

Combining the two estimates, for $i = 0, \dots, j-1$, the coefficient of the block

$$\sum_{\substack{q=t_{k-i-1}+1 \\ q\notin S_{\le t_k}}}^{t_{k-i}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}$$

becomes

$$C_j^{i+1}(t_k, t_{k-j}, s) + C_j^{i+1}(t_k, t_{k-j}, t_{k-j-1}) \frac{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j}}}{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_k}} Z_{[s+1,t_{k-j-1}]}^{k-j\to k-j-1},$$

which is exactly

$$C_{j+1}^{i+1}(t_k, t_{k-j-1}, s), \qquad i = 0, \dots, j-1.$$

For the new block

$$\{t_{k-j-1}+1, \dots, t_{k-j}\},$$

the coefficient becomes

$$C_j^j(t_k, t_{k-j}, t_{k-j-1}) + \frac{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j}}}{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_k}} Z_{[s+1,t_{k-j-1}]}^{k-j\to k-j-1},$$

which is exactly

$$C_{j+1}^{j+1}(t_k, t_{k-j-1}, s).$$

Therefore,

$$\|D_{j+1}(t_k,s)\| \le \left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a^{t_k}\right)\left[\sum_{\substack{q=s+1\\ q\notin S_{\le t_k}}}^{t_{k-j-1}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} + \sum_{i=0}^{j}\sum_{\substack{q=t_{k-i-1}+1\\ q\notin S_{\le t_k}}}^{t_{k-i}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} C_{j+1}^{i+1}(t_k,t_{k-j-1},s)\right].$$

This is the desired estimate with j replaced by $j+1$. Hence the proof is complete by induction.

For $B_{j,j}$, we again prove the claim by induction on j.

First, by definition,

$$B_{0,0}(t_k,s) = \prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \widetilde{P}_a^{t_k}\widetilde{G}_s^{t_k} - \prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \widehat{P}_a^{t_k}\widehat{G}_s^{t_k}.$$

By Lemma D.4, we have

$$\|B_{0,0}(t_k,s)\| \le \left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a^{t_k}\right)\sum_{\substack{q=s+1\\ q\notin S_{\le t_k}}}^{t_k} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} \hat{g}_s^{t_k} + \left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \rho_a^{t_k}\right)\bar{g}_s^{t_k}.$$

Now consider $j=1$. By Lemma D.4, we have

$$\begin{aligned} B_{1,1}(t_k,s) =& B_{0,0}(t_k,s) - D_0(t_k,t_{k-1})b_{0,0}(t_{k-1},s) && (106)\\ =& B_{0,0}(t_k,s) - D_0(t_k,t_{k-1})\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-1}} \widehat{P}_a^{t_k}\hat{G}_s^{t_k}\\ &+ D_0(t_k,t_{k-1})\left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-1}} \widehat{P}_a^{t_k}\hat{G}_s^{t_k} - \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-1}}}}^{t_{k-1}} \widehat{P}_a^{t_{k-1}}\hat{G}_s^{t_{k-1}}\right) && (107)\\ =& \left(\prod_{\substack{a=t_{k-1}+1\\ a\notin S_{\le t_k}}}^{t_k} \widetilde{P}_a^{t_k}\right)\left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-1}} \widetilde{P}_a^{t_k}\widetilde{G}_s^{t_k} - \prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-1}} \widehat{P}_a^{t_k}\hat{G}_s^{t_k}\right)\\ &+ D_0(t_k,t_{k-1})\left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-1}} \widehat{P}_a^{t_k}\hat{G}_s^{t_k} - \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-1}}}}^{t_{k-1}} \widehat{P}_a^{t_{k-1}}\hat{G}_s^{t_{k-1}}\right). && (108) \end{aligned}$$

Taking norms gives

$$\begin{aligned} \|B_{1,1}(t_k,s)\| \le& \left\|\prod_{\substack{a=t_{k-1}+1\\ a\notin S_{\le t_k}}}^{t_k} \widetilde{P}_a^{t_k}\right\| \left\|\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-1}} \widetilde{P}_a^{t_k}\widetilde{G}_s^{t_k} - \prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-1}} \widehat{P}_a^{t_k}\hat{G}_s^{t_k}\right\|\\ &+ \|D_0(t_k,t_{k-1})\|\left\|\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_{k-1}} \widehat{P}_a^{t_k}\hat{G}_s^{t_k} - \prod_{\substack{a=s+1\\ a\notin S_{\le t_{k-1}}}}^{t_{k-1}} \widehat{P}_a^{t_{k-1}}\hat{G}_s^{t_{k-1}}\right\|. \end{aligned}$$

The first term is bounded by

$$\left(\prod_{\substack{a=t_{k-1}+1\\a\notin S_{\le t_k}}}^{t_k}\rho_a^{t_k}\right)\left(\left(\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_{k-1}}\widehat{\rho}_a^{t_k}\right)\sum_{\substack{q=s+1\\q\notin S_{\le t_k}}}^{t_{k-1}}\frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}}\prod_{\substack{b=q+1\\b\notin S_{\le t_k}}}^{t_{k-1}}\frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}\hat{g}_s^{t_k}+\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_{k-1}}\rho_a^{t_k}\bar{g}_s^{t_k}\right)$$
$$=\left(\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_k}\widehat{\rho}_a^{t_k}\right)\sum_{\substack{q=s+1\\q\notin S_{\le t_k}}}^{t_{k-1}}\frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}}\prod_{\substack{b=q+1\\b\notin S_{\le t_k}}}^{t_k}\frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}\hat{g}_s^{t_k}+\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_k}\rho_a^{t_k}\bar{g}_s^{t_k}.$$

For the second term, the bound for $D_0(t_k,t_{k-1})$ is in (105). Moreover, by Lemma D.6,

$$\left\|\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_{k-1}}\widehat{P}_a^{t_k}\hat{G}_s^{t_k}-\prod_{\substack{a=s+1\\a\notin S_{\le t_{k-1}}}}^{t_{k-1}}\widehat{P}_a^{t_{k-1}}\hat{G}_s^{t_{k-1}}\right\|\le\left(\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_{k-1}}\widehat{\rho}_a^{t_k}\right)Z_{[s+1,t_{k-1}]}^{k\to k-1}\hat{g}_s^{t_k}$$
$$+\prod_{\substack{a=s+1\\a\notin S_{\le t_{k-1}}}}^{t_{k-1}}\widehat{\rho}_a^{t_{k-1}}\zeta_s^{t_{k-1}\to t_k}.$$

Hence the second term is bounded by

$$\left(\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_k}\widehat{\rho}_a^{t_k}\right)\sum_{\substack{q=t_{k-1}+1\\q\notin S_{\le t_k}}}^{t_k}\frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}}\prod_{\substack{b=q+1\\b\notin S_{\le t_k}}}^{t_k}\frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}\left(Z_{[s+1,t_{k-1}]}^{k\to k-1}\hat{g}_s^{t_k}+\frac{\prod_{\substack{a=s+1\\a\notin S_{\le t_{k-1}}}}^{t_{k-1}}\widehat{\rho}_a^{t_{k-1}}}{\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_{k-1}}\widehat{\rho}_a^{t_k}}\zeta_s^{t_{k-1}\to t_k}\right)$$

Combining the two estimates, we obtain

$$\|B_{1,1}(t_k,s)\| \tag{109}$$
$$\le\left(\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_k}\widehat{\rho}_a^{t_k}\right)\times\left[\sum_{\substack{q=s+1\\q\notin S_{\le t_k}}}^{t_{k-1}}\frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}}\prod_{\substack{b=q+1\\b\notin S_{\le t_k}}}^{t_k}\frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}\hat{g}_s^{t_k}\right.$$
$$\left.+\sum_{\substack{q=t_{k-1}+1\\q\notin S_{\le t_k}}}^{t_k}\frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}}\prod_{\substack{b=q+1\\b\notin S_{\le t_k}}}^{t_k}\frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}\left(Z_{[s+1,t_{k-1}]}^{k\to k-1}\hat{g}_s^{t_k}+\frac{\prod_{\substack{a=s+1\\a\notin S_{\le t_{k-1}}}}^{t_{k-1}}\widehat{\rho}_a^{t_{k-1}}}{\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_{k-1}}\widehat{\rho}_a^{t_k}}\zeta_s^{t_{k-1}\to t_k}\right)\right]$$
$$+\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_k}\rho_a^{t_k}\bar{g}_s^{t_k}. \tag{110}$$

Since

$$\hat{C}_1^1(t_k,t_{k-1},s)=Z_{[s+1,t_{k-1}]}^{k\to k-1}\hat{g}_s^{t_k}+\frac{\prod_{\substack{a=s+1\\a\notin S_{\le t_{k-1}}}}^{t_{k-1}}\widehat{\rho}_a^{t_{k-1}}}{\prod_{\substack{a=s+1\\a\notin S_{\le t_k}}}^{t_{k-1}}\widehat{\rho}_a^{t_k}}\zeta_s^{t_{k-1}\to t_k},$$

this is exactly the desired bound for $j=1$.

Now assume the estimate holds for some $j \geq 1$. We prove it for $j+1$. We use the decomposition

$$B_{j+1,j+1}(t_k, s) = B'_{j,j}(t_k, s) + D_j(t_k, t_{k-j-1})\Bigg(\prod_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j}}}}^{t_{k-j-1}} \widehat{P}_a^{t_{k-j}} \hat{G}_s^{t_{k-j}} - \prod_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j-1}}}}^{t_{k-j-1}} \widehat{P}_a^{t_{k-j-1}} \hat{G}_s^{t_{k-j-1}} \Bigg),$$

where

$$\begin{aligned} B'_{j,j}(t_k, s) = &\sum_{x=1}^{j} D_{j-x}(t_k, t_{k-j-1}) \\ &\times \Bigg(\prod_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j+x}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+x}} \hat{G}_s^{t_{k-j+x}} - \prod_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j+x-1}}}}^{t_{k-j-1}} \hat{P}_a^{t_{k-j+x-1}} \hat{G}_s^{t_{k-j+x-1}} \Bigg) \\ &+ B_{0,0}(t_k, s) \\ &- D_0(t_k, t_{k-j-1}) \prod_{\substack{a=s+1 \\ a \notin S_{\leq t_k}}}^{t_{k-j-1}} \hat{P}_a^{t_k} \hat{G}_s^{t_k}. \end{aligned}$$

For the first term, the induction hypothesis gives the following bound after replacing t_{k-j} with t_{k-j-1}:

$$\begin{aligned} \|B'_{j,j}(t_k, s)\| \leq &\left(\prod_{\substack{a=s+1 \\ a \notin S_{\leq t_k}}}^{t_k} \widehat{\rho}_a^{t_k} \right) \Bigg[\sum_{\substack{q=s+1 \\ q \notin S_{\leq t_k}}}^{t_{k-j-1}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b \notin S_{\leq t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} \hat{g}_s^{t_k} \\ &+ \sum_{\substack{q=t_{k-j-1}+1 \\ q \notin S_{\leq t_k}}}^{t_{k-j}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b \notin S_{\leq t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} \hat{C}_j^{j}(t_k, t_{k-j-1}, s) \\ &+ \sum_{i=0}^{j-1} \sum_{\substack{q=t_{k-i-1}+1 \\ q \notin S_{\leq t_k}}}^{t_{k-i}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b \notin S_{\leq t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} \hat{C}_j^{i+1}(t_k, t_{k-j-1}, s) \Bigg] \\ &+ \prod_{\substack{a=s+1 \\ a \notin S_{\leq t_k}}}^{t_k} \rho_a^{t_k} \bar{g}_s^{t_k}. \end{aligned}$$

For the second term, Lemma D.6 gives the following bound:

$$\left\| \prod_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j}}}}^{t_{k-j-1}} \widehat{P}_a^{t_{k-j}} \hat{G}_s^{t_{k-j}} - \prod_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j-1}}}}^{t_{k-j-1}} \widehat{P}_a^{t_{k-j-1}} \hat{G}_s^{t_{k-j-1}} \right\| \tag{111}$$

$$\leq \left(\prod_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j}} \right) \left(Z^{k-j \to k-j-1}_{[s+1, t_{k-j-1}]} \hat{g}_s^{t_{k-j}} + \frac{\prod_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j-1}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j-1}}}{\prod_{\substack{a=s+1 \\ a \notin S_{\leq t_{k-j}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j}}} \zeta_s^{t_{k-j-1} \to t_{k-j}} \right). \tag{112}$$

Applying the proven results on $D_j(t_k, t_{k-j-1})$, and then multiplying the two bounds, yields

$$\left\| D_j(t_k, t_{k-j-1}) \left(\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j-1}} \widehat{P}_a^{t_{k-j}} \hat{G}_s^{t_{k-j}} - \prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j-1}}}}^{t_{k-j-1}} \widehat{P}_a^{t_{k-j-1}} \hat{G}_s^{t_{k-j-1}} \right) \right\|$$

$$\le \left(\prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a \right) \frac{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j}}}{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_k}} \left(Z^{k-j\to k-j-1}_{[s+1,t_{k-j-1}]} \hat{g}_s^{t_{k-j}} + \frac{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j-1}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j-1}}}{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j}}} \zeta_s^{t_{k-j-1}\to t_{k-j}} \right)$$

$$\left[\sum_{\substack{q=t_{k-j-1}+1 \\ q\notin S_{\le t_k}}}^{t_{k-j}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} + \sum_{i=0}^{j-1} \sum_{\substack{q=t_{k-i-1}+1 \\ q\notin S_{\le t_k}}}^{t_{k-i}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}} C_j^{i+1}(t_k, t_{k-j}, t_{k-j-1}) \right].$$

Combining the two estimates, the coefficient of the following block with $i = 0, \dots, j-1$

$$\sum_{\substack{q=t_{k-i-1}+1 \\ q\notin S_{\le t_k}}}^{t_{k-i}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1 \\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}$$

becomes

$$\begin{aligned}
&\hat{C}_j^{i+1}(t_k, t_{k-j}, s) \\
&+ C_j^{i+1}(t_k, t_{k-j}, t_{k-j-1}) \Bigg(\\
&\quad \frac{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j}}}{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_k}} Z^{k-j\to k-j-1}_{[s+1,t_{k-j-1}]} \hat{g}_s^{t_{k-j}} \\
&\quad + \frac{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j-1}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j-1}}}{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_k}} \zeta_s^{t_{k-j-1}\to t_{k-j}} \Bigg),
\end{aligned}$$

which is exactly

$$\hat{C}_{j+1}^{i+1}(t_k, t_{k-j-1}, s), \qquad i = 0, \dots, j-1.$$

For the new block

$$\{t_{k-j-1}+1, \dots, t_{k-j}\},$$

the coefficient becomes

$$\begin{aligned}
&\hat{C}_j^{j}(t_k, t_{k-j}, t_{k-j-1}) \\
&+ \Bigg(\frac{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j}}}{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_k}} Z^{k-j\to k-j-1}_{[s+1,t_{k-j-1}]} \hat{g}_s^{t_{k-j}} \\
&\quad + \frac{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_{k-j-1}}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_{k-j-1}}}{\prod_{\substack{a=s+1 \\ a\notin S_{\le t_k}}}^{t_{k-j-1}} \widehat{\rho}_a^{t_k}} \zeta_s^{t_{k-j-1}\to t_{k-j}} \Bigg),
\end{aligned}$$

which is exactly

$$\hat{C}_{j+1}^{j+1}(t_k, t_{k-j-1}, s).$$

Therefore,

$$\begin{aligned}
&\|B_{j+1,j+1}(t_k,s)\| \\
&\le \left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \widehat{\rho}_a^{t_k}\right)\Bigg[\sum_{\substack{q=s+1\\ q\notin S_{\le t_k}}}^{t_{k-j-1}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}\hat{g}_s^{t_k} \\
&\quad + \sum_{i=0}^{j}\sum_{\substack{q=t_{k-i-1}+1\\ q\notin S_{\le t_k}}}^{t_{k-i}} \frac{\bar{\rho}_q^{t_k}}{\widehat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\ b\notin S_{\le t_k}}}^{t_k} \frac{\rho_b^{t_k}}{\widehat{\rho}_b^{t_k}}\hat{C}_{j+1}^{i+1}(t_k,t_{k-j-1},s)\Bigg] \\
&\quad + \prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \rho_a^{t_k}\bar{g}_s^{t_k}.
\end{aligned}$$

This is the desired estimate with j replaced by $j+1$. Hence the proof of Lemma D.10 is complete by induction.

For Corollary D.11, first observe from the definitions of D_0 and $B_{0,0}$ that

$$\begin{aligned}
&D_0(t_k,s)\Delta_s - B_{0,0}(t_k,s) \\
&= \left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \widetilde{P}_a^{t_k} - \prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \widehat{P}_a^{t_k}\right)(\Delta_s - \widehat{G}_s^{t_k}) \\
&\quad + \left(\prod_{\substack{a=s+1\\ a\notin S_{\le t_k}}}^{t_k} \widetilde{P}_a^{t_k}\right)(\widehat{G}_s^{t_k} - \widetilde{G}_s^{t_k}).
\end{aligned}$$

Thus, the product-perturbation bound used above applies to the first term with $\hat{g}_s^{t_k}$ replaced by $\hat{\kappa}_s^{t_k}$, while the second term remains bounded by $\big(\prod_a \rho_a^{t_k}\big)\bar{g}_s^{t_k}$. At successive request times, the same induction applies because

$$\left\|\Delta_s - \widehat{G}_s^{\tau}\right\| \le \hat{\kappa}_s^{\tau}, \qquad \left\|(\Delta_s - \widehat{G}_s^{t_i}) - (\Delta_s - \widehat{G}_s^{t_j})\right\| \le \zeta_s^{t_i\to t_j}.$$

Consequently, replacing $\hat{C}_j^i$ by $\tilde{C}_j^i$ in the preceding induction gives exactly $\mathbf{DB}_j(t_k,s)$, proving the corollary. □

Lemma D.12. *Suppose that $k-i>0$ and $k-i+1 \le m \le k-1$. For any $s \in I_{m,m-k+i-1}^k$, we have*

$$B_{m,m-k+i-1}(t_k,s) = -\sum_{n=k-i+1}^{m} D_{k-i}(t_k,t_{k-n})\, b_{m-n,m-n}(t_{k-n},s). \tag{113}$$

Proof. Let

$$j := m-k+i-1.$$

By the definition of the partition sets,

$$I_{m,j}^k = \big([\underline{s}_{k-m}+1, \underline{s}_{k-m-1}-1] \cup [t_{k-m-1}+1, t_{k-m}]\big) \cap S_{t_i}.$$

Thus, before the unlearning time t_i, the elements in $I_{m,j}^k$ are still retained; from time t_i to time t_k, they have been removed. Therefore, using the Hessian update in (14), we have the initial identity

$$B_{m,j}(t_k,s) = -\sum_{h=1}^{k-i} D_0(t_k,t_{k-h})\, b_{m-h,j}(t_{k-h},s) - \sum_{h=k-i+1}^{m} D_0(t_k,t_{k-h})\, b_{m-h,m-h}(t_{k-h},s). \tag{114}$$

We prove by induction on $\ell = 0, \dots, k-i-1$ that

$$B_{m,j}(t_k, s) = -\sum_{h=\ell+1}^{k-i} D_\ell(t_k, t_{k-h})\, b_{m-h,j}(t_{k-h}, s) - \sum_{h=k-i+1}^{m} D_\ell(t_k, t_{k-h})\, b_{m-h,m-h}(t_{k-h}, s). \tag{115}$$

The case $\ell = 0$ is exactly (114).

Assume that (115) holds for some $\ell < k-i-1$. Isolating the first term in the first summation gives

$$B_{m,j}(t_k, s) = -D_\ell(t_k, t_{k-\ell-1})\, b_{m-\ell-1,j}(t_{k-\ell-1}, s) - \sum_{h=\ell+2}^{k-i} D_\ell(t_k, t_{k-h})\, b_{m-h,j}(t_{k-h}, s) - \sum_{h=k-i+1}^{m} D_\ell(t_k, t_{k-h})\, b_{m-h,m-h}(t_{k-h}, s).$$

By the corresponding recursion for the b-terms at time $t_{k-\ell-1}$,

$$b_{m-\ell-1,j}(t_{k-\ell-1}, s) = -\sum_{r=1}^{k-i-\ell-1} d_0(t_{k-\ell-1}, t_{k-\ell-1-r})\, b_{m-\ell-1-r,j}(t_{k-\ell-1-r}, s) - \sum_{r=k-i-\ell}^{m-\ell-1} d_0(t_{k-\ell-1}, t_{k-\ell-1-r})\, b_{m-\ell-1-r,m-\ell-1-r}(t_{k-\ell-1-r}, s).$$

Substituting this relation and reindexing with $h = \ell + 1 + r$, we obtain

$$B_{m,j}(t_k, s) = -\sum_{h=\ell+2}^{k-i} \left[D_\ell(t_k, t_{k-h}) - D_\ell(t_k, t_{k-\ell-1}) d_0(t_{k-\ell-1}, t_{k-h})\right] b_{m-h,j}(t_{k-h}, s) - \sum_{h=k-i+1}^{m} \left[D_\ell(t_k, t_{k-h}) - D_\ell(t_k, t_{k-\ell-1}) d_0(t_{k-\ell-1}, t_{k-h})\right] b_{m-h,m-h}(t_{k-h}, s).$$

By Lemma D.8,

$$D_{\ell+1}(t_k, t_{k-h}) = D_\ell(t_k, t_{k-h}) - D_\ell(t_k, t_{k-\ell-1}) d_0(t_{k-\ell-1}, t_{k-h}).$$

Hence (115) holds with ℓ replaced by $\ell + 1$. This proves the induction claim.

Taking $\ell = k-i-1$ in (115) gives

$$B_{m,j}(t_k, s) = -D_{k-i-1}(t_k, t_i)\, b_{m-k+i,j}(t_i, s) - \sum_{h=k-i+1}^{m} D_{k-i-1}(t_k, t_{k-h})\, b_{m-h,m-h}(t_{k-h}, s).$$

At time t_i, the remaining off-diagonal term satisfies

$$b_{m-k+i,j}(t_i, s) = -\sum_{r=1}^{m-k+i} d_0(t_i, t_{i-r})\, b_{m-k+i-r,m-k+i-r}(t_{i-r}, s).$$

Substituting this identity and reindexing by $h = k - i + r$, we get

$$\begin{aligned} B_{m,j}(t_k, s) &= \sum_{h=k-i+1}^{m} D_{k-i-1}(t_k, t_i) d_0(t_i, t_{k-h})\, b_{m-h,m-h}(t_{k-h}, s) - \sum_{h=k-i+1}^{m} D_{k-i-1}(t_k, t_{k-h})\, b_{m-h,m-h}(t_{k-h}, s) \\ &= -\sum_{h=k-i+1}^{m} \left[D_{k-i-1}(t_k, t_{k-h}) - D_{k-i-1}(t_k, t_i) d_0(t_i, t_{k-h})\right] b_{m-h,m-h}(t_{k-h}, s). \end{aligned}$$

Using Lemma D.8 one more time, the bracketed term is $D_{k-i}(t_k, t_{k-h})$. Therefore,

$$B_{m,j}(t_k, s) = -\sum_{h=k-i+1}^{m} D_{k-i}(t_k, t_{k-h})\, b_{m-h,m-h}(t_{k-h}, s).$$

Since $j = m-k+i-1$, this is exactly (113). □

Lemma D.13. *Suppose that $k-i>0$ and $k-i+1 \le m \le k-1$. For any $s \in I^k_{m,m-k+i-1}$, we have*

$$B_{m,m-k+i-1}(t_k,s) = -\sum_{n=k-i+1}^{m} D_{n-1}(t_k,t_{k-n})\, b_{0,0}(t_{k-n},s) \tag{116}$$
$$= -\sum_{r=0}^{m-k+i-1} D_{k-i+r}(t_k,t_{i-r-1})\, b_{0,0}(t_{i-r-1},s).$$

Proof. Let

$$j := m-k+i-1.$$

By Lemma D.12,

$$B_{m,j}(t_k,s) = -\sum_{n=k-i+1}^{m} D_{k-i}(t_k,t_{k-n})\, b_{m-n,m-n}(t_{k-n},s). \tag{117}$$

We now expand the diagonal b-terms one at a time. We claim that for $\ell = 0,\dots,m-k+i$,

$$B_{m,j}(t_k,s) = -\sum_{r=0}^{\ell-1} D_{k-i+r}(t_k,t_{i-r-1})\, b_{0,0}(t_{i-r-1},s)$$
$$-\sum_{n=k-i+\ell+1}^{m} D_{k-i+\ell}(t_k,t_{k-n})\, b_{m-n,m-n}(t_{k-n},s), \tag{118}$$

where the first summation is empty when $\ell=0$. The case $\ell=0$ is exactly (117).

Assume that (118) holds for some $\ell < m-k+i$. Isolating the first term in the second summation gives

$$B_{m,j}(t_k,s) = -\sum_{r=0}^{\ell-1} D_{k-i+r}(t_k,t_{i-r-1})\, b_{0,0}(t_{i-r-1},s)$$
$$- D_{k-i+\ell}(t_k,t_{i-\ell-1})\, b_{m-k+i-\ell-1,m-k+i-\ell-1}(t_{i-\ell-1},s)$$
$$-\sum_{n=k-i+\ell+2}^{m} D_{k-i+\ell}(t_k,t_{k-n})\, b_{m-n,m-n}(t_{k-n},s).$$

Using the diagonal recursion for b at time $t_{i-\ell-1}$,

$$b_{m-k+i-\ell-1,m-k+i-\ell-1}(t_{i-\ell-1},s) = b_{0,0}(t_{i-\ell-1},s) - \sum_{q=1}^{m-k+i-\ell-1} \begin{matrix} d_0(t_{i-\ell-1},t_{i-\ell-1-q}) \\ \cdot\, b_{m-k+i-\ell-1-q,m-k+i-\ell-1-q}(t_{i-\ell-1-q},s) \end{matrix}.$$

Substituting this relation and reindexing the remaining terms with $n = k-i+\ell+1+q$, we obtain

$$B_{m,j}(t_k,s) = -\sum_{r=0}^{\ell} D_{k-i+r}(t_k,t_{i-r-1})\, b_{0,0}(t_{i-r-1},s)$$
$$-\sum_{n=k-i+\ell+2}^{m} [D_{k-i+\ell}(t_k,t_{k-n}) - D_{k-i+\ell}(t_k,t_{i-\ell-1}) d_0(t_{i-\ell-1},t_{k-n})]$$
$$\times\, b_{m-n,m-n}(t_{k-n},s).$$

By Lemma D.8,

$$D_{k-i+\ell+1}(t_k,t_{k-n}) = D_{k-i+\ell}(t_k,t_{k-n}) - D_{k-i+\ell}(t_k,t_{i-\ell-1}) d_0(t_{i-\ell-1},t_{k-n}).$$

Hence (118) holds with ℓ replaced by $\ell+1$. This proves the induction claim.

Taking $\ell = m-k+i$ in (118), the second summation is empty, and therefore

$$B_{m,j}(t_k,s) = -\sum_{r=0}^{m-k+i-1} D_{k-i+r}(t_k,t_{i-r-1})\, b_{0,0}(t_{i-r-1},s).$$

Finally, using the change of variables $n = k-i+1+r$, this is equivalently

$$B_{m,j}(t_k,s) = -\sum_{n=k-i+1}^{m} D_{n-1}(t_k,t_{k-n})\, b_{0,0}(t_{k-n},s).$$

Since $j = m - k + i - 1$, the proof is complete. □

E FORGETTING-ENHANCED HESSIAN UNLEARNING

The forgetting-enhanced Hessian update is defined as follows. At the first unlearning time t_1, set $t_- := \min S_{t_1}$; at every subsequent unlearning time t, set

$$t_- := \max\big(U_t \setminus \{t\}\big),$$

In both cases, define $S'_t := \{s \in S_t : s > t_-\}$. The forgetting-enhanced update is

$$w_t^{-\mathbf{S}_{1:t}} = w_t^{-\mathbf{S}_{1:t-1}} + \sum_{s\in S'_t} \left(\prod_{\substack{j=s+1\\ j\notin S_{\le t}}}^{t} \hat{P}_j^t\right) \Delta_s + \sum_{\substack{n=t_-+1\\ n\notin S_{\le t}}}^{t} \left(\prod_{\substack{j=n+1\\ j\notin S_{\le t}}}^{t} \hat{P}_j^t\right) \hat{G}_n^t. \tag{119}$$

After this update, the system discards the stored Hessian information and learning increments for all tasks preceding t.

Theorem E.1. *At time t, let $U_t = \{t_1, \dots, t_k\}$ be the ordered set of unlearning times up to t, where $k \ge 1$. Set $t_0 := \underline{s}_1$, $S'_{\le t_k} := \bigcup_{i=1}^k S'_{t_i}$, and $S'_{t_i} := \{s \in S_{t_i} : s > t_{i-1}\}$, with the conventions $Z_{[a,b]}^{k\to k} := 0$ and $\zeta_n^{t_k\to t_k} := 0$. Then*

$$\left\| w_t^{-S_{\le t}} - w_t^{-\mathbf{S}_{1:t}} \right\| \le \left(\prod_{j=t_k+1}^{t} \rho_j^{t_k}\right) \gamma_{\mathrm{EH}}(\mathbf{S}_{1:t_k}) + \sum_{i=t_k+1}^{t} \left(\prod_{a=i+1}^{t} \rho_a^{t_k}\right) g_i^{t_k}, \tag{120}$$

where $\gamma_{\mathrm{EH}}(\mathbf{S}_{1:t_k})$ equals

$$\begin{aligned}
\gamma_{\mathrm{EH}}(\mathbf{S}_{1:t_k}) := {} & \sum_{i=1}^{k} \sum_{s\in S'_{t_i}} \left(\prod_{\substack{j=t_i+1\\ j\notin S_{\le t_k}}}^{t_k} \rho_j^{t_k}\right) \left(\prod_{\substack{c=s+1\\ c\notin S_{\le t_k}}}^{t_i} \hat{\rho}_c^{t_k}\right) \\
& \times \left[Z_{[s+1,t_i]}^{k\to i} + \sum_{\substack{q=s+1\\ q\notin S_{\le t_k}}}^{t_i} \frac{\bar{\rho}_q^{t_k}}{\hat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\ b\notin S_{\le t_k}}}^{t_i} \frac{\rho_b^{t_k}}{\hat{\rho}_b^{t_k}} \right] \Gamma_s \\
& + \sum_{s\in S_{\le t_k}\setminus S'_{\le t_k}} \left(\prod_{\substack{j=s+1\\ j\notin S_{\le t_k}}}^{t_k} \rho_j^{t_k}\right) \Gamma_s \\
& + \sum_{i=1}^{k} \sum_{\substack{n=t_{i-1}+1\\ n\notin S_{\le t_k}}}^{t_i} \left(\prod_{\substack{j=t_i+1\\ j\notin S_{\le t_k}}}^{t_k} \rho_j^{t_k}\right) \\
& \times \Bigg[\left(\prod_{\substack{c=n+1\\ c\notin S_{\le t_k}}}^{t_i} \hat{\rho}_c^{t_k}\right) \left(Z_{[n+1,t_i]}^{k\to i} + \sum_{\substack{q=n+1\\ q\notin S_{\le t_k}}}^{t_i} \frac{\bar{\rho}_q^{t_k}}{\hat{\rho}_q^{t_k}} \prod_{\substack{b=q+1\\ b\notin S_{\le t_k}}}^{t_i} \frac{\rho_b^{t_k}}{\hat{\rho}_b^{t_k}} \right) \hat{g}_n^{t_k} \\
& + \left(\prod_{\substack{c=n+1\\ c\notin S_{\le t_k}}}^{t_i} \rho_c^{t_k}\right) \bar{g}_n^{t_k} + \left(\prod_{\substack{c=n+1\\ c\notin S_{\le t_i}}}^{t_i} \hat{\rho}_c^{t_i}\right) \zeta_n^{t_i\to t_k} \Bigg] \\
& + \sum_{\substack{n=\underline{s}_k+1\\ n\notin S_{\le t_k}}}^{t_0} \left(\prod_{\substack{j=n+1\\ j\notin S_{\le t_k}}}^{t_k} \rho_j^{t_k}\right) g_n^{t_k} \\
& + \sum_{i=1}^{k-1} \sum_{\substack{n=t_{i-1}+1\\ n\notin S_{\le t_i},\, n\in S_{\le t_k}}}^{t_i} \left(\prod_{\substack{j=t_i+1\\ j\notin S_{\le t_k}}}^{t_k} \rho_j^{t_k}\right) \left(\prod_{\substack{j=n+1\\ j\notin S_{\le t_i}}}^{t_i} \hat{\rho}_j^{t_i}\right) \hat{g}_n^{t_i}.
\end{aligned} \tag{121}$$

Proof of Theorem E.1. Firstly, $\gamma_{\mathrm{EH}}(\mathbf{S}_{1:t_k})$ is defined as the upper bound of $\left\| w_{t_k}^{-S_{\le t_k}} - w_{t_k}^{-\mathbf{S}_{1:t_k}} \right\|$ under forgetting-enhanced Hessian. Applying the triangle inequality and submultiplicativity to the first identity in Lemma D.7, and then using the definitions of $\rho_j^{t_k}$ and $g_i^{t_k}$, gives (120). We next prove for $\gamma_{\mathrm{EH}}(\mathbf{S}_{1:t_k})$. Lemma D.7 gives

$$\begin{aligned} & w_{t_k}^{-S_{\le t_k}} - w_{t_k}^{-\mathbf{S}_{1:t_k}} \\ &= \sum_{s \in S_{\le t_k}} \left(\prod_{\substack{j=s+1 \\ j \notin S_{\le t_k}}}^{t_k} \widetilde{P}_j^{t_k} \right) \Delta_s + \sum_{\substack{n=\underline{s}_k+1 \\ n \notin S_{\le t_k}}}^{t_k} \left(\prod_{\substack{j=n+1 \\ j \notin S_{\le t_k}}}^{t_k} \widetilde{P}_j^{t_k} \right) \widetilde{G}_n^{t_k} - \sum_{i=1}^{k} \left(\prod_{\substack{j=t_i+1 \\ j \notin S_{\le t_k}}}^{t_k} \widetilde{P}_j^{t_k} \right) \bar{\Delta}_{t_i}. \end{aligned} \tag{122}$$

At each unlearning time t_i, the forgetting-enhanced correction is

$$\bar{\Delta}_{t_i} = \sum_{s \in S'_{t_i}} \left(\prod_{\substack{j=s+1 \\ j \notin S_{\le t_i}}}^{t_i} \widehat{P}_j^{t_i} \right) \Delta_s + \sum_{\substack{n=t_{i-1}+1 \\ n \notin S_{\le t_i}}}^{t_i} \left(\prod_{\substack{j=n+1 \\ j \notin S_{\le t_i}}}^{t_i} \widehat{P}_j^{t_i} \right) \widehat{G}_n^{t_i}. \tag{123}$$

Substituting (123) into the preceding identity and grouping the deleted-task and regularizer-mismatch terms yield

$$\begin{aligned} & w_{t_k}^{-S_{\le t_k}} - w_{t_k}^{-\mathbf{S}_{1:t_k}} \\ &= \sum_{i=1}^{k} \sum_{s \in S'_{t_i}} \left(\prod_{\substack{j=t_i+1 \\ j \notin S_{\le t_k}}}^{t_k} \widetilde{P}_j^{t_k} \right) \left[\prod_{\substack{j=s+1 \\ j \notin S_{\le t_k}}}^{t_i} \widetilde{P}_j^{t_k} - \prod_{\substack{j=s+1 \\ j \notin S_{\le t_i}}}^{t_i} \widehat{P}_j^{t_i} \right] \Delta_s + \sum_{s \in S_{\le t_k} \setminus S'_{\le t_k}} \left(\prod_{\substack{j=s+1 \\ j \notin S_{\le t_k}}}^{t_k} \widetilde{P}_j^{t_k} \right) \Delta_s \\ &\quad + \sum_{i=1}^{k} \sum_{\substack{n=t_{i-1}+1 \\ n \notin S_{\le t_k}}}^{t_i} \left(\prod_{\substack{j=t_i+1 \\ j \notin S_{\le t_k}}}^{t_k} \widetilde{P}_j^{t_k} \right) \left[\left(\prod_{\substack{j=n+1 \\ j \notin S_{\le t_k}}}^{t_i} \widetilde{P}_j^{t_k} \right) \widetilde{G}_n^{t_k} - \left(\prod_{\substack{j=n+1 \\ j \notin S_{\le t_i}}}^{t_i} \widehat{P}_j^{t_i} \right) \widehat{G}_n^{t_i} \right] \\ &\quad + \sum_{\substack{n=\underline{s}_k+1 \\ n \notin S_{\le t_k}}}^{t_0} \left(\prod_{\substack{j=n+1 \\ j \notin S_{\le t_k}}}^{t_k} \widetilde{P}_j^{t_k} \right) \widetilde{G}_n^{t_k} - \sum_{i=1}^{k-1} \sum_{\substack{n=t_{i-1}+1 \\ n \notin S_{\le t_i},\, n \in S_{\le t_k}}}^{t_i} \left(\prod_{\substack{j=t_i+1 \\ j \notin S_{\le t_k}}}^{t_k} \widetilde{P}_j^{t_k} \right) \left(\prod_{\substack{j=n+1 \\ j \notin S_{\le t_i}}}^{t_i} \widehat{P}_j^{t_i} \right) \widehat{G}_n^{t_i}. \end{aligned} \tag{124}$$

Taking norms in this decomposition, we treat its two product differences separately. For each $s \in S'_{t_i}$, insert and subtract

$$\prod_{\substack{j=s+1 \\ j \notin S_{\le t_k}}}^{t_i} \widehat{P}_j^{t_k}.$$

Lemma D.3 bounds the difference between the $\widetilde{P}^{t_k}$- and $\widehat{P}^{t_k}$-products, while Lemma D.5 bounds the difference between the $\widehat{P}^{t_k}$- and $\widehat{P}^{t_i}$-products. Together with $\|\Delta_s\| \le \Gamma_s$, these give the first summation in (121).

Similarly, for each retained $n \in [t_{i-1}+1, t_i]$, insert and subtract

$$\left(\prod_{\substack{j=n+1 \\ j \notin S_{\le t_k}}}^{t_i} \widehat{P}_j^{t_k} \right) \widehat{G}_n^{t_k}.$$

Lemma D.4 controls the resulting $(\widetilde{P}^{t_k}, \widetilde{G}^{t_k})$-to- $(\widehat{P}^{t_k}, \widehat{G}^{t_k})$ difference, and Lemma D.6 controls the remaining request-time asymmetry. The other terms follow directly from the definitions of $\rho_n^{t_k}$, $g_n^{t_k}$, $\hat{\rho}_n^{t_i}$, $\hat{g}_n^{t_i}$, and Γ_s. Applying the triangle inequality therefore yields

$$\left\| w_{t_k}^{-S_{\le t_k}} - w_{t_k}^{-\mathbf{S}_{1:t_k}} \right\| \le \gamma_{\mathrm{EH}}(\mathbf{S}_{1:t_k}). \tag{125}$$

□

F Operator Bounds for Task-Additive Reg-CL

Throughout this section, we consider the task-additive regularizer in (11) under Assumptions 2.3–4.2.

Lemma F.1. *For* $\Gamma_n = \sup \|\Delta_n\|$,

$$\Gamma_n \le \frac{L}{\lambda + N_{S_{\le n-1}}(n-1)\mu_\Omega}. \tag{126}$$

Proof. By the first order condition. □

Lemma F.2. *For* $\rho_n^t = \sup \left\|\widetilde{P}_n^t\right\|$,

$$\rho_n^t \le \frac{N_{S_{\le t}}(n-1)M_\Omega + \lambda}{\mu + \lambda + N_{S_{\le t}}(n-1)\mu_\Omega}. \tag{127}$$

Proof.

$$\widetilde{P}_n^t = \left(\tilde{H}_n + \lambda I + \sum_{\substack{i=1 \\ i\notin S_{\le t}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i)\right)^{-1} \left(\lambda I + \sum_{\substack{i=1 \\ i\notin S_{\le t}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i)\right)$$

□

Lemma F.3. *For* $\bar{\rho}_n^t = \sup \left\|\widetilde{P}_n^t - \hat{P}_n^t\right\|$, *we have*

$$\begin{aligned}\bar{\rho}_n^t \le{}& \frac{N_{S_{\le t}}([\underline{s}^t, n-1])(M_\Omega - \mu_\Omega)}{\mu + \lambda + N_{S_{\le t}}(n-1)\mu_\Omega} \\ &+ \frac{N_{S_{\le t}}(n-1)M_\Omega + \lambda}{\big(\mu + \lambda + N_{S_{\le t}}(n-1)\mu_\Omega\big)} \\ &\cdot\big(\mu - \nu + \lambda + N_{S_{\le t}}(n-1)\mu_\Omega\big) \\ &\times \Big[M - \mu + \nu + N_{S_{\le t}}([\underline{s}^t, n-1])(M_\Omega - \mu_\Omega)\Big],\end{aligned} \tag{128}$$

$$\tag{129}$$

Proof. Write

$$\begin{aligned}\widetilde{P}_n^t - \hat{P}_n^t = {}& \left(\widetilde{H}_n + \lambda I + \sum_{\substack{i=1 \\ i\notin S_{\le t}}}^{n-1} F(w_i^{-S_{\le t}}, D_i)\right)^{-1} \left(\lambda I + \sum_{\substack{i=1 \\ i\notin S_{\le t}}}^{n-1} F(w_i^{-S_{\le t}}, D_i)\right) \\ &- \left(\hat{H}_n + \lambda I + \sum_{\substack{i=1 \\ i\notin S_{\le t}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i)\right)^{-1} \left(\lambda I + \sum_{\substack{i=1 \\ i\notin S_{\le t}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i)\right).\end{aligned} \tag{130}$$

□

Lemma F.4. *For* $\hat{\rho}_n^t = \sup\{\left\|\hat{P}_n^t\right\|\}$,

$$\hat{\rho}_n^t \le \frac{N_{S_{\le t}}(n-1)M_\Omega + \lambda}{\mu - \nu + \lambda + N_{S_{\le t}}(n-1)\mu_\Omega}. \tag{131}$$

Proof.

$$\hat{P}_n^t = \left(\hat{H}_n + \lambda I + \sum_{\substack{i=1 \\ i \notin S_{\leq t}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) \right)^{-1} \left(\lambda I + \sum_{\substack{i=1 \\ i \notin S_{\leq t}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) \right)$$

□

Lemma F.5. *For* $g_n^t = \sup\{\left\|\tilde{G}_n^t\right\|\}$,

$$g_n^t \leq \frac{(S_{\leq t}(n-1) - S_{\leq n-1}(n-1))M_\Omega + N_{S_{\leq t}}([\underline{s}^t, n-1])(M_\Omega - \mu_\Omega)}{\mu + \lambda + N_{S_{\leq t}}(n-1)\mu_\Omega} \frac{L}{\lambda + N_{S_{\leq n-1}}(n-1)\mu_\Omega}. \tag{132}$$

Proof.

$$\widetilde{G}_n^t = \left(\widetilde{H}_n + \lambda I + \sum_{\substack{i=1 \\ i \notin S_{\leq t}}}^{n-1} F(w_i^{-S_{\leq t}}, D_i) \right)^{-1} \left[\sum_{\substack{i=1 \\ i \notin S_{\leq n-1}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) - \sum_{\substack{i=1 \\ i \notin S_{\leq t}}}^{n-1} F(w_i^{-S_{\leq t}}, D_i) \right] \times (w_n^{-\mathbf{S}_{1:n-1}} - w_{n-1}^{-\mathbf{S}_{1:n-1}})$$

□

Lemma F.6. *For* $\hat{g}_n^t = \sup \left\|\hat{G}_n^t\right\|$,

$$\hat{g}_n^t \leq \frac{(S_{\leq t}(n-1) - S_{\leq n-1}(n-1))M_\Omega}{\mu - \nu + \lambda + N_{S_{\leq t}}(n-1)\mu_\Omega} \frac{L}{\lambda + N_{S_{\leq n-1}}(n-1)\mu_\Omega}. \tag{133}$$

Proof.

$$\hat{G}_n^t = \left(\hat{H}_n + \lambda I + \sum_{\substack{i=1 \\ i \notin S_{\leq t}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) \right)^{-1} \times \left[\sum_{\substack{i=1 \\ i \notin S_{\leq n-1}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) - \sum_{\substack{i=1 \\ i \notin S_{\leq t}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) \right] (w_n^{-\mathbf{S}_{1:n-1}} - w_{n-1}^{-\mathbf{S}_{1:n-1}})$$

□

Lemma F.7. *For* $\bar{g}_n^t = \sup\{\left\|\tilde{G}_n^t - \hat{G}_n^t\right\|\}$,

$$\begin{aligned} \bar{g}_n^t \leq \Bigg[& \frac{N_{S_{\leq t}}([\underline{s}^t, n-1])(M_\Omega - \mu_\Omega)}{\mu + \lambda + N_{S_{\leq t}}(n-1)\mu_\Omega} \\ & + \frac{(S_{\leq t}(n-1) - S_{\leq n-1}(n-1))M_\Omega}{\mu + \lambda + N_{S_{\leq t}}(n-1)\mu_\Omega} \\ & \times \frac{M - \mu + \nu + N_{S_{\leq t}}([\underline{s}^t, n-1])(M_\Omega - \mu_\Omega)}{\mu - \nu + \lambda + N_{S_{\leq t}}(n-1)\mu_\Omega} \Bigg] \\ & \times \frac{L}{\lambda + N_{S_{\leq n-1}}(n-1)\mu_\Omega}. \end{aligned} \tag{134}$$

Proof.

$$\begin{aligned}\widetilde{G}_n^t - \hat{G}_n^t = &\left(\widetilde{H}_n + \lambda I + \sum_{\substack{i=1\\ i\notin S_{\le t}}}^{n-1} F(w_i^{-S_{\le t}}, D_i)\right)^{-1}\\ &\times\left[\sum_{\substack{i=1\\ i\notin S_{\le n-1}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) - \sum_{\substack{i=1\\ i\notin S_{\le t}}}^{n-1} F(w_i^{-S_{\le t}}, D_i)\right](w_n^{-\mathbf{S}_{1:n-1}} - w_{n-1}^{-\mathbf{S}_{1:n-1}})\\ &-\left(\hat{H}_n + \lambda I + \sum_{\substack{i=1\\ i\notin S_{\le t}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i)\right)^{-1}\\ &\times\left[\sum_{\substack{i=1\\ i\notin S_{\le n-1}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) - \sum_{\substack{i=1\\ i\notin S_{\le t}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i)\right](w_n^{-\mathbf{S}_{1:n-1}} - w_{n-1}^{-\mathbf{S}_{1:n-1}})\end{aligned}$$

□

Lemma F.8. *For* $\eta_n^t = \sup\{\left\|I - \hat{P}_n^t\right\|\}$,

$$\eta_n^t \le \frac{\max\{|M|, |\mu|\} + \nu}{\mu - \nu + \lambda + N_{S_{\le t}}(n-1)\mu_\Omega}. \tag{135}$$

Proof.

$$I - \hat{P}_n^t = \left(\hat{H}_n + \lambda I + \sum_{\substack{i=1\\ i\notin S_{\le t}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i)\right)^{-1} \hat{H}_n$$

□

Lemma F.9. *For* $\hat{\kappa}_n^t = \sup\{\left\|\Delta_n - \hat{G}_n^t\right\|\}$,

$$\hat{\kappa}_n^t \le \left(\frac{(S_{\le t}(n-1) - S_{\le n-1}(n-1))M_\Omega}{\mu - \nu + \lambda + N_{S_{\le t}}(n-1)\mu_\Omega} + 1\right)\frac{L}{\lambda + N_{S_{\le n-1}}(n-1)\mu_\Omega}. \tag{136}$$

Proof.

$$\begin{aligned}&\Delta_n - \hat{G}_n^t = -(w_n^{-\mathbf{S}_{1:n-1}} - w_{n-1}^{-\mathbf{S}_{1:n-1}})\\ &-\left(\hat{H}_n + \lambda I + \sum_{\substack{i=1\\ i\notin S_{\le t}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i)\right)^{-1} \sum_{\substack{i=1\\ i\in S_{\le t}, i\notin S_{\le n-1}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i)(w_n^{-\mathbf{S}_{1:n-1}} - w_{n-1}^{-\mathbf{S}_{1:n-1}})\end{aligned}$$

□

Lemma F.10. *For* $\kappa_n^t = \sup\{\left\|\Delta_n - \tilde{G}_n^t\right\|\}$,

$$\kappa_n^t \le \left(\frac{(S_{\le t}(n-1) - S_{\le n-1}(n-1))M_\Omega + N_{S_{\le t}}([\underline{s}^t, n-1])(M_\Omega - \mu_\Omega)}{\mu + \lambda + N_{S_{\le t}}(n-1)\mu_\Omega} + 1\right)\frac{L}{\lambda + N_{S_{\le n-1}}(n-1)\mu_\Omega}. \tag{137}$$

Proof.

$$\Delta_n - \tilde{G}_n^t = -(w_n^{-\mathbf{S}_{1:n-1}} - w_{n-1}^{-\mathbf{S}_{1:n-1}})$$
$$- \left(\widetilde{H}_n + \lambda I + \sum_{\substack{i=1 \\ i \notin S_{\leq t}}}^{n-1} F(w_i^{-S_{\leq t}}, D_i) \right)^{-1} \left[\sum_{\substack{i=1 \\ i \notin S_{\leq n-1}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) - \sum_{\substack{i=1 \\ i \notin S_{\leq t}}}^{n-1} F(w_i^{-S_{\leq t}}, D_i) \right]$$
$$\times (w_n^{-\mathbf{S}_{1:n-1}} - w_{n-1}^{-\mathbf{S}_{1:n-1}})$$

□

Lemma F.11. *For* $\chi_n^{t_b \to t_a} = \sup\{\left\| \hat{P}_n^{t_a} - \hat{P}_n^{t_b} \right\|\}$ *with* $n \leq t_b < t_a$,

$$\chi_n^{t_b \to t_a} \leq \frac{(N_{S_{\leq t_a}}(n-1)M_\Omega + \lambda)(N_{S_{\leq t_b}}(n-1) - N_{S_{\leq t_a}}(n-1))M_\Omega}{(\mu - \nu + \lambda + N_{S_{\leq t_a}}(n-1)\mu_\Omega)} \cdot (\mu - \nu + \lambda + N_{S_{\leq t_b}}(n-1)\mu_\Omega) + \frac{(N_{S_{\leq t_b}}(n-1) - N_{S_{\leq t_a}}(n-1))M_\Omega}{\mu - \nu + \lambda + N_{S_{\leq t_b}}(n-1)\mu_\Omega}. \tag{138}$$

Proof.

$$\hat{P}_n^{t_a} - \hat{P}_n^{t_b} = \left(\hat{H}_n + \lambda I + \sum_{\substack{i=1 \\ i \notin S_{\leq t_a}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i}}, D_i) \right)^{-1} \left(\lambda I + \sum_{\substack{i=1 \\ i \notin S_{\leq t_a}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i}}, D_i) \right)$$
$$- \left(\hat{H}_n + \lambda I + \sum_{\substack{i=1 \\ i \notin S_{\leq t_b}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i}}, D_i) \right)^{-1} \left(\lambda I + \sum_{\substack{i=1 \\ i \notin S_{\leq t_b}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i}}, D_i) \right)$$

□

Lemma F.12. *For* $\zeta_n^{t_b \to t_a} = \sup\{\left\| \hat{G}_n^{t_a} - \hat{G}_n^{t_b} \right\|\}$,

$$\zeta_n^{t_b \to t_a} \leq \left[\frac{(N_{S_{\leq n-1}}(n-1) - N_{S_{\leq t_a}}(n-1))M_\Omega}{(\mu - \nu + \lambda + N_{S_{\leq t_a}}(n-1)\mu_\Omega)} \cdot (\mu - \nu + \lambda + N_{S_{\leq t_b}}(n-1)\mu_\Omega) + \frac{(N_{S_{\leq t_b}}(n-1) - N_{S_{\leq t_a}}(n-1))M_\Omega}{\mu - \nu + \lambda + N_{S_{\leq t_b}}(n-1)\mu_\Omega} \right] \times \frac{L}{\lambda + N_{S_{\leq n-1}}(n-1)\mu_\Omega}. \tag{139}$$

Proof.

$$
\begin{aligned}
\hat{G}_n^{t_a} - \hat{G}_n^{t_b} = &\left(\hat{H}_n + \lambda I + \sum_{\substack{i=1 \\ i \notin S_{\le t_a}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) \right)^{-1} \\
&\times \left[\sum_{\substack{i=1 \\ i \notin S_{\le n-1}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) - \sum_{\substack{i=1 \\ i \notin S_{\le t_a}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) \right] \\
&\times (w_n^{-\mathbf{S}_{1:n-1}} - w_{n-1}^{-\mathbf{S}_{1:n-1}}) \\
&- \left(\hat{H}_n + \lambda I + \sum_{\substack{i=1 \\ i \notin S_{\le t_b}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) \right)^{-1} \\
&\times \left[\sum_{\substack{i=1 \\ i \notin S_{\le n-1}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) - \sum_{\substack{i=1 \\ i \notin S_{\le t_b}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) \right] \\
&\times (w_n^{-\mathbf{S}_{1:n-1}} - w_{n-1}^{-\mathbf{S}_{1:n-1}}) \\
= &\left(\hat{H}_n + \lambda I + \sum_{\substack{i=1 \\ i \notin S_{\le t_a}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) \right)^{-1} \left[\sum_{\substack{i=1 \\ i \notin S_{\le n-1}, i \in S_{\le t_a}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) \right] \\
&\times (w_n^{-\mathbf{S}_{1:n-1}} - w_{n-1}^{-\mathbf{S}_{1:n-1}}) \\
&- \left(\hat{H}_n + \lambda I + \sum_{\substack{i=1 \\ i \notin S_{\le t_b}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) \right)^{-1} \left[\sum_{\substack{i=1 \\ i \notin S_{\le n-1}, i \in S_{\le t_b}}}^{n-1} F(w_i^{-\mathbf{S}_{1:i-1}}, D_i) \right] \\
&\times (w_n^{-\mathbf{S}_{1:n-1}} - w_{n-1}^{-\mathbf{S}_{1:n-1}})
\end{aligned}
$$

□

G Additional Experiments

G.1 Effects of Unlearning Timing and Request Order

We further examine how the timing and order of multiple unlearning requests affect the approximation error. This experiment complements the fixed multi-request sequence in the main paper by holding the task stream and the final deleted set fixed while changing only the request schedule. In particular, all schedules eventually delete $\{4, 5, 8, 10, 11, 13, 14\}$, but issue the requests in the following orders:

$$
\begin{aligned}
\mathcal{Q}_1 &= (7 : \{4, 5\};\ 12 : \{8, 10, 11\};\ 17 : \{13, 14\}), \\
\mathcal{Q}_2 &= (12 : \{4, 5\};\ 16 : \{8, 10, 11\};\ 18 : \{13, 14\}), \\
\mathcal{Q}_3 &= (15 : \{13, 14\};\ 16 : \{8, 10, 11\};\ 18 : \{4, 5\}).
\end{aligned}
$$

Here, $\mathcal{Q}_1$ is forward-synchronous: every later request deletes only tasks arriving after the preceding unlearning time. In contrast, $\mathcal{Q}_2$ and $\mathcal{Q}_3$ delete tasks from an earlier portion of the stream after a previous request, and are therefore asynchronous in the sense of Section 5.2. The reverse-ordered schedule $\mathcal{Q}_3$ provides the strongest test of this effect.

Experimental Setup. We use the same CIFAR-100 feature stream, pretrained ResNet-18 backbone, and three-layer neural-network head as in the main experiments. The stream contains 20 tasks,

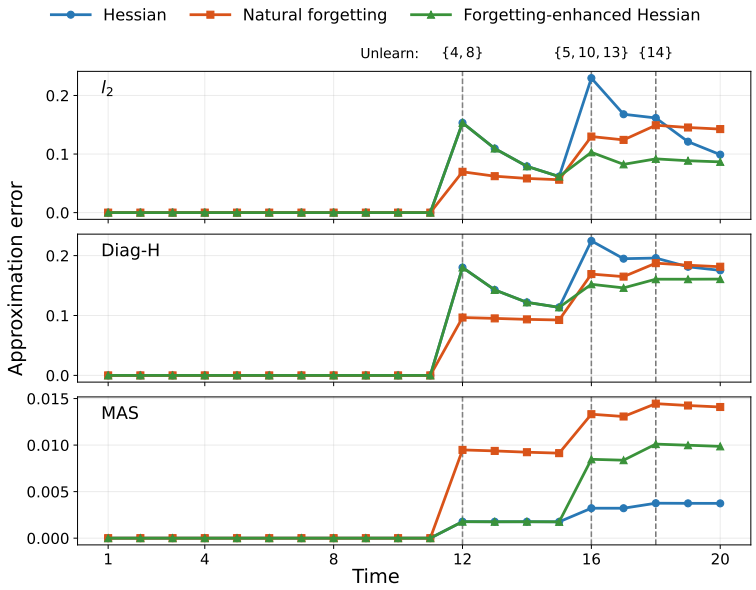


Figure 2: Approximation error during multi-request CLU. The requests at $t = 12, 16, 18$ delete $\{4, 8\}$, $\{5, 10, 13\}$, and $\{14\}$, respectively. Each request increases the error, while subsequent retained-task updates propagate or reduce its effect.

with 200 optimization steps per task. We evaluate EWC and MAS with regularization strengths 50, regularizer constants 8 and 5, respectively, and 50 diagonal-Hessian probe vectors. Hessian unlearning is executed only when a request is received. At every time t, we measure the approximation error

$$\left\| w_t^{-\mathbf{S}_{1:t}} - w_t^{-S_{\le t}} \right\|_2$$

against the checkpoint obtained by exact retraining on the corresponding retained data. Natural forgetting is evaluated against the same retraining reference.

Effect of Request Order. Figure 3 shows that the Hessian trajectory depends substantially on request order, even though all schedules have the same final deleted set. Under EWC, the forward-synchronous schedule $\mathcal{Q}_1$ gives the smallest final error, 0.1325, compared with 0.1702 for $\mathcal{Q}_2$ and 0.1973 for the reverse-ordered $\mathcal{Q}_3$. Thus, enforcing the forward-synchronous order reduces the final EWC error by 32.8% relative to $\mathcal{Q}_3$. This agrees with the analysis in Section 5.2: later deletion of tasks lying inside an earlier correction path can interrupt its propagation and introduce additional regularizer-mismatch terms.

MAS exhibits much smaller absolute errors for all three schedules, but its empirical ordering differs from EWC. The first request of $\mathcal{Q}_3$, which deletes the recent tasks $\{13, 14\}$ at $t = 15$, has error 7.98×10^{-5}; after the final deletion of the older tasks $\{4, 5\}$, its error increases to 0.003545. Nevertheless, this remains lower than the final errors of $\mathcal{Q}_1$ and $\mathcal{Q}_2$. Consequently, $\mathcal{Q}_3$ improves over $\mathcal{Q}_1$ by 42.6% under MAS. This does not contradict the asynchronous upper bound: the bound adds nonnegative worst-case terms but does not impose a deterministic ordering on realized errors. With the stronger and highly anisotropic MAS regularizer, task-dependent correction magnitudes and cancellation can dominate the conservative asynchrony penalty.

Hessian Unlearning Versus Natural Forgetting. Natural forgetting has the same final error across the three schedules for a fixed regularizer: 0.276465 for EWC and 0.011616 for MAS. This is expected because natural forgetting applies no request-time correction, and the final retained task set is identical. Request timing changes the intermediate retraining reference, but not the final natural-forgetting model or final retrained reference. In contrast, Hessian unlearning stores and propagates request-specific corrections, so its final state depends on the order in which requests arrive.

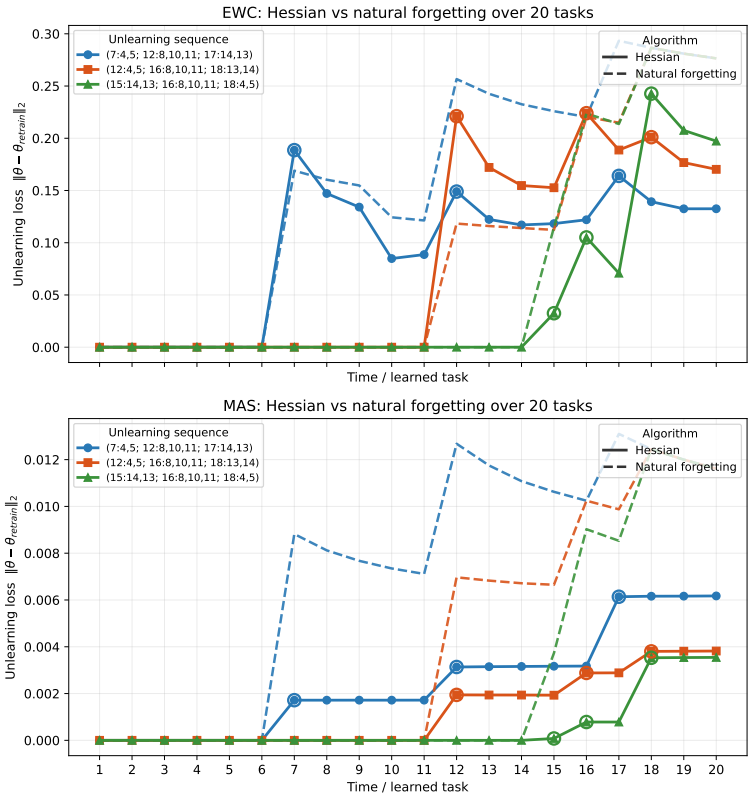


Figure 3: Approximation error under three multi-request schedules for EWC (top) and MAS (bottom). Solid curves denote Hessian unlearning and dashed curves denote natural forgetting. Hollow markers identify request times. The schedules delete the same final task set but differ in request timing and order.

Across all six regularizer-schedule combinations, Hessian unlearning improves on natural forgetting at $t = 20$. The reduction ranges from 28.6% to 52.1% under EWC and from 46.9% to 69.5% under MAS. These results support two conclusions. First, explicit second-order correction is consistently useful when several deletions accumulate. Second, the preferred request schedule is regularizer-dependent: forward-synchronous requests are particularly beneficial for EWC, whereas MAS is sufficiently stable that the identity and recency of the deleted tasks dominate the observed order effect.

G.2 Experiment Details

Dataset, Representation, and Model. We conduct all experiments on CIFAR-100. Each continual-learning (CL) task is formed by sampling a subset of the 100 classes with a fixed random seed; class subsets may overlap across tasks, yielding a non-IID stream rather than a disjoint class-incremental split. The accuracy and approximation-error runs train only a three-layer neural-network head,

$$d_{\text{in}} \longrightarrow 512 \longrightarrow 256 \longrightarrow 100,$$

with ReLU activations after the two hidden layers. The released runs load a frozen, precomputed CIFAR-100 representation containing $50{,}000$ training and $10{,}000$ test vectors of dimension 768. The implementation also provides a raw-data path in which each arriving image is transformed by an ImageNet-pretrained ResNet-18 with a frozen backbone and a 512-dimensional output. We use this raw-data path only for the end-to-end runtime and CPU data-storage benchmark; it prevents the feature-extraction cost from being hidden by the cache.

Optimization and Reproducibility. Unless specified otherwise, we use Adam, mini-batches of size 128, 200 optimization steps per task, and random seed 42. We make the CUDA/cuDNN exe-

cution deterministic when supported. Hessian-vector and diagonal-Hessian quantities are estimated with 50 stochastic probe vectors. The single-request experiment uses learning rate 2×10^{-3}; the 20-task experiments use 10^{-3}. A StepLR scheduler with step size 500 is instantiated for each task, so it does not decay the learning rate within a 200-step task. The strong-convexity damping parameter is 10^{-4}, and the additional Hessian damping parameter is zero.

Hyperparameter Settings. Table 3 lists the shared optimization hyperparameters. The Adam arguments not explicitly overridden by the implementation retain the PyTorch defaults: momentum coefficients $(\beta_1, \beta_2) = (0.9, 0.999)$, numerical-stability constant $\epsilon_{\mathrm{Adam}} = 10^{-8}$, and zero weight decay. Regularization is added explicitly to the training objective and is therefore not implemented through optimizer weight decay. Table 4 gives the experiment-specific task construction and regularization strengths. Here, λ multiplies the task-dependent EWC, MAS, or Diag-H statistic, whereas c denotes its isotropic constant; scalar ℓ_2 has only c.

Table 3: Optimization and estimation hyperparameters shared across runs.

Hyperparameter	Value
Optimizer	Adam
Adam (β_1, β_2)	$(0.9, 0.999)$
Adam ϵ_{Adam}	10^{-8}
Optimizer weight decay	0
Batch size	128
Updates per task	200
StepLR step size / multiplier	$500/0.5$
Random seed	42
Diagonal-Hessian probes	50
Strong-convexity damping	10^{-4}
Additional Hessian damping	0
Classifier hidden widths	$512, 256$

Regularization and Unlearning Baselines. We study scalar ℓ_2, accumulated ℓ_2, EWC, MAS, and diagonal-Hessian regularization. EWC uses the accumulated empirical-Fisher diagonal, MAS uses the accumulated output-energy sensitivity, and Diag-H uses a stochastic diagonal-Hessian estimate. For the shared Diag-H variant, every estimated diagonal is projected onto the nonnegative orthant before it is used for learning, propagation, unlearning, or storage; the same stored diagonal therefore serves as both the Hessian approximation and the regularizer. We compare Hessian unlearning (HU), natural forgetting (NF), forgetting-enhanced Hessian unlearning (FE-HU), and exact retraining (ER). HU applies a correction only when a request arrives. NF performs no explicit request-time correction. At a noninitial request, FE-HU keeps only the post-previous-request portion required by its recurrence and discards obsolete Hessian and increment information immediately after the update. The compressed NF implementation stores one accumulated regularizer rather than all task-wise regularizer vectors. ER reruns CL on the retained data and is used as the reference.

Metrics. At time t, approximation error is the Euclidean parameter distance

$$\left\| w_t^{-\mathbf{S}_{1:t}} - w_t^{-S_{\le t}} \right\|_2$$

between an approximate-unlearning checkpoint and the matching exact-retraining checkpoint. Test accuracy is evaluated on the retained-task test set using the corresponding final checkpoint. Runtime is wall-clock time for the complete experiment. Peak GPU allocation is measured with the CUDA allocator. Peak CPU data storage counts raw examples that must remain resident, the currently arriving raw task, and its temporary features. The reported total storage is the sum of peak GPU allocation and peak CPU data storage. ER keeps all undeleted historical raw datasets, whereas the online methods retain only the current raw task.

Single-Request Experiment. The main single-request study contains 10 tasks. Each task samples between 50 and 60 CIFAR-100 classes, and a request is issued after task 10. The archived canonical regimes delete an early task $\{2\}$, a middle task $\{5\}$, a late task $\{8\}$, an early pair $\{2, 3\}$, or a mixed triple $\{2, 5, 8\}$; the main table reports the three singleton cases. The canonical regularization strength is 100. For EWC, MAS, and Diag-H, the task-dependent term has multiplier 100 and the

Table 4: Experiment-specific hyperparameters. "Classes/task" is the number of CIFAR-100 labels sampled independently for each task.

Experiment	Tasks	Classes/task	Learning rate	Regularization
Single req.	10	50–60	2×10^{-3}	ℓ_2/Accum-ℓ_2 : $c = 100$; EWC/MAS/Diag-H: $\lambda = 100, c = 5$
Multi-req. figure	20	65	10^{-3}	ℓ_2 : $c = 50$; Diag-H: $\lambda = 50, c = 10$; MAS: $\lambda = 30, c = 1$
Cost/noise	20	65	10^{-3}	ℓ_2 : $c = 50$; Diag-H: $\lambda = 50, c = 8$; MAS: $\lambda = 30, c = 1$
Order appendix	20	65	10^{-3}	EWC: $\lambda = 50, c = 8$; MAS: $\lambda = 50, c = 5$

isotropic constant is 5. Scalar and accumulated ℓ_2 use the matched strength 100. We report the non-dominated configurations under joint minimization of final approximation error and maximization of retrained-model accuracy. The method label H or N records whether HU or NF provides the selected approximation error.

Multi-Request Experiment. The main multi-request stream contains 20 tasks, each sampling 65 classes, with requests

$$(12 : \{4, 8\};\ 16 : \{5, 10, 13\};\ 18 : \{14\}).$$

We compare HU, NF, FE-HU, and ER under ℓ_2, shared nonnegative Diag-H, and MAS. The ℓ_2 constant is 50. MAS uses $\lambda_{\mathrm{MAS}} = 30$ and isotropic constant 1. The current approximation-error figure uses $\lambda_{\mathrm{Diag-H}} = 50$ and isotropic constant 10. The separately measured cost and certified-noise results retain the preceding Diag-H setting with the same multiplier 50 and isotropic constant 8; we keep both configurations in the artifact to avoid conflating the two runs.

Cost and Certified Accuracy. For the raw-streaming benchmark, every arriving CIFAR-100 image passes through the frozen ImageNet-pretrained ResNet-18 before the head update. ER incurs additional CPU storage because it retains undeleted raw tasks and repeats the CL pass after each request. For certified accuracy, Gaussian output perturbation is calibrated from the final theoretical approximation-error bound with $(\varepsilon, \delta) = (1000, 10^{-5})$. We add independent zero-mean Gaussian noise to every released model parameter and report the mean and standard deviation of retained-task accuracy over 30 noise draws. FE-HU is reported with calibrated noise, while ER is reported without noise.

Empirical Bound-Parameter Estimation. We estimate the parameters appearing in the approximation-error bound along the realized neural-network trajectory. For task n and deletion history available at time t, let R_n^t be the retained-data regularizer, R_n^{all} its counterpart before removing the requested tasks, and let

$$A_n^t = H_n + R_n^t + \mu I, \qquad P_n^t = (A_n^t)^{-1} R_n^t, \qquad G_n^t = (A_n^t)^{-1}(R_n^{\mathrm{all}} - R_n^t)u_n, \tag{140}$$

where $\mu = 10^{-4}$, H_n is the empirical task-loss Hessian evaluated at the stored task checkpoint, and $u_n = w_n - w_{n-1} = -\Delta_n$ is the observed learning increment. The full-Hessian calculation does not materialize H_n. Instead, it evaluates Hessian-vector products by automatic differentiation, solves systems in A_n^t with diagonally preconditioned BiCGSTAB, and estimates

$$\rho_n^t = \|P_n^t\|_2, \qquad g_n^t = \|G_n^t\|_2$$

using power iteration on $(P_n^t)^\top P_n^t$. We use at most five mini-batches per local Hessian estimate, four power iterations, at most 40 linear-solver iterations, relative tolerance 10^{-4}, a preconditioner floor of 10^{-3}, and multiply the resulting ρ_n^t and g_n^t estimates by a 1.05 safety factor.

For diagonal propagation, the loss-Hessian diagonal is estimated with 50 Hutchinson probes and projected onto the nonnegative orthant in the Diag-H experiment. Because $\hat{P}_n^t$ is diagonal, we compute

$$\hat{\rho}_n^t = \|\hat{P}_n^t\|_2 = \|\operatorname{diag}(\hat{P}_n^t)\|_\infty, \quad \hat{g}_n^t = \|\hat{G}_n^t\|_2,$$
$$\eta_n^t = \|I - \hat{P}_n^t\|_2, \qquad \Gamma_n = \|u_n\|_2.$$

For every pair of request histories $t_i < t_k$, we additionally measure $\chi_n^{t_i \to t_k} = \|\hat{P}_n^{t_k} - \hat{P}_n^{t_i}\|_2$ and $\zeta_n^{t_i \to t_k} = \|\hat{G}_n^{t_k} - \hat{G}_n^{t_i}\|_2$. The composite factors Z are then evaluated directly from the products

and sums of these measured $\hat{\rho}$, η, and χ values using the same request intervals and deleted-task exclusions as in the theoretical bound. Finally, we substitute all measured quantities into the full forgetting-enhanced expression and propagate the request-time bound from $t = 18$ to the final time $T = 20$. The ℓ_2 and MAS noise experiments use the full-Hessian estimates of ρ and g together with the measured diagonal quantities; the shared nonnegative Diag-H experiment uses its directly measured diagonal operators throughout.

Noise Calibration. Let γ_T denote the resulting final plug-in bound. We release

$$\widetilde{w}_T = w_T + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2 I), \qquad \sigma = \gamma_T \frac{\sqrt{2\log(1.25/\delta)}}{\varepsilon}. \tag{141}$$

At $(\varepsilon, \delta) = (1000, 10^{-5})$, the multiplicative conversion factor from γ_T to per-parameter noise standard deviation is 0.004844805. Table 5 reports the exact bounds and standard deviations used for the accuracy experiment. For each method, we generate 30 independent noisy checkpoints with seeds $20260724, \ldots, 20260753$ and evaluate each checkpoint on the retained test labels.

Table 5: Trajectory-specific bounds and per-parameter Gaussian noise used for the reported accuracy experiment.

Reg.	UL method	γ_{20}	σ
ℓ_2	NF	0.996885	0.00482971
	FE-HU	0.504609	0.00244473
Diag-H	NF	1.254805	0.00607929
	FE-HU	0.962242	0.00466187
MAS	NF	0.076526	0.000370752
	FE-HU	0.040202	0.000194771

These quantities are local, trajectory-specific empirical estimates rather than uniform suprema over an entire parameter ball. Consequently, the reported noise is a plug-in implementation of the certified-unlearning calibration: its formal guarantee is conditional on the estimated γ_T upper-bounding the true approximation error.

Additional Request-Order Experiment. The appendix experiment holds the 20-task stream and final deleted set $\{4, 5, 8, 10, 11, 13, 14\}$ fixed while varying request timing and order:

$$\mathcal{Q}_1 = (7 : \{4, 5\};\ 12 : \{8, 10, 11\};\ 17 : \{13, 14\}),$$
$$\mathcal{Q}_2 = (12 : \{4, 5\};\ 16 : \{8, 10, 11\};\ 18 : \{13, 14\}),$$
$$\mathcal{Q}_3 = (15 : \{13, 14\};\ 16 : \{8, 10, 11\};\ 18 : \{4, 5\}).$$

We evaluate HU and NF for EWC and MAS. EWC uses $\lambda_{\mathrm{EWC}} = 50$ and constant 8; MAS uses $\lambda_{\mathrm{MAS}} = 50$ and constant 5. All other optimization settings match the 20-task experiment. HU executes only at request times, while approximation error is recorded after every task against the schedule-matched ER checkpoint. Because the final deleted set is the same, NF has the same final error across the three schedules; HU remains order-dependent through its request-specific propagated corrections.