Recovery Guarantees for Continual Learning of Dependent Tasks: Memory, Data-Dependent Regularization, and Data-Dependent Weights

Liangzu Peng* **Uday Kiran Reddy Tadipatri*** **Ziqing Xu** **Eric Eaton** **René Vidal**
University of Pennsylvania

Abstract

Continual learning (CL) is concerned with learning multiple tasks sequentially without forgetting previously learned tasks. Despite substantial empirical advances over recent years, the theoretical development of CL remains in its infancy. At the heart of developing CL theory lies the challenge that the data distribution varies across tasks, and we argue that properly addressing this challenge requires understanding this variation–dependency among tasks. To explicitly model task dependency, we consider nonlinear regression tasks and propose the assumption that these tasks are dependent in such a way that the data of the current task is a nonlinear transformation of previous data. With this model and under natural assumptions, we prove statistical recovery guarantees (more specifically, bounds on estimation errors) for several CL paradigms in practical use, including experience replay with data-independent regularization and data-independent weights that balance the losses of tasks, replay with data-dependent weights, and continual learning with data-dependent regularization (e.g., knowledge distillation). To the best of our knowledge, our bounds are informative in cases where prior work gives vacuous bounds.

1 INTRODUCTION

Continual learning (CL) aims to learn multiple tasks presented sequentially, with a key goal to address the situation of *catastrophic forgetting* (McCloskey and Cohen, 1989): learning new tasks risks performance degradation on previously learned tasks. To reduce forgetting, *memory-based methods* store some past data, to be used with new data for training the new task (Robins, 1993; Shin et al., 2017; Aljundi et al., 2019; Dokania et al., 2019; Prabhu et al., 2020; Verwimp et al., 2021; Bang et al., 2021; Wang et al., 2022); *regularization-based methods* optimize the current task with a regularization term that encourages proximity to the model previously learned (Li and Hoiem, 2017; Kirkpatrick et al., 2017; Zenke et al., 2017; Rebuffi et al., 2017; Park et al., 2019; Buzzega et al., 2020); *constrained optimization methods* enforce non-forgetting requirements as optimization constraints while solving the current task (Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2019; Peng et al., 2023a; Elenter et al., 2023; Li et al., 2024).

In contrast to the abundance of empirical advances, theoretical investigations in CL are relatively scarce. Existing theoretical contributions are of at least two types. The first type considers linear models or the kernel regime (Bennani and Sugiyama, 2020; Doan et al., 2021; Heckel, 2022; Evron et al., 2022; Lin et al., 2023; Li et al., 2023; Swartworth et al., 2023; Goldfarb et al., 2024; Zhao et al., 2024; Ding et al., 2024; Banayeeanzade et al., 2025; Evron et al., 2026; Karpel et al., 2026). With such specific settings, it is possible to derive meaningful and even tight bounds, but the resulting theory might not apply to existing CL methods that are specifically designed for deep networks or general nonlinear models. The other type of work considers general models (e.g., within a PAC-Bayes framework or based on the Rademacher complexity) (Pentina and Lampert, 2015; Yin et al., 2020; Ye and Bors, 2022; Friedman and Meir, 2024). With such general settings, it is possible to derive CL guarantees by leveraging tools from classic (PAC) learning theory (Shalev-Shwartz and Ben-David, 2014; Mohri et al., 2018; Alquier et al., 2024), but the resulting theory often fails to capture the benefits of learning multiple tasks sequentially, or the resulting bound does not necessarily tend to zero even in the presence of infinitely many samples.

In light of the above, we are motivated to seek a middle ground that combines the best of both worlds: derive meaningful error bounds for general nonlinear models and commonly used CL paradigms.

Importance of Task Dependency. We argue that attaining the above goal requires modeling how tasks are related to each other. We illustrate this by way of examples.

Example 1. The error bounds (specifically, generalization

*: Equal contribution.

bounds) of Ye and Bors (2022); Friedman and Meir (2024); Mansour et al. (2009) grow linearly with the "*distance*" between the data distributions of two tasks. For two zero-mean Gaussian distributions with variances 1 and s^2, respectively, this distance is proportional to $|s^2 - 1|$, and thus grows unbounded as s increases (cf. Remark 3). These are two basic distributions related in a simple way, i.e., one is a scalar multiple of the other, yet the bound depending linearly on s^2 becomes vacuous even for a mildly large s.

Example 2. In the work of Peng et al. (2023a), the data within each task are i.i.d., and data across tasks are dependent. The proof of Peng et al. (2023a) reduces from a complicated CL situation to the standard scenario of learning a single task from i.i.d. data. As a side effect of this reduction, the bound of *every* task in Peng et al. (2023a) depends logarithmically on the total number T of seen tasks and thus worsens as T grows; this is somehow counterintuitive as it indicates learning more tasks enlarges the error bound of *every* task. In hindsight, analyzing dependent data appears to be intractable under the general setting of Peng et al. (2023a) as they do not specify *how* the tasks depend on each other. This leaves open the challenge of identifying task dependency that avoids degrading, or ideally improves, continual learning performance.

The Proposed CL Model With Task Dependency. In order to address issues pertaining to Examples 1 and 2 and to furthermore develop meaningful error bounds, we propose a CL framework with explicit modeling of task dependency. Specifically, we consider the problem of *continual noisy nonlinear regression*: learn a function f^* from a sequence of noisy nonlinear regression tasks with f^* assumed to be the shared *true predictor* of all these tasks. This is a setting that generalizes previous works on *continual linear regression* to the nonlinear, noisy case. Recognizing that directly analyzing such continual nonlinear regression model might lead to unsatisfactory bounds (Examples 1 and 2), we then arm this model with *task dependency*, which posits that the data of the present task are obtained as a nonlinear yet unknown transformation of the data from previous tasks. This task dependency is motivated from several perspectives (Section 2.2). For example, it draws inspiration from the *rotated MNIST* and *permuted MNIST* datasets that have benchmarked many fundamental CL methods; there, the transformation is either some rotation or permutation. Also, it finds inspiration from *dynamical systems*, where the current state is obtained as a transformation of the previous state.

Implications of Our CL Model and Task Dependency. We now sketch how the issues in Examples 1 and 2 are resolved with the proposed model and task dependency. For the two tasks in Example 1, the nonlinear transformation is simply a scaling function that multiplies its input by s; and s is the *Lipschitz constant* of this transformation. Crucially, we find that all our theorems exhibit only a *logarithmic* dependency on s in the context of Example 1, thereby allowing s to grow polynomially with problem size (e.g., dimension, sample size), without drastically affecting our bounds. This is a significant improvement over the linear dependency of s^2 in Example 1 (Mansour et al., 2009; Ye and Bors, 2022; Friedman and Meir, 2024).

Unlike Example 2, our task dependency assumption makes the analysis tractable, though still non-trivial (cf. Fig. 2). In particular, it enables us to prove concentration bounds for dependent data, and it suffices to apply the bounds only once. By doing so, we eliminate the logarithmic dependency on T as discussed in Example 2 where the concentration inequalities are invoked T times.

More importantly, the proposed model and task dependency allow us to develop theoretical guarantees under a unifying framework and in a systematic fashion. Under basic assumptions, we bound the estimation errors of recovering f^* for the aforementioned CL paradigms, including memory-based methods, regularization-based methods, and constrained optimization methods. In more detail:

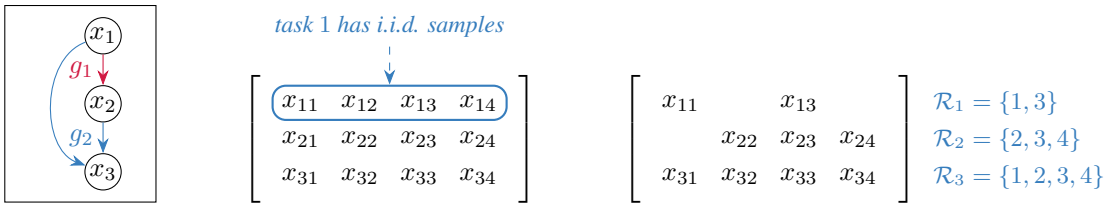
- We begin by analyzing *weighted experience replay with data-independent regularizers*, where the weights balance contributions of each task (Section 3.2). With a non-uniform choice of weights, our bound on the weighted estimation error is inversely proportional to the total number of available data; such a bound is optimal. With uniform weights, our bound is also optimal if the replay buffer stores a constant fraction of the full data of each task.

- Then, we analyze regularization-based methods with a *knowledge distillation* regularizer (Section 3.3). While such regularizer is *data-dependent*, we can still prove estimation error bounds via a non-trivial recursive argument, resulting in improvements over prior work in terms of generality (Heckel, 2022; Li et al., 2023; Zhao et al., 2024; Zhu et al., 2025) and tightness (Yin et al., 2020) (cf. Remark 5).

- Lastly, we draw motivations from the constrained learning framework (Chamon et al., 2022; Elenter et al., 2023) and derive error bounds for *replay with data-dependent weights*, where the weights might now be thought of as dual variables in primal dual optimization (Section 3.4). We obtain error bounds of a similar flavor by extending our previous results to account for the fact that the weights are now random variables as well, depending on the data.

2 PROBLEM SETUP

In Section 2.1, we introduce our data model with the task dependency specified in an autoregressive fashion, based upon which we will develop our theorems. In Section 2.2 we provide justifications and motivations on our proposed task dependency model.



(a) data dependency (b) full data (c) training data available at task 3

Figure 1: Example setup ($T = 3, m = 4$). Fig. 1a: task dependency (2); Fig. 1b: full data in a matrix, where each column is generated as per (2) and each row represents data of each task; Fig. 1c: index sets $\mathcal{R}_t$ and data available at task 3.

2.1 Data Model, Task Dependency, and Samples

Data Model with Autoregressive Task Dependency. Let T be the number of tasks seen thus far. Let $[T] := \{1, \dots, T\}$. We consider a nonlinear regression setting, where each task t shares a true predictor $f^* : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ that maps input x_t to output y_t up to some random noise v_t, similarly to prior work (Evron et al., 2022; Peng et al., 2023a; Li et al., 2023; Zhao et al., 2024; Elenter et al., 2023; Zhu et al., 2025). In addition, we specify the task dependency in an autoregressive manner through some deterministic, nonlinear transformation $g_t : \mathbb{R}^{(t-1)\times d_x} \to \mathbb{R}^{d_x}$. Concretely, we consider the model

$$y_t = f^*(x_t) + v_t, \qquad \forall t \geq 1; \tag{1}$$
$$x_t = g_t(x_1, \dots, x_{t-1}), \quad \forall t > 1. \tag{2}$$

Here, $x_1 \in \mathbb{R}^{d_x}$ is a random vector drawn from some probability distribution $\mathcal{D}_1$. Thus, (1) and (2) implicitly specify the distribution of y_t (conditioned on x_t) and of x_t (conditioned on $x_1, \dots, x_{t-1}$). By defining the relationship between data of different tasks, the transformation g_t captures the dependency among these tasks (cf. Fig. 1a). While we interpret (1) and (2) as a task dependency model in the CL context, our main motivation is from the control literature: (1) and (2) define a nonlinear dynamical system where x_t's are system states and f^* consists of system parameters to be identified. Different from standard dynamical systems, our model has no control input and (2) is in the absence of noise. This is for the sake of simplicity, and it is not difficult to extend our results to the case of *noisy* task dependency where x_t is equal to $g_t(x_1, \dots, x_{t-1})$ up to additive random noise. One more difference is that in (2) we require x_t to depend on all the past $x_1, \dots, x_{t-1}$, while a common special case is of the Markov type $x_t = g_t(x_{t-1})$. We discuss more examples and motivations in Section 2.2.

Samples and Memory. For task 1, we sample m i.i.d. input data points $\{x_{1,i}\}_{i=1}^m$ from distribution $\mathcal{D}_1$, that is

$$\{x_{1,i}\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} \mathcal{D}_1. \tag{3}$$

We will omit the comma and write x_{1i} for $x_{1,i}$ if it does not cause confusion. For task $t > 1$, we generate input-output pairs (x_{ti}, y_{ti}) as per (1) and (2), that is we have

$$x_{ti} = g_t(x_{1i}, \dots, x_{t-1,i}), \quad \forall t > 1, i \in [m], \tag{4}$$

and $y_{ti} = f^*(x_{ti}) + v_{ti}$ for some noise v_{ti}. Fig. 1b visualizes the input samples $\{x_{ti}\}_{t\in[T], i\in[m]}$ in a matrix form.

Motivated by memory-based methods in CL, we furthermore assume the availability of only part of the samples indexed by some fixed subset $\mathcal{M}$ of $[T] \times [m]$. That is, (x_{ti}, y_{ti}) is available for training if and only if $(t, i) \in \mathcal{M}$. The available inputs $(x_{ti})_{(t,i)\in\mathcal{M}}$ can be arranged into a $T \times m$ partial matrix (cf. Fig. 1c). The t-th row of this matrix corresponds to the stored data $\{(x_{ti}, y_{ti})\}_{i\in\mathcal{R}_t}$ of task t; we index them by $\mathcal{R}_t$, a subset of $[m]$ (Fig. 1c). Let $n_t := |\mathcal{R}_t|$. We have $n_t \leq m$. We have $|\mathcal{R}_T| = m$, as we assume full access to the samples of the current task T.

Given the above problem setup, our goal is to learn f^* from the available samples indexed by $\mathcal{M}$.

Remark 1. Constructing $\mathcal{M}$ is to select samples to store and is an interesting CL topic. This can be done via some information-theoretical criterion or optimization (Borsos et al., 2020; Sun et al., 2022; Elenter et al., 2023). On the other hand, simple strategies such as random sampling or *reservoir sampling* (Vitter, 1985) actually work very well (Dokania et al., 2019; Araujo et al., 2022). We assume $\mathcal{M}$ is fixed, while our results apply directly to the case where $\mathcal{M}$ is constructed via random or reservoir sampling.

Comparison to Other Modeling Assumptions. To highlight the advantages of our modeling assumptions, we first consider the following alternatives:

- A common setting is that the samples within each task are i.i.d., but the samples across tasks be dependent in an arbitrary way; namely, x_{ti} depends on $x_{\tau j}$ for any $t \neq \tau$ and any i, j (cf. Fig. 2a). This is a practical and general setup as considered in Example 2, but it is also a challenging situation for which prior work does not shed light on the benefits of learning from multiple dependent tasks.

- In another setting, one has T distributions $\{\mathcal{D}_t\}_{t\in[T]}$. For each t, one draws i.i.d. samples x_{ti} from $\mathcal{D}_t$. Thus, the samples of all tasks are independent, though not identically distributed (cf. Fig. 2b). This setting is intuitive but of less

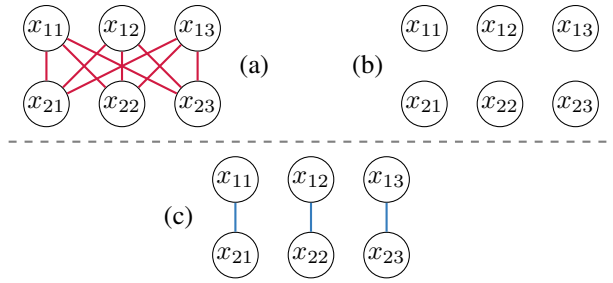


Figure 2: Illustrating the sample-level dependency assumptions on 3 tasks, each with 3 samples. In (a, b, c), samples are i.i.d. within each task (i.e., no horizontal edges), and samples across tasks are distributed differently; (a): samples across different tasks are dependent; (b): samples across tasks are independent; (c) samples across tasks exhibit one-to-one dependency, as specified by (2).

technical interest, as the proof of estimation error bounds is significantly simplified due to the full independence.

In comparison, our proposed setting strikes a balance by assuming that the data across tasks are dependent in a specific way (cf. Fig. 2c). While the proof is still nontrivial under our setting, the intuitive benefit is clear: it reduces the potential dependency across tasks, making it possible to derive better bounds (Fig. 2a versus Fig. 2c).

Comparison to Task Dependency Assumptions. Learning from the full data (or partial data) as shown in Fig. 1 can be viewed as a multitask learning problem where each task t has m (or n_t) samples (Maurer et al., 2016; Tripuraneni et al., 2020, 2021; Du et al., 2021). The differences from recent multitask learning papers are twofold: Our samples across tasks are *dependent*, as specified in (2); our definition of task dependency is different. In fact, pursuing a suitable definition of task dependency has led to fruitful developments in *multitask learning* and *transfer learning*. Ben-David and Borbely (2008) capture the notion of task dependency by an equivalence relation between functions in the hypothesis space. This allows empirical risk minimization to be conducted in a smaller (quotient) space, which translates to a tighter theoretical bound. But this benefit is compromised as optimization over such a space is difficult. A practical task dependency assumption, as considered by Maurer et al. (2016); Tripuraneni et al. (2020, 2021); Du et al. (2021) and even in earlier work (Caruana, 1997; Baxter, 2000; Argyriou et al., 2006), posits that each task t admits a predictor of the form $f_t \circ \phi$, where ϕ is called the shared representation and f_t a task-specific predictor often assumed to be linear. Applications of this setting include deep networks with a shared backbone representing ϕ and multiple task-specific heads representing f_t. However, such a network needs knowledge of the task corresponding to an input in order to select the corresponding head, which makes it inapplicable to some CL scenarios where no such task identity is available at test time (e.g., *class-incremental* or *domain-incremental* learning) (van de Ven et al., 2022; Ramesh and Chaudhari, 2022).

2.2 Motivation and Context

We now discuss several lines of prior work that motivate our problem settings and task dependency assumptions.

CL Context. Our nonlinear transformation g_t in (2) covers as special cases the rotations or permutations that arise in CL datasets known respectively as *rotated* MNIST and *permuted* MNIST, where task 1 is to classify the MNIST images x_1, and the images x_t of subsequent tasks are obtained by rotating or permuting the pixels of x_1. While these datasets appear artificial, they have been used to benchmark many CL methods, which make themselves practically significant. The task dependency with rotations was theoretically analyzed recently by Goldfarb and Hand (2023); Goldfarb et al. (2024) in the case of two linear regression tasks. In these works, the rotation acts only on the input data and does not change the labels, similarly to the setting of Ben-David and Borbely (2008).

Motivation from Machine Learning and Control. Note that g_t defines a type of autoregressive models which have played major roles in machine learning for computing with sequential data (Bishop, 2006, Chapter 13). Classic such models include (hidden) Markov models and linear dynamical systems, and they find ample applications in speech recognition, language modeling, and control systems where g_t is often to be learned. Furthermore, recent empirical research shows that deep generative models such as diffusion models can transform one distribution or dataset into another. In diffusion models the transformation between consecutive time steps is of the type *identity-plus-noise*. Different from our deterministic g_t, this is a random transformation due to the use of random noise.

Our data generation protocol (1) and (2) also find motivations in recent research on learning from *many trajectories* or *dependent data* (Tu et al., 2024; Ziemann et al., 2023; Ziemann and Tu, 2022; Tadipatri et al., 2025). A major difference is that these works assume the availability of full data (Fig. 1b). Instead, we are motivated by CL scenarios (memory-based methods), and consider a more general situation of learning from many *incomplete* or *partial* trajectories (Fig. 1c). At a technical level, we build upon the proof ideas of Ziemann et al. (2023); Tadipatri et al. (2025) but also make extensions for cases such as partial trajectories, weighted objectives, data-dependent regularization, and data-dependent weights (cf. Section 3).

Geometric Vision Context. In *geometric vision* (Hartley and Zisserman, 2004), g_t also arises. There, a classical scenario is *point cloud registration* (Arun et al., 1987): Given two point clouds $\{x_{1i}\}_{i\in[m]}$ and $\{x_{2i}\}_{i\in[m]}$ of the same

scene captured by a moving camera, the goal is to find some Euclidean transformation g_2 that aligns them; ideally we have $x_{2i} = g_2(x_{1i})$ for all $i \in [m]$, while it is possible that not all points of the scene are captured and some points are missing (cf. Fig. 1c). Here, g_2 models how the camera moves and is independent of the point clouds. Finally, a sequence of multiple point clouds arise naturally, as the camera moves and continually measures the scene.

3 THEORETICAL CONTRIBUTIONS

We first describe the notations and technical assumptions (Section 3.1). Then we introduce our theory for *weighted experience replay with regularization* (Section 3.2), *data-dependent regularization* (Section 3.3), and *data-dependent weights* (Section 3.4).

3.1 Notations and Technical Assumptions

Notations. Let $\mathbb{B}_s(d)$ be the ball centered at 0 in $\mathbb{R}^d$ of radius $s > 0$, that is $\mathbb{B}_s(d) := \{z \in \mathbb{R}^d : \|z\|_2 \le s\}$. Denote by $\mathsf{poly}(\cdot)$ some polynomial function of its input. The notation $\tilde{\mathcal{O}}(\cdot)$ suppresses additive lower-order terms and multiplicative logarithmic terms of the form $\ln(\cdot)$. In any inequality of the form $a \lesssim b$, the symbol $\lesssim$ means that the inequality holds up to a constant; define $\gtrsim$ similarly. We work with the class of functions $\mathcal{F} := \{f_\theta : \theta \in \Theta\}$ parameterized by elements of Θ. With $q \ge 1$ we assume $\mathcal{F}$ is in the L^q space with respect to the joint distribution of $x_1, \dots, x_T$. Let $\|\cdot\|_{\mathcal{F}}$ be the q-norm of a function; namely, for $f \in \mathcal{F}$ we define $\|f\|_{\mathcal{F}} := \frac{1}{T}\sum_{t\in[T]} \mathbb{E}_{x_t}[\|f(x_t)\|_q^q]^{1/q}$ (where x_t is independent of f). Since the precise values of q are not very relevant to our development, we hide the dependency on q in the notation $\|\cdot\|_{\mathcal{F}}$.

Data. We now describe our assumptions (see Table 1 for a summary). First, we assume our input x_1 and noise v_t are *sub-Gaussian* (see the appendix):

Assumption 1. x_1 is a d_x-dimensional *sub-Gaussian* vector with independent coordinates and *proxy variance* σ^2/d_x. Furthermore, noise v_t, when conditioned on $x_1, \dots, x_t$, is a d_y-dimensional sub-Gaussian vector with independent coordinates and proxy variance ν^2.

Bounded random variables and Gaussian random variables are all sub-Gaussian. It is harmless, though less general, to think of x_1 (*resp.* v_t) as a vector with i.i.d. Gaussian entries, each with mean 0 and variance σ^2/d_x (*resp.* ν^2).

Parameter and Function Space. We assume f^* is realizable in a bounded parameter space Θ:

Assumption 2. Parameter space $\Theta \subset \mathbb{R}^p$ is bounded with $\mathrm{diam}(\Theta) := \sup\{\|\theta - \theta'\|_2 : \theta, \theta' \in \Theta\} = \tilde{\mathcal{O}}(\mathsf{poly}(p))$, and the true predictor f^* is realizable in Θ, that is $f^* \in \mathcal{F}$.

Direct Difference Map. In (1), y_t is a function of input x_t. In (2), we have x_t as a function of x_1. We now describe the function that maps x_1 *directly* to y_t. With the identity mapping g_1, write $g_t \circ g_{t-1} \circ \cdots \circ g_1$ for the function that maps x_1 to x_t; this is for convenience and should not cause confusion, even though the output dimension of g_{i-1} does not match the input dimension of g_i. Then, the composition $f^* \circ g_t \circ g_{t-1} \circ \cdots \circ g_1$ is the direct input-output map in our model. Since we are interested in finding some function f that is close to f^*, we consider the following function, which we call *direct difference map*:

$$G_{f,t} := (f - f^*) \circ g_t \circ g_{t-1} \circ \cdots \circ g_1. \tag{5}$$

It maps x_1 to x_t and then takes the difference $f(x_t) - f^*(x_t)$. Note that if all g_i's are such that $x_1 = \cdots = x_t$, then $G_{f,t}$ is simply the difference $f - f^*$. More generally, $G_{f,t}$ takes task relationship into account as it is defined with respect to g_i's. This $G_{f,t}$ is crucial to our analysis, and we assume it has the following properties:

Assumption 3. We have $\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^4\right] < \infty$ and $\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^2\right] > 0$ $(\forall f \in \mathcal{F}\setminus\{f^*\}, t \in [T])$.

Assumption 4. For constants $L_{\mathcal{F}} > 0$, $L_G > 0$ we have

$$\|f_\theta - f_{\theta'}\|_{\mathcal{F}} \le L_{\mathcal{F}} \cdot \|\theta - \theta'\|_2, \quad \forall \theta, \theta' \in \Theta; \tag{6}$$
$$\|G_{f,t}(z) - G_{f,t}(z')\|_2 \le L_G \cdot \|z - z'\|_2, \ \ \forall z, z' \in \mathbb{R}^{d_x};$$

There is a *sufficiently large* number r_x such that for every $f, f' \in \mathcal{F}$ the following hold with $K_G = \tilde{\mathcal{O}}(\mathsf{poly}(r_x))$:

$$\sup_{z\in\mathbb{B}_{r_x}(d_x)} \|G_{f,t}(z) - G_{f',t}(z)\|_2 \le K_G \cdot \|f - f'\|_{\mathcal{F}};$$
$$\sup_{z\in\mathbb{B}_{r_x}(d_x)} \left|\|G_{f,t}(z)\|_2^2 - \|G_{f',t}(z)\|_2^2\right| \le K_G \|f - f'\|_{\mathcal{F}}.$$

Assumptions 1 to 3 are standard. Assumption 4 imposes Lipschitz-type conditions on our parameterized function f_θ and direct difference map $G_{f,t}$ and it requires r_x to be *sufficiently large*; we will show the precise values of r_x in the full version of our theorems in the appendix. Roughly speaking, (6) holds for deep networks with bounded parameters and bounded inputs, while the rest inequalities hold true as soon as all f and g_t's are Lipschitz continuous. Note that rotations and permutations aforementioned are Lipschitz continuous.

3.2 Recovery Guarantee 1: Weighted Replay With Data-Independent Regularization

In this and next two sections, we introduce our main theorems (see Table 2 for a summary).

Inspired by memory-based and regularization-based methods, we formulate the following problem:

$$\min_{\theta\in\Theta} \frac{1}{T} \sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t} \cdot \mathcal{L}\Big(y_{ti}, f_\theta(x_{ti})\Big) + \lambda \cdot \Omega_T(\theta). \tag{7}$$

Table 1: Summary of assumptions that we describe in Section 3.1 and use throughout the paper.

Assumption	Description
Assumption 1	Input and noise are sub-Gaussian with independent coordinates
Assumption 2	The parameter space is bounded and true predictor is realizable
Assumption 3	The direct difference map (5) has bounded and nondegenerate moments
Assumption 4	The direct difference map and parametrized predictors satisfy Lipschitz-type conditions

In (7), we use the squared loss $\mathcal{L}(y, \hat{y}) = \|\hat{y} - y\|_2^2$, as it is empirically useful in CL practice (McDonnell et al., 2023; Peng et al., 2025) and is a common objective of interest in CL theory. The factor $1/n_t$ normalizes the loss of each task. Each w_t is some non-negative weight for task t, and we assume $w_T > 0$, as we have full data for the current task T. Moreover, we assume that w_t's are hyperparameters, chosen by a user and independent of data. The function $\Omega_T(\cdot)$ serves the purpose of regularization, and it is weighted by some non-negative number λ. Thus, (7) amounts to minimizing a multitask loss with regularization $\Omega_T(\cdot)$. The generality of (7) stems from its flexibility to choose weights w_t and regularization $\Omega_T(\cdot)$. For example, *experience replay* or *rehearsal* amounts to solving (7) with $\lambda = 0$, while regularization-based methods often set $w_T = 1$ and set all previous weights to 0. Here, we assume $\Omega_T(\theta)$ is independent of data and noise, as is often the case in machine learning (Goodfellow et al., 2016, Chapter 7), and as considered in some CL algorithms (Kumar et al., 2025; Lewandowski et al., 2025). For our results on data-dependent regularizers, see Section 3.3.

Remark 2. To unify memory-based and regularization-based methods, Wang et al. (2024) considers a general formulation similar to (7). Wang et al. (2024) has a specific algorithmic focus and its formulation is not leveraged in full generality for theoretical developments, while we develop statistical recovery results for (7) and its variants.

We are ready to state our main result of this section:

Theorem 1. *Fix $\delta \in (0,1)$. Suppose Assumptions 1 to 4 hold. Recall that noise v_t is conditionally sub-Gaussian with proxy variance ν^2. Let $\hat{\theta}_T \in \Theta \subset \mathbb{R}^p$ be a global minimizer of (7) with regularization parameter λ satisfying $\lambda \lesssim \frac{1}{T}\max_{t\in[T]} \frac{w_t}{n_t}$. Define*

$$\kappa := \sup_{f\in\mathcal{F}\setminus\{f^*\}} \sup_{t\in[T], w_t>0} \frac{\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^4\right]^{1/2}}{\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^2\right]}. \tag{8}$$

Assume $n_t \geq \kappa^2 \cdot \tilde{\mathcal{O}}\left(p\ln(T) + \ln(1/\delta)\right)$ for all $t \in [T]$. With probability at least $1-\delta$ the weighted estimation error $\frac{1}{T}\sum_{t\in[T]} w_t \cdot \mathbb{E}\|f^*(x_t) - f_{\hat{\theta}_T}(x_t)\|_2^2$ is bounded above by

$$\tilde{\mathcal{O}}\left(\frac{\nu^2(p + \ln(1/\delta)) + \mathsf{poly}(\sigma)}{T} \cdot \max_{t\in[T]} \frac{w_t}{n_t}\right). \tag{9}$$

Theorem 1 is a finite sample guarantee with bound (9) depending on variances of data ($\mathsf{poly}(\sigma)$) and noise (ν^2), the number of tasks T, the number of of available samples n_t and weight w_t of task t (note that the bounds hides dependency on the dimensions d_x, d_y among other terms). We discuss these quantities next.

Proxy Variance of Data. Since both x_{it}'s and v_{it}'s are random, the estimation error depends on their proxy variances σ^2 and ν^2. First, the dependency on σ is $\mathsf{poly}(\sigma)$, which is because Assumption 4 introduces a polynomial dependency of r_x whose precise values depend on σ.

Weights and Sample Complexity. While prior work sets uniform weights $w_1 = \cdots = w_T > 0$ to derive statistical bounds in their CL settings (Lin et al., 2023; Friedman and Meir, 2024), we prove our bound in (9) with arbitrary non-negative weights w_t (independent of data). This allows us to set different weights and acquire different theoretical insights. For example, if we set all weights but w_t to 0, then (9) becomes a single-task bound

$$\mathbb{E}\|f^*(x_t) - f_{\hat{\theta}_T}(x_t)\|_2^2 \leq \begin{cases} \tilde{\mathcal{O}}\left(\frac{1}{m}\right) & T = t, \\ \tilde{\mathcal{O}}\left(\frac{1}{n_t}\right) & T > t. \end{cases}$$

The case $T = t$ holds as we have m samples at task t. We now see that when transiting from task t to task $t+1$, forgetting occurs as the bound transits from $\tilde{\mathcal{O}}(1/m)$ to $\tilde{\mathcal{O}}(1/n_t)$. However, starting from task $t+1$, the bound remains the same, thus no significant forgetting is furthermore entailed.

If we set $w_1 = \cdots = w_T > 0$, then Theorem 1 upper bounds the average error $\frac{1}{T}\sum_{t\in[T]} \mathbb{E}\|f^*(x_t) - f_{\hat{\theta}_T}(x_t)\|_2^2$ by $\tilde{\mathcal{O}}\left(\frac{1}{T\cdot\min_{t\in[T]} n_t}\right)$. For this bound to be informative, we need a *balanced* memory that stores approximately the same number of samples for each task, which is indeed the case when we construct the memory by random sampling or reservoir sampling (Remark 1). Furthermore, this is in agreement with common CL practice, where balanced memory tends to perform better than the imbalanced ones (Dokania et al., 2019; Prabhu et al., 2020; Araujo et al., 2022). Then, given balanced memory and as T increases, our result gets better, partly because the upper bound decreases, and partly because it provides bounds on the average error of more tasks. Finally, we note that, with non-uniform weights $w_t = \frac{Tn_t}{n_1+\cdots+n_T}$ that sum to 1, Theorem 1 gives the bound $\tilde{\mathcal{O}}\left(\frac{1}{n_1+\cdots+n_T}\right)$, which is with the optimal sample complexity $\tilde{\mathcal{O}}(1/(n_1+\cdots+n_T))$.

Are the above bounds too good to be true? Indeed, we have at most m samples starting (the samples of all tasks are deterministic transformations of the m samples from task 1); at first glance one might conclude that the sample complexity lower bound would be $\tilde{\mathcal{O}}(1/m)$, but Theorem 1 gives a tighter result. This is not a contradiction, the reason being that Theorem 1 assumes sufficiently many samples with $m \geq n_t \geq \kappa^2 \cdot \tilde{\mathcal{O}}\left(p\ln(T) + \ln(1/\delta)\right)$, all tasks share a common predictor f^*, and we consider the regression formulation. For example, if f^* is linear, then with $m \geq p$ linearly independent (random) samples the regression problem has a unique global minimizer f^*. In this case, we can recover the true f^* exactly, which gives us the zero bound instead of $\tilde{\mathcal{O}}(1/m)$.

Condition on Sufficiently Many Samples. Theorem 1 assumes $n_t \geq \kappa^2 \cdot \tilde{\mathcal{O}}\left(p\ln(T) + \ln(1/\delta)\right)$, that is, every task stores sufficiently many samples. To understand what this assumption imposes, it suffices to analyze the role of κ. While computing the numerical values of κ is difficult, from Assumption 3 we know κ is finite. More can be said if we make extra assumptions on how tasks are dependent (note that κ depends on $G_{f,t}$ and $G_{f,t}$ encodes the information about task dependency):

Example 3. Assume the function class $\mathcal{F}$ consists of linear functions and all entries of x_1 are i.i.d. Gaussian, sampled from $\mathcal{N}(0,1)$. Consider two tasks with $g_2 : x_1 \mapsto 10^{10}x_1$. With two functions $f, f^* \in \mathcal{F}$ fixed and parameterized by θ, θ^*, respectively, we have $G_{f,t}(x_1) = (\theta - \theta^*)^\top x_1$. Since $G_{f,t}(x_1)$ is also a zero-mean Gaussian, a standard calculation of its fourth and second moments gives $\kappa^2 = 3$. Note that the scaling factor 10^{10} has no effect in κ.

We have $\kappa^2 \leq 3$ if Gaussianity of Example 3 is replaced with *strict sub-Gaussianity*. Finally, we note that κ is upper bounded by some universal constant for sub-Gaussian data.

Robustness to Distant Distributions. Example 3 generalizes to multiple tasks with g_t's being of the form $x_1 \mapsto s_t x_t$ for some s_t. Such a simple example of g_t's can in fact make a difficult case for analysis. Indeed, it affects the Lipschitz parameter L_G and radius r_x in Assumption 4. That said, all of our bounds exhibit only a logarithmic dependency on L_G, r_x, so our theorems allow s_t to be a polynomial function of problem parameters. The second difficulty Example 3 brings manifests itself, not in our theorem, but in prior work. Indeed, for large s_t, the data distributions of two tasks, $\mathcal{N}(0, I_{d_x})$ and $\mathcal{N}(0, s_t^2 I_{d_x})$, could be very *distant*; here I_{d_x} is the $d_x \times d_x$ identity matrix. This distance has fundamental impacts on many statistical bounds that depend *linearly* on this distance:

Remark 3. The theory of Mansour et al. (2009) on *domain adaptation* bounds statistical errors of the second task (i.e., *target task*) by the *discrepancy distance* between distributions π_1, π_2 of the first task (*source task*) and the second task. The discrepancy $\mathrm{dist}(\pi_1, \pi_2)$, defined as

$$\sup_{\theta_1, \theta_2 \in \Theta} \left| \mathbb{E}_{\pi_1} \mathcal{L}\left(f_{\theta_1}(z), f_{\theta_2}(z)\right) - \mathbb{E}_{\pi_2} \mathcal{L}\left(f_{\theta_1}(z), f_{\theta_2}(z)\right) \right|,$$

measures the similarities between π_1, π_2. If $\mathrm{dist}(\pi_1, \pi_2)$ is small, we might expect some benefits of learning task 1 prior to task 2 (e.g., small errors on task 2). This basic intuition is extended to various settings, including online multitask learning, transfer learning, and continual learning (Mohri and Muñoz Medina, 2012; Wang et al., 2023; Ye and Bors, 2022). That said, the discrepancy can be large in Example 3 with the squared loss $\mathcal{L}(\cdot,\cdot)$. Indeed, set $\pi_1 = \mathcal{N}(0, I_{d_x})$ and $\pi_2 = \mathcal{N}(0, s^2 I_{d_x})$ with $s \gg 1$, and one verifies $\mathrm{dist}(\pi_1, \pi_2) = (s^2 - 1) \cdot \mathrm{diam}(\Theta)^2$, that is, the distance grows unbounded as s increases; their corresponding bound is vacuous in this scenario. In consequence, the proof based on such a discrepancy is unable to handle simple yet distant tasks. For a similar reason, the recent PAC-Bayes bound of Friedman and Meir (2024) easily becomes vacuous due to the presence of a similar discrepancy term.

Our final note on Theorem 1 is a comparison to the experience replay theory of Peng et al. (2023a).

Remark 4. Theorem 3 of Peng et al. (2023a) provides the bound $\tilde{\mathcal{O}}(\frac{1}{n_t})$ on the excess risk of (7) with $\lambda = 0$ for each task t. This directly implies the bound $\tilde{\mathcal{O}}(\frac{\log T}{T} \sum_{t \in [T]} \frac{1}{n_t})$ for the average joint loss. This bound becomes $\tilde{\mathcal{O}}(\log T / n_t)$ when all n_t's are of the same order. In contrast, we directly work with the average estimation error, and our bound has an $\tilde{\mathcal{O}}(1/T)$ dependency on T.

3.3 Recovery Guarantee 2: Data-Dependent Regularization and Knowledge Distillation

Theorem 1 in Section 3.2 assumes the regularizer is independent of data. On the other hand, it is not uncommon to utilize data-dependent regularizers for CL. A typical data-dependent regularizer used in CL is *knowledge distillation*, that is to match the (intermediate) outputs of the current network f_θ and the previously learned network $f_{\hat{\theta}_t}$ at stored samples. This motivates the following CL method:

- (*Step 1*) Set hyperparameter $\beta_T > 0$. Then solve

$$\hat{\theta}_T \in \operatorname*{argmin}_{\theta \in \Theta} \beta_T \sum_{i \in [m]} \mathcal{L}\Big(y_{Ti}, f_\theta(x_{Ti})\Big) + \Omega_T(\theta), \quad (10)$$

where $\Omega_T(\cdot)$ is either of the following two regularizers:

$$\Omega_T(\theta) = \sum_{t \in [T-1]} \sum_{i \in \mathcal{R}_t} \beta_t \cdot \|f_\theta(x_{ti}) - f_{\hat{\theta}_{T-1}}(x_{ti})\|_2^2 \quad (11)$$

$$\Omega_T(\theta) = \sum_{t \in [T-1]} \sum_{i \in \mathcal{R}_t} \beta_t \cdot \|f_\theta(x_{ti}) - f_{\hat{\theta}_t}(x_{ti})\|_2^2 \quad (12)$$

- (*Step 2*) Choose indices $\mathcal{R}_T \subset [m]$. Store $\{(x_{Ti}, y_{Ti})\}_{i \in \mathcal{R}_T}$. Increase T. Go back to *Step 1*.

The above method are with regularizer (11) or (12) to compute $\hat{\theta}_1, \ldots, \hat{\theta}_T$ continually. Regularizer (11) aims to match f_θ and $f_{\hat{\theta}_{T-1}}$ on all previous tasks, while regularizer (12) aims to match f_θ and $f_{\hat{\theta}_t}$ on each task t ($\forall t \in [T-1]$). In both cases, it is understood that task 1 is solved without regularization and that the algorithm consistently chooses either (11) or (12). Computationally, for (12) we can store $f_{\hat{\theta}_t}(x_{ti})$ after training task t, while for (11) we need to compute $f_{\hat{\theta}_{T-1}}(x_{ti})$ after training task $T-1$. For both regularizers, we have the following theorem.

Theorem 2. *Fix $\delta \in (0,1)$. Let κ be as in (8). Suppose Assumptions 1 to 4 hold. Recall that noise v_t is conditionally sub-Gaussian with proxy variance ν^2. Assume $n_t \geq \kappa^2 \cdot \tilde{\mathcal{O}}\left(p \ln(T) + \ln(1/\delta)\right)$ for all $t \in [T]$. If $\hat{\theta}_1, \ldots, \hat{\theta}_T \in \Theta \subset \mathbb{R}^p$ are global minimizers of (11) with $\beta_t = 1/4^{T-t}$. With probability at least $1-\delta$ the weighted error $\sum_{t\in[T]} \frac{\beta_t n_t}{\sum_{t\in[T]} \beta_t n_t} \cdot \mathbb{E}\|f^*(x_t) - f_{\hat{\theta}_T}(x_t)\|_2^2$ is bounded above by*

$$\tilde{\mathcal{O}}\left(\frac{\nu^2(pT + \ln(1/\delta)) + \mathsf{poly}(\sigma)}{\sum_{t\in[T]} \beta_t n_t}\right). \tag{13}$$

If $\hat{\theta}_1, \ldots, \hat{\theta}_T$ are solutions to (12), then with probability at least $1-\delta$ we have (13) holds as well.

In Theorem 2, the error term that we bound has its weights on distances $\|f^*(x_t) - f_{\hat{\theta}_T}(x_t)\|_2^2$ decay exponentially as t decreases, meaning that (13) carries exponentially less control over the distances on past tasks. Therefore, Theorem 2 is informative for a small number of the most recent tasks, and to some extent, it illuminates the limits of regularization-based methods in combating forgetting. In fact, this bound is worse than that of Theorem 1 for experience replay. This aligns with empirical observations that experience replay typically outperforms regularization-based approaches; for example, see Table 4 of (Prabhu et al., 2020), which compares *LwF* (Li and Hoiem, 2017) and *GDumb* (Prabhu et al., 2020).

On a technical note, bound (13) itself could be sub-optimal for two reasons. First, since the algorithm analyzed involves all predictors $f_{\hat{\theta}_1}, \ldots, f_{\hat{\theta}_T}$, all of which depend on data, we have to run an ε-net argument on the product space Θ^T in $\mathbb{R}^{pT}$, which brings the dependency pT. Second, the exponential forgetting phenomenon is due to our choice $\beta_t = 1/4^{T-t}$. This choice is crucial, as it prevents the bound from getting exponentially large. Improving Theorem 2 in these aspects is left for future work.

We finish the section by remarking on existing theoretical contributions on regularization-based methods:

Remark 5. The analysis of Heckel (2022); Li et al. (2023); Zhao et al. (2024); Levinstein et al. (2025) relies on linear models or the kernel regime. The analysis of Zhu et al. (2025) relies on either the assumption of shared global minimizers or a noisy linear regression model. The setting of Yin et al. (2020) is general, but the error bound of their Theorem 4 contains the distance of the form $\|\hat{\theta}_T - \hat{\theta}_t\|_2$. This is a random variable whose dependency on problem parameters is unclear. In contrast, our bound in (13) contains only parameters related to the problem configuration. We achieve this via several key inequalities that give rise to a recurrence relation, which we unroll to eliminate all random variables related to $\hat{\theta}_t$.

3.4 Recovery Guarantee 3: Data-Dependent Weights and Constrained Optimization

Our formulation (7) in Section 3.2 assumes the weights are independent of data. Here, we consider

$$\hat{\theta}_T \in \operatorname*{argmin}_{\theta\in\Theta} \frac{1}{T} \sum_{(t,i)\in\mathcal{M}} \frac{\tilde{w}_t}{n_t} \cdot \mathcal{L}\Big(y_{ti}, f_\theta(x_{ti})\Big), \tag{14}$$

where weight $\tilde{w}_t$ is a random variable depending on all data and noise. While we allowed zero weights in Section 3.2 to study regularization-based methods without replay, here we assume $\tilde{w}_t \neq 0$ ($\forall t \in [T]$), or otherwise formulation (14) reduces to a case with fewer tasks. Then, dividing the smallest weight if necessary, we assume $\tilde{w}_t \in [1, W]$.

There are multiple ways to choose $\hat{w}_t$ in a data-dependent fashion. For example, set $\hat{w}_t$ to be proportional (or inversely proportional) to the loss of task t (in light of *boosting* (Schapire, 1990; Ramesh and Chaudhari, 2022; Wang et al., 2023) or *iteratively reweighted least-squares* (Daubechies et al., 2010; Kümmerle et al., 2021; Peng et al., 2022, 2023b)), or solve a bilevel program in weight variables that represent *coresets* (Borsos et al., 2020), or set the weights to the values of dual variables which arise in primal-dual methods for constrained learning (Lopez-Paz and Ranzato, 2017; Chamon et al., 2022; Peng et al., 2023a; Elenter et al., 2023; Li et al., 2024).

We now give recovery guarantees for (14):

Theorem 3. *Fix $\delta \in (0,1)$. Let κ be as in (8). Suppose Assumptions 1 to 4 hold. Recall that noise v_t is conditionally sub-Gaussian with proxy variance ν^2 and that $\tilde{w}_t \in [1, W]$. Let $\hat{\theta}_T \in \Theta \subset \mathbb{R}^p$ be a global minimizer of (14). Assume $n_t \geq \kappa^2 \cdot \tilde{\mathcal{O}}\left(p \ln(T) + \ln(1/\delta)\right)$ for all $t \in [T]$ and $W \leq 1 + \frac{1}{Tn_t}$. With probability at least $1-\delta$, the average estimation error $\frac{1}{T}\sum_{t\in[T]} \mathbb{E}\|f^*(x_t) - f_{\hat{\theta}_T}(x_t)\|_2^2$ is bounded above by*

$$\tilde{\mathcal{O}}\left(\frac{\nu^2(p + \ln(1/\delta)) + \max\{\mathsf{poly}(\sigma), \ln(\sigma T)\}}{T \min_{t\in[T]} n_t}\right). \tag{15}$$

The bound of Theorem 3 follows from that of Theorem 1, and in fact they are identical except assumption $W \leq 1 + \frac{1}{Tn_t}$. This extra condition is what we pay for the case of data-dependent weights. Due to a number of differences in the settings, Theorem 3 is not directly comparable to

theoretical results of Li et al. (2024); Peng et al. (2023a); Chamon et al. (2022). However, a distinguishing aspect is that, as T gets large, our bound would in general get better, while their bounds deteriorate. This is because we explicitly model the dependency between tasks and take advantage of it in the proofs, while their proofs do not consider task dependency and need instead to apply T concentration inequalities, one for each task (Example 2).

4 CONCLUSION

Inspired by nonlinear dynamical systems and several other topics, we formalized a notion of task dependency for continual learning of nonlinear regression tasks. Building upon it, we developed some recovery guarantees for commonly used CL methods that involve replay, regularization (e.g., knowledge distillation), and data-dependent weighting. The estimation error bounds we derived are well-behaved, as they diminish as the number of samples tends to infinity or for sufficiently small noise. The key to proving our theorems is a careful balance we maintain between the generality of the problem formulation and the tightness of the resulting bounds; put differently, this can also be viewed as a limitation: our proof framework does not support deriving stronger guarantees for more general scenarios where different tasks are dependent arbitrarily (Fig. 2). The other limitation is this: Our results are based on the assumption that the number of stored samples per task exceeds the dimension of the parameter space. It will be of interest to relax this assumption and furthermore extend our results to the overparameterized regime.

In the future, we plan to improve the bounds for regularization-based methods (Theorem 2) by exploring more specific settings such as linear models (cf. Remark 5), to extend our Theorem 3 with data-dependent weights within the constrained learning framework of Chamon et al. (2022) that guarantees the feasibility of constraints, and to extend the insights of the paper here for other CL paradigms (e.g., expansion-based methods). Also, note that our analysis is independent of the algorithmic choice, and in the future we plan to analyze the effect of the algorithm on the recovery errors as well.

Acknowledgement

This work is supported by the National Science Foundation (grants 2031985), the Simons Foundation (grant 814201), and the Office of Naval Research (grant 503405-78051).

References

Aljundi, R., Lin, M., Goujaud, B., and Bengio, Y. (2019). Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems*. 1

Alquier, P. et al. (2024). User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303. 1

Araujo, V., Balabin, H., Hurtado, J., Soto, A., and Moens, M.-F. (2022). How relevant is selective memory population in lifelong language learning? In *Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 3, 6

Argyriou, A., Evgeniou, T., and Pontil, M. (2006). Multi-task feature learning. *Advances in Neural Information Processing Systems*. 4

Arun, K. S., Huang, T. S., and Blostein, S. D. (1987). Least-squares fitting of two 3D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):698–700. 4

Banayeeanzade, M., Soltanolkotabi, M., and Rostami, M. (2025). Theoretical insights into overparameterized models in multi-task and replay-based continual learning. *Transactions on Machine Learning Research*. 1

Bang, J., Kim, H., Yoo, Y., Ha, J.-W., and Choi, J. (2021). Rainbow memory: Continual learning with a memory of diverse samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1

Baxter, J. (2000). A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198. 4

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, 79:151–175. 33

Ben-David, S. and Borbely, R. S. (2008). A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine Learning*, 73(3):273–287. 4

Bennani, M. A. and Sugiyama, M. (2020). Generalisation guarantees for continual learning with orthogonal gradient descent. In *4th Lifelong Machine Learning Workshop at ICML 2020*. 1

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. 4

Borsos, Z., Mutny, M., and Krause, A. (2020). Coresets via bilevel optimization for continual learning and streaming. *Advances in Neural Information Processing Systems*. 3, 8

Buzzega, P., Boschini, M., Porrello, A., Abati, D., and Calderara, S. (2020). Dark experience for general continual learning: a strong, simple baseline. *Advances in Neural Information Processing Systems*. 1

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75. 4

Chamon, L. F., Paternain, S., Calvo-Fullana, M., and Ribeiro, A. (2022). Constrained learning with non-convex losses. *IEEE Transactions on Information Theory*. 2, 8, 9

Chaudhry, A., Ranzato, M., Rohrbach, M., and Elhoseiny, M. (2019). Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*. 1

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49. 39

Dar, Y., LeJeune, D., and Baraniuk, R. G. (2024). The common intuition to transfer learning can win or lose: Case studies for linear regression. *SIAM Journal on Mathematics of Data Science*, 6(2):454–480. 14

Daubechies, I., DeVore, R., Fornasier, M., and Güntürk, C. S. (2010). Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38. 8

Ding, M., Ji, K., Wang, D., and Xu, J. (2024). Understanding forgetting in continual learning with linear regression. In *International Conference on Machine Learning*. 1

Doan, T., Bennani, M. A., Mazoure, B., Rabusseau, G., and Alquier, P. (2021). A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. In *International Conference on Artificial Intelligence and Statistics*. 1

Dokania, P., Torr, P., and Ranzato, M. (2019). Continual learning with tiny episodic memories. In *Workshop on Multi-Task and Lifelong Reinforcement Learning*. 1, 3, 6

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. (2021). Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*. 4

Elenter, J., NaderiAlizadeh, N., Javidi, T., and Ribeiro, A. (2023). Primal dual continual learning: Balancing stability and plasticity through adaptive memory allocation. Technical report, arXiv:2310.00154v2 [cs.LG]. 1, 2, 3, 8, 14

Evron, I., Levinstein, R., Schliserman, M., Sherman, U., Koren, T., Soudry, D., and Srebro, N. (2026). From continual learning to SGD and back: Better rates for continual linear models. In *International Conference on Algorithmic Learning Theory*. 1

Evron, I., Moroshko, E., Ward, R., Srebro, N., and Soudry, D. (2022). How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*. 1, 3, 14

Friedman, L. and Meir, R. (2024). Data-dependent and oracle bounds on forgetting in continual learning. Technical report, arXiv:2406.09370v2 [cs.LG]. 1, 2, 6, 7

Goldfarb, D., Evron, I., Weinberger, N., Soudry, D., and Hand, P. (2024). The joint effect of task similarity and overparameterization on catastrophic forgetting — an analytical model. In *International Conference on Learning Representations*. 1, 4

Goldfarb, D. and Hand, P. (2023). Analysis of catastrophic forgetting for random orthogonal transformation tasks in the overparameterized regime. In *International Conference on Artificial Intelligence and Statistics*. 4

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. 6

Hartley, R. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press. 4

Heckel, R. (2022). Provable continual learning via sketched Jacobian approximations. In *International Conference on Artificial Intelligence and Statistics*. 1, 2, 8

Karpel, G., Moroshko, E., Levinstein, R., Meir, R., Soudry, D., and Evron, I. (2026). Optimal L2 regularization in high-dimensional continual linear regression. In *International Conference on Algorithmic Learning Theory*. 1

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of The National Academy of Sciences*, 114(13):3521–3526. 1

Kumar, S., Marklund, H., and Van Roy, B. (2025). Maintaining plasticity in continual learning via regenerative regularization. In *Conference on Lifelong Learning Agents*, pages 410–430. PMLR. 6

Kümmerle, C., Mayrink Verdun, C., and Stöger, D. (2021). Iteratively reweighted least squares for basis pursuit with global linear convergence rate. *Advances in Neural Information Processing Systems*. 8

Levinstein, R., Attia, A., Schliserman, M., Sherman, U., Soudry, D., Koren, T., and Evron, I. (2025). Optimal rates in continual linear regression via increasing regularization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 8

Lewandowski, A., Bortkiewicz, M., Kumar, S., György, A., Schuurmans, D., Ostaszewski, M., and Machado, M. C. (2025). Learning continually by spectral regularization. In *International Conference on Learning Representations*. 6

Li, G., Yu, W., Yao, Y., Tong, W., Liang, Y., Lin, Q., and Yang, T. (2024). Model developmental safety: A retention-centric method and applications in vision-language models. Technical report, arXiv:2410.03955 [cs.LG]. 1, 8, 9

Li, H., Wu, J., and Braverman, V. (2023). Fixed design analysis of regularization-based continual learning. In *Conference on Lifelong Learning Agents*. 1, 2, 3, 8, 14

Li, Z. and Hoiem, D. (2017). Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947. 1, 8

Lin, S., Ju, P., Liang, Y., and Shroff, N. (2023). Theory on forgetting and generalization of continual learning. In *International Conference on Machine Learning*. 1, 6

Lopez-Paz, D. and Ranzato, M. (2017). Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems*. 1, 8

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. In *Proceedings of The 22nd Annual Conference on Learning Theory (COLT 2009)*, Montréal, Canada. 2, 7

Matni, N. and Tu, S. (2019). A tutorial on concentration bounds for system identification. In *2019 IEEE 58th conference on decision and control (CDC)*, pages 3741–3749. IEEE. 33

Maurer, A. and Pontil, M. (2021). Concentration inequalities under sub-gaussian and sub-exponential conditions. *Advances in Neural Information Processing Systems*, 34:7588–7597. 33

Maurer, A., Pontil, M., and Romera-Paredes, B. (2016). The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32. 4

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier. 1

McDonnell, M. D., Gong, D., Parvaneh, A., Abbasnejad, E., and van den Hengel, A. (2023). RanPAC: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*. 6

Mohri, M. and Muñoz Medina, A. (2012). New analysis and algorithm for learning with drifting distributions. In *International Conference on Algorithmic Learning Theory*. 7, 34

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*. MIT Press. 1

Park, D., Hong, S., Han, B., and Lee, K. M. (2019). Continual learning by asymmetric loss approximation with single-side overestimation. In *IEEE/CVF International Conference on Computer Vision*. 1

Peng, B. and Risteski, A. (2022). Continual learning: A feature extraction formalization, an efficient algorithm, and fundamental obstructions. In *Advances in Neural Information Processing Systems*. 14

Peng, L., Elenter, J., Agterberg, J., Ribeiro, A., and Vidal, R. (2025). TSVD: Bridging theory and practice in continual learning with pre-trained models. In *International Conference on Learning Representations*. 6

Peng, L., Giampouras, P., and Vidal, R. (2023a). The ideal continual learner: An agent that never forgets. In *International Conference on Machine Learning*. 1, 2, 3, 7, 8, 9, 14

Peng, L., Kümmerle, C., and Vidal, R. (2022). Global linear and local superlinear convergence of IRLS for non-smooth robust regression. In *Advances in Neural Information Processing Systems*. 8

Peng, L., Kümmerle, C., and Vidal, R. (2023b). On the convergence of IRLS and its variants in outlier-robust estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8

Peng, L. and Vidal, R. (2025). Mathematics of continual learning. Technical report, arXiv:2504.17963 [cs.LG]. 14

Pentina, A. and Lampert, C. H. (2015). Lifelong learning with non-iid tasks. *Advances in Neural Information Processing Systems*. 1

Prabhu, A., Torr, P. H., and Dokania, P. K. (2020). GDumb: A simple approach that questions our progress in continual learning. In *European Conference on Computer Vision*. 1, 6, 8

Ramesh, R. and Chaudhari, P. (2022). Model zoo: A growing brain that learns continually. In *International Conference on Learning Representations*. 4, 8

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. (2017). iCaRL: Incremental classifier and representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1

Rigollet, P. and Hütter, J.-C. (2023). High-dimensional statistics. Technical report, arXiv:2310.19244 [math.ST]. 37

Robins, A. (1993). Catastrophic forgetting in neural networks: The role of rehearsal mechanisms. In *The First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*. 1

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, 5:197–227. 8

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. 1

Shin, H., Lee, J. K., Kim, J., and Kim, J. (2017). Continual learning with deep generative replay. *Advances in Neural Information Processing Systems*. 1

Shui, C., Abbasi, M., Robitaille, L.-É., Wang, B., and Gagné, C. (2019). A principled approach for learning task similarity in multitask learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3446–3452. 33

Sun, S., Calandriello, D., Hu, H., Li, A., and Titsias, M. (2022). Information-theoretic online memory selection for continual learning. In *International Conference on Learning Representations*. 3

Swartworth, W., Needell, D., Ward, R., Kong, M., and Jeong, H. (2023). Nearly optimal bounds for cyclic forgetting. *Advances in Neural Information Processing Systems*. 1

Tadipatri, U. K. R., Haeffele, B. D., Agterberg, J., Ziemann, I., and Vidal, R. (2025). Nonconvex linear system identification with minimal state representation. In Ozay, N., Balzano, L., Panagou, D., and Abate, A., editors, *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pages 1286–1299. PMLR. 4

Tripuraneni, N., Jin, C., and Jordan, M. (2021). Provable meta-learning of linear representations. In *International Conference on Machine Learning*. 4

Tripuraneni, N., Jordan, M., and Jin, C. (2020). On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*. 4

Tu, S., Frostig, R., and Soltanolkotabi, M. (2024). Learning from many trajectories. *Journal of Machine Learning Research*, 25(216):1–109. 4

van de Ven, G. M., Tuytelaars, T., and Tolias, A. S. (2022). Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197. 4

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press. 37, 38, 39

Verwimp, E., De Lange, M., and Tuytelaars, T. (2021). Rehearsal revealed: The limits and merits of revisiting samples in continual learning. In *IEEE/CVF International Conference on Computer Vision*. 1

Vitter, J. S. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11(1):37–57. 3

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press. 37

Wang, B., Mendez, J. A., Shui, C., Zhou, F., Wu, D., Xu, G., Gagné, C., and Eaton, E. (2023). Gap minimization for knowledge sharing and transfer. *Journal of Machine Learning Research*, 24(33):1–57. 7, 8, 34

Wang, L., Zhang, X., Yang, K., Yu, L., Li, C., Hong, L., Zhang, S., Li, Z., Zhong, Y., and Zhu, J. (2022). Memory replay with data compression for continual learning. In *International Conference on Learning Representations*. 1

Wang, Z., Li, Y., Shen, L., and Huang, H. (2024). A unified and general framework for continual learning. In *International Conference on Learning Representations*. 6

Ye, F. and Bors, A. G. (2022). Task-free continual learning via online discrepancy distance learning. In *Advances in Neural Information Processing Systems*. 1, 2, 7

Yin, D., Farajtabar, M., Li, A., Levine, N., and Mott, A. (2020). Optimization and generalization of regularization-based continual learning: a loss approximation viewpoint. Technical report, arXiv:2006.10974v3 [cs.LG]. 1, 2, 8

Zenke, F., Poole, B., and Ganguli, S. (2017). Continual learning through synaptic intelligence. In *International Conference on Machine Learning*. 1

Zhao, X., Wang, H., Huang, W., and Lin, W. (2024). A statistical theory of regularization-based continual learning. In *International Conference on Machine Learning*. 1, 2, 3, 8, 14

Zhu, F., Liu, Y., Liu, W., and Zhang, Z. (2025). Global convergence of continual learning on non-iid data. Technical report, arXiv:2503.18511 [cs.LG]. 2, 3, 8, 14

Ziemann, I., Tsiamis, A., Lee, B., Jedra, Y., Matni, N., and Pappas, G. J. (2023). A tutorial on the non-asymptotic theory of system identification. In *IEEE Conference on Decision and Control*. 4

Ziemann, I., Tsiamis, A., Lee, B., Jedra, Y., Matni, N., and Pappas, G. J. (2024a). A Tutorial on the Non-Asymptotic Theory of System Identification. arXiv:2309.03873. 27

Ziemann, I. and Tu, S. (2022). Learning with little mixing. *Advances in Neural Information Processing Systems*. 4

Ziemann, I., Tu, S., Pappas, G. J., and Matni, N. (2024b). Sharp rates in dependent learning theory: Avoiding sample size deflation for the square loss. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62779–62802. PMLR. 33

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

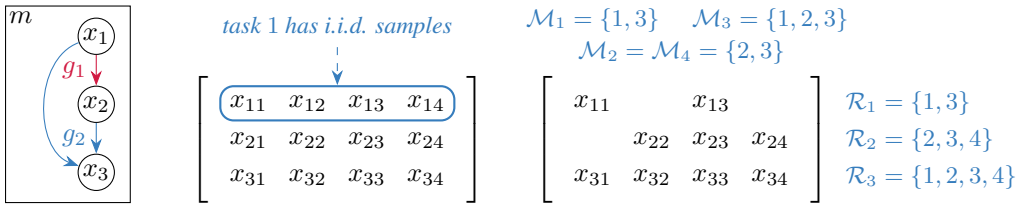
Recovery Guarantees for Continual Learning of Dependent Tasks: Memory, Data-Dependent Regularization, and Data-Dependent Weights: Supplementary Materials

In this material, we will present proofs of results, examples, related work, extra discussion on extensions, and preliminaries for the main draft. The below is the table of contents of this material:

Contents

Table 2: Summary of theorems that we describe in Sections 3.2 to 3.4.

Theorem	Description
Theorem 1	Recovery error bound for weighted replay with data-independent regularization
Theorem 2	Recovery error bound for data-dependent regularization
Theorem 3	Recovery error bound for data-dependent weights



(a) data dependency (b) full data (c) training data available at task 3

Figure 3: Example of our setup ($T = 3, m = 4$): 1a shows m i.i.d. trajectories generated by $x_t = g_t(x_1, \dots, x_{t-1})$; 1b shows full data in a matrix, where columns represent trajectories and each row represents data of each task; 1c shows the index sets $\mathcal{R}_t, \mathcal{M}_i$ and data available at task 3.

A Extra Details on Main Paper

A.1 Extra Notations and Figures

In our proof, we will need to index our data by columns, thus we extend Fig. 1 and include examples of the index sets $\mathcal{M}_i \subset [T]$ for each column i. Specifically, the i-th column of this partial matrix in Fig. 3c corresponds to a partial *trajectory* of samples $\{(x_{ti}, y_{ti})\}_{t \in \mathcal{M}_i}$.

A.2 On Task-Specific Predictors

At first glance, (1) seems to ask all tasks to share a common true predictor f^*, similarly to prior work (Evron et al., 2022; Peng and Risteski, 2022; Peng et al., 2023a; Li et al., 2023; Zhao et al., 2024; Elenter et al., 2023; Zhu et al., 2025). But the extra presence of (2) on top of (1) recovers a case where each task t possesses its own predictor f_t^*:

Example 4. Consider the following model:

$$\begin{aligned} y_t &= f_t^*(\overline{x}_t) + v_t, \;\; \overline{x}_t = \overline{g}_t(\overline{x}_1, \dots, \overline{x}_{t-1}), \\ f_t^* &= f_{t-1}^* \circ h_t. \end{aligned} \tag{16}$$

In (16), the first two equations are identical to (1) and (2), despite the different notations $\overline{x}_t$ and $\overline{g}_t$. Moreover, (16) contains an extra equation $f_t^* = f_{t-1}^* \circ h_t$, where h_t is some *known* invertible mapping capturing a relationship between the predictors of two consecutive tasks. Let h_1 be the identity mapping, and write $h_{1:t} := h_1 \circ \cdots \circ h_t$ and $x_t := h_{1:t}(\overline{x}_t)$. By basic algebra we eliminate the presence of $f_t^*, \overline{x}_t$ in (16), and obtain $y_t = f_1^*(x_t) + v_t$ and $x_t = h_{1:t}(\overline{g}_t(h_1^{-1}(x_1), \dots, h_{1:t-1}^{-1}(x_{t-1})))$. The former equation is identical to (1) with $f^* = f_1^*$, and the latter equation on x_t defines g_t in (2). Thus, estimating f_1^* in (16) reduces to estimating f^* in (1) and (2) with a particular choice of g_t. Moreover, once f_1^* is estimated, we estimate all other f_t's via $f_t^* = f_{t-1}^* \circ h_t$. Conversely, our model reduces to (16) when h_t's are all identity mappings. Thus, when h_t is known and invertible, imposing task dependency on predictors (16) is equivalent to imposing such dependency on input samples (2).

Remark 6. Consider two tasks with linear predictors f_1^*, f_2^* parameterized by θ_1^*, θ_2^* respectively, and assume they are related by some invertible matrix H_2, that is $\theta_2^* = H_2\theta_1^*$. Since such H_2 always exists, if H_2 is unknown, then the relationship $\theta_2^* = H_2\theta_1^*$ gives us no extra information for learning either θ_1^* or θ_2^*. Thus, to understand the benefit of having $\theta_2^* = H_2\theta_1^*$, we consider H_2 is known. Furthermore, in Peng and Vidal (2025); Dar et al. (2024), the relationship $\theta_2^* = H_2\theta_1^*$ is assumed to hold up to some additive Gaussian noise. In Peng and Vidal (2025), it is shown that Kalman filtering and smoothing improve the performance on task 1 after learning task 2, an example of *positive backward transfer*. In Dar et al. (2024), a transfer learning setup is considered, and it is shown that learning task 1 would benefit learning task 2 in the underparameterized case but might hurt otherwise (Dar et al., 2024, Fig. 1).

A.3 On Assumption 4

We claimed in the main paper that Assumption 4 holds true as soon as all f and g_t's are Lipschitz continuous. A more precise statement of it is shown below in Lemma 1.

Lemma 1. *Suppose that each transformation g_i is Lipschitz continuous with constant $L_{g,i}$ ($i \in [T]$). Assume $\mathcal{F}$ consist of functions $f : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ such that $f(z) < \infty$ for any finite input z. Consider the ball $\mathbb{B}_{r'}(d_x)$ of radius $r' := r \cdot \max_{t\in[T]} \prod_{i=1}^{t} L_{g,i}$. Suppose Assumption 2 and (6) hold with Lipschitz parameter $L_{\mathcal{F}}$ and with the following norm $\|\cdot\|_{\mathcal{F}}$:*

$$\|f\|_{\mathcal{F}} := \sup_{z\in\mathbb{B}_{r'}(d_x)} \|f(z)\|_2.$$

Furthermore, assume f is Lipschitz continuous with respect to its input with also constant $L_{\mathcal{F}}$. Then rest conditions in Assumption 4 hold with

$$L_G = 2L_{\mathcal{F}} \max_{t\in[T]} \prod_{i=1}^{t} L_{g,i}, \qquad K_G = \max\left(1, 2rL_G + 2C_{\max}\right).$$

Proof. First note that we have (for all $z, z' \in \mathbb{R}^{d_x}$, for all $f \in \mathcal{F}$, and for all $t \in [T]$)

$$\begin{aligned} \|G_{f,t}(z) - G_{f,t}(z')\|_2 &= \|(f - f^*) \circ g_t \circ \cdots \circ g_1(z) - (f - f^*) \circ g_t \circ \cdots \circ g_1(z')\|_2 \\ &\le 2L_{\mathcal{F}} \cdot \|g_t \circ \cdots \circ g_1(z) - g_t \circ \cdots \circ g_1(z')\|_2 \\ &\le 2L_{\mathcal{F}} L_{g,1} \cdot \|g_{t-1} \circ \cdots \circ g_1(z) - g_{t-1} \circ \cdots \circ g_1(z')\|_2 \\ &\le 2L_{\mathcal{F}} \prod_{i=1}^{t} L_{g,i} \cdot \|z - z'\|_2 . \end{aligned} \tag{17}$$

Therefore, $L_G = 2L_{\mathcal{F}} \max_{t\in[T]} \prod_{i=1}^{t} L_{g,i}$ in this case.

On the other hand, for any $f, f' \in \mathcal{F}$ parametrized by θ, θ' respectively, and for any $z \in \mathbb{B}_r(d_x)$, we have $g_t \circ \cdots \circ g_1(z) \in \mathbb{B}_{r'}(d_x)$ and thus

$$\begin{aligned} \sup_{z\in\mathbb{B}_r(d_x)} \|G_{f,t}(z) - G_{f',t}(z)\|_2 &= \sup_{z\in\mathbb{B}_r(d_x)} \|f_\theta \circ g_t \circ \cdots \circ g_1(z) - f_{\theta'} \circ g_t \circ \cdots \circ g_1(z)\|_2 \\ &\le \|f_\theta - f_{\theta'}\|_{\mathcal{F}} \end{aligned}$$

Finally, to bound $\left|\|G_{f,t}(z)\|_2^2 - \|G_{f',t}(z)\|_2^2\right|$, let us define

$$C_{\max} = \max\left\{\sup_{\theta\in\Theta} \|f_\theta \circ g_t \circ \cdots \circ g_1(0)\|_2, \|f^* \circ g_t \circ \cdots \circ g_1(0)\|_2\right\}.$$

Since Θ is bounded, $\sup_{\theta\in\Theta} \|f_\theta \circ g_t \circ \cdots \circ g_1(0)\|_2$ is finite, and thus $C_{\max}$ is a finite constant. It then follows from the triangular inequality and (17) that

$$\|G_{f,t}(z)\|_2 \le \|G_{f,t}(z) - G_{f,t}(0)\|_2 + \|G_{f,t}(z)\|_2 \le L_G \cdot \|z\|_2 + C_{\max}.$$

We then obtain

$$\begin{aligned} &\sup_{z\in\mathbb{B}_r(d_x)} \left|\|G_{f,t}(z)\|_2^2 - \|G_{f',t}(z)\|_2^2\right| \\ \le &\sup_{z\in\mathbb{B}_r(d_x)} \|G_{f,t}(z) - G_{f',t}(z)\|_2 \cdot \|G_{f,t}(z) + G_{f',t}(z)\|_2 \\ \le &\sup_{z\in\mathbb{B}_r(d_x)} \|f - f'\|_{\mathcal{F}} \cdot (2L_G \cdot \|z\|_2 + 2C_{\max}) \\ = &\|f - f'\|_{\mathcal{F}} \cdot (2rL_G + 2C_{\max}) \end{aligned}$$

Set $K_G = \max\left(1, 2rL_G + 2C_{\max}\right)$ and we finish the proof. □

B Full Statement of Theorem 1 and Its Proof

Note that the values of K_G depend on r_x in Assumption 4. Sometimes we make this dependency explicit by writing $K_G(r)$ for any radius $r > 0$.

Theorem 4 (Full Version of Theorem 1)**.** *Recall the definitions of M_2 and κ in (65) and (8):*

$$M_2 = \sup_{f\in\mathcal{F}}\sup_{t\in[T],w_t>0} \mathbb{E}\left[\|G_{f,t}(x_1)\|_2^2\right], \quad \kappa = \sup_{f\in\mathcal{F}\setminus\{f^*\}}\sup_{t\in[T],w_t>0} \frac{\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^4\right]^{1/2}}{\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^2\right]}.$$

Let $\delta \in (0,1]$. Let $w_{avg} := \frac{1}{T}\sum_{t\in[T]} w_t$. Let C be some constant with $C > 1$. Define

$$\underline{r}_x := \sigma\left[3 + 16\sqrt{\frac{\ln(4m/\delta)}{d_x}}\right], \quad \underline{r}_v := 2\nu\left[\sqrt{d_y} + 8\sqrt{2\ln(4m/\delta)}\right].$$

Define $n' := \min_{t\in[T]} n_t$, and $n'' := \min_{t:w_t>0}(n_t/w_t)$. Suppose Assumptions 1 to 4 hold and let $\sigma, \nu, r_x, K_G, L_G$ be defined therein. Let $K_G(r) \le k_G r^\alpha$ for some $\alpha > 0$ and $k_G > 0$. Assume that

$$\begin{aligned} n' &\ge \frac{2\kappa^2 C^2}{C-1}\left[\ln(4/\delta) + p\ln\left(1 + \frac{2B(C+1)}{C(1+\underline{r}_v)}Tn''\right)\right], \\ m &\ge \frac{\sqrt{2}L_G\sigma\delta\sqrt{(d_x+64)}}{e^{d_x/256}\sqrt{M_2}}. \end{aligned} \tag{18}$$

Let $\hat{\theta}_T \in \Theta \subset \mathbb{R}^p$ be a global minimizer of (7) with the regularization parameter $\lambda \le 4\nu^2/Tn''$. Then, with probability at least $1-\delta$ the quantity $\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} w_t \cdot \|f^*(x_t) - f_{\hat{\theta}_T}(x_t)\|_2^2\right]$ is upper bounded by

$$\begin{aligned} &\left\{2\alpha C w_{avg} k_G\left(1 + 8\nu\left[1 + 8\sqrt{2\ln(4m/\delta)}\right]\right)\left[\left(\left[3 + 16\sqrt{\frac{\ln(4m/\delta)}{d_x}}\right]^\alpha + \frac{1}{16^\alpha d_x^{\alpha/2}}\right)\right.\right. \\ &\left.\left. + \left(\max\left\{\frac{256}{d_x}\ln\left(\frac{32L_G\sqrt{2M_2}}{\alpha 16^\alpha (C+1)w_{avg}k_G}\frac{d_x^{\alpha/2-1}\sqrt{d_x+64}}{\sigma^{\alpha-1}}m^2\right), 0\right\}\right)^{\alpha/2}\right]\right\}\frac{\sigma^\alpha}{Tn''} \\ &+ 8C\nu^2\frac{p\ln\left(1 + \frac{2B(C+1)}{C(1+8\nu\left[1+8\sqrt{2\ln(4m/\delta)}\right])}Tn''\right) + \Omega_T(f^*)}{Tn''} + 8C\nu^2\frac{\ln(4/\delta)}{Tn''}. \end{aligned} \tag{19}$$

Corollary 1. *Under the setting of Theorem 4, if $\nu = 0$ and*

$$\begin{aligned} n' &> \frac{2\kappa^2 C}{C-1}\left[\ln(4/\delta) + p\ln\left(1 + \frac{2B(C+1)}{C(1+8\nu\left[1+8\sqrt{2\ln(4m/\delta)}\right])}Tn''\right)\right], \\ m &\ge \frac{\sqrt{2}L_G\sigma\delta\sqrt{(d_x+64)}}{e^{d_x/256}\sqrt{M_2}}, \end{aligned} \tag{20}$$

then with probability at least $1-\delta$ the quantity $\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} w_t \cdot \|f^*(x_t) - f_{\hat{\theta}_T}(x_t)\|_2^2\right]$ is upper bounded by

$$\begin{aligned} &\left\{2\alpha C w_{avg} k_G\left[\left(\left[3 + 16\sqrt{\frac{\ln(4m/\delta)}{d_x}}\right]^\alpha + \frac{1}{16^\alpha d_x^{\alpha/2}}\right)\right.\right. \\ &\left.\left. + \left(\max\left\{\frac{256}{d_x}\ln\left(\frac{32L_G\sqrt{2M_2}}{\alpha 16^\alpha (C+1)w_{avg}k_G}\frac{d_x^{\alpha/2-1}\sqrt{d_x+64}}{\sigma^{\alpha-1}}m^2\right), 0\right\}\right)^{\alpha/2}\right]\right\}\frac{\sigma^\alpha}{Tn''}. \end{aligned} \tag{21}$$

Corollary 2. *Under the setting of Theorem 4, if $\sigma = 0$ and*

$$n' > \frac{2\kappa^2 C^2}{C-1}\left[\ln(4/\delta) + p\ln\left(1 + \frac{2B(C+1)}{C(1+8\nu\left[1+8\sqrt{2\ln(4m/\delta)}\right])}Tn''\right)\right], \tag{22}$$

then with probability at least $1-\delta$ the quantity $\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} w_t \cdot \|f^*(x_t) - f_{\hat{\theta}_T}(x_t)\|_2^2\right]$ is upper bounded by

$$8C\frac{\nu^2}{Tn''}\left[p\ln\left(1 + \frac{2B(C+1)}{C(1+8\nu\left[1+8\sqrt{2\ln(4m/\delta)}\right])}Tn''\right) + \Omega_T(f^*) + \ln(4/\delta)\right]. \tag{23}$$

Corollary 3. *Under the setting of Theorem 4, and if T or n_t is large enough such that $n' > \kappa^2 Cp\ln(Tn'')/(C-1)$, then with probability at least $1-\delta$ we have*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} w_t \cdot \|f^*(x_t) - f_{\hat{\theta}_T}(x_t)\|_2^2\right] \le \tilde{\mathcal{O}}\left(\nu^2\frac{p+\Omega_T(f^*)+\ln(4/\delta)}{Tn''} + \sigma^\alpha\frac{1}{Tn''}\right). \tag{24}$$

Proof of Theorem 4. First note that κ and M_2 are finite under Assumption 3 (see Lemma 12). We then derive some basic inequalities. Suppose $\hat{f} := f_{\hat{\theta}_T}$ is a global minimizer for (7). Since f^* is realizable by Assumption 2, the optimality of $\hat{f}$ implies

$$\begin{aligned}
&\frac{1}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\cdot\mathcal{L}\Big(y_{ti},\hat{f}(x_{ti})\Big) + \lambda\Omega_T(\hat{f}) \le \frac{1}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\cdot\mathcal{L}\Big(y_{ti},f^*(x_{ti})\Big) + \lambda\Omega_T(f^*)\\
\Leftrightarrow\ &\frac{1}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\cdot\|y_{ti}-\hat{f}(x_{ti})\|_2^2 + \lambda\Omega_T(\hat{f}) \le \frac{1}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\cdot\|y_{ti}-f^*(x_{ti})\|_2^2 + \lambda\Omega_T(f^*).
\end{aligned}$$

Using the fact $y_{ti} = f^*(x_{ti}) + v_{ti}$, expanding the terms, and simplifying, we obtain

$$\frac{1}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\cdot\|f^*(x_{ti})+v_{ti}-\hat{f}(x_{ti})\|_2^2 + \lambda\Omega_T(\hat{f}) \le \frac{1}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\cdot\|v_{ti}\|_2^2 + \lambda\Omega_T(f^*),$$

which is equivalent to

$$\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{Tn_t}\cdot\|f^*(x_{ti})-\hat{f}(x_{ti})\|_2^2 \le \sum_{(t,i)\in\mathcal{M}}\frac{2w_t}{Tn_t}\cdot\langle v_{ti},\hat{f}(x_{ti})-f^*(x_{ti})\rangle + \lambda\left[\Omega_T(f^*)-\Omega_T(\hat{f})\right].$$

Multiplying both sides by 2, rearranging the terms, and with the definition

$$M(f) := \frac{1}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\left[4\cdot\langle v_{ti},f(x_{ti})-f^*(x_{ti})\rangle - \|f^*(x_{ti})-f(x_{ti})\|_2^2\right], \tag{25}$$

we obtain

$$\frac{1}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\cdot\|f^*(x_{ti})-\hat{f}(x_{ti})\|_2^2 \le M(\hat{f}) + 2\lambda\left[\Omega_T(f^*)-\Omega_T(\hat{f})\right].$$

Multiplying both sides by C (recall $C>1$) yields

$$C\cdot\frac{1}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\cdot\|f^*(x_{ti})-\hat{f}(x_{ti})\|_2^2 \le C\cdot M(\hat{f}) + 2C\lambda\left[\Omega_T(f^*)-\Omega_T(\hat{f})\right].$$

We can furthermore rewrite the above inequality with the definition

$$Q(f,C) := \frac{1}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\left(\mathbb{E}\left[\|f(x_t)-f^*(x_t)\|_2^2\right] - C\cdot\|f(x_{ti})-f^*(x_{ti})\|_2^2\right), \tag{26}$$

arriving at

$$\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} w_t \cdot \|f^*(x_t) - \hat{f}(x_t)\|_2^2\right] \leq Q(\hat{f}, C) + C \cdot M(\hat{f}) + 2C\lambda\left[\Omega_T(f^*) - \Omega_T(\hat{f})\right]. \tag{27}$$

Taking any $r_x, r_v, \varepsilon, u > 0$, and combining Propositions 1 and 2, we see that

$$\begin{aligned}\mathbb{P}\left(Q(f, C) > 2B_{\text{sq}} + (C+1)w_{\text{avg}}K_G\varepsilon^2, \forall f \in \mathcal{F}\right) \leq |\mathcal{F}(\varepsilon)| \cdot \exp\left(-\frac{(C-1)\min_{t\in[T]} n_t}{2C^2\kappa^2}\right) \\ + m \cdot \exp\left(-\frac{d_x(r_x - 2\sigma)^2}{128\sigma^2}\right),\end{aligned}$$

$$\begin{aligned}\mathbb{P}\left(M(f) > w_{\text{avg}}K_G\varepsilon\left(1 + 4r_v\right) + u/2C,\ \forall f \in \mathcal{F}\right) \leq |\mathcal{F}(\varepsilon)| \cdot \exp\left(-u \cdot \frac{T}{16C\nu^2}\min_{t\in[T], w_t>0}\frac{n_t}{w_t}\right) \\ + m \cdot \exp\left(-\frac{d_x(r_x - 2\sigma)^2}{128\sigma^2}\right) + m \cdot \exp\left(-\frac{(r_v - 2\nu\sqrt{d_y})^2}{128\nu^2}\right).\end{aligned}$$

Since the common term $m \cdot \exp\left(-\frac{d_x(r_x-2\sigma)^2}{128\sigma^2}\right)$ in the above two probability bounds arise from bounding the norm of x_{1i}'s, we apply the union bound and obtain that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} w_t \cdot \|f^*(x_t) - \hat{f}(x_t)\|_2^2\right] > (C+1)w_{\text{avg}}K_G(r_x)\varepsilon^2 + Cw_{\text{avg}}K_G(r_x)\left(1 + 4r_v\right)\varepsilon + \frac{u}{2} + 2B_{\text{sq}}(r_x) \\ + 2C\lambda\left[\Omega_T(f^*) - \Omega_T(\hat{f})\right],\end{aligned}$$

holds with probability at most

$$\begin{aligned}|\mathcal{F}(\varepsilon)| \cdot \left(\exp\left(-\frac{(C-1)\min_{t\in[T]} n_t}{2C^2\kappa^2}\right) + \exp\left(-u \cdot \frac{T}{16C\nu^2}\min_{t\in[T], w_t>0}\frac{n_t}{w_t}\right)\right) \\ + m \cdot \exp\left(-\frac{d_x(r_x - 2\sigma)^2}{128\sigma^2}\right) + m \cdot \exp\left(-\frac{(r_v - 2\nu\sqrt{d_y})^2}{128\nu^2}\right),\end{aligned} \tag{28}$$

Recall $n' = \min_{t\in[T]} n_t$, $n'' = \min_{t\in[T], w_t>0}\frac{n_t}{w_t}$. For any $\delta \in (0, 1]$, consider the following problem (which essentially minimizes the bound while maintaining high probability $1 - \delta$):

$$u^* = \min_{r_x, r_v, \varepsilon, u \geq 0} \quad u \tag{O}$$

$$\text{s.t.} \quad (C+1)w_{\text{avg}}K_G(r_x)\varepsilon^2 + Cw_{\text{avg}}K_G(r_x)\left(1 + 4r_v\right)\varepsilon + 2B_{\text{sq}}(r_x) - \frac{u}{2} \leq 0, \tag{C1}$$

$$\ln(4/\delta) + \ln(|\mathcal{F}(\varepsilon)|) - \frac{(C-1)n'}{2C^2\kappa^2} \leq 0, \tag{C2}$$

$$\ln(4/\delta) + \ln(|\mathcal{F}(\varepsilon)|) - u \cdot \frac{Tn''}{16C\nu^2} \leq 0, \tag{C3}$$

$$\ln(4m/\delta) - \frac{d_x(r_x - 2\sigma)^2}{128\sigma^2} \leq 0, \tag{C4}$$

$$\ln(4m/\delta) - \frac{(r_v - 2\nu\sqrt{d_y})^2}{128\nu^2} \leq 0. \tag{C5}$$

Choosing r_x, r_v, ϵ, u:

Recall

$$\underline{r}_x := \sigma\left[3 + 16\sqrt{\frac{\ln(4m/\delta)}{d_x}}\right], \quad \underline{r}_v := 2\nu\left[\sqrt{d_y} + 8\sqrt{2\ln(4m/\delta)}\right].$$

Then from (C4), and (C5) we have that $r_x \geq \underline{r}_x$, and $r_v \geq \underline{r}_v$. From (C1), and (C3) we have

$$\frac{u}{2} \geq (C+1)w_{\text{avg}}K_G(r_x)\varepsilon^2 + Cw_{\text{avg}}K_G(r_x)\left(1+4r_v\right)\varepsilon + 2B_{\text{sq}}(r_x) \geq 2B_{\text{sq}}(r_x),$$
$$\frac{u}{2} \geq 8C\nu^2\frac{\ln(4/\delta)+\ln(|\mathcal{F}(\varepsilon)|)}{Tn''} \geq 0.$$

We can upper bound the optimization program by setting u to be sum of the lower bounds obtained earlier (by non-negativity), this gives us

$$\begin{aligned} u^* &\leq \min_{r_x\geq\underline{r}_x, r_v\geq\underline{r}_v, \varepsilon\geq 0}\left[(C+1)w_{\text{avg}}K_G(r_x)\varepsilon^2 + Cw_{\text{avg}}K_G(r_x)\left(1+4r_v\right)\varepsilon + 2B_{\text{sq}}(r_x) + 8C\nu^2\frac{\ln(4/\delta)+\ln(|\mathcal{F}(\varepsilon)|)}{Tn''}\right], \\ &= 8C\nu^2\frac{\ln(4/\delta)}{Tn''} + \min_{r_x\geq\underline{r}_x,\varepsilon\geq 0}\left[\left\{(C+1)w_{\text{avg}}\varepsilon^2 + Cw_{\text{avg}}\left(1+4\underline{r}_v\right)\varepsilon\right\}K_G(r_x) + 2B_{\text{sq}}(r_x) + 8C\nu^2\frac{\ln(|\mathcal{F}(\varepsilon)|)}{Tn''}\right], \end{aligned}$$

Rewrite the right-hand side of the above and we have:

$$\begin{aligned} u^* \leq \min_{\epsilon>0}\Big[8C\nu^2\frac{\ln(4/\delta)}{Tn''} + \Big[8C\nu^2\frac{\ln(|\mathcal{F}(\varepsilon)|)}{Tn''} + \left\{(C+1)w_{\text{avg}}\varepsilon^2 + Cw_{\text{avg}}\left(1+4\underline{r}_v\right)\varepsilon\right\} \\ \times \min_{r_x\geq\underline{r}_x}\left(K_G(r_x) + \frac{2}{\left\{(C+1)w_{\text{avg}}\varepsilon^2 + Cw_{\text{avg}}\left(1+4\underline{r}_v\right)\varepsilon\right\}}B_{\text{sq}}(r_x)\right)\Big]. \end{aligned} \tag{29}$$

We recall Lemma 4 which states that when $r_x > 3\sigma$ we have

$$\frac{B_{\text{sq}}(r_x)}{128w_{\text{avg}}} \leq \left[\frac{L_G\sigma}{d_x}\sqrt{(d_x+64)}\exp\left(-\frac{d_x(r_x-2\sigma)^2}{256\sigma^2}\right)\right]^2 + \sqrt{\frac{M_2}{32}}\frac{L_G\sigma}{d_x}\sqrt{(d_x+64)}\exp\left(-\frac{d_x(r_x-2\sigma)^2}{256\sigma^2}\right).$$

We now use the inequality

$$(r_x-2\sigma)^2 \geq (\underline{r}_x-2\sigma)^2 + (r_x-\underline{r}_x)^2 > (r_x-\underline{r}_x)^2 + \sigma^2 + 256\sigma^2\frac{\ln(4m/\delta)}{d_x},$$

to obtain

$$\frac{L_G\sigma}{d_x}\sqrt{(d_x+64)}\exp\left(-\frac{d_x(r_x-2\sigma)^2}{256\sigma^2}\right) \leq \frac{L_G\sigma}{4d_xe^{d_x/256}}\sqrt{(d_x+64)}\frac{\delta}{m}\exp\left(-\frac{d_x(r_x-\underline{r}_x)^2}{256\sigma^2}\right).$$

Therefore, when $m \geq \frac{\sqrt{2}L_G\sigma}{\sqrt{M_2}}\delta e^{-d_x/256}\frac{\sqrt{d_x+64}}{d_x}$, it is true that

$$\begin{aligned} \frac{L_G\sigma}{d_x}\sqrt{(d_x+64)}\exp\left(-\frac{d_x(r_x-2\sigma)^2}{256\sigma^2}\right) &\leq \frac{L_G\sigma}{4d_xe^{d_x/256}}\sqrt{(d_x+64)}\frac{\delta}{m}\exp\left(-\frac{d_x(r_x-\underline{r}_x)^2}{256\sigma^2}\right), \\ &\leq \sqrt{\frac{M_2}{32}}\exp\left(-\frac{d_x(r_x-\underline{r}_x)^2}{256\sigma^2}\right) \leq \sqrt{\frac{M_2}{32}}. \end{aligned} \tag{30}$$

Based on the inequality $x^2 + ax \leq 2ax$ when $0 < x \leq a$, this gives the bound

$$B_{\text{sq}}(r_x) \leq 32\frac{w_{\text{avg}}L_G\sigma}{d_x}\sqrt{2M_2(d_x+64)}\exp\left(-\frac{d_x(r_x-2\sigma)^2}{256\sigma^2}\right). \tag{31}$$

We substitute the upper bound (31) to (29) and obtain

$$\begin{aligned} u^* \leq 8C\nu^2\frac{\ln(4/\delta)}{Tn''} + \min_{\epsilon>0}\Big[8C\nu^2\frac{\ln(|\mathcal{F}(\varepsilon)|)}{Tn''} + \left\{(C+1)w_{\text{avg}}\varepsilon^2 + Cw_{\text{avg}}\left(1+4\underline{r}_v\right)\varepsilon\right\} \\ \times \min_{r_x\geq\underline{r}_x}\left(K_G(r_x) + \frac{64w_{\text{avg}}L_G\sigma\sqrt{2M_2(d_x+64)}}{d_x\left\{(C+1)w_{\text{avg}}\varepsilon^2 + Cw_{\text{avg}}\left(1+4\underline{r}_v\right)\varepsilon\right\}}\exp\left(-\frac{d_x(r_x-2\sigma)^2}{256\sigma^2}\right)\right)\Big]. \end{aligned}$$

Let $K_G(r) \leq k_G r^\alpha$, then we have the bound $K_G(r_x - \underline{\mathrm{r}}_x + \underline{\mathrm{r}}_x) \leq \alpha k_G (r_x - \underline{\mathrm{r}}_x)^\alpha + \alpha k_G (\underline{\mathrm{r}}_x)^\alpha$. Now define $r := r_x - \underline{\mathrm{r}}_x$ this gives us

$$\begin{aligned} u^* \leq & 8C\nu^2 \frac{\ln(4/\delta)}{Tn''} + \min_{\epsilon>0} \Big[\alpha k_G \left\{ (C+1) w_{\text{avg}} \varepsilon^2 + C w_{\text{avg}} \left(1 + 4\underline{\mathrm{r}}_v\right) \varepsilon \right\} \\ & \times \min_{r \geq 0} \left((\underline{\mathrm{r}}_x)^\alpha + r^\alpha + \frac{64 w_{\text{avg}} L_G \sigma \sqrt{2M_2(d_x+64)}}{\alpha k_G d_x \left\{ (C+1) w_{\text{avg}} \varepsilon^2 + C w_{\text{avg}} \left(1 + 4\underline{\mathrm{r}}_v\right) \varepsilon \right\}} \mathsf{exp}\left(- \frac{d_x (r - (2\sigma - \underline{\mathrm{r}}_x))^2}{256\sigma^2} \right) \right) \\ & + 8C\nu^2 \frac{\ln(|\mathcal{F}(\varepsilon)|)}{Tn''} \Big]. \end{aligned} \tag{32}$$

Moreover, define the quantities:

$$\begin{aligned} & \tau := 64 w_{\text{avg}} L_G \sigma \sqrt{2M_2(d_x+64)}/d_x, \beta := \tau \left(\alpha k_G \left\{ (C+1) w_{\text{avg}} \varepsilon^2 + C w_{\text{avg}} \left(1 + 4\underline{\mathrm{r}}_v\right) \varepsilon \right\} \right)^{-1}, \\ & \gamma := d_x/(256\sigma^2), \text{ and } \zeta := 2\sigma - \underline{\mathrm{r}}_x < 0, \end{aligned}$$

Rewriting (32) gives us

$$u^* \leq \min_{\epsilon>0} \left[\frac{\tau}{\beta} \times \min_{r \geq 0} \left((\underline{\mathrm{r}}_x)^\alpha + r^\alpha + \beta \mathsf{exp}\left(-\gamma (r-\zeta)^2 \right) \right) + 8C\nu^2 \frac{\ln(|\mathcal{F}(\varepsilon)|)}{Tn''} \right] + 8C\nu^2 \frac{\ln(4/\delta)}{Tn''}. \tag{33}$$

We can further upper bound this via setting $r = \sqrt{\max\{\ln(\beta\gamma^{\alpha/2})/\gamma, 0\}}$, and from Proposition 4 we obtain:

$$u^* \leq \min_{\epsilon>0} \left[\frac{\tau}{\beta} \times \left[\underline{\mathrm{r}}_x^\alpha + \frac{1}{\gamma^{\alpha/2}} + \left(\max\{\ln(\beta\gamma^{\alpha/2})/\gamma, 0\} \right)^{\alpha/2} \right] + 8C\nu^2 \frac{\ln(|\mathcal{F}(\varepsilon)|)}{Tn''} \right] + 8C\nu^2 \frac{\ln(4/\delta)}{Tn''}. \tag{34}$$

When $0 < \epsilon < C(1+4\underline{\mathrm{r}}_v)/(C+1)$ we have

$$\frac{\tau}{2C w_{\text{avg}} \alpha k_G (1+4\underline{\mathrm{r}}_v)\epsilon} \leq \beta \leq \frac{\tau}{2(C+1) w_{\text{avg}} \alpha k_G \epsilon^2}.$$

Restricting to this domain gives us,

$$\begin{aligned} u^* \leq \min_{\epsilon' \in (0,1]} \Big[& 2\alpha C w_{\text{avg}} k_G (1+4\underline{\mathrm{r}}_v)\epsilon' \times \left[\underline{r}_x^\alpha + \frac{1}{\gamma^{\alpha/2}} + \left(\max\left\{ \ln\left(\frac{\tau\gamma^{\alpha/2}}{2(C+1) w_{\text{avg}} \alpha k_G \epsilon'^2} \right)/\gamma, 0 \right\} \right)^{\alpha/2} \right] \\ & + 8C\nu^2 \frac{\ln(|\mathcal{F}(\varepsilon' C(1+4\underline{\mathrm{r}}_v)/(C+1))|)}{Tn''} \Big] + 8C\nu^2 \frac{\ln(4/\delta)}{Tn''}, \end{aligned} \tag{35}$$

where $\epsilon' = \frac{(C+1)\epsilon}{C(1+4\underline{\mathrm{r}}_v)}$.

Now set $\epsilon' = 1/Tn''$ this gives us the bound

$$\begin{aligned} u^* \leq & 2\alpha C w_{\text{avg}} k_G (1+4\underline{\mathrm{r}}_v) \left[\underline{r}_x^\alpha + \frac{1}{\gamma^{\alpha/2}} + \left(\max\left\{ \ln\left(\frac{\tau\gamma^{\alpha/2}}{2(C+1) w_{\text{avg}} \alpha k_G} T^2 n''^2 \right)/\gamma, 0 \right\} \right)^{\alpha/2} \right] \frac{1}{Tn''} & (36) \\ & + 8C\nu^2 \frac{p \ln\left(1 + \frac{2B(C+1)}{C(1+\underline{\mathrm{r}}_v)} Tn''\right)}{Tn''} + 8C\nu^2 \frac{\ln(4/\delta)}{Tn''}, & (37) \end{aligned}$$

plugging back the values and retaining $\sigma, \nu, d_x, n', n'', m$ and T gives us

$$\begin{aligned} u^* \leq & 2\alpha C w_{\text{avg}} k_G \left(1 + 8\nu \left[1 + 8\sqrt{2\ln(4m/\delta)}\right]\right) \times \Big[\sigma^\alpha \left(\left[3 + 16\sqrt{\frac{\ln(4m/\delta)}{d_x}} \right]^\alpha + \frac{16^\alpha}{d_x^{\alpha/2}} \right) \\ & + \left(\max\left\{ 256\frac{\sigma^2}{d_x} \ln\left(\frac{32T^2 (n'')^2 L_G \sqrt{2M_2}}{\alpha 16^\alpha (C+1) k_G} \frac{d_x^{\alpha/2-1}\sqrt{d_x+64}}{\sigma^{\alpha-1}} \right), 0 \right\} \right)^{\alpha/2} \Big] \frac{1}{Tn''} \\ & + 8C\nu^2 \frac{p \ln\left(1 + \frac{2B(C+1)}{C(1+\underline{\mathrm{r}}_v)} Tn''\right)}{Tn''} + 8C\nu^2 \frac{\ln(4/\delta)}{Tn''}, \end{aligned} \tag{38}$$

when $m \geq \frac{L_G \sigma}{\sqrt{2M_2}} \delta e^{-d_x/256} \frac{\sqrt{d_x+64}}{d_x}$ and $n' \geq \frac{2\kappa^2 C^2}{C-1}\left[\ln(4/\delta) + p\ln\left(1+\frac{2B(C+1)}{C(1+\mathfrak{r}_v)}Tn''\right)\right]$. Finally we upper bound the term $2C\lambda\left[\Omega_T(f^*) - \Omega_T(\hat{f})\right]$ by $8C\nu^2\Omega_T(f^*)/Tn''$. Combining this upper bound, (38), and (27), finishes the proof.

□

B.1 Proposition 1 and Its Proof

Proposition 1. *Fix $u > 0$. Let C be some constant larger than 1, and write $w_{avg} := \frac{1}{T}\sum_{t\in[T]} w_t$. Recall the quadratic term $Q(f, C)$ defined in (26):*

$$Q(f,C) := \frac{1}{T}\sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t}\left(\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^2\right] - C\cdot\|G_{f,t}(x_{1i})\|_2^2\right).$$

Suppose Assumptions 3 and 4 hold, and let K_G, L_G be defined therein. Let M_2 and κ be defined in (65) and (8) respectively. For any $r_x > 2\sigma$, define $\tilde{G}_{f,t} := G_{f,t} \circ \mathcal{P}_{\mathbb{B}_{r_x}(d_x)}$ and

$$B_{sq} := \sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} w_t\left(\|\tilde{G}_{f,t}(x_1)\|_2^2 - \|G_{f,t}(x_1)\|_2^2\right)\right]\right|. \tag{39}$$

Let $\mathcal{F}(\varepsilon)$ be the smallest ε-net of $\mathcal{F}\setminus\{f^*\}$, then we have

$$\begin{aligned}\mathbb{P}\left(Q(f,C) > 2B_{sq} + (C+1)w_{avg}K_G\varepsilon^2, \forall f\in\mathcal{F}\right) \leq |\mathcal{F}(\varepsilon)|\cdot\exp\left(-\frac{(C-1)\min_{t\in[T]} n_t}{2C^2\kappa^2}\right)\\ + m\cdot\exp\left(-\frac{d_x(r_x-2\sigma)^2}{128\sigma^2}\right).\end{aligned}$$

Proof. For any $f\in\mathcal{F}$, let $f'\in\mathcal{F}(\varepsilon)$ be such that $\|f'-f\|_{\mathcal{F}}\leq\varepsilon$. It follows from Assumption 4 that

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} w_t\cdot\left(\|\tilde{G}_{f,t}(x_1)\|_2^2 - \|\tilde{G}_{f',t}(x_1)\|_2^2\right)\right] \leq w_{\text{avg}}K_G\|f-f'\|_{\mathcal{F}}^2;\\ &\frac{C}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\cdot\left(\|\tilde{G}_{f',t}(x_1)\|_2^2 - \|\tilde{G}_{f,t}(x_1)\|_2^2\right) \leq Cw_{\text{avg}}K_G\|f-f'\|_{\mathcal{F}}^2.\end{aligned} \tag{40}$$

We consider the following two events that hold with high probability:

- For any $r_x > 2\sigma$, using Lemma 17, we have

$$\mathbb{P}\left(\|x_{1i}\|_2 \leq r_x,\ \forall i\in[m]\right) \geq 1 - m\cdot\exp\left(-\frac{d_x(r_x-2\sigma)^2}{128\sigma^2}\right).$$

- In Lemma 2 we have established for a fixed $f\in\mathcal{F}$ and $C>1$ that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} w_t\cdot\|G_{f,t}(x_1)\|_2^2\right] \leq \frac{C}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\cdot\|G_{f,t}(x_{1i})\|_2^2$$

holds with probability at least $1-\exp\left(-\frac{(C-1)\min_{t\in[T]} n_t}{2C^2\kappa^2}\right)$. Applying the union bound with the ε-net $\mathcal{F}(\varepsilon)$, we see that the following holds with probability at least $1-|\mathcal{F}(\varepsilon)|\cdot\exp\left(-\frac{(C-1)\min_{t\in[T]} n_t}{2C^2\kappa^2}\right)$:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} w_t\cdot\|G_{f',t}(x_1)\|_2^2\right] \leq \frac{C}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\cdot\|G_{f',t}(x_{1i})\|_2^2,\quad \forall f'\in\mathcal{F}(\varepsilon). \tag{41}$$

Under these two events, we seek to establish the upper bound $Q(f, C)$ for general f:

$$\begin{aligned}
Q(f,C) &= \frac{1}{T}\sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t}\left(\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^2\right] - C\cdot\|G_{f,t}(x_{1i})\|_2^2\right) \\
&\overset{\text{(i)}}{\leq} B_{\text{sq}} + \frac{1}{T}\sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t}\left(\mathbb{E}\left[\|\tilde{G}_{f,t}(x_1)\|_2^2\right] - C\cdot\|G_{f,t}(x_{1i})\|_2^2\right) \\
&\overset{\text{(ii)}}{\leq} B_{\text{sq}} + \frac{1}{T}\sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t}\left(\mathbb{E}\left[\|\tilde{G}_{f,t}(x_1)\|_2^2\right] - C\cdot\|\tilde{G}_{f,t}(x_{1i})\|_2^2\right) \\
&\overset{\text{(iii)}}{\leq} B_{\text{sq}} + \frac{1}{T}\sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t}\left(\mathbb{E}\left[\|\tilde{G}_{f',t}(x_1)\|_2^2\right] - C\cdot\|\tilde{G}_{f',t}(x_{1i})\|_2^2\right) \\
&\quad + (C+1)w_{\text{avg}}K_G\|f-f'\|_{\mathcal{F}}^2 \\
&\overset{\text{(iv)}}{\leq} B_{\text{sq}} + \frac{1}{T}\sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t}\left(\mathbb{E}\left[\|\tilde{G}_{f',t}(x_1)\|_2^2\right] - C\cdot\|G_{f',t}(x_{1i})\|_2^2\right) \\
&\quad + (C+1)w_{\text{avg}}K_G\|f-f'\|_{\mathcal{F}}^2 \\
&\overset{\text{(v)}}{\leq} 2B_{\text{sq}} + \frac{1}{T}\sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t}\left(\mathbb{E}\left[\|G_{f',t}(x_1)\|_2^2\right] - C\cdot\|G_{f',t}(x_{1i})\|_2^2\right) \\
&\quad + (C+1)w_{\text{avg}}K_G\|f-f'\|_{\mathcal{F}}^2 \\
&\overset{\text{(vi)}}{\leq} 2B_{\text{sq}} + (C+1)w_{\text{avg}}K_G\|f-f'\|_{\mathcal{F}}^2 \\
&\overset{\text{(vii)}}{\leq} 2B_{\text{sq}} + (C+1)w_{\text{avg}}K_G\varepsilon^2.
\end{aligned}$$

Here, (i) follows from the definitions of B_{sq} and $Q(f, C)$, (iii) follows by summing the two inequalities in (40) and rearranging, (ii) and (iv) follow from the event $x_{1i} \leq r_x$ ($\forall i \in [m]$), which implies $G_{f,t}(x_{1i}) = \tilde{G}_{f,t}(x_{1i})$ and $G_{f',t}(x_{1i}) = \tilde{G}_{f',t}(x_{1i})$, (v) follows again from the definition of B_{sq}, (vi) follows from the event that (41) holds, and (vii) follows from the definition of the ε-net $\mathcal{F}(\varepsilon)$.

By an application of the union bound, we arrive at

$$\begin{aligned}
\mathbb{P}\left(Q(f,C) > 2B_{\text{sq}} + (C+1)w_{\text{avg}}K_G\varepsilon^2, \forall f\in\mathcal{F}\right) \leq |\mathcal{F}(\varepsilon)|\cdot\exp\left(-\frac{(C-1)\min_{t\in[T]} n_t}{2C^2\kappa^2}\right) \\
+ m\cdot\exp\left(-\frac{d_x(r_x-2\sigma)^2}{128\sigma^2}\right).
\end{aligned}$$

The proof is now complete. □

Lemma 2. *Recall the definition of κ in (8):*

$$\kappa = \sup_{f\in\mathcal{F}\setminus\{f^*\}}\sup_{t\in[T],w_t>0}\frac{\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^4\right]^{1/2}}{\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^2\right]}.$$

For any fixed $f \in \mathcal{F}\setminus\{f^*\}$ and $C > 1$ we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} w_t\cdot\|f(x_t)-f^*(x_t)\|_2^2\right] \leq \frac{C}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\cdot\|f^*(x_{ti})-f(x_{ti})\|_2^2$$

with probability at least

$$1-\exp\left(-\frac{(C-1)\min_{t\in[T],w_t>0} n_t}{2C^2\kappa^2}\right).$$

Proof. In the proof we assume $w_t > 0$ for all $t \in [T]$, as the case with some of the w_t's being zero follows immediately. Define $F := f - f^*$. Then we need to compute the probability of the event

$$\left\{ \mathbb{E}\left[\frac{1}{T} \sum_{t\in[T]} w_t \cdot \|F(x_t)\|_2^2 \right] \leq \frac{C}{T} \sum_{(t,i)\in\mathcal{M}} \frac{w_t \cdot \|F(x_{ti})\|_2^2}{n_t} \right\}, \tag{42}$$

or equivalently

$$\left\{ \mathbb{E}\left[\sum_{t\in[T]} w_t \cdot \|F(x_t)\|_2^2 \right] \leq C \cdot \sum_{(t,i)\in\mathcal{M}} \frac{w_t \cdot \|F(x_{ti})\|_2^2}{n_t} \right\}. \tag{43}$$

Applying the Chernoff-Cramér bound gives

$$\begin{aligned}
&\mathbb{P}\left(C \sum_{(t,i)\in\mathcal{M}} \frac{w_t \cdot \|F(x_{ti})\|_2^2}{n_t} \leq \sum_{t\in[T]} \mathbb{E}\Big[w_t \cdot \|F(x_t)\|_2^2 \Big] \right) \\
&\leq \inf_{\alpha>0} \exp\left(\alpha \cdot \sum_{t\in[T]} \mathbb{E}\Big[w_t \cdot \|F(x_t)\|_2^2 \Big] \right) \cdot \mathbb{E}\left[\exp\left(-\alpha \cdot C \sum_{(t,i)\in\mathcal{M}} \frac{w_t \cdot \|F(x_{ti})\|_2^2}{n_t} \right) \right].
\end{aligned}$$

We proceed by bounding the rightmost term:

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(-\alpha \cdot C \sum_{(t,i)\in\mathcal{M}} \frac{w_t \cdot \|F(x_{ti})\|_2^2}{n_t} \right) \right] \\
&= \mathbb{E}\left[\exp\left(-C\alpha \sum_{i\in[m]} \sum_{t\in\mathcal{M}_i} \frac{w_t \cdot \|F(x_{ti})\|_2^2}{n_t} \right) \right] \\
&\overset{\text{(i)}}{=} \prod_{i\in[m]} \mathbb{E}\left[\exp\left(-C\alpha \sum_{t\in\mathcal{M}_i} \frac{w_t \cdot \|F(x_{ti})\|_2^2}{n_t} \right) \right] \\
&\overset{\text{(ii)}}{\leq} \prod_{i\in[m]} \exp\left(-C\alpha \sum_{t\in\mathcal{M}_i} \frac{\mathbb{E}\left[w_t \cdot \|F(x_{ti})\|_2^2 \right]}{n_t} + \frac{C^2\alpha^2}{2}\kappa^2 \left(\sum_{t\in\mathcal{M}_i} \frac{\mathbb{E}\left[w_t \cdot \|F(x_{ti})\|_2^2 \right]}{n_t} \right)^2 \right) \\
&= \exp\left(-C\alpha \sum_{t\in[T]} \mathbb{E}\Big[w_t \cdot \|F(x_t)\|_2^2 \Big] + \frac{C^2\alpha^2}{2}\kappa^2 \sum_{i\in[m]} \left(\sum_{t\in\mathcal{M}_i} \frac{\mathbb{E}\left[w_t \cdot \|F(x_t)\|_2^2 \right]}{n_t} \right)^2 \right) \\
&\overset{\text{(iii)}}{\leq} \exp\left(-C\alpha \sum_{t\in[T]} \mathbb{E}\Big[w_t \cdot \|F(x_t)\|_2^2 \Big] + \frac{C^2\alpha^2\kappa^2}{2\min_{t\in[T]} n_t} \left(\sum_{t\in[T]} \mathbb{E}\Big[w_t \cdot \|F(x_t)\|_2^2 \Big] \right)^2 \right)
\end{aligned}$$

where (i) follows from the assumption that the data indexed by $\mathcal{R}_1$ are i.i.d., and therefore the trajectories are independent (Lemma 7), (ii) follows from Lemma 10, and (iii) follows from Lemma 11. Combining the above gives

$$\begin{aligned}
&\mathbb{P}\left(C \sum_{(t,i)\in\mathcal{M}} \frac{w_t \cdot \|F(x_{ti})\|_2^2}{n_t} \leq \sum_{t\in[T]} \mathbb{E}\Big[w_t \cdot \|F(x_t)\|_2^2 \Big] \right) \\
&\leq \exp\left(-(C-1)\alpha \sum_{t\in[T]} \mathbb{E}\Big[w_t \cdot \|F(x_t)\|_2^2 \Big] + \frac{C^2\alpha^2\kappa^2}{2\min_{t\in[T]} n_t} \left(\sum_{t\in[T]} \mathbb{E}\Big[w_t \cdot \|F(x_t)\|_2^2 \Big] \right)^2 \right).
\end{aligned}$$

Since the above holds for any $\alpha > 0$, we take

$$\alpha = \frac{C-1}{C^2 \sum_{t\in[T]} \mathbb{E}\Big[w_t \cdot \|F(x_t)\|_2^2 \Big]} \cdot \frac{\min_{t\in[T]} n_t}{\kappa^2}$$

and plugging this value back we obtain the desired result. □

Lemma 3 (Projection Difference). *Let $z = [z_1, \ldots, z_d]^\top$ be a sub-Gaussian vector with proxy variance σ^2/d and independent coordinates. Let h be a vector-valued function that is Lipschitz continuous with constant L. For any $r > 2\sigma$ we have*

$$\mathbb{E}\left[\left\|h(\mathcal{P}_{\mathbb{B}_r(d)}(z)) - h(z)\right\|_2^2\right] \leq L^2\left(\frac{128\sigma^2}{d} + \frac{2\cdot 64^2\sigma^4}{d^2(r-2\sigma)^2}\right)\exp\left(-\frac{d(r-2\sigma)^2}{128\sigma^2}\right).$$

Furthermore, if $r > 3\sigma$, then we have $\sigma^2/(r-2\sigma)^2 \leq 1$ and thus

$$\mathbb{E}\left[\left\|h(\mathcal{P}_{\mathbb{B}_r(d)}(z)) - h(z)\right\|_2^2\right] \leq 128\frac{L^2\sigma^2}{d^2}(d+64)\exp\left(-\frac{d(r-2\sigma)^2}{128\sigma^2}\right).$$

Proof. Since h is L-Lipschitz, we have

$$\mathbb{E}\left[\left\|h(\mathcal{P}_{\mathbb{B}_r(d)}(z)) - h(z)\right\|_2^2\right] \leq L^2\cdot\mathbb{E}\left[\left\|\mathcal{P}_{\mathbb{B}_r(d)}(z) - z\right\|_2^2\right],$$

and therefore it remains to bound the expectation $\mathbb{E}\left[\left\|\mathcal{P}_{\mathbb{B}_r(d)}(z) - z\right\|_2^2\right]$. If $z \in \mathbb{B}_r(d)$, then we have this expectation equal to 0. Hence we only need to consider $z \notin \mathbb{B}_r(d)$, which means

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathcal{P}_{\mathbb{B}_r(d)}(z) - z\right\|_2^2\right] &= \mathbb{E}\left[\left\|\mathcal{P}_{\mathbb{B}_r(d)}(z) - z\right\|_2^2 \;\middle|\; z \notin \mathbb{B}_r(d)\right] \\
&= \mathbb{E}\left[\left\|r\cdot\frac{z}{\|z\|_2} - z\right\|_2^2 \;\middle|\; z \notin \mathbb{B}_r(d)\right] \\
&= \mathbb{E}\left[\left|r - \|z\|_2\right|^2 \;\middle|\; z \notin \mathbb{B}_r(d)\right]
\end{aligned}$$

Applying the *Integral Identity* (Lemma 20) with some basic probability calculation yields

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathcal{P}_{\mathbb{B}_r(d)}(z) - z\right\|_2^2\right] &= \int_0^\infty \mathbb{P}\left(\left|r - \|z\|_2\right|^2 > \tau \;\middle|\; z \notin \mathbb{B}_r(d)\right)\, d\tau \\
&= \int_0^\infty \mathbb{P}\left(\|z\|_2 - r > \sqrt{\tau} \;\middle|\; z \notin \mathbb{B}_r(d)\right)\, d\tau \\
&\leq \int_0^\infty \mathbb{P}\left(\|z\|_2 - r > \sqrt{\tau}\right)\, d\tau \\
&= \int_0^\infty \mathbb{P}\left(\|z\|_2 - 2\sigma > \sqrt{\tau} + r - 2\sigma\right)\, d\tau \\
&\leq \int_0^\infty \exp\left(-\frac{d(\sqrt{\tau}+r-2\sigma)^2}{128\sigma^2}\right)\, d\tau.
\end{aligned}$$

Here, the last inequality follows from Lemma 16 with the assumption that $r > 2\sigma$. It remains to evaluate this integral. Observe that

$$\begin{aligned}
\exp\left[-d\cdot\frac{(\sqrt{\tau}+r-2\sigma)^2}{128\sigma^2}\right] &= \exp\left[-\frac{d}{128\sigma^2}\left(\tau + 2(r-2\sigma)\sqrt{\tau} + (r-2\sigma)^2\right)\right] \\
&= \exp\left(-\frac{d(r-2\sigma)^2}{128\sigma^2}\right)\cdot\exp\left[-\frac{d}{128\sigma^2}\left(\tau + 2(r-2\sigma)\sqrt{\tau}\right)\right]
\end{aligned}$$

and that by a change of variable $u^2 = \tau$ we have

$$\begin{aligned}\int_0^\infty \exp\left[-\frac{d}{128\sigma^2}\left(\tau + 2(r-2\sigma)\sqrt{\tau}\right)\right] d\tau &= 2\int_0^\infty \exp\left[-\frac{d}{128\sigma^2}\left(u^2 + 2(r-2\sigma)u\right)\right] u\, du \\ &= \left(-\frac{128\sigma^2}{d}\exp\left[-\frac{d}{128\sigma^2}u^2\right]\right)\Big|_0^\infty \\ &\quad + 2\int_0^\infty \exp\left[-\frac{d}{64\sigma^2}(r-2\sigma)u\right] u\, du \\ &= \frac{128\sigma^2}{d} + 2\int_0^\infty \exp\left[-\frac{d}{64\sigma^2}(r-2\sigma)u\right] u\, du \\ &= \frac{128\sigma^2}{d} + \frac{2\cdot 64^2\sigma^4}{d^2(r-2\sigma)^2},\end{aligned}$$

where the last equality is due to Lemma 8. In summary, we have shown

$$\mathbb{E}\left[\left\|\mathcal{P}_{\mathbb{B}_r(d)}(z) - z\right\|_2^2\right] \le \left(\frac{128\sigma^2}{d} + \frac{2\cdot 64^2\sigma^4}{d^2(r-2\sigma)^2}\right)\exp\left(-\frac{d(r-2\sigma)^2}{128\sigma^2}\right).$$

Adding the Lipschitz parameter L back finishes the proof. □

Lemma 4. *Consider the ball $\mathbb{B}_{r_x}(d_x)$ in $\mathbb{R}^{d_x}$ of radius r_x with $r_x > 3\sigma$, and recall that $\mathcal{P}_{\mathbb{B}_{r_x}(d_x)}$ denotes projection onto $\mathbb{B}_{r_x}(d_x)$. Write $\tilde{G}_{f,t} := G_{f,t} \circ \mathcal{P}_{\mathbb{B}_{r_x}(d_x)}$. Recall $w_{avg} := \frac{1}{T}\sum_{t\in[T]} w_t$. and the definition of B_{sq} in (39):*

$$B_{sq}(r_x) = \sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} w_t\left(\|\tilde{G}_{f,t}(x_1)\|_2^2 - \|G_{f,t}(x_1)\|_2^2\right)\right]\right|.$$

Under Assumptions 3 and 4 we have

$$\begin{aligned}B_{sq}(r_x) \le\ & 128 w_{avg}\frac{L_G^2\sigma^2}{d_x^2}(d_x+64)\exp\left(-\frac{d_x(r_x-2\sigma)^2}{128\sigma^2}\right) \\ &+ 16 w_{avg}\frac{L_G\sigma}{d_x}\sqrt{2M_2(d_x+64)}\exp\left(-\frac{d_x(r_x-2\sigma)^2}{256\sigma^2}\right).\end{aligned} \tag{44}$$

Proof. Note that for two random vectors a, b we apply Jensen's and Cauchy-Schwartz inequality and obtain

$$\begin{aligned}\left|\mathbb{E}[\|a\|_2^2 - \|b\|_2^2]\right| &= \mathbb{E}[\|a-b\|_2^2] + 2\cdot\left|\mathbb{E}[b^\top(a-b)]\right| \\ &\le \mathbb{E}[\|a-b\|_2^2] + 2\cdot\mathbb{E}[\|b\|_2^2]^{1/2}\cdot\mathbb{E}[\|a-b\|_2^2]^{1/2}.\end{aligned}$$

With $a = \tilde{G}_{f,t}(x_1)$ and $b = G_{f,t}(x_1)$, the above inequality and Lemma 3 with the assumption $r \ge 3\sigma$ give

$$\begin{aligned}\mathbb{E}\left[\|\tilde{G}_{f,t}(x_1)\|_2^2 - \|G_{f,t}(x_1)\|_2^2\right] \le\ & 128\frac{L_G^2\sigma^2}{d_x^2}(d_x+64)\exp\left(-\frac{d_x(r-2\sigma)^2}{128\sigma^2}\right) \\ &+ 16\frac{L_G\sigma}{d_x}\sqrt{2M_2(d_x+64)}\exp\left(-\frac{d_x(r-2\sigma)^2}{256\sigma^2}\right).\end{aligned}$$

Weighting the above inequality by w_t, averaging it over t, and taking the supremum over $\mathcal{F}$, we finish the proof. □

B.2 Proposition 2 and Its Proof

Proposition 2. *Recall that $M(f)$ is defined in (25) as*

$$M(f) = \frac{1}{T}\sum_{(t,i)\in\mathcal{M}}\frac{w_t}{n_t}\left[4\cdot\langle v_{ti}, f(x_{ti}) - f^*(x_{ti})\rangle - \|f^*(x_{ti}) - f(x_{ti})\|_2^2\right],$$

where v_{ti} conditioned on $x_{1i}, \ldots, x_{ti}$ is sub-Gaussian with proxy variance ν^2. Fix $u > 0$. Let $r_x > 2\sigma, r_v > 2\nu\sqrt{d_y}$, and ε-net $\mathcal{F}(\varepsilon)$ be the smallest ε-net of $\mathcal{F} \setminus \{f^*\}$. Then we have

$$\mathbb{P}\left(M(f) > w_{avg} K_G \varepsilon\,(1+4r_v) + u/2,\ \forall f \in \mathcal{F}\right) \le |\mathcal{F}(\varepsilon)| \cdot \exp\left(-u \cdot \frac{T}{16\nu^2} \min_{t\in[T], w_t>0} \frac{n_t}{w_t}\right) + m \cdot \exp\left(-\frac{d_x(r_x-2\sigma)^2}{128\sigma^2}\right) + m \cdot \exp\left(-\frac{(r_v - 2\nu\sqrt{d_y})^2}{128\nu^2}\right).$$

Proof. For any $f \in \mathcal{F}$, let $f' \in \mathcal{F}(\varepsilon)$ be such that $\|f' - f\|_{\mathcal{F}} \le \varepsilon$. Write $\tilde{G}_{f,t} := G_{f,t} \circ \mathcal{P}_{\mathbb{B}_r(d_x)}$. Using Assumption 4, the triangular inequality, and the assumption $\varepsilon \le 1$, we arrive at

$$\begin{aligned}
\langle v_{ti}, \tilde{G}_{f,t}(x_{1i})\rangle - \langle v_{ti}, \tilde{G}_{f',t}(x_{1i})\rangle &\le K_G \|f - f'\|_{\mathcal{F}} \cdot \max_{(t,i)\in\mathcal{M}} \|v_{ti}\|_2 \le K_G \varepsilon \cdot \max_{(t,i)\in\mathcal{M}} \|v_{ti}\|_2,\\
\|\tilde{G}_{f',t}(x_{ti})\|_2^2 - \|\tilde{G}_{f,t}(x_{ti})\|_2^2 &\le K_G \|f - f'\|_{\mathcal{F}}^2 \le K_G \varepsilon,
\end{aligned}$$

which in turn implies

$$\begin{aligned}
\frac{1}{T}\sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t} \cdot \left(\langle v_{ti}, \tilde{G}_{f,t}(x_{1i})\rangle - \langle v_{ti}, \tilde{G}_{f',t}(x_{1i})\rangle\right) &\le w_{\text{avg}} K_G \varepsilon \cdot \max_{(t,i)\in\mathcal{M}} \|v_{ti}\|_2,\\
\frac{1}{T}\sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t} \cdot \left(\|\tilde{G}_{f',t}(x_{ti})\|_2^2 - \|\tilde{G}_{f,t}(x_{ti})\|_2^2\right) &\le w_{\text{avg}} K_G \varepsilon.
\end{aligned} \tag{45}$$

We now consider three events that hold with high probability:

- From Lemma 5, we see that for a fixed $f' \in \mathcal{F}$ we have

  $$\mathbb{P}\left(M(f') \ge \frac{u}{2}\right) \le \exp\left(-u \cdot \frac{T}{16\nu^2} \min_{t\in[T], w_t>0} \frac{n_t}{w_t}\right).$$

  Applying the union bound with the ε-net $\mathcal{F}(\varepsilon)$, we obtain

  $$\mathbb{P}\left(M(f') \le \frac{u}{2}, \forall f' \in \mathcal{F}(\varepsilon)\right) \le 1 - |\mathcal{F}(\varepsilon)| \cdot \exp\left(-u \cdot \frac{T}{16\nu^2} \min_{t\in[T], w_t>0} \frac{n_t}{w_t}\right).$$

- For any $r_x > 2\sigma$ from Lemma 17, we have

  $$\mathbb{P}\left(\|x_{1i}\|_2 \le r_x,\ \forall i \in [m]\right) \ge 1 - m \cdot \exp\left(-\frac{d_x(r_x-2\sigma)^2}{128\sigma^2}\right).$$

- Similarly, for any $r_v > 2\mu\sqrt{d_y}$, using Lemma 17, we obtain

  $$\mathbb{P}\left(\|v_{ti}\|_2 \le r_v,\ \forall (t,i) \in \mathcal{M}\right) \ge 1 - m \cdot \exp\left(-\frac{(r_v - 2\nu\sqrt{d_y})^2}{128\nu^2}\right).$$

Under these three events, we derive that

$$\begin{aligned}
M(f) &= \frac{1}{T}\sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t}\left[4 \cdot \langle v_{ti}, \tilde{G}_{f,t}(x_{1i})\rangle - \|\tilde{G}_{f,t}(x_{1i})\|_2^2\right]\\
&\overset{\text{(i)}}{\le} w_{\text{avg}} K_G \varepsilon \left(1 + 4 \max_{(t,i)\in\mathcal{M}} \|v_{ti}\|_2\right) + \frac{1}{T}\sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t}\left[4 \cdot \langle v_{ti}, \tilde{G}_{f',t}(x_{1i})\rangle - \|\tilde{G}_{f',t}(x_{1i})\|_2^2\right]\\
&\overset{\text{(ii)}}{\le} w_{\text{avg}} K_G \varepsilon\,(1+4r_v) + \frac{1}{T}\sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t}\left[4 \cdot \langle v_{ti}, G_{f',t}(x_{1i})\rangle - \|G_{f',t}(x_{1i})\|_2^2\right]\\
&\overset{\text{(iii)}}{=} w_{\text{avg}} K_G \varepsilon\,(1+4r_v) + M(f')\\
&\overset{\text{(iv)}}{\le} w_{\text{avg}} K_G \varepsilon\,(1+4r_v) + u/2.
\end{aligned}$$

Above, (i) follows by weighting the inequalities in (45) properly and summing them together; (ii) follows as $\|x_{1i}\|_2 \leq r_x$ and $\|v_{ti}\|_2 \leq r_v$, the former indicating $G_{f,t}(x_{1i}) = \tilde{G}_{f,t}(x_{1i})$ and $G_{f',t}(x_{1i}) = \tilde{G}_{f',t}(x_{1i})$; (iii) follows from the definition of $M(f)$; (iv) follows from the event $M(f') \leq u/2$. Now, applying the union bound, we obtain

$$\mathbb{P}\left(M(f) > w_{\text{avg}} K_G \varepsilon\left(1+4r_v\right) + u/2,\ \forall f \in \mathcal{F}\right) \leq |\mathcal{F}(\varepsilon)| \cdot \exp\left(-u \cdot \frac{T}{16\nu^2} \min_{t\in[T], w_t>0} \frac{n_t}{w_t}\right) + m \cdot \exp\left(-\frac{d_x(r_x - 2\sigma)^2}{128\sigma^2}\right) + m \cdot \exp\left(-\frac{(r_v - 2\nu\sqrt{d_y})^2}{128\nu^2}\right).$$

The proof is now complete. □

Lemma 5. *Fix $f \in \mathcal{F}$. Recall that $M(f)$ is defined in (25) as*

$$M(f) = \frac{1}{T} \sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t} \left[4 \cdot \langle v_{ti}, f(x_{ti}) - f^*(x_{ti})\rangle - \|f^*(x_{ti}) - f(x_{ti})\|_2^2\right],$$

where v_{ti} conditioned on $x_{1i}, \ldots, x_{ti}$ is sub-Gaussian with proxy variance ν^2. Then

$$\mathbb{P}\left(M(f) \geq u\right) \leq \exp\left(-u \cdot \frac{T}{8\nu^2} \min_{t\in[T], w_t>0} \frac{n_t}{w_t}\right).$$

Proof. In the proof we assume $w_t > 0$ for all $t \in [T]$, as the case with some of the w_t's being zero follows immediately. The proof follows the strategy of Proposition F.2 of Ziemann et al. (2024a) but extends it for the case with weights w_t and partial trajectories. Defining

$$a := \frac{T}{8\nu^2} \min_{t\in[T]} \frac{n_t}{w_t}, \tag{46}$$

we need to prove $\mathbb{P}\left(M(f) \geq u\right) \leq \exp\left(-ua\right)$. Since Markov's inequality implies

$$\begin{aligned}\mathbb{P}\left(M(f) \geq u\right) &= \mathbb{P}\left(\exp\left(a \cdot M(f)\right) \geq \exp(au)\right) \\ &\leq \mathbb{E}\left[\exp\left(a \cdot M(f)\right)\right] \cdot \exp(-au),\end{aligned}$$

it suffices to show $\mathbb{E}\left[\exp\left(a \cdot M(f)\right)\right] \leq 1$. To do so, we first bound each summand of $M(f)$. For any fixed $f \in \mathcal{F}$, write $F := f - f^*$. Note that v_{ti} is conditioned on $x_{1i}, \ldots, x_{ti}$ is sub-Gaussian, thus, for any $(t,i) \in \mathcal{M}$ we have

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\frac{a}{T} \cdot \frac{w_t}{n_t}\left[4 \cdot \langle v_{ti}, F(x_{ti})\rangle - \|F(x_{ti})\|_2^2\right]\right) \;\middle|\; x_{1i}, \ldots, x_{ti}\right] \\ &\overset{\text{(i)}}{\leq} \exp\left(\frac{\nu^2}{2} \cdot \frac{16a^2 w_t^2}{T^2 n_t^2}\|F(x_{ti})\|_2^2 - \frac{a w_t}{T n_t} \cdot \|F(x_{ti})\|_2^2\right) \\ &\overset{\text{(ii)}}{\leq} 1,\end{aligned} \tag{47}$$

where (i) follows from the property of sub-Gaussian vectors v_{ti} as shown in (72), and (ii) follows from the definition of a, that is $a \leq \frac{T}{8\nu^2} \cdot \frac{n_t}{w_t}$, which implies $\frac{\nu^2}{2T} \cdot \frac{16a^2 w_t^2}{T^2 n_t^2} - \frac{a w_t}{T n_t} \leq 0$. Define

$$R_t := \sum_{i\in\mathcal{R}_t} \frac{a}{T} \cdot \frac{w_t}{n_t}\left[4 \cdot \langle v_{ti}, F(x_{ti})\rangle - \|F(x_{ti})\|_2^2\right]$$

and write $X_{1:t} := \{x_{1i}, \ldots, x_{ti}\}_{i\in[m]}$. Since for any $i \in \mathcal{R}_t$ we have that v_{ti}'s are independent of each other, from (47) it follows that

$$\mathbb{E}\left[\exp\left(R_t\right) \;\middle|\; X_{1:t}\right] \leq 1. \tag{48}$$

Note that $a \cdot M(f) = \sum_{t\in[T]} R_t$, and we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(a \cdot M(f)\right)\right] &= \mathbb{E}\left[\exp\left(\sum_{t\in[T]} R_t\right)\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}\left[\mathbb{E}\left[\exp\left(\sum_{t\in[T]} R_t\right) \;\middle|\; X_{1:T}\right]\right] \\
&\overset{\text{(ii)}}{=} \mathbb{E}\left[\mathbb{E}\left[\exp\left(\sum_{t\in[T-1]} R_t\right) \;\middle|\; X_{1:T}\right]\right] \cdot \mathbb{E}\left[\mathbb{E}\left[\exp\left(R_T\right) \;\middle|\; X_{1:T}\right]\right] \\
&\overset{\text{(iii)}}{\leq} \mathbb{E}\left[\mathbb{E}\left[\exp\left(\sum_{t\in[T-1]} R_t\right) \;\middle|\; X_{1:T}\right]\right] \\
&\overset{\text{(iv)}}{=} \mathbb{E}\left[\exp\left(\sum_{t\in[T-1]} R_t\right)\right] \\
&\leq \cdots \leq 1.
\end{aligned} \tag{49}$$

Above, (i) and (iv) follow from the *law of total expectation*, (ii) follows from the fact that v_{Ti}'s are independent of other noise vectors v_{ti}'s when conditioned on $X_{1:T}$, and (iii) follows from (48), and the rest of the derivations follows a recursive application of the previous steps. □

C Full Statement of Theorem 2 and Its Proof (Data-Dependent Regularization)

Theorem 5 (Data-Dependent Regularization, Full Version of Theorem 2)**.** *Recall the definitions of M_2 and κ in (65) and (8):*

$$M_2 = \sup_{f\in\mathcal{F}} \sup_{t\in[T],w_t>0} \mathbb{E}\left[\|G_{f,t}(x_1)\|_2^2\right], \quad \kappa = \sup_{f\in\mathcal{F}\setminus\{f^*\}} \sup_{t\in[T],w_t>0} \frac{\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^4\right]^{1/2}}{\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^2\right]}.$$

Let $\delta \in (0,1]$. Let $\hat{\theta}_1, \ldots, \hat{\theta}_T \in \Theta \subset \mathbb{R}^p$ be global minimizers of (11) with $\beta_t = 1/4^{T-t}$. Define $w_t := 4^{T-t} \cdot \frac{n_t}{m} \cdot \beta_t$ and $w_{avg} := \frac{1}{T}\sum_{t\in[T]} w_t$. Let C be some constant with $C > 1$. Define $n' := \min_{t\in[T]} n_t$, and $n'' := \min_{t:w_t>0}(n_t/w_t)$. Suppose Assumptions 1 to 4 hold and let $\sigma, \nu, r_x, K_G, L_G$ be defined therein. Let $K_G(r) \leq k_G r^\alpha$ for some $\alpha > 0$, $k_G > 0$, and $\underline{r}_v := 2\nu\left[\sqrt{d_y} + 8\sqrt{2\ln(4m/\delta)}\right]$.

Assume that

$$\begin{aligned}
n' &\geq \frac{2\kappa^2 C^2}{C-1}\left[\ln(4/\delta) + p\ln\left(1 + \frac{2B(C+1)}{C(1+\underline{r}_v)}Tn''\right)\right], \\
m &\geq \frac{\sqrt{2}L_G\sigma\delta\sqrt{(d_x+64)}}{e^{d_x/256}\sqrt{M_2}}.
\end{aligned} \tag{50}$$

Then, with probability at least $1-\delta$ the quantity $\mathbb{E}\left[\frac{\sum_{t\in[T]} n_t\beta_t\cdot\|f^*(x_t)-\hat{f}_T(x_t)\|_2^2}{n_1+\cdots+n_T}\right]$ is bounded above by

$$\Big\{2\alpha C w_{avg} k_G \left(1+8\nu\left[1+8\sqrt{2\ln(4m/\delta)}\right]\right)\Big[\left(\left[3+16\sqrt{\frac{\ln(4m/\delta)}{d_x}}\right]^{\alpha}+\frac{1}{16^{\alpha}d_x^{\alpha/2}}\right)$$
$$+\left(\max\left\{\frac{256}{d_x}\ln\left(\frac{32L_G\sqrt{2M_2}}{\alpha 16^{\alpha}(C+1)w_{avg}k_G}\frac{d_x^{\alpha/2-1}\sqrt{d_x+64}}{\sigma^{\alpha-1}}m^2\right),0\right\}\right)^{\alpha/2}\Big]\Big\}\frac{\sigma^{\alpha}}{n_1+\cdots+n_T}\max_{t\in[T]}\{2^{2(T-t)}\beta_t\} \quad (51)$$
$$+8C\nu^2\frac{pT\ln\left(1+\frac{2B(C+1)}{C(1+8\nu\left[1+8\sqrt{2\ln(4m/\delta)}\right])}\frac{Tm}{\max_{t\in T}2^{2(T-t)}\beta_t}\right)}{n_1+\cdots+n_T}+8C\nu^2\frac{\ln(4/\delta)}{n_1+\cdots+n_T}\max_{t\in[T]}\{2^{2(T-t)}\beta_t\},$$

Proof for Global Minimizers of (11). We first analyze the case with objective (12) that is associated with the regularizer $\Omega_T(\theta)=\sum_{t\in[T-1]}\sum_{i\in\mathcal{R}_t}\beta_t\cdot\|f_\theta(x_{ti})-f_{\hat\theta_{T-1}}(x_{ti})\|_2^2$. Suppose we have just computed $\hat\theta_T$ in *Step 1*. At that moment, we have access to all the m samples of task T, but only to part of the samples from previous tasks. To unify the notation, we define $\mathcal{R}_T:=[m]$ and $n_T:=m$. Write $\hat f_t:=f_{\hat\theta_t}$. Applying the inequality $\|a+b\|_2^2-\|b\|_2^2\geq(1-\frac{1}{s})\cdot\|a\|_2^2-s\cdot\|b\|_2^2$ with $a=\hat f_T(x_{ti})-f^*(x_{ti})$ and $b=f^*(x_{ti})-\hat f_{T-1}(x_{ti})$, we obtain

$$\begin{aligned}&\|\hat f_T(x_{ti})-\hat f_{T-1}(x_{ti})\|_2^2-\|f^*(x_{ti})-\hat f_{T-1}(x_{ti})\|_2^2\\ \geq&\left(1-\frac{1}{s}\right)\cdot\|\hat f_T(x_{ti})-f^*(x_{ti})\|_2^2-s\cdot\|f^*(x_{ti})-\hat f_{T-1}(x_{ti})\|_2^2.\end{aligned}$$

Applying this inequality to the regularization term $\Omega_T(\theta)=\sum_{t\in[T-1]}\sum_{i\in\mathcal{R}_t}\beta_t\cdot\|f_\theta(x_{ti})-f_{\hat\theta_{T-1}}(x_{ti})\|_2^2$, we obtain

$$\begin{aligned}&\Omega_T(f^*)-\Omega_T(\hat f_T)\\ =&\sum_{t\in[T-1]}\sum_{i\in\mathcal{R}_t}\beta_t\left(\|f^*(x_{ti})-\hat f_{T-1}(x_{ti})\|_2^2-\|\hat f_T(x_{ti})-\hat f_{T-1}(x_{ti})\|_2^2\right)\\ \leq&\sum_{t\in[T-1]}\sum_{i\in\mathcal{R}_t}\beta_t\left(s\cdot\|f^*(x_{ti})-\hat f_{T-1}(x_{ti})\|_2^2-\left(1-\frac{1}{s}\right)\cdot\|\hat f_T(x_{ti})-f^*(x_{ti})\|_2^2\right).\end{aligned}$$

Similarly to the beginning of the proof of Theorem 4, we have

$$\sum_{i\in\mathcal{R}_T}\beta_T\cdot\|f^*(x_{Ti})-\hat f_T(x_{Ti})\|_2^2\leq\sum_{i\in[m]}2\beta_T\cdot\langle v_{Ti},\hat f_T(x_{Ti})-f^*(x_{Ti})\rangle+\Omega_T(f^*)-\Omega_T(\hat f_T).$$

Defining

$$M_t(f):=\sum_{i\in[m]}\left[2s\cdot\langle v_{ti},f(x_{ti})-f^*(x_{ti})\rangle-\|f^*(x_{ti})-f(x_{ti})\|_2^2\right], \quad (52)$$

and combining the above with some rearrangements of terms and rescaling by $\frac{s}{s-1}$ yields

$$\begin{aligned}&\sum_{t\in[T]}\sum_{i\in\mathcal{R}_t}\beta_t\cdot\|f^*(x_{ti})-\hat f_T(x_{ti})\|_2^2\\ \leq&\frac{\beta_T}{s-1}\cdot M_T(\hat f_T)+\frac{s^2}{s-1}\sum_{t\in[T-1]}\sum_{i\in\mathcal{R}_t}\beta_{t-1}\cdot\|f^*(x_{ti})-\hat f_{T-1}(x_{ti})\|_2^2\end{aligned} \quad (53)$$

We can now unroll the above recurrence relation and arrive at

$$\sum_{t\in[T]}\sum_{i\in\mathcal{R}_t}\beta_t\cdot\|f^*(x_{ti})-\hat f_T(x_{ti})\|_2^2\leq\sum_{t\in[T]}\frac{s^{2(T-t)}}{(s-1)^{T-t+1}}\beta_t\cdot M_t(\hat f_t). \quad (54)$$

Set $s=2$, and the above becomes

$$\sum_{t\in[T]}\sum_{i\in\mathcal{R}_t}\beta_t\cdot\|f^*(x_{ti})-\hat f_T(x_{ti})\|_2^2\leq\sum_{t\in[T]}2^{2(T-t)}\beta_t\cdot M_t(\hat f_t). \quad (55)$$

Write $\beta_t' := \frac{n_t}{m} \cdot \beta_t$, and we obtain

$$\sum_{t\in[T]} \sum_{i\in\mathcal{R}_t} \frac{\beta_t'}{n_t} \cdot \|f^*(x_{ti}) - \hat{f}_T(x_{ti})\|_2^2 \leq \sum_{t\in[T]} \frac{2^{2(T-t)}\beta_t'}{n_t} \cdot M_t\big(\hat{f}_t\big).$$

Multiplying both sides by constant $C > 1$, dividing it by T, adding $\mathbb{E}\left[\sum_{t\in[T]} \beta_t' \cdot \|f^*(x_{ti}) - \hat{f}_T(x_{ti})\|_2^2\right]$, we obtain

$$\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} \beta_t' \cdot \|f^*(x_t) - \hat{f}_T(x_t)\|_2^2\right] \leq Q(\hat{f}_T, C, \beta_t') + C \cdot \sum_{t\in[T]} \frac{2^{2(T-t)}\beta_t'}{Tn_t} \cdot M_t\big(\hat{f}_t\big),$$

where $Q(f, C, \beta_t')$ is defined as

$$Q(f, C, \beta_t') := \frac{1}{T} \sum_{(t,i)\in\mathcal{M}} \frac{\beta_t'}{n_t} \left(\mathbb{E}\left[\|f(x_t) - f^*(x_t)\|_2^2\right] - C \cdot \|f(x_{ti}) - f^*(x_{ti})\|_2^2\right).$$

To obtain a high probability bound of $Q(f, C, \beta_t')$, we can now invoke Proposition 1, and to obtain a probabilistic bound of $\sum_{t\in[T]} \frac{2^{2(T-t)}\beta_t'}{Tn_t} \cdot M_t\big(\hat{f}_t\big)$ we invoke Proposition 3. These respectively give

$$\begin{aligned}\mathbb{P}\left(Q(f, C, \beta_t') > 2B_{\text{sq}} + (C+1)w_{\text{avg}}K_G\varepsilon^2, \forall f \in \mathcal{F}\right) \leq |\mathcal{F}(\varepsilon)| \cdot \exp\left(-\frac{(C-1)\min_{t\in[T]} n_t}{2C^2\kappa^2}\right) \\ + m \cdot \exp\left(-\frac{d_x(r_x - 2\sigma)^2}{128\sigma^2}\right)\end{aligned}$$

and

$$\sum_{t\in[T]} \frac{2^{2(T-t)}\beta_t'}{Tn_t} \cdot M_t(f_t) \leq w_{\text{avg}}K_G\varepsilon\left(1 + 4r_v\right) + u/2, \qquad \forall f_1, \ldots, f_T \in \mathcal{F},$$

with probability at least

$$|\mathcal{F}(\varepsilon)|^T \cdot \exp\left(-u \cdot \frac{T}{16\nu^2} \min_{t\in[T]} \frac{n_t}{2^{2(T-t)}\beta_t'}\right) + m \cdot \exp\left(-\frac{d_x(r_x - 2\sigma)^2}{128\sigma^2}\right) + m \cdot \exp\left(-\frac{(r_v - 2\nu\sqrt{d_y})^2}{128\nu^2}\right).$$

Then, following the identical idea of proving Theorem 4 we obtain that, with probability at least $1 - \delta$, the quantity $\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} \beta_t' \cdot \|f^*(x_t) - f_{\hat{\theta}_T}(x_t)\|_2^2\right]$ is upper bounded by

$$\begin{aligned}&\left\{2\alpha C w_{\text{avg}} k_G \left(1 + 8\nu\left[1 + 8\sqrt{2\ln(4m/\delta)}\right]\right)\right\}\left[\left(\left[3 + 16\sqrt{\frac{\ln(4m/\delta)}{d_x}}\right]^\alpha + \frac{1}{16^\alpha d_x^{\alpha/2}}\right)\right. \\ &\left. + \left(\max\left\{\frac{256}{d_x}\ln\left(\frac{32L_G\sqrt{2M_2}}{\alpha 16^\alpha (C+1)w_{\text{avg}}k_G}\frac{d_x^{\alpha/2-1}\sqrt{d_x+64}}{\sigma^{\alpha-1}}m^2\right), 0\right\}\right)^{\alpha/2}\right]\Big\}\frac{\sigma^\alpha}{Tn''} \\ &+ 8C\nu^2\frac{pT\ln\left(1 + \frac{2B(C+1)}{C(1+8\nu\left[1+8\sqrt{2\ln(4m/\delta)}\right])}Tn''\right)}{Tn''} + 8C\nu^2\frac{\ln(4/\delta)}{Tn''},\end{aligned} \tag{56}$$

with n'' now defined as $n'' = \min_{t\in[T]} \frac{n_t}{2^{2(T-t)}\beta_t'}$.

Substituting the identity $\beta_t' = \frac{n_t}{m}\beta_t$ with $n_T = m$ to the above result and rescaling by $T/(n_1 + \cdots + n_T)$, we obtain that $\mathbb{E}\left[\frac{\sum_{t\in[T]} n_t\beta_t \cdot \|f^*(x_t) - \hat{f}_T(x_t)\|_2^2}{n_1 + \cdots + n_T}\right]$ is bounded above by

$$\Big\{2\alpha C w_{\text{avg}} k_G \left(1+8\nu\left[1+8\sqrt{2\ln(4m/\delta)}\right]\right)\Big[\left(\left[3+16\sqrt{\frac{\ln(4m/\delta)}{d_x}}\right]^{\alpha}+\frac{1}{16^{\alpha}d_x^{\alpha/2}}\right)$$
$$+\left(\max\left\{\frac{256}{d_x}\ln\left(\frac{32L_G\sqrt{2M_2}}{\alpha 16^{\alpha}(C+1)w_{\text{avg}}k_G}\frac{d_x^{\alpha/2-1}\sqrt{d_x+64}}{\sigma^{\alpha-1}}m^2\right),0\right\}\right)^{\alpha/2}\Big]\Big\}\frac{\sigma^{\alpha}}{n_1+\cdots+n_T}\max_{t\in[T]}\{2^{2(T-t)}\beta_t\} \quad (57)$$
$$+8C\nu^2\frac{pT\ln\left(1+\frac{2B(C+1)}{C(1+8\nu\left[1+8\sqrt{2\ln(4m/\delta)}\right])}\frac{Tm}{\max_{t\in T}2^{2(T-t)}\beta_t}\right)}{n_1+\cdots+n_T}+8C\nu^2\frac{\ln(4/\delta)}{n_1+\cdots+n_T}\max_{t\in[T]}\{2^{2(T-t)}\beta_t\},$$

We finish the proof with $\beta_t = 1/4^{T-t}$. □

Proof for Global Minimizers of (12). We now consider objective (12) that is associated with the regularizer $\Omega_T(\theta) = \sum_{t\in[T-1]}\sum_{i\in\mathcal{R}_t}\beta_t\cdot\|f_\theta(x_{ti})-f_{\hat{\theta}_t}(x_{ti})\|_2^2$. Write $\hat{f}_t := f_{\hat{\theta}_t}$. Similarly to the proof of Theorem 5, we unify the notation by defining $\mathcal{R}_T := [m]$ and $n_T := m$. Similarly to (53), we have

$$\begin{aligned}&\sum_{t\in[T]}\sum_{i\in\mathcal{R}_t}\beta_t\cdot\|f^*(x_{ti})-\hat{f}_T(x_{ti})\|_2^2\\ &\le\frac{\beta_T}{s-1}\cdot M_T(\hat{f}_T)+\frac{s^2}{s-1}\sum_{t\in[T-1]}\sum_{i\in\mathcal{R}_t}\beta_t\cdot\|f^*(x_{ti})-\hat{f}_t(x_{ti})\|_2^2.\end{aligned} \quad (58)$$

Note here that the rightmost term is with $\hat{f}_t$, not $\hat{f}_{T-1}$.

Similarly to the beginning of the proof of Theorem 4, we have for every $t\in[T-1]$ that

$$\begin{aligned}\sum_{i\in\mathcal{R}_t}\beta_t\cdot\|f^*(x_{ti})-\hat{f}_t(x_{ti})\|_2^2&\le\sum_{i\in[m]}\beta_t\cdot\|f^*(x_{ti})-\hat{f}_t(x_{ti})\|_2^2\\ &\le\sum_{i\in[m]}2\beta_t\cdot\langle v_{ti},\hat{f}_t(x_{ti})-f^*(x_{ti})\rangle+\Omega_t(f^*)-\Omega_t(\hat{f}_t)\\ &\le\sum_{i\in[m]}2\beta_t\cdot\langle v_{ti},\hat{f}_t(x_{ti})-f^*(x_{ti})\rangle+\Omega_t(f^*)\\ &=\sum_{i\in[m]}2\beta_t\cdot\langle v_{ti},\hat{f}_t(x_{ti})-f^*(x_{ti})\rangle\\ &\quad+\sum_{\tau\in[t-1]}\sum_{i\in\mathcal{R}_t}\beta_\tau\cdot\|f^*(x_{\tau i})-\hat{f}_\tau(x_{\tau i})\|_2^2.\end{aligned}$$

Multiplying both sides by $s/(s-1)$ and rearranging the terms with the definition of $M_t(f)$ in (52), we get

$$\sum_{i\in\mathcal{R}_t}\beta_t\cdot\|f^*(x_{ti})-\hat{f}_t(x_{ti})\|_2^2\le\frac{\beta_t}{s-1}\cdot M_t(\hat{f}_t)+\frac{s}{s-1}\sum_{\tau\in[t-1]}\sum_{i\in\mathcal{R}_\tau}\beta_\tau\cdot\|f^*(x_{\tau i})-\hat{f}_\tau(x_{\tau i})\|_2^2.$$

This inequality is of the form $a_t\le c_t+\frac{s}{s-1}\sum_{\tau\in[t-1]}a_\tau$, and it implies

$$\sum_{t\in[T-1]}a_t\le c_{T-1}+\left(\frac{s}{s-1}+1\right)\sum_{t\in[T-2]}a_t=c_{T-1}+\frac{2s-1}{s-1}\sum_{t\in[T-2]}a_t.$$

Unrolling this recurrence relation gives

$$\sum_{t\in[T-1]}a_t\le\sum_{t\in[T-1]}\left(\frac{2s-1}{s-1}\right)^{T-t-1}c_t.$$

We have just shown that

$$\sum_{t\in[T-1]}\sum_{i\in\mathcal{R}_t}\beta_t\cdot\|f^*(x_{ti})-\hat{f}_t(x_{ti})\|_2^2\le\sum_{t\in[T-1]}\left(\frac{2s-1}{s-1}\right)^{T-t-1}\beta_t\cdot M_t(\hat{f}_t).$$

Substituting this back to (58) with $s=2$ yields

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{i\in\mathcal{R}_t}\beta_t\cdot\|f^*(x_{ti})-\hat{f}_T(x_{ti})\|_2^2\\
&\le\frac{\beta_T}{s-1}\cdot M_T\big(\hat{f}_T\big)+\frac{s^2}{s-1}\sum_{t\in[T-1]}\left(\frac{2s-1}{s-1}\right)^{T-t-1}\beta_t\cdot M_t(\hat{f}_t)\\
&=\beta_T\cdot M_T\big(\hat{f}_T\big)+4\sum_{t\in[T-1]}3^{T-t-1}\beta_t\cdot M_t(\hat{f}_t)\\
&\le\sum_{t\in[T]}4^{T-t}\beta_t\cdot M_t(\hat{f}_t).
\end{aligned}$$

This is identical to (55), thus the rest of the steps follow directly from the previous proof. □

C.1 Proposition 3 and Its Proof

Proposition 3. *Recall that $M_t(f)$ is defined in (52) as*

$$M_t(f)=\sum_{i\in[m]}\left[2s\cdot\langle v_{ti},f(x_{ti})-f^*(x_{ti})\rangle-\|f^*(x_{ti})-f(x_{ti})\|_2^2\right]$$

where $s=2$ and v_{ti} is conditionally sub-Gaussian with proxy variance ν^2. Fix $u>0$. Recall $\beta_t>0$ and define $w_t:=s^{2(T-t)}\beta_t$ ($\forall t\in[T]$). Let $r_x>2\sigma, r_v>2\nu\sqrt{d_y}$, and ε-net $\mathcal{F}(\varepsilon)$ be the smallest ε-net of $\mathcal{F}\setminus\{f^*\}$. Then we have

$$\sum_{t\in[T]}\frac{w_t}{Tn_t}\cdot M_t(f_t)\le w_{avg}K_G\varepsilon\,(1+4r_v)+u/2,\qquad\forall f_1,\dots,f_T\in\mathcal{F},$$

with probability at least

$$|\mathcal{F}(\varepsilon)|^T\cdot\exp\left(-u\cdot\frac{T}{4\nu^2s^2}\min_{t\in[T]}\frac{n_t}{s^{2(T-t)}\beta_t}\right)+m\cdot\exp\left(-\frac{d_x(r_x-2\sigma)^2}{128\sigma^2}\right)+m\cdot\exp\left(-\frac{(r_v-2\nu\sqrt{d_y})^2}{128\nu^2}\right).$$

Proof. The proof is almost the same as Proposition 2, except that we apply union bound for each $M_t(f)$ using ε-net; this gives us the exponent T in the failure probability. □

Lemma 6. *Fix $f_1,\dots,f_T\in\mathcal{F}$. Recall $\beta_t>0$ and that $M_t(f)$ is defined in (52) as*

$$M_t(f)=\sum_{i\in[m]}\left[2s\cdot\langle v_{ti},f(x_{ti})-f^*(x_{ti})\rangle-(s-1)\cdot\|f^*(x_{ti})-f(x_{ti})\|_2^2\right],$$

where v_{ti} conditioned on $x_{1i},\dots,x_{ti}$ is sub-Gaussian with proxy variance ν^2. Then

$$\mathbb{P}\left(\sum_{t\in[T]}\frac{s^{2(T-t)}\beta_t}{Tn_t}\cdot M_t\big(f_t\big)\ge u\right)\le\exp\left(-u\cdot\frac{T(s-1)}{2\nu^2s^2}\min_{t\in[T]}\frac{n_t}{s^{2(T-t)}\beta_t}\right).$$

Proof. This follows from a similar proof to Lemma 5. □

D Discussion on Extension of Results

In this section, we will discuss modifications of existing proofs to accommodate noise in (2), i.e., noisy transformation of tasks rather than deterministic transformation. Finally, our Assumption 2 requires that the true predictor f^* to be realizable in the class of learnable functions; here we will provide pointer to extend our results.

Noisy tasks setting. Consider the following model

$$y_t = f^*(x_t) + v_t, \qquad \forall t \geq 1; \tag{59}$$
$$x_t = g_t(x_1, \ldots, x_{t-1}) + \eta_t, \quad \forall t > 1, \tag{60}$$

where η_t is conditionally independent sub-Gaussian noise with proxy variance σ_η^2. Our analysis relies primarily on invoking Proposition 1, Proposition 2, Proposition 3 and **??**. The crucial step in establishing these intermediate results is obtaining tail bounds for

$$Q(f, C) = \frac{1}{T} \sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t} \left(\mathbb{E}\left[\|f(x_t)\|_2^2 \right] - C \cdot \|f(x_{ti})\|_2^2 \right). \tag{61}$$

The key difficulty arises from the temporal dependencies in the noisy setting. In the noiseless case analyzed previously, we could deterministically relate $f(x_t)$ to x_1, and directly apply Lipschitz concentration bounds. For the case here with additive noise η_t, we can in principle make use of and extend the results extensively studied in the system identification literature (see Matni and Tu (2019) for a detailed survey).

Non-realizable setting. We now consider the case where f^* does not belong to the realizable function class, meaning our model is misspecified. In this setting, (27) becomes

$$\mathbb{E}\left[\frac{1}{T} \sum_{t\in[T]} w_t \cdot \|\tilde{f}(x_t) - \hat{f}(x_t)\|_2^2 \right] \leq \frac{1}{T} \sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t} \left(\mathbb{E}\left[\|f(x_t) - \tilde{f}(x_t)\|_2^2 \right] - C \cdot \|f(x_{ti}) - \tilde{f}(x_{ti})\|_2^2 \right) \tag{62}$$
$$+ \frac{C}{T} \sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t} \left[4 \cdot \langle v_{ti} - f^*(x_{ti}), f(x_{ti}) - \tilde{f}(x_{ti}) \rangle - \|\tilde{f}(x_{ti}) - f(x_{ti})\|_2^2 \right] \tag{63}$$
$$+ 2C\lambda \left[\Omega_T(\tilde{f}) - \Omega_T(\hat{f}) \right], \tag{64}$$

where $\tilde{f} = f_{\tilde{\theta}}$ and $\tilde{\theta} \in \operatorname{argmin}_{\theta\in\Theta} \mathbb{E}\left[\frac{1}{T} \sum_{(t,i)\in\mathcal{M}} \frac{w_t}{n_t} \cdot \mathcal{L}\Big(y_{ti}, f_\theta(x_{ti}) \Big) + \lambda \cdot \Omega_T(\theta) \right]$.

The key challenge arises in the term $\langle v_{ti} - f^*(x_{ti}), f(x_{ti}) - \tilde{f}(x_{ti}) \rangle - \|\tilde{f}(x_{ti}) - f(x_{ti})\|_2^2$, where the first argument in the inner product is no longer independent of x_t. This dependence is critical because the proof of Proposition 2 relied crucially on the independence between the first and second arguments of the inner product.

To obtain concentration bounds for such complex dependent terms, we require sophisticated mixed-tail concentration results. The framework developed by Maurer and Pontil (2021) provides such tools, and Ziemann et al. (2024b) present concrete results that appear well-suited to our setting. Adapting their approach to handle the dependence structure in our problem remains a promising direction for future work.

E Related Work on Distance Measures

Here we review some relevant works on multitask learning, transfer learning, and domain adaptation, with a focus on the distance measures that arise in generalization bounds. Such bounds typically depend linearly on the distance $\mathrm{dist}(\pi_1, \pi_2)$ between two distributions, π_1 and π_2. In the main paper, we discussed the discrepancy distance in Example 1 and Remark 3. Furthermore, there have been multiple choices of such distance. Here, we examine a few other choices and evaluate whether they are suitable for our case.

The first distance we consider is the so-called $\mathcal{H}$-divergence (Ben-David et al., 2010; Shui et al., 2019). In Ben-David et al. (2010) the $\mathcal{H}$-divergence is defined for classification tasks, while here we consider regression losses. In Shui et al. (2019), the $\mathcal{H}$-divergence is defined exactly the same as the discrepancy distance in Remark 3, with a difference that the ℓ_1 loss is considered in Shui et al. (2019). Similarly to Remark 3, the discrepancy with the ℓ_1 loss could grow unbounded,

e.g., if the two distributions are $\mathcal{N}(0,1)$ and $\mathcal{N}(a,1)$, respectively, with a growing polynomially with the problem size (e.g., dimension, number of samples). The second distance is called $\mathcal{Y}$-*discrepancy* (Mohri and Muñoz Medina, 2012; Wang et al., 2023). It is similar to the discrepancy distance in Remark 3, and we can show it is unbounded for simple distributions as well.

F Basic Definitions and Auxiliary Lemmas

Here we present some basic definitions and lemmas that serve as fundamental building blocks for our proof. In Section F.1 we develop several lemmas tailored for our proof. In Section F.2, we collect notations and results from high dimensional statistics.

F.1 Notations and Lemmas

Denote by $\mathcal{P}_{\mathbb{B}_r(d)}(\cdot)$ the projection of its input onto $\mathbb{B}_r(d)$. Define the supremum of the second order moment of $G_{f,t}(x_1)$:

$$M_2 := \sup_{f\in\mathcal{F}} \sup_{t\in[T],w_t>0} \mathbb{E}\left[\|G_{f,t}(x_1)\|_2^2\right]. \tag{65}$$

Note that M_2 is finite under Assumption 3 (see Lemma 12).

Lemma 7. *Let $\{x_{1i}\}_{i=1}^n$ be n independent random variables in $\mathbb{R}^{d_x}$, each distributed according to π_1. Suppose we have deterministic functions $g_2,\dots,g_t$ and generate for each $i\in\{1,\dots,n\}$:*

$$x_{ki} = g_k\big(x_{1i}, x_{2i}, \dots, x_{k-1,i}\big) \quad \textit{for } k=2,\dots,t. \tag{66}$$

Then, for each i, we have trajectories $\big(x_{1i},x_{2i},\dots,x_{ti}\big)$. These trajectories are also mutually independent, that is, with $i\neq j$ we have

$$(x_{1i},x_{2i},\dots,x_{ti}) \;\perp\; (x_{1j},x_{2j},\dots,x_{tj}). \tag{67}$$

Proof. By assumption, the random variables x_{1i} are mutually independent for $i=1,\dots,n$. Then, for each i, note that the trajectory $\big(x_{1i},x_{2i},\dots,x_{ti}\big)$ is obtained via a deterministic transformation of x_{1i}, namely,

$$x_{2i}=g_2(x_{1i}),\quad x_{3i}=g_3(x_{1i},x_{2i}),\quad \dots,\quad x_{ti}=g_t\big(x_{1i},\dots,x_{t-1,i}\big). \tag{68}$$

Hence the trajectory $\big(x_{1i},x_{2i},\dots,x_{ti}\big)$ is a measurable (deterministic) function of the initial variable x_{1i}. Since trajectory $\big(x_{1i},\dots,x_{ti}\big)$ depends *only* on x_{1i}, and $\big(x_{1j},\dots,x_{tj}\big)$ depends *only* on x_{1j}, the mutual independence of x_{1i} and x_{1j} for $i\neq j$ directly implies

$$\big(x_{1i},\dots,x_{ti}\big) \;\perp\; \big(x_{1j},\dots,x_{tj}\big). \tag{69}$$

This finishes the proof. □

Lemma 8. *For all $c\neq 0$ we have the following identity:*

$$\int \tau\cdot\exp(c\tau)\,d\tau = \frac{c\tau-1}{c^2}\cdot\exp(c\tau).$$

Moreover, if $c<0$ then we have

$$\int_0^\infty \tau\cdot\exp(c\tau)\,d\tau = \frac{1}{c^2}.$$

Proof. The first identity follows from *integration by parts*, or by observing that the derivative of $\frac{c\tau-1}{c^2}\cdot\exp(c\tau)$ with respect to τ is equal to $\tau\cdot\exp(c\tau)$. The second identity follows by evaluating the integral at infinity and at 0. □

Lemma 9. *For k non-negative random variables $z_1,\dots,z_k$ with $\mathbb{E}[z_i]\neq 0$ we have*

$$\mathbb{E}\left[\left(\sum_{i=1}^k z_i\right)^2\right] \le \left(\sum_{i=1}^k \mathbb{E}[z_i]\right)^2 \cdot \max_{i\in[k]}\frac{\mathbb{E}[z_i^2]}{\mathbb{E}[z_i]^2}.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}(z_1+\cdots+z_k)^2 &= \sum_{i=1}^k \mathbb{E}[z_i^2] + \sum_{i\neq j} 2\mathbb{E}[z_i z_j] \\
&\overset{\text{(i)}}{\leq} \sum_{i=1}^k \mathbb{E}[z_i^2] + \sum_{i\neq j} 2\sqrt{\mathbb{E}[z_i^2]\mathbb{E}[z_j^2]} \\
&\leq \sum_{i=1}^k \mathbb{E}[z_i]^2 \cdot \max_{i\in[k]} \frac{\mathbb{E}[z_i^2]}{\mathbb{E}[z_i]^2} + \sum_{i\neq j} 2\sqrt{\mathbb{E}[z_i]^2\mathbb{E}[z_j]^2} \cdot \max_{i\in[k]} \frac{\mathbb{E}[z_i^2]}{\mathbb{E}[z_i]^2} \\
&= \sum_{i=1}^k \mathbb{E}[z_i]^2 \cdot \max_{i\in[k]} \frac{\mathbb{E}[z_i^2]}{\mathbb{E}[z_i]^2} + \sum_{i\neq j} 2\mathbb{E}[z_i]\mathbb{E}[z_j] \cdot \max_{i\in[k]} \frac{\mathbb{E}[z_i^2]}{\mathbb{E}[z_i]^2} \\
&= \left(\sum_{i=1}^k \mathbb{E}[z_i]\right)^2 \cdot \max_{i\in[k]} \frac{\mathbb{E}[z_i^2]}{\mathbb{E}[z_i]^2}
\end{aligned}$$

where (i) follows from the Cauchy-Schwarz inequality. □

Lemma 10. *For k non-negative random variables $z_1,\ldots,z_k$ with $\mathbb{E}[z_i]\neq 0$, and for any fixed positive constant $\alpha > 0, C > 1$ we have*

$$\mathbb{E}\left[\exp\left(-C\alpha\sum_{i=1}^k z_i\right)\right] \leq \exp\left(-C\alpha\sum_{i=1}^k \mathbb{E}[z_i] + \frac{C^2\alpha^2}{2}\left(\sum_{i=1}^k \mathbb{E}[z_i]\right)^2 \cdot \max_{i\in[k]} \frac{\mathbb{E}[z_i^2]}{\mathbb{E}[z_i]^2}\right),$$

and moreover we have

$$\mathbb{P}\left(C\sum_{i=1}^k z_i \leq \sum_{i=1}^k \mathbb{E}[z_i]\right) \leq \exp\left(-\frac{C-1}{2C^2\max_{i\in[k]}\frac{\mathbb{E}[z_i^2]}{\mathbb{E}[z_i]^2}}\right).$$

Proof. For any $\alpha > 0$, Markov's inequality gives

$$\begin{aligned}
\mathbb{P}\left(C\sum_{i=1}^k z_i \leq \sum_{i=1}^k \mathbb{E}[z_i]\right) &= \mathbb{P}\left(\exp\left(-C\alpha\sum_{i=1}^k z_i\right) \geq \exp\left(-\alpha\sum_{i=1}^k \mathbb{E}[z_i]\right)\right) \\
&\leq \exp\left(\alpha\sum_{i=1}^k \mathbb{E}[z_i]\right)\cdot\mathbb{E}\left[\exp\left(-C\alpha\sum_{i=1}^k z_i\right)\right].
\end{aligned}$$

We upper bound the rightmost terms:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(-C\alpha\sum_{i=1}^k z_i\right)\right] &\overset{\text{(i)}}{\leq} \mathbb{E}\left[1 - C\alpha\sum_{i=1}^k z_i + \frac{C^2\alpha^2}{2}\left(\sum_{i=1}^k z_i\right)^2\right] \\
&\overset{\text{(ii)}}{\leq} \exp\left(-C\alpha\sum_{i=1}^k \mathbb{E}[z_i] + \frac{C^2\alpha^2}{2}\mathbb{E}\left[\left(\sum_{i=1}^k z_i\right)^2\right]\right) \\
&\overset{\text{(iii)}}{\leq} \exp\left(-C\alpha\sum_{i=1}^k \mathbb{E}[z_i] + \frac{C^2\alpha^2}{2}\left(\sum_{i=1}^k \mathbb{E}[z_i]\right)^2 \cdot \max_{i\in[k]}\frac{\mathbb{E}[z_i^2]}{\mathbb{E}[z_i]^2}\right)
\end{aligned}$$

where (i) follows from the inequality $\exp(-a) \leq 1 - a + a^2/2$ ($\forall a \geq 0$), (ii) from the inequality $1 + a \leq \exp(a)$ ($\forall a$), and (iii) from Lemma 9. Combining the above gives

$$\mathbb{P}\left(C\sum_{i=1}^k z_i \leq \sum_{i=1}^k \mathbb{E}[z_i]\right) \leq \exp\left(-(C-1)\alpha\sum_{i=1}^k \mathbb{E}[z_i] + \frac{C^2\alpha^2}{2}\left(\sum_{i=1}^k \mathbb{E}[z_i]\right)^2 \cdot \max_{i\in[k]}\frac{\mathbb{E}[z_i^2]}{\mathbb{E}[z_i]^2}\right).$$

Since the above holds for any α and since $C > 1$, we set α to

$$\alpha = \frac{C-1}{C^2 \sum_{i=1}^{k} \mathbb{E}[z_i]} \cdot \frac{1}{\max_{i\in[k]} \frac{\mathbb{E}[z_i^2]}{\mathbb{E}[z_i]^2}}$$

and plug this back into the above inequality to obtain the desired result. □

Lemma 11. *Let $S_1, \ldots, S_m$ be subsets of $[T]$, and denote by n_t the number of times that the index t is contained in $S_1, \ldots, S_m$, that is*

$$n_t := \sum_{i\in[m]} \mathbb{1}(t \in S_i).$$

where $\mathbb{1}(\cdot)$ is the binary indicator function that outputs 1 if its input statement is true, or outputs 0 otherwise. Then, for T non-negative numbers $a_1, \ldots, a_T$, we have

$$\sum_{i\in[m]} \left(\sum_{t\in S_i} \frac{a_t}{n_t} \right)^2 \leq \frac{1}{\min_{t\in[T]}\{n_t\}} \cdot \left(\sum_{i\in[T]} a_t \right)^2.$$

Proof. We expand the square on the left-hand side and make the following observations:

- The left-hand side has the term a_t^2/n_t, as each square gives a_t^2/n_t^2 and there are n_t such sets that contains index t. Therefore we have
$$\frac{a_t^2}{n_t} \leq \frac{a_t}{\min_{t\in[T]}\{n_t\}}.$$
- For any $t_1, t_2 \in [T]$ The left-hand side has the term $\frac{2a_{t_1}a_{t_2}}{n_{t_1}n_{t_2}} \cdot c_{t_1,t_2}$, where c_{t_1,t_2} denotes the number of times that both indices t_1, t_2 are contained in the sets $S_1, \ldots, S_m$, that is
$$c_{t_1,t_2} := \sum_{i\in[m]} \mathbb{1}\left(t_1 \in S_i \text{ and } t_2 \in S_i\right).$$
By definition we have $c_{t_1,t_2} \leq \min\{n_{t_1}, n_{t_2}\}$. Therefore
$$\frac{2a_{t_1}a_{t_2}}{n_{t_1}n_{t_2}} \cdot c_{t_1,t_2} \leq \frac{2a_{t_1}a_{t_2}}{\max\{n_{t_1}, n_{t_2}\}} \leq \frac{2a_{t_1}a_{t_2}}{\min_{t\in[T]}\{n_t\}}.$$

These prove the desired inequality. □

Lemma 12. *Under Assumption 3, we have $\kappa < \infty$ and $M_2 < \infty$, where κ is defined in (8) and M_2 in (65).*

Proof. From Jensen's inequality $\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^2\right] \leq \sqrt{\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^4\right]}$ and Assumption 3 that $\mathbb{E}\left[\|G_{f,t}(x_1)\|_2^4\right]$ is finite for all $f \in \mathcal{F}$ and $t \in [T]$, we see that M_2 is indeed finite. That κ is finite also follows from Assumption 3. □

Proposition 4. *Suppose $\alpha, \beta, \gamma > 0$. Whenever $r = \max\{\zeta, 0\} + \sqrt{\max\{\ln(\beta\gamma^{\alpha/2})/\gamma, 0\}}$ it holds that:*

$$\min_r r^\alpha + \beta \mathsf{exp}(-\gamma(r-\zeta)^2) \leq \min\left\{ \beta + \max\{\zeta^\alpha, 0\}, \frac{1}{\gamma^{\alpha/2}} + \left[\max\{\zeta, 0\} + \sqrt{\max\left\{ \frac{\ln\left(\beta\gamma^{\alpha/2}\right)}{\gamma}, 0 \right\}} \right]^\alpha \right\}. \quad (70)$$

Proof. We first consider the case when $\zeta > 0$, we obtain:

$$\left(r^\alpha + \beta \mathsf{exp}(-\gamma(r-\zeta)^2)\right) = \left[\zeta + \sqrt{\max\{\ln(\beta\gamma^{\alpha/2})/\gamma, 0\}} \right]^\alpha + \frac{1}{\gamma^{\alpha/2}}.$$

Whenever, $\zeta \leq 0$ we can upper bound the function on r by dropping ζ:

$$\left(r^\alpha + \beta \mathsf{exp}(-\gamma(r-\zeta)^2)\right) \leq \left(r^\alpha + \beta \mathsf{exp}(-\gamma r^2)\right).$$

Now this reduces to the case when $\zeta = 0$. This completes the proof. □

F.2 Definitions and Lemmas from High-Dimensional Statistics

Here we collect some basic definitions used and auxiliary lemmas from high-dimensional statistics and probability. For a more detailed account on this subject, see, e.g., Rigollet and Hütter (2023); Vershynin (2018); Wainwright (2019). We mostly follow the presentation of Rigollet and Hütter (2023). For a few lemmas that we did not find exact statements or proofs in Rigollet and Hütter (2023); Vershynin (2018); Wainwright (2019), we provide independent proofs.

Sub-Gaussian Random Variables. A random variable ξ is called *sub-Gaussian* with proxy variance σ^2 if it has mean zero and satisfies

$$\mathbb{E}[\exp(u\xi)] \leq \exp\left(\frac{\sigma^2 u^2}{2}\right). \tag{71}$$

A random vector $z = [z_1, \ldots, z_d]^\top$ is called sub-Gaussian with proxy variance σ^2 if it has mean zero and if $z^\top \omega$ is sub-Gaussian with proxy variance σ^2 for every unit vector $\omega \in \mathbb{R}^d$ (with $\|\omega\|_2 = 1$). It follows from (71) that, if a sub-Gaussian vector z has independent coordinates then

$$\mathbb{E}[\exp(z^\top \omega)] \leq \exp\left(\frac{\sigma^2 \cdot \|\omega\|_2^2}{2}\right), \quad \forall \omega \in \mathbb{R}^d. \tag{72}$$

In particular, taking ω to be standard basis vectors of $\mathbb{R}^d$, we see that every entry z_i is a sub-Gaussian random variable with proxy variance σ^2.

Sub-exponential Random Variables. A random variable ξ is called *sub-exponential* with parameter s if it has zero mean and satisfies

$$\mathbb{E}[\exp(u\xi)] \leq \exp\left(\frac{s^2 u^2}{2}\right), \quad \forall u \in \left[-\frac{1}{s}, \frac{1}{s}\right] \tag{73}$$

This defining property is identical to (71), with one difference that u is now constrained to lie in the interval $\left[-\frac{1}{s}, \frac{1}{s}\right]$.

We now state a few lemmas about sub-Gaussian and sub-exponential random variables.

Lemma 13 (see, e.g., Lemmas 1.3 and 1.4 of Rigollet and Hütter (2023))**.** *For a sub-Gaussian random variable ξ with proxy variance σ^2, it holds that*

$$\mathbb{P}(\xi > u) \leq \exp\left(-\frac{u^2}{2\sigma^2}\right), \forall u \in \mathbb{R}.$$

Furthermore, we have $\mathbb{E}[\xi^2] \leq 4\sigma^2$.

Lemma 14 (see, e.g., Lemma 1.12 of Rigollet and Hütter (2023))**.** *If ξ is sub-Gaussian with proxy variance σ^2, then $\xi^2 - \mathbb{E}[\xi^2]$ is sub-exponential with parameter $16\sigma^2$.*

Lemma 15 (Bernstein's inequality, see, e.g., Theorem 1.13 of Rigollet and Hütter (2023))**.** *Let $\xi_1, \ldots, \xi_d$ be independent sub-exponential random variables, each with parameter s. Then, for any $u > 0$, we have*

$$\mathbb{P}\left[\left|\sum_{i=1}^{d} \xi_i\right| \geq u\right] \leq \exp\left[-\frac{1}{2}\min\left(\frac{u^2}{s^2 d}, \frac{u}{s}\right)\right].$$

Lemma 16. *Suppose $z = [z_1, \ldots, z_d]^\top$ is a sub-Gaussian vector with proxy variance σ^2/d and independent coordinates. Then we have*

$$\mathbb{P}\left[\|z\|_2 - 2\sigma \geq u\right] \leq \exp\left[-\frac{du^2}{128\sigma^2}\right], \quad \forall u > 0.$$

Proof. From Lemma 14 we know that $z_i^2 - \mathbb{E}[z_i^2]$ is sub-exponential with parameter $16\sigma^2/d$. From Lemma 13 we know that $\mathbb{E}[\|z\|_2^2] \leq 4\sigma^2$. Moreover, by our assumption, $z_1^2 - \mathbb{E}[z_1^2], \ldots, z_d^2 - \mathbb{E}[z_d^2]$ are independent. Then we apply Lemma 15

(Bernstein's inequality) and obtain

$$\begin{aligned}\mathbb{P}\left[\frac{\|z\|_2^2}{4\sigma^2} - 1 \geq \frac{u}{4\sigma^2}\right] &= \mathbb{P}\left[\|z\|_2^2 - 4\sigma^2 \geq u\right] \\ &\leq \mathbb{P}\left[\|z\|_2^2 - \mathbb{E}[\|z\|_2^2] \geq u\right] \\ &\leq \mathbb{P}\left[\sum_{i=1}^{d}\left(z_i^2 - \mathbb{E}[z_i^2]\right) \geq u\right] \\ &\leq \mathbb{P}\left[\Big|\sum_{i=1}^{d}\left(z_i^2 - \mathbb{E}[z_i^2]\right)\Big| \geq u\right] \\ &\leq \exp\left[-\frac{d}{2}\min\left(\frac{u^2}{256\sigma^4}, \frac{u}{16\sigma^2}\right)\right].\end{aligned}$$

We now use the technique of (Vershynin, 2018, Theorem 3.1.1) to obtain a bound for $\|z\|_2$. Since $\frac{\|z\|_2}{2\sigma} - 1 \geq \frac{u_1}{2\sigma}$ implies $\frac{\|z\|_2^2}{4\sigma^2} - 1 \geq \frac{u_1^2}{4\sigma^2} + \frac{u_1}{\sigma} \geq \max\{\frac{u_1^2}{4\sigma^2}, \frac{u_1}{2\sigma}\}$ for all $u_1 > 0$, we choose u such that $\frac{u}{4\sigma^2} = \max\{\frac{u_1^2}{4\sigma^2}, \frac{u_1}{2\sigma}\}$ and obtain

$$\begin{aligned}\mathbb{P}\left[\|z\|_2 - 2\sigma \geq u_1\right] &= \mathbb{P}\left[\frac{\|z\|_2}{2\sigma} - 1 \geq \frac{u_1}{2\sigma}\right] \\ &\leq \mathbb{P}\left[\frac{\|z\|_2^2}{4\sigma^2} - 1 \geq \max\left\{\frac{u_1^2}{4\sigma^2}, \frac{u_1}{2\sigma}\right\}\right] \\ &\leq \exp\left[-\frac{d}{2}\min\left(\frac{\max\left\{\frac{u_1^4}{16\sigma^4}, \frac{u_1^2}{4\sigma^2}\right\}}{16}, \frac{\max\left\{\frac{u_1^2}{4\sigma^2}, \frac{u_1}{2\sigma}\right\}}{4}\right)\right] \\ &\leq \exp\left[-\frac{d}{8}\min\left(\max\left\{\frac{u_1^4}{16\sigma^4}, \frac{u_1^2}{4\sigma^2}\right\}, \max\left\{\frac{u_1^2}{4\sigma^2}, \frac{u_1}{2\sigma}\right\}\right)\right] \\ &= \exp\left[-\frac{d}{32}\cdot\frac{u_1^2}{4\sigma^2}\right].\end{aligned}$$

The last equality follows from a simple discussion on whether $\frac{u_1}{2\sigma}$ is greater than 1 or not. The proof is now complete. □

Lemma 17. *Let $z_1, \ldots, z_n$ be d-dimensional sub-Gaussian vectors, each with proxy variance σ^2/d and independent coordinates. Then for any $r > 2\sigma$ and some universal constant $c > 0$ we have*

$$\mathbb{P}\left(\|z_i\|_2 \leq r, \forall i \in [n]\right) \geq 1 - n \cdot \exp\left(-\frac{d(r-2\sigma)^2}{128\sigma^2}\right).$$

Proof. In Lemma 16, take $u = \sigma(a-2)$ with $a > 2$, and we obtain

$$\mathbb{P}\left[\|z_i\|_2 \geq a\sigma\right] \leq \exp\left[-\frac{d(a-2)^2}{128}\right], \quad \forall a > 2.$$

Then take $r = a\sigma$, and we obtain

$$\mathbb{P}\left[\|z_i\|_2 \geq r\right] \leq \exp\left[-\frac{d(r-2\sigma)^2}{128\sigma^2}\right], \quad \forall r > 2\sigma.$$

Applying the union bound finishes the proof. □

Nets. In our proof, we will often consider some specific subset of the function class $\mathcal{F}$, called ϵ-net. The defining property of an ϵ-net $\mathcal{F}(\epsilon)$ is that, for every $f \in \mathcal{F}$, there exists some $f' \in \mathcal{F}(\varepsilon)$ satisfying $\|f' - f\|_{\mathcal{F}} \leq \varepsilon$; thus, the definition of $\mathcal{F}(\varepsilon)$ implicitly depends on the norm $\|\cdot\|_{\mathcal{F}}$.

Lemma 18. *Suppose (6) holds. Let $\Theta(\varepsilon) := \{\theta_1, \ldots, \theta_N\}$ be the smallest ε-net of Θ. Then $\{f_{\theta_1}, \ldots, f_{\theta_N}\}$ is an $(L_{\mathcal{F}}\varepsilon)$-net of $\mathcal{F}$. Thus, the smallest ε-net of $\mathcal{F}\setminus\{f^*\}$ has smaller size than the smallest $(\varepsilon/L_{\mathcal{F}})$-net of Θ.*

Proof. For any $f_\theta \in \mathcal{F} \setminus \{f^*\}$, there is some $\theta_i \in \Theta$ such that $\|\theta - \theta_i\|_2 \leq \varepsilon$. By (6) we have

$$\|f_\theta - f_{\theta_i}\|_{\mathcal{F}} \leq L_{\mathcal{F}} \cdot \|\theta - \theta_i\|_2 \leq L_{\mathcal{F}} \cdot \varepsilon.$$

The proof is finished. □

Lemma 19. *Assume Θ is a bounded subset of $\mathbb{R}^p$ with bounded diameter $\operatorname{diam}(\Theta)$ (Assumption 2). Then the smallest ε-net of Θ is at most of size*

$$\exp\left(p \cdot O\left(\ln \frac{\operatorname{diam}(\Theta)}{\varepsilon}\right)\right).$$

Proof. Note that Θ can be covered by a ball of radius $\operatorname{diam}(\Theta)$. Then invoke, e.g., Proposition 5 of Cucker and Smale (2002). □

Lemma 20 (Integral Identity, cf. Lemma 1.2.1 of Vershynin (2018))**.** *For a non-negative random variable ξ we have*

$$\mathbb{E}[\xi] = \int_0^\infty \mathbb{P}(\xi > \tau)\, d\tau.$$