PAC-Bayesian Bounds for Learning Partially Observed Stochastic Linear Time-Invariant State-Space Systems with Inputs and Sub-Gaussian Noise

Mihaly Petreczky
Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille,

Mohamad Al Ahab
Department of Electronic Systems, Aalborg University, Denmark

John Leth
Department of Electronic Systems, Aalborg University, Denmark

Abstract

In this paper we derive a Probably Approximately Correct (PAC)-Bayesian error bound for partially observed linear time-invariant (LTI) stochastic dynamical systems in state-space form with inputs. Such bounds are widespread in machine learning, and they are useful for characterizing the predictive power of models learned from finitely many data points. The bound derived in this paper relates the expectation of prediction errors with the prediction error generated by the model on the data used for learning. In addition, we show that it can also be used to derive bounds for the parameter estimation error. In turn, this allows us to provide finite-sample error bounds for the prediction error and parameter estimation error for a wide class of system identification algorithms.

1 Introduction

PAC and PAC-Bayesian bounds are fundamental tools for analyzing learning algorithms. They provide generalization guarantees that relate the *true error* (performance on unseen data) to the *empirical error* (performance on training data), by bounding the *generalization gap*, the difference between true and empirical error, in a way that is independent of the specific learning algorithm. In this framework, learning is viewed as selecting a model from a data-dependent posterior distribution over hypotheses, which is a modification of a prior based on the observed data. Different learning algorithms correspond to different choices of priors, posteriors, and selection mechanisms (e.g., sampling, averaging, or maximizing the posterior). PAC-Bayesian bounds then provide a uniform upper bound on the average generalization gap with respect to any posterior, thus offering generalization guarantees for a broad class of learning and model selection methods.

This generality makes PAC-Bayesian approaches especially useful for analyzing complex models, including neural networks, where they can yield non-vacuous generalization bounds [7]. Although PAC and PAC-Bayesian theory is well developed for i.i.d. settings [36, 2, 17], there are comparatively few results for dynamical systems. This paper presents PAC-Bayesian bounds for learning discrete-time stochastic linear time-invariant (LTI) state-space systems with sub-Gaussian noise and partial observations (the full states is not observed) from a single time-series. LTI systems are widely used in control theory to model physical systems. Learning such models is known as *system identification* in control theory. The core problem of the former is the same as that of machine learning: learning

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

predictive models from data. The main difference is that system system identification prioritizes model classes useful for control design. For this reason, it often focuses on parameter estimation error over generalization error, as the former is more relevant for control. Learning theory of LTI systems is also relevant for machine learning, since LTI systems are special cases of such popular architectures as recurrent neural networks (RNNs) or selective state-space models (SSMs). A detailed discussion on machine learning and control, the relationship between the learning problem of this paper and classical learning theory, and possible extensions of the current work to more general settings (e.g., RNNs) is presented in Appendix B.1.

Contribution: We derive PAC-Bayesian error bounds for both the generalization gap and the parameter estimation error for learning LTI models, rendering these bounds are relevant for both system identification and machine learning. Following standard practice in system identification, we view LTI systems as predictors that generate predictions for the current output based on past inputs and outputs. Our main contribution is a PAC-Bayesian bound showing that, with high probability (at least $1-\delta$), for any posterior distribution over predictors, the average generalization gap is $O(\frac{(\ln N)^r}{\sqrt{N}})$, where N is the sample size, $r = 0$ for bounded noise, $r = 1$ for Lipschitz loss, and $r = 3/2$ for quadratic loss. The bound depends on the Kullback–Leibler (KL) divergence between prior and posterior distributions and on $\ln(1/\delta)$. We subsequently derive generalization gap and parameter estimation error bounds for concrete learning algorithms (e.g., maximum likelihood, regularized prediction-error methods (PEM), and Bayesian methods).

Related work: PAC and PAC-Bayesian bounds for autoregressive models with bounded signals or Lipschitz loss have been studied in [5, 44, 3, 1, 27, 29]. In contrast, our results address state-space models with partially observed states and with inputs, which are more general than autoregressive models, and we allow for unbounded (sub-Gaussian) noise, and handle quadratic loss. We also build on techniques from [3, 1].

In contrast to the literature on finite-sample parameter estimation bounds for specific least-squares based learning algorithms for LTI systems, e.g., [38, 37, 32, 24, 14, 19, 42, 20, 35], this papers presents bounds on both the parameter estimation and the generalization gap for a wide-range of algorithms, see Remark F.12, Appendix F.5 for a detailed comparison.

PAC bounds for recurrent neural networks, of which LTI state-space representations are a subclass, have been developed using VC dimension [23, 41], Rademacher complexity [23, 41, 6], and PAC-Bayesian approaches [7]. However, existing results assume noiseless models, fixed prediction horizons, i.i.d. training data samples, and bounded signals. In contrast, this paper considers noisy models, infinite-horizon prediction errors, a single observed time series, and unbounded signals with sub-Gaussian noise.

In [12], bounds based on Rényi divergence were proposed, but the resulting probabilistic guarantees scale as $O(1/\delta)$ with the confidence level, rather than the standard $O(\ln(1/\delta))$, and it applies only to quadratic losses as opposed to general Lipschitz losses, and it provides parameter estimation bounds only in terms of the empirical error. In contrast, the bound of this paper depends on the confidence level $O(\ln(1/\delta))$ and it provides an absolute bound on the parameter estimation error. The works [11, 8] provide PAC-Bayesian generalization gap bounds for RNNs and LTI systems with bounded noise, respectively. In contrast, the present paper addresses unbounded (sub-Gaussian) noise and also derives parameter estimation error bounds. Even for bounded signals, our results refine the constants used in [11, 8] (see Remark F.11, Appendix F.5). PAC-Bayesian bounds for autonomous LTI systems without inputs were considered in [9] but those bounds did not converge to zero as $N \to \infty$.

Paper Outline: We start by defining the problem formulation in Section 2, followed by the main results in Section 3 and discussion of the bounds in Section 4. The technical proofs and numerical examples are presented in the appendices.

2 Problem formulation

2.1 Notation and terminology

We occasionally use $\triangleq$ to denote "defined by". Let $\mathcal{F}$ denote a σ-algebra on the set Ω and $\mathbf{P}$ be a probability measure on $\mathcal{F}$. Unless otherwise stated, all probabilistic considerations will be with respect to the probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and we let $\mathbf{E}[\mathbf{z}]$ denote expectation of the stochastic

variable $\mathbf{z}$. We use bold face letters to indicate stochastic signal vectors and processes relative to $(\Omega, \mathcal{F}, \mathbf{P})$. Each euclidean space is associated with the topology generated by the 2-norm $\|\cdot\|_2$, and the Borel σ-algebra generated by the open sets. The induced matrix 2-norm is also denoted $\|\cdot\|_2$ and $\|\cdot\|_\infty$ denote the infinity norm.

Let B_Θ be the σ-algebra of Lebesgue-measurable subsets of $\Theta \subset \mathbb{R}^{n_\theta}$, and m the Lebesgue measure. For a probability density ρ on (Θ, B_Θ, m) and a measurable, integrable function $g : \Theta \to \mathbb{R}$, define $E_{\theta\sim\rho}[g(\theta)] \triangleq \int_\Theta \rho(\theta)g(\theta)dm(\theta)$ as the expectation of g under ρ. For a density π on Θ, let $\mathcal{M}_\pi$ denote the set of all densities on Θ whose probability measure is absolutely continuous w.r.t. the probability measure of π.

Let $\mathrm{vol}(A) = \int_A dm(\theta)$ denote the Lebesgue measure of $A \subseteq \mathbb{R}^{n_\theta}$. The uniform distribution on a measurable set $A \in B_\Theta$ is $\rho(\theta) = \frac{\chi_A(\theta)}{\mathrm{vol}(A)}$, where χ_A is the indicator function. For $A = \Theta$, this is the uniform distribution on Θ.

The phrase *probability over data* refers to probability under $\mathbf{P}$ on $(\Omega, \mathcal{F})$. Specifically, for measurable functions $\{X_i, Y_i\}_{i\in I}$ and a binary relation $\diamond$, we say $\forall i \in I : X_i \diamond Y_i$ holds with probability at least $1-\delta$ over data if the set $F = \{\omega \in \Omega : \forall i \in I,\ X_i(\omega) \diamond Y_i(\omega)\}$ satisfies $\mathbf{P}(F) \geq 1-\delta$.

A complete list of symbols can be found in Appendix A.

2.2 Problem formulation

The problem formulation follows [11, 8, 12].

Let us fix stationary stochastic processes $\mathbf{y}(t) \in \mathbb{R}^{n_\mathbf{y}}$, and $\mathbf{u}(t) \in \mathbb{R}^{n_\mathbf{u}}$, that share a time axis $t \in \mathbb{Z}$, that is, for any $t \in \mathbb{Z}$, $\mathbf{y}(t) : \Omega \to \mathbb{R}^{n_\mathbf{y}}; \omega \mapsto \mathbf{y}(t)(\omega)$, and $\mathbf{u}(t) : \Omega \to \mathbb{R}^{n_\mathbf{u}}; \omega \mapsto \mathbf{u}(t)(\omega)$ are random vectors on $(\Omega, \mathcal{F}, \mathbf{P})$: $\mathbf{y}$ represents outputs (labels), and $\mathbf{u}$ represents inputs (features). The goal is to learn models from the finite sample $\mathcal{D} = \{\mathbf{y}(t)(\omega), \mathbf{u}(t)(\omega)\}_{t=0}^{N-1}$, $\omega \in \Omega$, such that the learned model predicts $\mathbf{y}$ from past measurements of $(\mathbf{y}, \mathbf{u})$ sufficiently well.

Predictor class: To simplify notation, let us define a stochastic process $\mathbf{w}(t) \in \mathbb{R}^{n_\mathbf{w}}$ by

$\mathbf{w}(t) = \left[\mathbf{y}^T(t), \quad \mathbf{u}^T(t)\right]^T$ when $n_\mathbf{w} = n_\mathbf{y} + n_\mathbf{u}$ and both past $\mathbf{y}$ and $\mathbf{u}$ are used for prediction,
$\mathbf{w}(t) = \mathbf{u}(t)$ when $n_\mathbf{w} = n_\mathbf{u}$, when only past $\mathbf{u}$ is used for prediction.

Note that with a trivial extension of the results of this paper one can define $\mathbf{w}(t)$ to consist of only some of the components of $\mathbf{y}(t)$, extending the results to not only prediction but also filtering.

Assumption 2.1 (Predictors)**.** *The class of predictors is the class of LTI systems parametrized by $\theta \in \Theta \subseteq \mathbb{R}^{n_\theta}$ which generate predictions $\hat{\mathbf{y}}_\theta(t|t_0)$ of $\mathbf{y}(t)$ based on $\{\mathbf{w}(s)\}_{s=t_0}^{t}$*

$$\hat{\mathbf{x}}(t+1) = \hat{A}_\theta\hat{\mathbf{x}}(t) + \hat{B}_\theta\mathbf{w}(t),\ \hat{\mathbf{x}}(t_0) = 0, \quad \hat{\mathbf{y}}_\theta(t|t_0) = \hat{C}_\theta\hat{\mathbf{x}}(t) + \hat{D}_\theta\mathbf{w}(t), \quad t \geq t_0, \tag{1}$$

where $\hat{A}_\theta, \hat{B}_\theta, \hat{C}_\theta, \hat{D}_\theta$ are $\hat{n} \times \hat{n}$, $\hat{n} \times n_\mathbf{w}$, $n_\mathbf{y} \times \hat{n}$ and $n_\mathbf{y} \times n_\mathbf{w}$ matrices respectively, they are are continuous in θ, and $\hat{A}_\theta$ is Schur (all its eigenvalues are inside the unit disk). Moreover, if $\mathbf{w}(t) = [\mathbf{y}^T(t)\ \mathbf{u}^T(t)]^T$, then $\hat{D}_\theta = [0\ \hat{D}_{\mathbf{u},\theta}]$ for some $n_\mathbf{y} \times n_\mathbf{u}$ matrix $\hat{D}_{\mathbf{u},\theta}$. [1] *Moreover, there exist constants $M, C > 0$, $\gamma \in (0,1)$ such that for all $\theta \in \Theta$, $\|\hat{A}_\theta^k\|_2 \leq M\gamma^k$, $\max\{\|\hat{B}_\theta\|_2, \|\hat{C}_\theta\|_2, \|\hat{D}_\theta\|_2\} \leq C$. In addition, we assume that the Markov parameteres of* (1) *are not all zero, i.e., either $\hat{D}_{\mathbf{u},\theta} \neq 0$ or $\hat{C}_\theta\hat{A}_\theta^k\hat{B}_\theta \neq 0$ for some $k \geq 0$.*

Intuitively, our assumption expresses that the predictors are stable LTIs and we have a uniform upper bound on the matrices and on the spectral radius of $\hat{A}_\theta$. These assumptions are standard in system identification, see Remark C.3, Appendix C. Note that the predictors are started at the zero initial state at inital time t_0. The latter represents the time instant at which the predictor was applied to generate predictions, or when learning started. Most of the time, we will set $t_0 = 0$ to simplify notation.

Data generator: Intuitively, we would like the data $\mathbf{y}$ to be generated by an LTI system

$$\hat{\mathbf{s}}_g(t+1) = A_0\hat{\mathbf{s}}_g(t) + B_0\mathbf{u}(t) + K_0\mathbf{e}^s(t), \quad \mathbf{y}(t) = C_0\hat{\mathbf{s}}_g(t) + D_0\mathbf{u}(t) + \mathbf{e}^s(t), \quad t \in \mathbb{Z}, \tag{2}$$

[1] The latter assumption is necessary, as it guarantees that $\hat{D}_\theta\mathbf{w}(t)$ depends only on $\mathbf{u}(t)$ and hence it does not depend on $\mathbf{y}(t)$, since otherwise we would be using $\mathbf{y}(t)$ to predict $\mathbf{y}(t)$, which is not meaningful.

where A_0, B_0, K_0, C_0, D_0 are matrices of suitable dimension, A_0 and $A_0 - K_0C_0$ are both Schur, and $\mathbf{e}^s(t)$ is a zero-mean i.i.d. sub-Gaussian process. The Schur conditions ensure stationarity of $\mathbf{x}$ and that $\mathbf{e}^s(t)$ is the innovation process (see [26, Chapter 4, p. 87]), i.e., the difference between $\mathbf{y}(t)$ and its best linear predictor based on past inputs and outputs. Thus, (2) is in steady-state Kalman filter form. Unlike predictors, the data generator (2) is assumed to operate in steady state, i.e., it has run for infinite time prior to prediction. This standard assumption in system identification (e.g., [26]) reflects that the physical system has been subject to noise and inputs long before learning begins. This subtle difference has significant theoretical implications.

Assuming the data is generated by a system of the same form as the predictors is a realizability assumption, as (2) yields the optimal predictor.

$$\hat{\mathbf{s}}_g(t+1 \mid t_0) = \hat{A}_0 \hat{\mathbf{s}}_g(t) + \hat{B}_0 \mathbf{w}(t),\ \hat{\mathbf{s}}_g(t_0|t_0) = 0, \quad \hat{\mathbf{y}}(t \mid t_0) = \hat{C}_0 \hat{\mathbf{s}}_g(t \mid t_0) + \hat{D}_0 \mathbf{w}(t), \quad (3)$$

where $t \geq t_0$, $\hat{A}_0 = A_0, \hat{B}_0 = B_0, \hat{C}_0 = C_0, \hat{D}_0 = D_0$ if $n_w = n_u$ and $\hat{A}_0 = A_0 - K_0C_0$, $\hat{B}_0 = [K_0 \quad B_0 - K_0D_0]$, $\hat{C}_0 = C_0$, $\hat{D}_0 = [0 \quad D_0]$ if $n_w = n_u + n_y$. When $\mathbf{w}(s) = [\mathbf{y}^T(s),\ \mathbf{u}^T(s)]^T$, the predictor (3) corresponds to stationary Kalman-filter and it is an optimal predictor of $\mathbf{y}(t)$ in the mean square error sense. Formally, instead of assuming the existence of (2), for the data generator we use the following assumption, which is essentially equivalent to the existence of (2).

Assumption 2.2. *Let $\mathbf{y}(t)$ and $\mathbf{u}(t)$ be generated by an autonomous stochastic LTI system*

$$\mathbf{x}(t+1) = A_g \mathbf{x}(t) + K_g \mathbf{e}_g(t), \quad \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{u}(t) \end{bmatrix} = C_g \mathbf{x}(t) + \mathbf{e}_g(t) \quad (4)$$

with $A_g \in \mathbb{R}^{n \times n}, K_g \in \mathbb{R}^{n \times n_\mathbf{e}}, C_g \in \mathbb{R}^{n_\mathbf{e} \times n}$, for some $n > 0$, and $n_\mathbf{e} = n_\mathbf{y} + n_\mathbf{u} \geq 2$, A_g and $A_g - K_gC_g$ are Schur. Moreover, $\mathbf{x}(t)$ has finite variance, $\mathbf{e}_g(t)$ is independent of $\mathbf{x}(t)$ for all $t \in \mathbb{Z}$, and $\mathbf{e}_g$ is square integrable, i.i.d process, $\|\mathbf{e}_g(t)\|_2$ is sub-Gaussian and the sub-Gaussian norm of $\|\mathbf{e}_g(t)\|_2$ [43, Definition 2.6.4] is denoted by $\|\mathbf{e}_g\|_{\psi_2}$.

If $\mathbf{e}_g(t)$ is sub-Gaussian, then $\|\mathbf{e}_g(t)\|_2$ is also sub-Gaussian, see [21, Lemma 1]. Under Assumption 2.2, $\mathbf{y}$ and $\mathbf{u}$ are jointly stationary and square integrable processes [4]. Under mild conditions, Assumption 2.2 is equivalent to the existence of an LTI system (2), see Remark C.2 in Appendix C. Assumption 2.2 is more convenient, and it is often used in system identification [22, Chapter 9].

Measure of prediction accuracy: Following standard practice in system identification [26], we measure the predictive accuracy using a *loss function* $\ell : \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to [0, \infty)$ which is continuous and satisfies $\ell(y, y) = 0$. Typical choices are $\ell(y, \hat{y}) = \|y - \hat{y}\|_2$, to which we refer to as ℓ_2 loss, and $\ell(y, \hat{y}) = \|y - \hat{y}\|_2^2$ to which we refer to as the *quadratic loss*.

Assumption 2.3 (Qudaratic or Lipschitz loss functions)**.** *The loss function ℓ is either quadratic or it is L_ℓ-Lipschitz, i.e., for all $y_1, y_2, \hat{y}_1, \hat{y}_2 \in \mathbb{R}^{n_y}$, $|\ell(y_1, \hat{y}_1) - \ell(y_2, \hat{y}_2)| \leq L_\ell(\|y_1 - y_2\|_2 + \|\hat{y}_1 - \hat{y}_2\|_2)$.*

That is, we consider the loss to be Lipschitz or quadratic, which, for the bounded case is also Lipschitz. Lipschitzness of the loss is standard assumptions for PAC bounds. Intuitively, this allows us to reduce the case of a general loss to the case when the loss is the norm of the prediction error. The quadratic case is considered due to its popularity in system identification. For more details on the role of losses in the proof of the results see Appendix B.2

We define the *true error* as the limit

$$\mathcal{L}(\theta) \triangleq \lim_{t_0 \to -\infty} \mathbf{E}[\ell(\mathbf{y}(t), \hat{\mathbf{y}}_\theta(t \mid t_0))]. \quad (5)$$

The true error captures the long-term, steady-state prediction error obtained during the deployment of the predictor on an increasing number of past inputs.

Lemma 2.4. *Let Assumptions 2.2, 2.1, and 2.3 hold. Then the limits $\hat{\mathbf{y}}_\theta(t) = \lim_{t_0 \to -\infty} \hat{\mathbf{y}}_\theta(t \mid t_0)$ and $\lim_{t \to \infty}(\hat{\mathbf{y}}_\theta(t) - \hat{\mathbf{y}}_\theta(t \mid 0)) = 0$ exist in the mean-square sense for all t and θ, the process $\hat{\mathbf{y}}_\theta(t)$ is stationary, and $\mathcal{L}(\theta) = \mathbf{E}[\ell(\mathbf{y}(t), \hat{\mathbf{y}}_\theta(t))] = \lim_{t \to \infty} \mathbf{E}[\ell(\mathbf{y}(t), \hat{\mathbf{y}}_\theta(t \mid 0))]$.*

The proof of Lemma 2.4 is presented in Appendix C. Intuitively, $\hat{\mathbf{y}}_\theta(t)$ is the output predicted by the LTI system corresponding to θ, if the latter is started from zero at $-\infty$, and as t grows, $\hat{\mathbf{y}}_\theta(t)$ and $\hat{\mathbf{y}}_\theta(t \mid 0)$ converge to each other.

The Learning Problem is to find a parameter $\theta_\star$ from sampled data $\mathcal{D}$ of the sequence of random variables $\{\mathbf{y}(t), \mathbf{u}(t)\}_{t=0}^{N-1}$ which minimizes the true error, i.e., $\theta_\star \in \arg\min \mathcal{L}(\theta)$. Note that we do not assume the knowledge of the matrices A_g, K_g, C_g and noise process $\mathbf{e}_g$ when learning.

Since the true error is unknown, a standard practice is to use the *empirical error*

$$\hat{\mathcal{L}}_N(\theta) = \frac{1}{N}\sum_{t=0}^{N-1} \ell(\mathbf{y}(t), \hat{\mathbf{y}}_\theta(t|0)) \tag{6}$$

as a proxy for the true error, and try to find a model $\hat{\theta}$ for which the empirical error $\hat{\mathcal{L}}_N(\hat{\theta})$ is small. While doing so, we would like to keep the *generalization gap* $\mathcal{L}(\theta) - \hat{\mathcal{L}}_N(\theta)$, i.e., the difference between a model's performance on the training data (empirical error) and its expected performance (true error) on unseen test data, small. The PAC-Bayesian approach provides a framework to derive such guarantees.

2.3 Constants used in PAC-bounds

In order to present the main result, we will need several constants, many of which have an interesting system-theoretic interpretation.

Definition 2.5 (Convolution and ℓ_1 norm)**.** *Let $\alpha = \{\alpha(k)\}_{k=0}^{\infty}$ be a sequence of matrices. Define the ℓ_1 norm $\|\alpha\|_{\ell_1}$ and the* weakly dependent mixing coefficient (WDMC) *$\theta_\infty(\alpha)$ of α as*

$$\|\alpha\|_{\ell_1} = \sum_{k=0}^{\infty} \|\alpha(k)\|_2, \quad \theta_\infty(\alpha) = \sum_{k=0}^{\infty} k\|\alpha(k)\|_2.$$

If $\|\alpha\|_{\ell_1} < \infty$ and $\mathbf{r}$ is a stationary square integrable stochastic process, then define the convolution *$(\alpha \star \mathbf{r})$ of α with $\mathbf{r}$ as the stochastic process $(\alpha \star \mathbf{r})(t) = \sum_{k=0}^{\infty} \alpha(k)\mathbf{r}(t-k)$. By [4, Theorem 1.1], $(\alpha \star \mathbf{r})$ is well-defined and stationary.*

Note that the matrices $\alpha = \{\alpha(k)\}_{k=0}^{\infty}$ are deterministic quantities, but their convolution $(\alpha \star \mathbf{r})$ is a stochastic process. Intuitively, the ℓ_1 norm of a sequence is the extension of the standard norm on the space of absolutely summable scalar sequences to the case of matrix valued ones. The convolution $(\alpha \star \mathbf{r})$ represents the response of a linear filter with impulse response α to the noise $\mathbf{r}$, and $\|\alpha\|_{\ell_1}$ represents the ℓ_1 norm of the filter. If $\mathbf{r}$ is bounded, i.e., $\|\mathbf{r}(t)\|_2 < C$, then the convolution $(\alpha \star \mathbf{r})(t)$ is bounded, i.e., $\|(\alpha \star \mathbf{r})\|_2 < \|\alpha\|_{\ell_1} C$, and it is a weakly dependent process, in the sense of [3]; the corresponding mixing coefficients $\{\theta_{\infty,n}(1)\}_{n=1}^{\infty}$ which is defined in [3], and whose definition is recalled in Appendix B.4, are bounded by the $2\theta_\infty(\alpha)C$, [3, Proposition 4.12]. Intuitively, the smaller $\theta_\infty(\alpha)$ is, the closer $(\alpha \star \mathbf{r})(t)$ is to being i.i.d., if $\mathbf{r}$ is i.i.d. If α is the sequence of Markov parameters of a stable LTI system, then both $\|\alpha\|_{\ell_1}$ and $\theta_\infty(\alpha)$ are finite and can be upper bounded in terms of bounds on the norms of the system matrices and the spectral radius of the system matrix A, see Remark D.1.

Definition 2.6 (Auxiliary definitions)**.** *Let us consider the Markov-parameters α_g and α_θ of the data generator* (4) *and predictor* (1) *corresponding to $\theta \in \Theta$ respectively:*

$$\alpha_g(k) = \begin{cases} C_g A_g^{k-1} K_g & k > 0 \\ I & k = 0 \end{cases}, \quad \alpha_\theta(k) = \begin{cases} \widehat{C}_\theta \widehat{A}_\theta^{k-1} \widehat{B}_\theta & k > 0 \\ \widehat{D}_\theta & k = 0 \end{cases} \tag{7}$$

Let $\bar{\alpha}(k) = \sup_{\theta\in\Theta} \|\alpha_\theta(k)\|_2$. With notation as above

$$G_e(\theta) \triangleq (1 + \|\alpha_\theta\|_{\ell_1})\|\alpha_g\|_{\ell_1}, \quad G_{e,1}(\theta) \triangleq \theta_\infty(\alpha_\theta)\|\alpha_g\|_{\ell_1} + \theta_\infty(\alpha_g)(\|\alpha_\theta\|_{\ell_1} + 1)$$
$$G_e \triangleq (1 + \|\bar{\alpha}\|_{\ell_1})\|\alpha_g\|_{\ell_1}, \quad G_{e,1} \triangleq \theta_\infty(\bar{\alpha})\|\alpha_g\|_{\ell_1} + \theta_\infty(\alpha_g)(\|\bar{\alpha}\|_{\ell_1} + 1)$$

For the definition of the upcoming bounds, we are going to use the quantities $\|\alpha\|_{\ell_1}$, $\theta_\infty(\alpha)$ for $\alpha \in \{\alpha_g, \bar{\alpha}\} \cup \{\alpha_\theta \mid \theta \in \Theta\}$ and the constants $G_e(\theta), G_{e,1}(\theta), G_e, G_{e,1}$ defined above. Note that all these quantities are well-defined, see Lemma D.2. In addition, they are related to the true outputs $\mathbf{y}(t)$ and predicted outputs $\hat{\mathbf{y}}_\theta(t)$, $\hat{\mathbf{y}}_\theta(t \mid t_0)$ as follows: $\begin{bmatrix}\mathbf{y}^T(t) & \mathbf{u}^T(t)\end{bmatrix}^T = (\alpha_g \star \mathbf{e}_g)(t)$, $\hat{\mathbf{y}}_\theta(t) = (\alpha_\theta \star \mathbf{w})(t)$, $\hat{\mathbf{y}}_\theta(t \mid t_0) = (\alpha_{\theta,t-t_0} \star \mathbf{w})(t)$. The formal relationship between the constants is provided in Lemma D.3, Appendix D. The interpretation of the various terms appearing in Definition 2.6 is as follows

$\|\alpha_g\|_{\ell_1}$**,** $\theta_\infty(\alpha_g)$ are the ℓ_1 norm and the mixing coefficients of Markov-parameters of the data generator: both quantities decrease with stability of the data generator, see Remark D.1.

$\|\alpha_\theta\|_{\ell_1}$ is the ℓ_1 norm of the predictor (1). This term decreases with the spectral radius of the predictor, see Remark D.1.

$\theta_\infty(\alpha_\theta)$ describes the mixing properties of the output generated by the predictor, and it also bounds the sum of differences between the infinite and finite past prediction $\sum_{t=0}^{N-1} \|\hat{\mathbf{y}}_\theta(t) - \hat{\mathbf{y}}_\theta(t \mid 0)\|$. Intuitively, the smaller $\theta_\infty(\alpha_\theta)$ is, the closer $\hat{\mathbf{y}}_\theta(t \mid 0)$ is to $\hat{\mathbf{y}}_\theta(t)$. This quantity decreases with the spectral radius of the predictor by Remark D.1.

$G_e(\theta)$ characterizes the gain of the system mapping $\mathbf{e}_g$ to the prediction error $\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)$ in the standard ℓ_∞ norm.

$G_{e,1}(\theta)$ characterizes mixing properties (short memory condition [1]) of the prediction error $\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)$. Intuitively, the smaller $G_{e,1}(\theta)$ is, the more the prediction error behaves like an i.i.d process. This constant is related to stability of the generator and the predictor, the smaller the spectral radius of those systems are , the smaller this constant is.

$\|\bar{\alpha}\|_\ell, \theta_\infty(\bar{\alpha}), G_e, G_{e,1}$ are uniform upper bounds on $\|\alpha_\theta\|_{\ell_1}, \theta_\infty(\alpha_\theta), G_e(\theta), G_{e,1}(\theta)$ respectively.

3 Main result

We start by presenting the main result of the paper, which is the following theorem.

Theorem 3.1 (Main result)**.** *Let Assumptions 2.2, 2.3 and 2.1 hold. For any a prior density π over Θ, confidence level $\delta \in (0, 0.5)$, integer $\epsilon \in \{-1, 1\}$, and $N \geq K_\delta$, the inequality*

$$\begin{aligned} &\forall \rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}\mathcal{L}(\theta) \leq \epsilon E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \\ &\frac{1}{\lambda}\Big(\mathrm{KL}(\rho\|\pi) + \ln\frac{C_\delta}{\delta} + \frac{1}{2}\sum_{i=1}^{2}\ln E_{\theta\sim\pi}\exp\left(\mathcal{C}_i(2\lambda,\theta,N,\delta,\epsilon)\right)\Big) \end{aligned} \tag{8}$$

holds with probability at least $1 - 2\delta$ over data, with $\mathrm{KL}(\rho\|\pi) = E_{\theta\sim\rho}\ln\frac{\rho(\theta)}{\pi(\theta)}$ the Kullback–Leibler divergence and the constants K_δ, $\mathcal{C}_i(\lambda,\theta,N,\delta,\epsilon)$, $i = 1, 2$ and C_δ are defined as follows:

Bounded noise: *If $\|\mathbf{e}_g(t)\|_\infty \leq c_{\mathbf{e}}$, then $C_\delta = 1, K_\delta = 2$, $\lambda > 0$, and*

$$\mathcal{C}_i(\lambda,\theta,N,\delta,\epsilon) \triangleq \mathcal{C}_{b,i}(\lambda,\theta,L(\theta),c_{\mathbf{e}},N), \quad \mathcal{C}_{b,i}(\lambda,\theta,L,c,N) \triangleq \frac{\mathcal{C}_{m,b,i}(\lambda,\alpha_\theta,G_e(\theta)+2G_{e,1}(\theta),L,c)}{N}$$

$$L(\theta) \triangleq \begin{cases} 2G_e(\theta)c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}} & \ell \text{ is quadratic loss} \\ L_\ell & \ell \text{ is } L_\ell\text{-Lipschitz} \end{cases}$$

$$\mathcal{C}_{m,b,i}(\lambda,\alpha,X,L,c) \triangleq \begin{cases} \lambda L c\sqrt{n_{\mathbf{e}}}\theta_\infty(\alpha)\|\alpha_g\|_{\ell_1} & i = 1 \\ \lambda^2\frac{L^2}{2}X(\sqrt{n_{\mathbf{e}}}c)^2 & i = 2 \end{cases}$$

Lipschitz loss: *If ℓ is L_ℓ-Lipschitz, then $C_\delta = 1, K_\delta = 2$, $0 < \lambda < \frac{N}{32L_\ell}$, and*

$$\begin{aligned} &\mathcal{C}_i(\lambda,\theta,N,\delta,\epsilon) \triangleq \mathcal{C}_{ub,i}(\lambda,\theta,L_\ell,N), \\ &\mathcal{C}_{ub,i}(\lambda,\theta,L,N) \triangleq \frac{1}{2}\mathcal{C}_{b,i}\left(2\lambda,\theta,L,\frac{2\ln(N)}{\|\alpha_g\|_{\ell_1}K},N\right) + \bar{\mathcal{C}}_{ub}(\lambda,L,N) \\ &\bar{\mathcal{C}}_{ub}(\lambda,L,N) \triangleq \frac{\lambda^2 32L^2C_{u,1}}{N} + \frac{4\lambda LC_{u,2}\ln(N)}{N} = \frac{\lambda^2 32L^2C_{u,1}}{N} + \frac{4\lambda LC_{u,2}\ln(N)}{N} \end{aligned}$$

where the constants, K,$C_{u,j}$, $j = 1, 2$ are defined as follows:

$$\begin{aligned} C_{u,1} &\triangleq \bar{\Psi}_{\mathbf{e}_g}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1}) + \bar{\Psi}_{\mathbf{e}_g}(\|\alpha_g\|_{\ell_1}), \quad \bar{\Psi}_{\mathbf{e}_g}(s) = e^{14^2s^2\|\mathbf{e}_g\|_{\psi_2}^2 + 3s\|\mathbf{e}_g\|_{\psi_2}} \\ C_{u,2} &\triangleq \bar{\Psi}_{\mathbf{e}}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1})\frac{\|\bar{\alpha}\|_{\ell_1}}{K} + \bar{\Psi}_{\mathbf{e}_g}(\|\alpha_g\|_{\ell_1})\frac{1}{K}, \qquad K = \min\{1, \|\bar{\alpha}\|_{\ell_1}\} \end{aligned} \tag{9}$$

Quadratic loss: *if ℓ is the quadratic loss, then $C_\delta = \frac{3}{2}$, $K_\delta = 3$, $0 < \lambda < \frac{N}{32L_{qd}}$,*

$$\mathcal{C}_i(\lambda,\theta,N,\delta,\epsilon) \triangleq \frac{\lambda L_{qd}^2\max\{0,\epsilon\}}{2N} + \mathcal{C}_{ub,i}\left(\lambda,\theta,L_{qd},N\right), \quad L_{qd} = 12G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{\ln\left(\frac{3N^2}{2\delta}\right)}$$

For $\epsilon = 1$, (8) bounds the average generalization gap $E_{\theta\sim\rho}(\mathcal{L}(\theta) - \hat{\mathcal{L}}_N(\theta))$; for $\epsilon = -1$, it bounds $E_{\theta\sim\rho}(\hat{\mathcal{L}}_N(\theta) - \mathcal{L}(\theta))$. The latter is useful for deriving oracle inequalities. By applying the union bound to both cases, we obtain a bound on $\left|E_{\theta\sim\rho}(\mathcal{L}(\theta) - \hat{\mathcal{L}}_N(\theta))\right|$. Note that except for the quadratic loss case with potentially unbounded data, $\mathcal{C}_i(\lambda, \theta, N, \delta, \epsilon)$ is independent of δ and ϵ. For a detailed discussion on the statement of Theorem 3.1 (bounded/unbounded data, Lipschitz/quadratic loss) see Appendix B.6.

Intuition behind the proof: The proof of Theorem 3.1 is in Appendix E. The key idea is to decompose the generalization gap using an auxiliary empirical error $V_N(\theta) = \frac{1}{N}\sum_{t=0}^{N-1} \ell(\mathbf{y}(t), \hat{\mathbf{y}}_\theta(t))$ where the predictor has run for infinite time (started at $t_0 = -\infty$). Since $\mathbf{E}[V_N(\theta)] = \mathcal{L}(\theta)$, we have $\epsilon(\mathcal{L}(\theta) - \hat{\mathcal{L}}_N(\theta)) = \epsilon(\mathcal{L}(\theta) - V_N(\theta)) + \epsilon(V_N(\theta) - \hat{\mathcal{L}}_N(\theta))$. The constants $\exp(\mathcal{C}_1)$ and $\exp(\mathcal{C}_2)$ bound the moment generating functions of these two differences. The PAC-Bayesian bound then follows by standard concentration arguments [2, Chapter 2].

Bounded case: $\mathcal{C}_1$ is obtained by direct calculation, and $\mathcal{C}_2$ via Rios's concentration inequality for weakly-dependent processes [1, Prop. 4.2], improving upon constants in [11, 8].

Unbounded case: We employ a truncation approach [3]. Define truncated data $\bar{\mathbf{y}}(t), \bar{\mathbf{u}}(t)$ generated from truncated noise $\bar{\mathbf{e}}_g(t) = \mathbf{e}_g(t)\chi(\|\mathbf{e}_g(t)\| \leq C)$ with $C = \frac{2\ln(N)}{K\|\alpha_g\|_{\ell_1}}$, where $K = \min\{1, \|\bar{\alpha}\|_{\ell_1}\}$. We bound the moment generating functions of the differences $\sup_{\theta\in\Theta} |\hat{\bar{\mathcal{L}}}_N(\theta) - \hat{\mathcal{L}}_N(\theta)|$ and $\sup_{\theta\in\Theta} |V_N(\theta) - \bar{V}_N(\theta)|$ by $\exp(\bar{\mathcal{C}}_{ub}(\lambda, L_\ell, N))$. Combining with the bounded-case bounds yields PAC-Bayesian guarantees for unbounded data.

Quadratic case: We show that with probability at least $1 - \frac{2\delta}{3}$ over data, the prediction errors satisfy $\|\mathbf{y}(t) - \mathbf{y}_\theta(t)\|_2 \leq C$ and $\|\mathbf{y}(t) - \mathbf{y}_\theta(t \mid 0)\|_2 \leq C$ for $C = 6G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{\ln(\frac{3N^2}{2\delta})}$. The empirical quadratic loss then coincides with the empirical error for the truncated Lipschitz loss $\ell_C(y, y') = \min\{\|y - y'\|_2^2, C^2\}$. The result follows by union bound and the Lipschitz case, noting that the true error difference between the quadratic loss and the truncated loss is $O(C^2/N)$.

Interpretation and $O(\frac{\ln(N)^r}{N})$ rates: The interpretation of Theorem 3.1 is similar to other PAC-Bayesian bounds [2]: we view learning algorithms as choices of parameters from a data-dependent probability density on model parameters, referred to as *posterior* $\rho^{\mathcal{D}}$, for which the right-hand side of (8) is small for $\epsilon = 1$. One option is to choose $\rho^{\mathcal{D}}$ as the *Gibbs posterior* which minimizes of the right-hand side of (8) and which is given by [2, Definition 2.1]:

$$\rho_{\text{Gibbs}} \triangleq \operatorname{argmin}_{\rho\in\mathcal{M}_\pi} \left(E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\mathrm{KL}(\rho\|\pi) \right), \quad \rho_{\text{Gibbs}}(\theta) = \frac{\pi(\theta)\exp(-\lambda\hat{\mathcal{L}}_N(\theta))}{E_{\theta\sim\pi}\exp(-\lambda\hat{\mathcal{L}}_N(\theta))},$$

Once a posterior $\rho^{\mathcal{D}}$ is obtained, the learned model $\theta_\star$ can be chosen as follows:

Mean posterior (MEP): as the mean of the posterior $\rho^{\mathcal{D}}$, i.e. $\theta_\star = E_{\theta\sim\rho^{\mathcal{D}}}\theta$,

Single draw: draw $\theta_\star$ randomly from $\rho^{\mathcal{D}}$, and

Maximum posterior (MAP): choose $\theta_\star$ as $\theta_\star = \operatorname{argmax}\rho^{\mathcal{D}}(\theta)$. For $\rho^{\mathcal{D}} = \rho^{\text{Gibbs}}$, it minimizes the empirical error regularized using the prior, i.e., $\theta_\star = \arg\min_{\theta\in\Theta}\left(\hat{\mathcal{L}}_N(\theta) - \frac{1}{\lambda}\ln\pi(\theta)\right)$,

These choices represent well-known classes of system identification algorithms, such as regularized PEM, maximimum likelihood estimation for state-space models, and Bayesian methods, see Remarks F.8, F.5, and F.6 in Appendix F.5.

Theorem 3.1 bounds the average generalization gap $E_{\theta\sim\rho}\left(\mathcal{L}(\theta) - \hat{\mathcal{L}}_N(\theta)\right)$ by a sum of two terms: the KL divergence $\frac{1}{\lambda}\mathrm{KL}(\rho\|\pi)$ between the posterior and the prior, and a complexity term $\mathcal{C}_i(2\lambda, \theta, N, \delta, \epsilon)$ that depends on the data and the prior. For bounded noise, $\mathcal{C}_i = O(1/N)$; for Lipschitz loss, $\mathcal{C}_i = O(\frac{\ln N)}{N})$; and for quadratic loss, $\mathcal{C}_i = O(\frac{(\ln N)^3}{N})$. By choosing $\lambda = \frac{\sqrt{N}}{L(\ln N)^r}$ with appropriately selected L and r, and thus forcing $\frac{1}{\lambda}\mathrm{KL}(\rho\|\pi)$ to tend to zero as $N \to \infty$, we obtain an $O\left(\frac{(\ln N)^r}{\sqrt{N}}\right)$:

Corollary 3.2 ($O((\ln(N))^r/\sqrt{N})$ bound)**.** *With the notation and assumptions of Theorem 3.1,*

$$\forall \rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}\mathcal{L}(\theta) \leq \epsilon E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\mathrm{KL}(\rho||\pi) + \ln\frac{C_\delta}{\delta} + \mathcal{C}(\delta,\epsilon)\right) \tag{10}$$

holds with probability at least $1-2\delta$ *over data, for any prior* π*,* $\delta \in (0, 0.5)$*, and for any* $\epsilon \in \{-1, 1\}$*, and* $N \geq 4$*, where* $\mathcal{C}(\delta,\epsilon)$*,* L*,* r *are defined as follows.*
Bounded noise: *If* $\|e_g(t)\|_\infty \leq c_{\mathbf{e}}$*, then* $r \triangleq 0$*,* $L \triangleq 1$*, and*

$$\mathcal{C}(\delta,\epsilon) \triangleq \frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}(2,\bar{\alpha},G_e+2G_{e,1},L_b,c_{\mathbf{e}}), \quad L_b \triangleq \begin{cases} 2G_e c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}} & \ell \text{ is the quadratic loss} \\ L_\ell & \ell \text{ is } L_\ell\text{-Lipschitz} \end{cases}$$

Lipschitz loss: *If* ℓ *is* L_ℓ*-Lipschitz, then* $r \triangleq 1$*,* $L \triangleq 16L_\ell$*, and*

$$\mathcal{C}(\delta,\epsilon) \triangleq \bar{\mathcal{C}}_{ub} = \left(C_{u,1} + C_{u,2} + \frac{1}{4}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}\left(\frac{1}{2},\bar{\alpha},G_e+2G_{e,1},1,\frac{1}{\|\alpha_g\|_{\ell_1}K}\right)\right),$$

Quadaratic loss: *If* ℓ *is quadratic, then* $L \triangleq 192G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{2}\sqrt{1+\ln\left(\frac{3}{2\delta}\right)}$*,* $r \triangleq 3/2$*,*

$$\mathcal{C}(\delta,\epsilon) \triangleq \frac{L\max\{0,\epsilon\}}{128} + C_{u,1} + C_{u,2} + \frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}\left(\frac{1}{4},\bar{\alpha},G_e+2G_{e,1},1,\frac{1}{\|\alpha_g\|_{\ell_1}K}\right)$$

The proof of Corollary 3.2 follows by substituting $\lambda = \frac{\sqrt{N}}{L(\ln N)^r}$ into Theorem 3.1 and simplifying the resulting bounds (see Appendix E.4 for the proof). For additional explanation of the statement of Corollary 3.2 see Appendix B.6. Note that while Corollary 3.2 makes the $O\left(\frac{(\ln N)^r}{\sqrt{N}}\right)$ rates explicit, Theorem 3.1 may give a tighter bound even for $\lambda = \frac{\sqrt{N}}{L(\ln N)^r}$ [2].

The previous results bound the average generalization gap with respect to a posterior $\rho^{\mathcal{D}}$. For practical learning, however, we are interested in the generalization gap for a specific choice of $\theta_\star$ (e.g., MEP, MAP, or a single draw from $\rho^{\mathcal{D}}$). Below, we provide such bounds when $\rho^{\mathcal{D}}$ is the Gibbs posterior with $\lambda = \frac{\sqrt{N}}{L(\ln N)^r}$, where L, r are as in Corollary 3.2. This yields a generalization gap for $\theta_\star$ of order $O\left(\frac{(\ln N)^{r+1}}{\sqrt{N}}\right)$. We now introduce the necessary assumptions for these results.

Assumption 3.3. *The following conditions are used in the sequel.*

(Opt) *An optimal model* $\theta_{true} = \arg\min_{\theta\in\Theta}\mathcal{L}(\theta)$ *exists.*

(LipModel) *The map* $\theta \mapsto \alpha_\theta$ *is* L_α*-Lipschitz w.r.t. to the* $\|\cdot\|_{\ell_1}$ *norm, i.e.,*

$$\|\alpha_{\theta_1} - \alpha_{\theta_2}\|_{\ell_1} \leq L_\alpha\|\theta_1 - \theta_2\|_2, \quad \forall\theta_1,\theta_2 \in \Theta.$$

(Gauss) $\Theta = \mathbb{R}^{n_\theta}$ *and* π *is Gaussian* $\mathcal{N}(\theta_m, P)$*.*

(Uni) Θ *is a closed ball* $\{\theta \in \mathbb{R}^{n_\theta} : \|\theta-\theta_m\|_2 \leq R\}$ *of radius* R *and center* θ_m *for some* $\theta_m \in \mathbb{R}^{n_\theta}$*, and* π *is the uniform density on* Θ*.*

(SConv) *The true error* $\mathcal{L}(\theta)$ *is smooth and strongly convex in* θ *and its Hessian is bounded from below* $m_\Theta I \preceq \nabla^2_\theta\mathcal{L}(\theta)$ *for all* $\theta \in \Theta$*.*

(Real) *The predictor* $\Sigma(\theta_{true})$ *equals the predictor* (3) *corresponding to the data generator.*

(PE) *The spectral density* $\Phi_{\mathbf{w}}$ *of* $\mathbf{w}$ *satisfies* $\Phi_{\mathbf{w}}(iz) \geq m_{\mathbf{w}}I$ *for all* $z \in [-\pi,\pi]$*.*

Corollary 3.4. *Assume the following*
(1) *Assumptions 2.2, 2.1, 2.3 and* ***(Opt)*** *and* ***(LipModel)*** *of Assumption 3.3 hold, and*
(2) *either* ***(Gauss)*** *and* $\sigma > 0$ *is arbitrary or* ***(Uni)*** *of Assumption 3.3 holds and* $\sigma < 2R$*, and*
(3) $\theta_\star$ *be selected from* ρ_{Gibbs} *using MEP, MAP or a single draw, and* ρ_{Gibbs} *is computed using* $\lambda = \lambda_N = \frac{\sqrt{N}}{L(\ln(N))^r}$ *with* r, L *are as defined in Corollary 3.2 and* $N \geq 4$*.*

(4) *Furthermore, if $\theta_\star$ is chosen using MEP, then assume that* ***(SConv)*** *of Assumption 3.3 holds. Consider the constants $\mathcal{C}_{em}(\delta,\sigma)$, $\mathcal{C}_{or}(\delta,\sigma)$, $r_{Lip,1}, r_{Lip,2}$ as in Table 1, and consider the inequalities*

$$\textbf{\textit{Empirical bound:}} \quad \mathcal{L}(\theta_\star) - \hat{\mathcal{L}}_N(\theta_\star) \leq \frac{L(\ln(N))^{r+1}}{\sqrt{N}} C_{em}(\theta_\star,\delta) \tag{11}$$

$$\textbf{\textit{Oracle bound:}} \quad \mathcal{L}(\theta_\star) - \mathcal{L}(\theta_{true}) \leq \frac{L(\ln(N))^{r+1}}{\sqrt{N}} \mathcal{C}_{or}(\delta,\sigma) \tag{12}$$

$$\textbf{\textit{Parameter estimation error:}} \quad \|\theta_\star - \theta_{true}\|_2^2 \leq \frac{L(\ln(N))^{r+1}}{\sqrt{N} m_\Theta} \mathcal{C}_{or}(\delta,\sigma), \tag{13}$$

$$H_2 \textbf{\textit{ error bound:}} \quad \|H_{\theta_{true}} - H_{\theta_\star}\|_{H_2}^2 \leq \frac{L(\ln(N))^{r+1}}{m_{\mathbf{w}}\sqrt{N}} \mathcal{C}_{or}(\delta,\sigma) \tag{14}$$

Then the following holds: ***(A)*** (11) *holds with probability at least $1-(2+r_{Lip,1})\delta$,*
(B) (12) *holds with probability at least $1-2\delta$ for single draw and MEP, and with probability at least $1-(2+r_{Lip,1})\delta$ for MAP,*
(C) (13) *holds with probability at least $1-2\delta$ for single draw and MEP, and with probability at least $1-(2+r_{Lip,1})\delta$ for MAP, if* ***(SConv)*** *of Assumption 3.3 holds,*
(D) (14) *holds with probability at least $1-2\delta$ for single draw and MEP, and with probability at least $1-(2+r_{Lip,1})\delta$ for MAP, if* ***(Real)*** *and* ***(PE)*** *of Assumption 3.3 holds, ℓ is the quadratic loss, and H_θ the transfer function of the predictor corresponding to θ, and denote by $\|H\|_{H_2}$ the H_2 norm of the transfer function H.*
Here, the probability is taken over data for MEP and MAP, and over the data and over all samples $\theta_\star$ drawn from ρ_{Gibbs} for the single draw. [2]

Table 1: Constants for learning bounds under different algorithm choices (Corollary 3.4)

Constant	**MEP / Single Draw**	**MAP**
$L_\Theta(\delta)$	$L_\alpha L_\ell \|\alpha_g\|_{\ell_1}\sqrt{n_{\mathbf{e}}}c_{\mathbf{e}}$ bounded noise $\frac{1}{16}L_\alpha\sqrt{n_{\mathbf{e}}}\frac{1}{K}$ Lipschitz/quadratic	
$C(\theta)$	$\frac{1}{2}\big(\sigma^2\mathrm{trace}(P^{-1}) - n_\theta + (\theta_m-\theta)^T P^{-1}(\theta_m-\theta) + \ln\frac{\det(P)}{\sigma^{2n_\theta}} + n_\theta\big)$ **(Gauss)** $\ln\frac{vol(\Theta)}{(\sqrt{\pi\sigma^2/4}^{n_\theta}\Gamma(n_\theta/2+1)} + \frac{n_\theta}{2}$ **(Uni)**	
$r_{\mathrm{Lip},1}, r_{\mathrm{Lip},2}$	$r_{\mathrm{Lip},1} = \begin{cases} 0 & \mathbf{e}_g \text{ is bounded} \\ 1 & \text{otherwise} \end{cases}$	$r_{\mathrm{Lip},2} = \begin{cases} 2 & \ell \text{ is quadratic and} \\ & \text{and } \mathbf{e}_g \text{ is unbounded} \\ 0 & \text{otherwise} \end{cases}$
$\mathcal{C}_{L,emp}$	$L_\Theta(\delta)\sigma + 2r_{\mathrm{Lip},1}(C_{u,1}+C_{u,2})$	
$\mathcal{C}_{L,true}$	$\mathcal{C}_{L,emp} + 2r_{\mathrm{Lip},2}\frac{L}{512}$	
$\mathcal{C}_L(\delta,\epsilon,\theta)$	$\mathcal{C}_{L,emp} + \ln\left(\frac{C_\delta}{\delta}\right) + C(\theta) + \mathcal{C}(\delta,\epsilon)$	
$C_{em}(\theta_\star,\delta)$	$\mathcal{C}_L(\delta,1,\theta_\star) + 2r_{\mathrm{Lip},1}\ln\left(\frac{C_\delta}{\delta}\right)$	$\mathcal{C}_L(\delta,1,\theta_\star) + \mathcal{C}_{L,true} + 2r_{\mathrm{Lip},1}\ln\left(\frac{C_\delta}{\delta}\right)$
$\mathcal{C}_{or}(\delta,\sigma)$	$\mathcal{C}_L(\frac{\delta}{2},1,\theta_{true}) + \mathcal{C}_L(\frac{\delta}{2},-1,\theta_{true})$ $+\mathcal{C}_{L,true} - 2\mathcal{C}_{L,emp}$	$\mathcal{C}_L(\frac{\delta}{2},1,\theta_\star) + \mathcal{C}_L(\frac{\delta}{2},-1,\theta_{true})$ $+4r_{\mathrm{Lip},1}\ln\left(\frac{2C_\delta}{\delta}\right) + 2\mathcal{C}_{L,true}$ $+\ln\left(\frac{\pi(\theta_{true})}{\pi(\theta_\star)}\right)$

In summary, for any outcome $\theta_\star$ obtained via MAP, MEP, or a single draw from the Gibbs posterior, Corollary 3.4 yields $O\left(\frac{(\ln N)^{r+1}}{\sqrt{N}}\right)$ bounds on: **(1)** the generalization gap (11), **(2)** the excess true error over the optimal model (12), and **(3)** the parameter estimation error (13)–(14). The constants for the learning bounds, as given by Corollary 3.4, are in Table 1. The influence of system properties and convergence rates is discussed in Section 4. For additional intuition on the relationship between Corollary 3.2 and Corollary 3.4 see Appendix B.6.

Intuition behind the proof: The proof of Corollary 3.4 is in Appendix F. For MAP, (11)–(12) are derived from Corollary 3.2 by constructing suitable data-dependent posteriors and applying oracle inequality techniques from [2, Theorem 4.1]. For single draw and MEP, similar arguments apply using [2, Theorems 2.7 and Corollary 2.6] combined with the moment generating function bounds from the proof of Theorem 3.1. Finally, the parameter estimation bounds (13)–(14) follow from the

[2] The precise meaning of probability over data and samples drawn from ρ_{Gibbs} is explained in Remark F.1

oracle inequality (12) by relating the parameter error to the loss difference through strong convexity (for (13)) or persistence of excitation (for (14)).

4 Discussion on the bound and conclusions

In this paper, we derived PAC-Bayesian error bounds for stochastic LTI systems with inputs and sub-Gaussian noise.

The convergence rates are fastest for bounded noise, slower for Lipschitz losses, and slowest for quadratic losses with general sub-Gaussian noise (see Table 2). For the unbounded case with Lipschitz loss it is $O((\ln N)/\sqrt{N})$ and the logarithmic factor stems from using the truncation argument which replaces the unbounded noise of the data generator with a bounded noise of magnitude $O(\ln(N))$. For the case of unbounded data and quadratic loss, the bound is $O((\ln N)^{3/2}/\sqrt{N})$ and the extra $O((\ln N)^{1/2})$ factor comes by replacing the quadratic loss with the truncated quadratic loss whose Lipschitz constant is $O((\ln N)^{1/2})$. For the parameter estimation errors all these bounds are extended with an additional $O((\ln N))$ factor, which arises from the uniform PAC bound (11) of Corollary 3.4. In turn, the factor $O((\ln N))$ in the uniform PAC bound comes by instantiating the original PAC-Bayesian bound using uniform or Gaussian priors and choosing a suitable posterior which is centered around the estimated model, such that KL-divergence between prior and posterior is $O(\ln(N))$

All bounds in Theorem 3.1, Corollary 3.2, and Corollary 3.4 improve with increased system stability, lower sub-Gaussian noise norm, and depend logarithmically on the confidence level δ via $\ln(1/\delta)$. Specifically, the bounds increase monotonically with the sub-Gaussian norm $\|\mathbf{e}_g\|_{\psi_2}$ of the noise, and with the ℓ_1 norms and mixing coefficients of the Markov parameters of both the predictors and the data generator ($\|\alpha_\theta\|_{\ell_1}$, $\|\alpha_g\|_{\ell_1}$, $\theta_\infty(\alpha_\theta)$, $\theta_\infty(\alpha_g)$). Smaller ℓ_1 norms and mixing coefficients correspond to more stable systems and weaker dependence on past inputs, making learning easier. Similarly, a smaller sub-Gaussian norm (i.e., lower noise variance) yields tighter bounds.

In summary, the bounds are tighter for less noisy and more stable systems, aligning with intuition: learning is easier when the system is less noisy and more stable. The bound of Theorem 3.1 and Corollary 3.2 depend on the prior π, both via the KL-divergence between the prior and posterior and via the terms $\ln E_{\theta\sim\pi} \exp\left(\mathcal{C}_i(2\lambda, \theta, N, \delta, \epsilon)\right)$. Strategies for choosing priors remains work for future research. Intuitively, one expects that a good prior is the one which is concentrated around the true parameter values, which renders the optimal (Gibbs posterior) close to the prior, which then renders the KL-divergence term small. Additionally, if the goal is to bound not only the generalization gap but also the parameter estimation error (in terms of the H_2 distance between the true and learned systems (14), or the Euclidean distance between parameters (13)), the bounds also depend on the *persistence of excitation* of the input $\mathbf{w}$ ($m_{\mathbf{w}}$) and the *identifiability* constant (m_Θ). Larger values of $m_{\mathbf{w}}$ (richer input) and m_Θ (greater separation between models) yield tighter bounds, see Appendix B.5 for a more detailed discussion. The bound (14) can also be used to bound the system matrices (up to similarity), see Remark F.10, Appendix F.5.

Table 2: Summary of convergence rates for different noise and loss function types

Scenario	Empirical PAC-Bayesian Bound	PAC/Oracle/parameter est./H_2 bounds
Bounded noise	$O\left(\frac{1}{\sqrt{N}}\right)$	$O\left(\frac{\ln(N)}{\sqrt{N}}\right)$
Lipschitz loss	$O\left(\frac{\ln(N)}{\sqrt{N}}\right)$	$O\left(\frac{(\ln(N))^2}{\sqrt{N}}\right)$
Quadratic loss	$O\left(\frac{(\ln(N))^{3/2}}{\sqrt{N}}\right)$	$O\left(\frac{(\ln(N))^{5/2}}{\sqrt{N}}\right)$

For parameter estimation error, our obtained rate $O((\ln N)^{r+1}/\sqrt{N})$ is more conservative than classical asymptotic theory [26, 4] (which yields $O(1/\sqrt{N})$ asymptotic consistency) and existing finite-sample bounds [32, 24, 39, 42](which achieve $O(\ln N/\sqrt{N})$) . However, this is the expected price for generality: our bound applies to a broad class of learning algorithms, provides non-asymptotic guarantees with high probability, and simultaneously bounds the generalization gap (properties not offered by prior results). In addition, most existing finite-sample bounds are algorithm-specific (e.g., for least-squares only) and do not address the generalization gap.

References

[1] P. Alquier, X Li, and O. Wintenberger. Prediction of time series by statistical learning: general losses and fast rates. *Dependence Modeling*, 1(2013):65–93, 2013.

[2] Pierre Alquier. User-friendly introduction to PAC-bayes bounds. *arXiv:2110.11216*, 2021.

[3] Pierre Alquier and Olivier Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883 – 913, 2012.

[4] P. E. Caines. *Linear Stochastic Systems*. John Wiley and Sons, 1988.

[5] Marco C Campi and Erik Weyer. Finite sample properties of system identification methods. *IEEE Transactions on Automatic Control*, 47(8):1329–1334, 2002.

[6] Minshuo Chen, Xingguo Li, and Tuo Zhao. On generalization bounds of a family of recurrent neural networks. In *Proceedings of AISTATS 2020*, volume 108 of *PMLR*, pages 1233–1243, 8 2020.

[7] G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *UAI*. AUAI Press, 2017.

[8] Deividas Eringis, John Leth, Zheng-Hua Tan, Rafael Wisniewski, and Mihaly Petreczky. PAC-bayesian bounds for learning lti-ss systems with input from empirical loss. *arXiv:2303.16816*, 2023.

[9] Deividas Eringis, John Leth, Zheng-Hua Tan, Rafal Wisniewski, Alireza Fakhrizadeh Esfahani, and Mihaly Petreczky. PAC-Bayesian theory for stochastic LTI systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 6626–6633, 2021.

[10] Deividas Eringis, John Leth, Zheng Hua Tan, Rafal Wisniewski, and Mihaly Petreczky. Explicit construction of the minimum error variance estimator for stochastic LTI-ss systems. *Automatica*, 153, July 2023.

[11] Deividas Eringis, John Leth, Zheng-Hua Tan, Rafal Wisniewski, and Mihaly Petreczky. PAC-Bayes Generalisation Bounds for Dynamical Systems Including Stable RNNs. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, number 11, pages 11901–11909, March 2024.

[12] Deividas Eringis, John Leth, Zheng-Hua Tan, Rafal Wisniewski, and Mihály Petreczky. PAC-Bayesian Error Bound, via Rényi Divergence, for a Class of Linear Time-Invariant State-Space Models. In *41st International Conference on Machine Learning*, volume 235, pages 12560–12587, Vienne, Austria, July 2024. PMLR.

[13] Herbert Federer. *Geometric Measure Theory*. Classics in Mathematics. Springer Berlin Heidelberg, Berlin, Heidelberg, 1 edition, 1996. eBook ISBN: 978-3-642-62010-2. Originally published as volume 153 in the series Grundlehren der mathematischen Wissenschaften. IV, 677 pages. Springer-Verlag Berlin Heidelberg, 1996.

[14] Dylan Foster and Max Simchowitz. Logarithmic regret for adversarial online control. In *Proceedings of the 37th ICML*, volume 119 of *PMLR*, page 3211–3221. PMLR, 7 2020.

[15] P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. PAC-bayesian theory meets bayesian inference. In *NIPS*, pages 1876–1884, 2016.

[16] I. I. Gikhman and A. V. Skorokhod. *Introduction to the Theory of Random Processes*. Dover, 1st edition, 1996.

[17] B. Guedj. A Primer on PAC-Bayesian Learning. *arXiv:1901.05353*, 2019.

[18] B. Hanzon. On the differentiable manifold of fixed order stable linear systems. *Systems & Control Letters*, 13(4):345–352, 1989.

[19] Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[20] Jiabao He, Ingvar Ziemann, Cristian R. Rojas, S. Joe Qin, and Håkan Hjalmarsson. Finite sample analysis of open-loop subspace identification methods. *IEEE Transactions on Automatic Control*, pages 1–16, 2026.

[21] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm, 2019.

[22] Tohru Katayama. *Subspace methods for system identification*. Communications and control engineering. Springer, Berlin, London, 2005.

[23] Pascal Koiran and Eduardo D. Sontag. Vapnik-chervonenkis dimension of recurrent neural networks. *Discrete Applied Mathematics*, 86(1):63–79, 1998.

[24] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888, 2020.

[25] A. Lindquist and G. Picci. *Linear Stochastic Systems: A Geometric Approach to Modeling, Estimation and Identification*. Springer, 2015.

[26] L. Ljung. *System Identification: Theory for the user (2nd Ed.)*. PTR Prentice Hall., Upper Saddle River, USA, 1999.

[27] Louis Massucci, Fabien Lauer, and Marion Gilson. A statistical learning perspective on switched linear system identification. *Automatica*, 145:110532, 2022.

[28] Tomas McKelvey, Anders Helmersson, and Thomas Ribarits. Data driven local coordinates for multivariable linear systems and their application to system identification. *Autom.*, 40(9):1629–1635, 2004.

[29] Abdelkader Metakalard, Fabien Lauer, Kevin Colin, and Marion Gilson. Uniform error bounds for quantized dynamical models. *IFAC Journal of Systems and Control*, 35:100373, 2026.

[30] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2004.

[31] Brett Ninness and Soren Henriksen. Bayesian system identification via markov chain monte carlo techniques. *Automatica*, 46(1):40–51, 2010.

[32] Samet Oymak and Necmiye Ozay. Revisiting ho–kalman-based system identification: Robustness and finite-sample analysis. *IEEE Transactions on Automatic Control*, 67(4):1914–1928, 4 2022.

[33] R. L. M. Peeters. *System Identification Based on Riemannian Geometry: Theory and Algorithms*. PhD thesis, Free University, Amsterdam, 1994.

[34] Gianluigi Pillonetto, Tianshi Chen, Alessandro Chiuso, Giuseppe De Nicolao, and Lennart Ljung. *Regularized system identification: Learning dynamic models from data*. Springer Nature, 2022.

[35] Tuhin Sarkar, Alexander Rakhlin, and Munther A. Dahleh. Finite time LTI system identification. *J. Mach. Learn. Res.*, 22:26:1–26:61, 2021.

[36] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[37] Max Simchowitz. *Statistical Complexity and Regret in Linear Control*. University of California, Berkeley, 2021.

[38] Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pages 2714–2802. PMLR, 2019.

[39] Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *Proceedings of the 37th ICML*, volume 119 of *PMLR*, page 8937–8948. PMLR, 7 2020.

[40] David J. Smith and Mavina K. Vamanamurthy. How small is a unit ball? *Mathematics Magazine*, 62(2):101–107, 1989.

[41] Eduardo D Sontag. A learning result for continuous-time recurrent neural networks. *Systems & control letters*, 34(3):151–158, 1998.

[42] A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas. Statistical learning theory for control: A finite-sample perspective. *IEEE Control Systems Magazine*, 43(6):67–97, 2023.

[43] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2 edition, 2026.

[44] Mathukumalli Vidyasagar and Rajeeva L Karandikar. A learning theory approach to system identification and stochastic adaptive control. *Probabilistic and randomized methods for design under uncertainty*, pages 265–302, 2006.

A Table of symbols

Table 3: Notation, Descriptions, and Dimensions

Notation	Description	Dimension
Sets and Spaces		
Ω	Sample space	Set
$\mathcal{F}$	σ-algebra on Ω	Collection of sets
$\mathbf{P}$	Probability measure on $\mathcal{F}$	Function
$\Theta \subseteq \mathbb{R}^{n_\theta}$	Parameter space	Set
B_Θ	σ-algebra of Lebesgue-measurable subsets of Θ	Collection of sets
$\mathbb{R}^n$	n-dimensional Euclidean space	Set
$\mathcal{M}_\pi$	Set of densities absolutely continuous w.r.t. π	Set of functions
Stochastic Processes and Random Variables		
$\mathbf{y}(t) \in \mathbb{R}^{n_\mathbf{y}}$	Output (label) stochastic process at time t	n_y-dimensional vector
$\mathbf{u}(t) \in \mathbb{R}^{n_\mathbf{u}}$	Input (feature) stochastic process at time t	n_u-dimensional vector
$\mathbf{w}(t) \in \mathbb{R}^{n_\mathbf{w}}$	Combined process: $[\mathbf{y}^T(t), \mathbf{u}^T(t)]^T$ or $\mathbf{u}(t)$	n_w-dimensional vector
$\mathbf{e}_g(t) \in \mathbb{R}^{n_e}$	Noise/innovation process, i.i.d., zero-mean, sub-Gaussian	n_e-dimensional vector
$\mathbf{e}^s(t)$	Innovation process in Kalman filter form	Vector
$\mathbf{x}(t) \in \mathbb{R}^n$	State of autonomous LTI data generator	n-dimensional vector
$\hat{\mathbf{x}}(t) \in \mathbb{R}^{\hat{n}}$	State of predictor LTI system	$\hat{n}$-dimensional vector
$\hat{\mathbf{y}}_\theta(t \mid t_0) \in \mathbb{R}^{n_\mathbf{y}}$	Predicted output based on past data from t_0 to t	n_y-dimensional vector
$\hat{\mathbf{y}}_\theta(t)$	Predicted output with infinite past ($t_0 \to -\infty$)	n_y-dimensional vector
$\hat{\mathbf{s}}_g(t) \in \mathbb{R}^n$	State of data generator in innovations form	n-dimensional vector
$\hat{\mathbf{s}}_g(t \mid t_0)$	State of predictor in innovations form	n-dimensional vector
System Matrices		
$\hat{A}_\theta \in \mathbb{R}^{\hat{n} \times \hat{n}}$	State matrix of predictor	$\hat{n} \times \hat{n}$ matrix
$\hat{B}_\theta \in \mathbb{R}^{\hat{n} \times n_\mathbf{w}}$	Input matrix of predictor	$\hat{n} \times n_w$ matrix
$\hat{C}_\theta \in \mathbb{R}^{n_\mathbf{y} \times \hat{n}}$	Output matrix of predictor	$n_y \times \hat{n}$ matrix
$\hat{D}_\theta \in \mathbb{R}^{n_\mathbf{y} \times n_\mathbf{w}}$	Direct feedthrough matrix of predictor	$n_y \times n_w$ matrix
$\hat{D}_{\mathbf{u},\theta} \in \mathbb{R}^{n_\mathbf{y} \times n_\mathbf{u}}$	Sub-matrix of $\hat{D}_\theta$ when $\mathbf{w} = [\mathbf{y}^T, \mathbf{u}^T]^T$	$n_y \times n_u$ matrix
A_0, B_0, K_0, C_0, D_0	Matrices of data generator in innovations form	Various dimensions
$A_g \in \mathbb{R}^{n \times n}$	State matrix of autonomous LTI system	$n \times n$ matrix
$K_g \in \mathbb{R}^{n \times n_e}$	Kalman gain matrix	$n \times n_e$ matrix
$C_g \in \mathbb{R}^{n_e \times n}$	Output matrix of autonomous system	$n_e \times n$ matrix
H_θ	Transfer function of predictor	Matrix/rational function
Parameters and Scalars		

Continued on next page

Table 3 – Continued

Notation	Description	Dimension
$\theta \in \mathbb{R}^{n_\theta}$	Parameter vector	n_θ-dimensional vector
$\theta_\star$	Learned/estimated parameter	n_θ-dimensional vector
θ_{true}	True parameter minimizing $\mathcal{L}(\theta)$	n_θ-dimensional vector
n_θ	Dimension of parameter space	Scalar
$n_{\mathbf{y}}$	Output dimension	Scalar
$n_{\mathbf{u}}$	Input dimension	Scalar
$n_{\mathbf{w}}$	Dimension of $\mathbf{w}(t)$: $n_{\mathbf{y}} + n_{\mathbf{u}}$ or $n_{\mathbf{u}}$	Scalar
$\hat{n}$	State dimension of predictor	Scalar
n	State dimension of data generator	Scalar
$n_e = n_{\mathbf{y}} + n_{\mathbf{u}}$	Dimension of noise/innovation	Scalar
$t \in \mathbb{Z}$	Time index	Scalar
t_0	Initial time for prediction	Scalar
N	Sample size (number of data points)	Scalar
$\delta \in (0, 0.5)$	Confidence level for PAC-Bayesian bound	Scalar
$\epsilon \in \{-1, 1\}$	Sign parameter in Theorem 3.1	Scalar
$\lambda > 0$	PAC-Bayesian trade-off parameter	Scalar
$\gamma \in (0, 1)$	Spectral radius bound: $\|\hat{A}_\theta^k\|_2 \leq M\gamma^k$	Scalar
$M, C > 0$	Constants bounding system matrices	Scalars
$c_{\mathbf{e}}$	Bound on $\|\mathbf{e}_g(t)\|_\infty$ (bounded noise case)	Scalar
L_ℓ	Lipschitz constant of loss function ℓ	Scalar
$L(\theta)$	Loss-dependent constant	Scalar
L_{qd}	Constant for quadratic loss case	Scalar
m_Θ	Identifiability constant (strong convexity lower bound)	Scalar
$m_{\mathbf{w}}$	Persistence of excitation constant	Scalar
K	$\min\{1, \|\bar{\alpha}\|_{\ell_1}\}$	Scalar
Functions and Operators		
$\mathbf{E}[\cdot]$	Expectation operator	Operator
$E_{\theta\sim\rho}[g(\theta)]$	Expectation of g under distribution ρ	Scalar
$\|\cdot\|_2$	Euclidean norm (vector) or induced matrix 2-norm	Scalar
$\|\cdot\|_\infty$	Infinity norm	Scalar
$\|\cdot\|_{\psi_2}$	Sub-Gaussian norm	Scalar
$\ell\colon \mathbb{R}^{n_{\mathbf{y}}} \times \mathbb{R}^{n_{\mathbf{y}}} \to [0, \infty)$	Loss function	Function
ℓ_2 loss	$\ell(y, \hat{y}) = \|y - \hat{y}\|_2$	Function
Quadratic loss	$\ell(y, \hat{y}) = \|y - \hat{y}\|_2^2$	Function
$\alpha(k)$	Sequence of matrices (Markov parameters)	Matrix-valued sequence
$\bar{\alpha}(k)$	$\sup_{\theta\in\Theta} \|\alpha_\theta(k)\|_2$	Scalar-valued sequence
$\chi_A(\cdot)$	Indicator function of set A	Function
$\text{vol}(A)$	Lebesgue measure of set A	Scalar
$\rho(\theta)$	Probability density function on Θ	Function
$\pi(\theta)$	Prior probability density on Θ	Function
$\mathcal{L}(\theta)$	True error (expected loss)	Scalar
$\hat{\mathcal{L}}_N(\theta)$	Empirical error (average loss on data)	Scalar
$V_N(\theta)$	Auxiliary empirical error with infinite past	Scalar

Continued on next page

Table 3 – Continued

Notation	Description	Dimension
$\mathrm{KL}(\rho\|\pi)$	Kullback–Leibler divergence between ρ and π	Scalar
$\star$	Convolution operator: $(\alpha \star \mathbf{r})(t) = \sum_{k=0}^{\infty} \alpha(k)\mathbf{r}(t-k)$	Operator
$\triangleq$	Defined by	Relation
$(\cdot)^T$	Matrix/vector transpose	Operator
$\arg\min$	Argument of minimum	Operator
Norms and Constants for PAC-Bounds		
$\|\alpha\|_{\ell_1}$	ℓ_1 norm of sequence: $\sum_{k=0}^{\infty} \|\alpha(k)\|_2$	Scalar
$\theta_\infty(\alpha)$	Weakly dependent mixing coefficient: $\sum_{k=0}^{\infty} k\|\alpha(k)\|_2$	Scalar
$\alpha_g(k)$	Markov parameters of data generator	Matrix
$\alpha_\theta(k)$	Markov parameters of predictor	Matrix
$G_e(\theta)$	$(1 + \|\alpha_\theta\|_{\ell_1})\|\alpha_g\|_{\ell_1}$	Scalar
$G_{e,1}(\theta)$	$\theta_\infty(\alpha_\theta)\|\alpha_g\|_{\ell_1} + \theta_\infty(\alpha_g)(\|\alpha_\theta\|_{\ell_1} + 1)$	Scalar
G_e	$(1 + \|\bar{\alpha}\|_{\ell_1})\|\alpha_g\|_{\ell_1}$	Scalar
$G_{e,1}$	$\theta_\infty(\bar{\alpha})\|\alpha_g\|_{\ell_1} + \theta_\infty(\alpha_g)(\|\bar{\alpha}\|_{\ell_1} + 1)$	Scalar
C_δ	Constant in PAC-Bayesian bound	Scalar
K_δ	Threshold for sample size N in PAC-Bayesian bound	Scalar
$\mathcal{C}_i(\lambda, \theta, N, \delta, \epsilon)$	Constants in PAC-Bayesian bound $(i = 1, 2)$	Scalar
$C_{u,1}, C_{u,2}$	Constants for unbounded noise case	Scalars
$\bar{\Psi}_{\mathbf{e}_g}(s)$	$e^{14^2 s^2 \|\mathbf{e}_g\|_{\psi_2}^2 + 3s\|\mathbf{e}_g\|_{\psi_2}}$	Scalar
Learning and Posterior Distributions		
ρ	Posterior distribution over Θ	Density function
ρ_{Gibbs}	Gibbs posterior minimizing PAC-Bayesian bound	Density function
$\mathcal{D}$	Training dataset: $\{\mathbf{y}(t)(\omega), \mathbf{u}(t)(\omega)\}_{t=0}^{N-1}$	Set of vectors

B Additional discussion

B.1 Relationship between the learning problem considered in this paper and control theory/machine learning

Control, system identification and machine learning. In control theory, a control problem involves designing inputs to achieve desired output/state properties using feedback laws. Learning a predictive model differs fundamentally: we aim to learn a model that predicts current outputs from past outputs and inputs – a time-series prediction problem.

The link between them is that solving control problems requires predictive models, which often have to be learned from data. In control theory, the latter learning task is known as system identification. That is, system identification and machine learning address the same problem: find a predictive model from data. The key distinction between them is that system identification prioritizes model classes useful for control design. For this reason, it often focuses on parameter estimation error over generalization error. Small parameter estimation error ensures the learned model approximates the true physical system across all inputs, including those outside the training distribution. This property is essential for control design, since controllers typically generate inputs very different from those used for model learning.

This paper addresses learning state-space models from data, and its bounds are relevant for both system identification and machine learning, as it bounds both the generalization error and the parameter estimation error, and in fact it shows that the two are related. Specifically, we learn

state-space models of the form (Assumption 2.1): $\hat{\mathbf{x}}(t+1) = \hat{A}_\theta \hat{\mathbf{x}}(t) + \hat{B}_\theta \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{u}(t) \end{bmatrix}$, $\hat{\mathbf{x}}(0) = 0$, $\hat{\mathbf{y}}(t \mid 0) = \hat{C}_\theta \hat{\mathbf{x}}(t) + \hat{D}_{\theta,u} \mathbf{u}(t)$, where $\hat{\mathbf{y}}(t \mid 0)$ is the predicted output at time t given all past outputs $\{\mathbf{y}(s)\}_{s=0}^{t-1}$ and past/present inputs $\{\mathbf{u}(s)\}_{s=0}^{t}$.

LTI systems as linear RNNs. This model class corresponds exactly to linear RNNs with inputs $\mathbf{w}(t) = \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{u}(t) \end{bmatrix}$ whose last layer is a completely linear layer $\hat{\mathbf{y}}(t \mid 0) = \hat{C}_\theta \hat{\mathbf{x}}(t) + \hat{D}_\theta \mathbf{w}(t)$, where $\hat{D}_\theta = \begin{bmatrix} 0 & \hat{D}_{\theta,u} \end{bmatrix}$ is such that only the component of $\mathbf{w}(t)$ corresponding to the input $\mathbf{u}(t)$ is used in the linear combination. A general RNN would then be of the form $\hat{\mathbf{x}}(t+1) = \vec{\sigma}\ (\hat{A}_\theta \hat{\mathbf{x}}(t) + \hat{B}_\theta [\mathbf{y}(t)\ \mathbf{u}(t)])$, $\hat{\mathbf{x}}(0) = 0$, $\hat{\mathbf{y}}(t \mid 0) = \hat{C}_\theta \hat{\mathbf{x}}(t) + \hat{D}_{\theta,u} \mathbf{u}(t)$ where $\vec{\sigma}$ is a component-wise application of nonlinear activation function σ, such as a sigmoid, ReLU or tanh function. Our case correspond to choosing σ to be the identity activation function, i.e., $\sigma(x) = x$. That is, learning linear RNNs is exactly the learning problem considered in this paper, and the proven bounds apply directly to linear RNNs.

Differences with respect to classical machine learning problems and time-series prediction. What perhaps distinguishes us from a more classical setting is that:

Autoregressive modeling: We also use past values of the output $\mathbf{y}$ to predict current value of $\mathbf{y}$: if we choose the parametrization so that $\hat{B}_\theta = \begin{bmatrix} 0 & \hat{B}_{\theta,u} \end{bmatrix}$, i.e., $\hat{B}_\theta \mathbf{w}(t) = \hat{B}_{\theta,u} \mathbf{u}(t)$ then we get the classical setting of learning models to predict the current output based on past and current inputs.

Learning from a single time series: We assume that the training data is a single sample $\{y(t), u(t)\}_{t=0}^{N}$ from the random variables $\{\mathbf{y}(t), \mathbf{u}(t)\}_{t=0}^{N}$, instead of having N independently drawn samples $\{y_i(t), u_i(t)\}_{t=0,i=1}^{T,N}$ of length T.

Realizability: We assume that $\mathbf{w}(t) = \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{u}(t) \end{bmatrix}$ is generated by a LTI state-space system.

The first two features are not unique to learning for control, but it is present in machine learning. The last assumption is more control-oriented but is used only for proving the bound, not for the learning algorithms themselves. As it was explained before, realizability ensures that the output can be predicted from past outputs and inputs using a linear state-space model (linear RNN), making the learning goal attainable. From a control perspective, this is a natural sanity check: if the underlying system cannot be modeled as a noisy LTI state-space system, then no LTI model will be good enough for control. From a learning perspective, realizability ensures data mixing, which is a standard requirement for the empirical loss to converge to the true loss (a law of large numbers). Note that i.i.d. data is mixing with zero mixing coefficients. While one could assume mixing explicitly, estimating mixing coefficients for partially known systems is difficult. By contrast, assuming the data is generated by a stable LTI system allows us to replace mixing coefficients with quantities depending on estimates on matrix norms and Lyapunov exponents of the data generator – information often available from physical considerations.

For this reason we believe that the problem considered in this paper is a particular case of time-series prediction/learning problem, and the results apply directly to learning linear RNNs from time-series data.

Challenges in extending our results to nonlinear RNNs. The results for LTI case could be useful for RNN case, as LTIs represent a special case of RNNs. However, extending the results to nonlinear RNNs faces several technical obstacles. First, extending the truncation argument for unbounded noise with Lipschitz loss to nonlinear state-space models is nontrivial because the output is not a simple convolution of inputs with Markov parameters. While [3] applied truncation to nonlinear autoregression, extending it to our setting would require bounding Lipschitz constants and decay rates of nonlinear predictors rather than relying on Markov parameters. Second, our truncation argument for quadratic loss exploits the fact that stable LTI systems driven by sub-Gaussian noise produce bounded responses with high probability. Whether this property extends to nonlinear models remains unclear. Third, our parameter estimation bounds (especially H_2 norm bounds) rely fundamentally on realization theory and identifiability of LTI systems – tools absent for general nonlinear systems. Extensions may be possible for restricted classes such as bilinear or switched systems with developed realization theory.

B.2 Assumptions on the loss

B.3 Proof of Theorem 3.1: quadratic loss

We assume the loss to be L-Lipschitz or quadratic For the bounded-noise case, L-Lipschitzness ensures that closeness in the predicted output translates to closeness in the loss: if $|\hat{\mathbf{y}}_\theta(t) - \mathbf{y}(t)|$ is small, then $\ell(\hat{\mathbf{y}}_\theta(t), \mathbf{y}(t))$ is also small. This is exploited in Lemmas E.5 and E.6. For the unbounded-noise case, Lipschitzness is used to reduce the problem to the bounded case. Specifically, we upper bound the moment generating function of the loss difference between original and truncated data (Lemma E.8), where the truncation bounds the noise. The key insight is that although quadratic loss is not globally Lipschitz, for LTI systems the prediction-error norm is high-probability bounded. We therefore replace quadratic loss with a truncated version that coincides with it in the high-probability region and is Lipschitz globally (Appendix E.3). Lipschitzness is standard in PAC bounds, even in the i.i.d., setting, whereas convexity is typically not required. Convexity and smoothness matter for optimizing the empirical loss, not for deriving generalization bounds. In system identification, quadratic loss is preferred because it is both convex and smooth, facilitating optimization. But for unbounded data, quadratic loss is not globally Lipschitz, motivating the truncation approach above.

One may wonder what the role of convexity is in the presented bounds. Similarly to other PAC bounds, convexity of the loss function does not play a role in our bounds, except in Corollary 3.4, in the part on the bound of parameter estimation error, we assume strong convexity (**SConv**, Assumption 3.3) of the true loss in the parameters, which is distinct from the loss function being convex. Strong convexity of the true loss depends on both the loss function and the predictor's dependence on parameters, and is not implied by convexity of the loss itself.

B.4 Weak dependence, mixing coefficients, convolution

Consider a bounded stationary process $\mathbf{z}$ taking values in $\mathbb{R}^{n_z}$ for some $n_z > 0$. Let $\mathcal{F}_{1-Lip,n}$ be the set of functions $f : (\mathbb{R}^{n_r})^n \to \mathbb{R}$ which are 1-Lipschitz in the following sense: $\|f(y_1, \dots, y_n) - f(y'_1, \dots, y'_n)\|_2 \leq \sum_{i=1}^n \|y_i - y'_i\|_2$ for all $y_1, \dots, y_n, y'_1, \dots, y'_n \in \mathbb{R}^n$. Let $\mathcal{F}_t$ be the σ-algebra generated by $\{\mathbf{z}(j)\}_{j \leq t}$. Define the γ-weak mixing coefficients $\theta_{\infty,n}(1) = \sup_{t \in \mathbb{Z}} \sup_{f \in \mathcal{F}_{1-Lip,n}} |\mathbb{E}[f(\mathbf{z}(t+1), \dots, \mathbf{z}(t+n)) \mid \mathcal{F}_t] - \mathbb{E}[f(\mathbf{z}(t+1), \dots, \mathbf{z}(t+n))]|$ of $\mathbf{z}$. We say that $\mathbf{z}$ is weakly dependent, if $\theta_\infty(1) = \sup_{n \geq 1} \theta_{\infty,n}(1) < +\infty$. It then follows that if $\mathbf{r}$ is a bounded i.i.d. process, and $\alpha = \{\alpha(k)\}_{k=0}^\infty$ is a square summable sequence of matrices, then $(\alpha \star \mathbf{r})$ is stationary and weakly dependent with $\theta_\infty(1) \leq 2\theta_\infty(\alpha)\mathcal{C}$, see [3, Proposition 4.1].

B.5 Persistence of excitation, identifiability

The PE condition amounts to the input being persistently exciting of order K for any K. Practically, it means that for any k, for large enough N, the following matrix formed from the observed sample path $w(t) = \begin{bmatrix} y^T(t) & u^T(t) \end{bmatrix}^T$ of the process $\mathbf{w}(t) = \begin{bmatrix} \mathbf{y}^T(t) & \mathbf{u}^T(t) \end{bmatrix}^T$ satisfies the following definiteness condition $m_{\mathbf{w}} I \leq \frac{1}{N} W_{0,|k,N} W_{0,|k,N}^T$, where $W_{0,|k,N} = \begin{bmatrix} w(0) & w(1) & \cdots & w(N-1) \\ w(1) & w(2) & \cdots & w(N) \\ \vdots & \vdots & \cdots & \vdots \\ w(k) & w(k+1) & \cdots & w(N+k-1) \end{bmatrix}$. The existence of a constant $m_{\mathbf{w}}$ is ensured by the assumption that the data generator is minimum-phase, or, alternatively, that the joint input-output process $\mathbf{w}(t) = \begin{bmatrix} \mathbf{y}^T(t) & \mathbf{u}^T(t) \end{bmatrix}^T$ is generated by an invertable ARMA system (see Lemma B.1 in [22]). This is a widespread assumption in system identification.

While the existence of $m_{\mathbf{w}}$ is ensured, its value indeed needs to be estimated. This can be done by taking $m_{\mathbf{w}}$ as a lower bound of the eigenvalues of $\frac{1}{N} W_{0|k,N} W_{0|k,N}^\top$ for large k and N: if the estimate is much lower that the true value of $m_{\mathbf{w}}$, the bound will be more conservative, if it is too optimistic (higher than the true value), the bound may be invalid.

The assumption (**SConv**) of Assumption 3.3 implies that $\mathcal{L}(\theta) - \mathcal{L}(\theta_{\text{true}}) \geq m_\Theta \|\theta - \theta_{\text{true}}\|_2^2$. In turn, the latter means *structural identifiability* around θ_{true}, i.e., that for any $\theta \neq \theta_{\text{true}}$, the input-output behavior of the corresponding models must be different. Otherwise, their true losses would be the same leading to $0 \geq m_\Theta \|\theta - \theta_{\text{true}}\|_2^2 > 0$, which is a contradiction. It also implies that $\mathbf{w}$ is

persistently exciting in the sense that if the response of two models parameterized by $\theta \neq \theta_{\text{true}}$ is the same for $\mathbf{w}$, then these models are identical and hence there output response is the same for any input. Indeed, if the output response of the two models to $\mathbf{w}$ are identical, then $\mathbf{y}_\theta(t) = \mathbf{y}_{\theta_{\text{true}}}(t)$ and hence $\mathcal{L}(\theta) = \mathcal{L}(\theta_{\text{true}})$.

That is, the value $m_\Theta > 0$ captures both identifiability and persistence of excitation, while $m_{\mathbf{w}}$ captures only persistence of excitation. However, $m_{\mathbf{w}}$ is used to bound the H_2 distance between the true and estimated systems, and the latter distance corresponds to the distance between the Markov-parameters. In turn, if we parametrize LTI systems by the infinite sequence of their Markov-parameters, the resulting parametrization is always structurally identifiable, as Markov-parameters determine the input-output behavior uniquely.

B.6 Additional explanation of the results for bounded, unbounded and Lipschitz loss and unbounded and quadratic loss cases

In a nutshell, the statements of Theorem 3.1 and Corollary 3.2 can be interpreted as follows

Bounded data This is the simplest case, the data is assumed to be generated by an LTI system driven by a noise whose norm is bounded by $c_{\mathbf{e}}$, i.e., $\|\mathbf{e}_g(t)\|_\infty \leq c_{\mathbf{e}}$. Then the learning data $\| \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{u}(t) \end{bmatrix} \|_2 \leq c_{\mathbf{e}}\sqrt{n_e}\|\alpha_g\|_2$ and $\max\{\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2, \|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t \mid 0)\|_2\} \leq G_e(\theta)c_{\mathbf{e}}\sqrt{n_e}$ and the statement of Theorem 3.1 is as follows:

Theorem B.1 (Theorem 3.1: bounded case)**.** *Let Assumptions 2.2, 2.3 and 2.1 hold and assume that the noise process of the data generator satisfies $\|\mathbf{e}_g(t)\|_\infty \leq c_{\mathbf{e}}$. For any a prior density π over Θ, confidence level $\delta \in (0, 0.5)$, integer $\epsilon \in \{-1, 1\}$, $N \geq 2$, and $\lambda > 0$, it holds that*

$$\mathbf{P}\Big(\forall \rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}\mathcal{L}(\theta) \leq \epsilon E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\Big(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \frac{1}{2}\sum_{i=1}^{2} \ln E_{\theta\sim\pi} \exp\left(\mathcal{C}_{b,i}(2\lambda, \theta, L(\theta), c_{\mathbf{e}}, N)\right)\ Big)\Big) \geq 1 - 2\delta$$

with $\mathrm{KL}(\rho\|\pi) = E_{\theta\sim\rho} \ln \frac{\rho(\theta)}{\pi(\theta)}$ *the Kullback–Leibler divergence and*

$$\mathcal{C}_{b,i}(\lambda, \theta, L(\theta), c_{\mathbf{e}}, N),\ \mathcal{C}_{b,i}(s, \theta, L, c, N) \triangleq \frac{\mathcal{C}_{m,b,i}(s, \alpha_\theta, G_e(\theta) + 2G_{e,1}(\theta), L, c)}{N}$$

$$L(\theta) \triangleq \begin{cases} 2G_e(\theta)c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}} & \ell \text{ is quadratic loss} \\ L_\ell & \ell \text{ is } L_\ell\text{-Lipschitz} \end{cases}$$

$$\mathcal{C}_{m,b,i}(s, \alpha, X, L, c) \triangleq \begin{cases} sLc\sqrt{n_{\mathbf{e}}}\theta_\infty(\alpha)\|\alpha_g\|_{\ell_1} & i = 1 \\ s^2\frac{L^2}{2}X(\sqrt{n_{\mathbf{e}}}c)^2 & i = 2 \end{cases}$$

In particular, for $\epsilon = 1$ the theorem states:

$$\mathbf{P}\Big(\forall \rho \in \mathcal{M}_\pi : E_{\theta\sim\rho}\mathcal{L}(\theta) \leq E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\Big(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \frac{1}{2}\sum_{i=1}^{2} \ln E_{\theta\sim\pi} \exp\left(\mathcal{C}_{b,i}(2\lambda, \theta, L(\theta), c_{\mathbf{e}}, N)\right)\ Big)\Big) \geq 1 - 2\delta$$

and for $\epsilon = -1$ it states:

$$\mathbf{P}\Big(\forall \rho \in \mathcal{M}_\pi : E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) \leq E_{\theta\sim\rho}\mathcal{L}(\theta) + \frac{1}{\lambda}\Big(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \frac{1}{2}\sum_{i=1}^{2} \ln E_{\theta\sim\pi} \exp\left(\mathcal{C}_{b,i}(2\lambda, \theta, L(\theta), c_{\mathbf{e}}, N)\right)\ Big)\Big) \geq 1 - 2\delta$$

and by using the union bound we get the following bound on the absolute value of the generalization gap:

$$\mathbf{P}\Big(\forall \rho \in \mathcal{M}_\pi : \Big|E_{\theta\sim\rho}\mathcal{L}(\theta) - E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta)\Big| \leq \frac{1}{\lambda}\Big(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \frac{1}{2}\sum_{i=1}^{2} \ln E_{\theta\sim\pi} \exp\left(\mathcal{C}_{b,i}(2\lambda, \theta, L(\theta), c_{\mathbf{e}}, N)\right)\ Big)\Big) \geq 1 - 4\delta$$

That is, the constant $\mathcal{C}_i(\lambda, \theta, N, \delta, \epsilon)$, $i = 1, 2$ in Theorem 3.1 in the case of bounded data equals $\mathcal{C}_{b,i}(\lambda, \theta, L(\theta), c_{\mathbf{e}}, N)$ and hence it does not depend on δ and ϵ for the bounded case.

Informally, the theorem says that for any prior π on models, uniformly over all posteriors, the average (w.r.t. to a posterior ρ on models) generalization gaps $E_{\theta\sim\rho}\mathcal{L}(\theta) - E_{\theta\sim\rho}\widehat{\mathcal{L}}_N(\theta)$ is bounded from above with probability $1-\delta$ over data, by the expression $\frac{1}{\lambda}(KL(\rho\|\pi) + \ln(\frac{1}{\delta}) + \frac{1}{2}\ln E_{\theta\sim\pi}\exp(\mathcal{C}_{b,1}(2\lambda, \theta, c, N)) + \frac{1}{2}\ln E_{\theta\sim\pi}\exp(\mathcal{C}_{b,2}(2\lambda, \theta, c_{\mathbf{e}}, N))$, where $\mathcal{C}_{b,i}(2\lambda, \theta, c_{\mathbf{e}}, N)$ are constants $O(\lambda^2/N)$. That is, the bound includes a KL divergence $KL(\rho\|\pi)$ between the prior and posterior, the term $\ln(\frac{1}{\delta})$ describing the dependence on the confidence level δ, and the average w.r.t. the prior π: $\frac{1}{2}\sum_{i=1}^{2}\ln E_{\theta\sim\pi}\exp(\mathcal{C}_{b,i}(2\lambda, \theta, N))$. Note that the bound in Theorem 3.1 depends only on the prior and the distance of the posterior from the prior and on the temperature parameter λ. Intuitively, if λ is small, we penalize the difference between the posterior and the prior, i.e., we would like the posterior to be close to the prior.

The constants $\mathcal{C}_{b,i}(2\lambda, \theta, c_{\mathbf{e}}, N)$ depend on $L(\theta)$, the Markov parameters α_θ of the predictor parametrized by θ and the Markov parameters α_g of the data generator. The dependence takes place via the constants $G_e(\theta), G_{e,1}(\theta)$ and $\theta_\infty(\alpha_\theta)$, $\|\alpha_g\|_{\ell_1}$ which are functions of the ℓ_1 norms $\|\alpha_\theta\|_{\ell_1}$, $\|\alpha_g\|\ell_1$ and WDMC $\theta_\infty(\alpha_\theta), \theta_\infty(\alpha_g)$ of these Markov parameters. The constant $L(\theta)$ is the Lipschitz loss of the loss function ℓ applied to the outputs of the data generator and the predictions generated by the predictor corresponding to the parameter θ. If the loss is L_ℓ-Lipschitz, then $L(\theta) = L_\ell$. If the loss ℓ is quadratic, then the restriction of ℓ to the set containing the output of the data generating system and the predictor corresponding to θ is $2G_e(\theta)c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}}$.

Corollary 3.2 is obtained from Theorem 3.1 by taking $\lambda = \sqrt{N}$ and replacing

$$\theta_\infty(\alpha_\theta),\ G_e(\theta),\ L(\theta),\ G_{1,\theta}$$

by their uniform upper bounds

$$\theta_\infty(\bar{\alpha}),\ G_e,\ G_{e,1},\ L_b = \begin{cases} 2G_e c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}} & \ell \text{ is the quadratic loss} \\ L_\ell & \ell \text{ is } L_\ell\text{-Lipschitz} \end{cases}$$

Then $\mathcal{C}_i(\lambda, \theta, N)$ is upper bounded by $\mathcal{C}_{b,m,i}(2\sqrt{N}, \bar{\alpha}, G_e + 2G_{e,1}, L_b, c_{\mathbf{e}})$ which is independent of θ, hence, $\frac{1}{2\lambda}\ln E_{\theta\sim\pi}\exp\left(\mathcal{C}_i(2\lambda, \theta, N, \delta, \epsilon)\right) \leq \frac{1}{2\lambda}\mathcal{C}_{b,m,i}(2\sqrt{N}, \bar{\alpha}, G_e + 2G_{e,1}, L_b, c_{\mathbf{e}}) \leq \frac{\mathcal{C}_{b,m,i}(2, \bar{\alpha}, G_e + 2G_{e,1}, L_b, c_{\mathbf{e}})}{\sqrt{N}}$, leading to the statement of Corollary 3.2:

Corollary B.2 (Corollary 3.2: bounded data case)**.** *For every prior π, $\delta(0, 0.5)$, $\epsilon \in \{-1, 1\}$,*

$$\mathbf{P}\Big(\forall\rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}\mathcal{L}(\theta) \leq \epsilon E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\sqrt{N}}\left(\mathrm{KL}(\rho||\pi) + \ln\frac{1}{\delta} + \frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}(2, \bar{\alpha}, G_e + 2G_{e,1}, L_b, c_{\mathbf{e}})\right)\Big) \geq 1 - 2\delta.$$

That is, for $\epsilon = 1$, it states that:

$$\mathbf{P}\Big(\forall\rho \in \mathcal{M}_\pi : E_{\theta\sim\rho}\mathcal{L}(\theta) \leq E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\sqrt{N}}\left(\mathrm{KL}(\rho||\pi) + \ln\frac{1}{\delta} + \frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}(2, \bar{\alpha}, G_e + 2G_{e,1}, L_b, c_{\mathbf{e}})\right)\Big) \geq 1 - 2\delta$$

and for $\epsilon = -1$:

$$\mathbf{P}\Big(\forall\rho \in \mathcal{M}_\pi : E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) \leq E_{\theta\sim\rho}\mathcal{L}(\theta) + \frac{1}{\sqrt{N}}\left(\mathrm{KL}(\rho||\pi) + \ln\frac{1}{\delta} + \frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}(2, \bar{\alpha}, G_e + 2G_{e,1}, L_b, c_{\mathbf{e}})\right)\Big) \geq 1 - 2\delta$$

and with the union bound

$$\mathbf{P}\Big(\forall\rho \in \mathcal{M}_\pi : \left|E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) - E_{\theta\sim\rho}\mathcal{L}(\theta)\right| \leq \frac{1}{\sqrt{N}}\left(\mathrm{KL}(\rho||\pi) + \ln\frac{1}{\delta} + \frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}(2, \bar{\alpha}, G_e + 2G_{e,1}, L_b, c_{\mathbf{e}})\right)\Big) \geq 1 - 4\delta.$$

That is, the upper bound of Corollary 3.2 is $\frac{1}{\sqrt{N}}(KL(\rho\|\pi) + \ln(\frac{1}{\delta}) + O(1))$ as $\mathcal{C}_{b,i}$ becomes $O(1)$. Bounding the generalization gap and its multiple by -1 allows us to bound **1)** the absolute value of the generalization gap using the union bound, **2)** to derive oracle inequalities, i.e., bound the gap between the true error of the learned model and the best possible true error. In turn, the latter is used for deriving bounds on the parameter estimation error.

The bound above assumes that the data is generated by an LTI system with bounded inputs and noise. However, in many applications the standard assumption is that the noise and inputs are (sub-)Gaussian, and hence potentially unbounded. This prompted us to extend the results to data generated by systems with sub-Gaussian noise.

Unbounded data and L-Lipschitz loss functions In this case, we approximate unbounded data with bounded data generated by the same LTI system, but with a truncated, and hence bounded noise: if $\mathbf{e}_g(t)$ is the noise of the data generating system, then we use the truncated noise $\mathbf{e}_{g,N}(t) = \mathbf{e}_g(t)$ if $\|\mathbf{e}_g(t)\|_2 < c_N$ and $\mathbf{e}_{g,N}(t) = 0$ otherwise, where $c_N = 2(\ln(N))/\max\{1, \|\bar{\alpha}\|_{\ell_1}\}$, and we consider the data $\bar{\mathbf{y}}(t)$, $\bar{\mathbf{u}}(t)$ generated by the data generator with $\mathbf{e}_g$ replaced by $\mathbf{e}_{g,N}$ and the predicted values $\bar{\hat{\mathbf{y}}}_\theta(t)$, $\bar{\hat{\mathbf{y}}}_\theta(t \mid 0)$ generated by the predictor corresponding to the parameter θ, with $\mathbf{y}, \mathbf{u}$ replaced by the bounded values $\bar{\mathbf{y}}, \bar{\mathbf{u}}$.

Then we can reduce the unbounded case to the bounded one., by adding extra terms. The resulting bound is similar to the bounded case, leading to the following.

Theorem B.3 (Theorem 3.1: unbounded data, Lipschitz loss case)**.** *Let Assumptions 2.2, 2.3 and 2.1 hold, and assume that ℓ is L_ℓ-Lipschitz. For any a prior density π over Θ, confidence level $\delta \in (0, 0.5)$, integer $\epsilon \in \{-1, 1\}$, $N \geq 2$, and $0 < \lambda < \frac{N}{32L_\ell}$, it holds that*

$$\mathbf{P}\Big(\forall \rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}\mathcal{L}(\theta) \leq \epsilon E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\Big(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \frac{1}{2}\sum_{i=1}^{2}\ln E_{\theta\sim\pi}\exp\left(\mathcal{C}_{ub,i}(2\lambda, \theta, L_\ell, N)\right)\Big)\Big) \geq 1 - 2\delta$$

with $\mathrm{KL}(\rho\|\pi) = E_{\theta\sim\rho}\ln\frac{\rho(\theta)}{\pi(\theta)}$ *the Kullback–Leibler divergence, and*

$$\mathcal{C}_{ub,i}(\lambda, \theta, L, N) \triangleq \frac{1}{2}\mathcal{C}_{b,i}\left(2\lambda, \theta, L, \frac{2\ln(N)}{\|\alpha_g\|_{\ell_1}K}, N\right) + \bar{\mathcal{C}}_{ub}(\lambda, L, N)$$
$$\bar{\mathcal{C}}_{ub}(\lambda, L, N) \triangleq \frac{\lambda^2 32L^2C_{u,1}}{N} + \frac{4\lambda L C_{u,2}\ln(N)}{N}$$

where $\mathcal{C}_{b,i}(s, \theta, L, c)$ is defined as in Theorem B.1, and the constants, K,$C_{u,j}$, $j = 1, 2$ are defined as follows:

$$C_{u,1} \triangleq \bar{\Psi}_{\mathbf{e}_g}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1}) + \bar{\Psi}_{\mathbf{e}_g}(\|\alpha_g\|_{\ell_1}), \quad \bar{\Psi}_{\mathbf{e}_g}(s) = e^{14^2s^2\|\mathbf{e}_g\|^2_{\psi_2} + 3s\|\mathbf{e}_g\|_{\psi_2}}$$
$$C_{u,2} \triangleq \bar{\Psi}_{\mathbf{e}}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1})\frac{\|\bar{\alpha}\|_{\ell_1}}{K} + \bar{\Psi}_{\mathbf{e}_g}(\|\alpha_g\|_{\ell_1})\frac{1}{K}, \qquad K = \min\{1, \|\bar{\alpha}\|_{\ell_1}\}$$

In particular, for $\epsilon = 1$ the theorem states:

$$\mathbf{P}\Big(\forall \rho \in \mathcal{M}_\pi : E_{\theta\sim\rho}\mathcal{L}(\theta) \leq E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\Big(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \frac{1}{2}\sum_{i=1}^{2}\ln E_{\theta\sim\pi}\exp\left(\mathcal{C}_{ub,i}(2\lambda, \theta, L_\ell, N)\right)\Big)\Big) \geq 1 - 2\delta$$

and for $\epsilon = -1$ it states:

$$\mathbf{P}\Big(\forall \rho \in \mathcal{M}_\pi : E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) \leq E_{\theta\sim\rho}\mathcal{L}(\theta) + \frac{1}{\lambda}\Big(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \frac{1}{2}\sum_{i=1}^{2}\ln E_{\theta\sim\pi}\exp\left(\mathcal{C}_{ub,i}(2\lambda, \theta, L_\ell, N)\right)\Big)\Big) \geq 1 - 2\delta$$

and by using the union bound we get the following bound on the absolute value of the generalization gap:

$$\mathbf{P}\Big(\forall\rho\in\mathcal{M}_\pi:\Big|E_{\theta\sim\rho}\mathcal{L}(\theta)-E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta)\Big|\leq$$
$$\frac{1}{\lambda}\Big(\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}+\frac{1}{2}\sum_{i=1}^{2}\ln E_{\theta\sim\pi}\exp\left(\mathcal{C}_{ub,i}(2\lambda,\theta,L_\ell,N)\right)\Big)\Big)\geq 1-4\delta$$

Note that the constant $\mathcal{C}_i(\lambda,\theta,N,\delta,\epsilon)=\mathcal{C}_{ub,i}(\lambda,\theta,L(\theta),N)$, $i=1,2$ of Theorem 3.1 again does not depend on δ and ϵ.

Informally, the right-hand side of Theorem 3.1 is now $\frac{1}{\lambda}[KL(\rho\|\pi)+\ln\frac{1}{\delta}+\frac{1}{2}\ln E_{\theta\sim\rho}\exp(\frac{1}{2}C_{b,1}(4\lambda,\theta,c_N,N)+C_{ub}(2\lambda,L,N))+\frac{1}{2}\ln E_{\theta\sim\rho}\exp(\frac{1}{2}C_{b,2}(4\lambda,\theta,c_N,N)+C_{ub}(2\lambda,L,N))]$ where $C_{ub}(\lambda,N)=O(\lambda^2L^2\ln(N)/N)$ is an additional term accounting for approximating unbounded data with bounded one. This additional term represents an upper bound on the moment generating function for the difference between the empirical losses for the original and truncated data.

The bound of Corollary 3.2 can then be obtained from this bound by taking $\lambda=16L_\ell\sqrt{N}/\ln(N)$, leading to the bound

Corollary B.4 (Corollary 3.2: unbounded, Lipschitz loss case)**.** *For every prior π, $\delta(0,0.5)$, $\epsilon\in\{-1,1\}$,*

$$\mathbf{P}\Big(\forall\rho\in\mathcal{M}_\pi:\epsilon E_{\theta\sim\rho}\mathcal{L}(\theta)\leq\epsilon E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta)+\frac{16L_\ell\ln(N)}{\sqrt{N}}\Big(\mathrm{KL}(\rho||\pi)+\ln\frac{1}{\delta}+$$
$$C_{u,1}+C_{u,2}+\frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}(2,\bar{\alpha},G_e+2G_{e,1},1,\frac{1}{\|\alpha_g\|_{\ell_1}K})\Big)\Big)\geq 1-2\delta$$

That is, for $\epsilon=1$, it states that:

$$\mathbf{P}\Big(\forall\rho\in\mathcal{M}_\pi:E_{\theta\sim\rho}\mathcal{L}(\theta)\leq E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta)+\frac{16L_\ell\ln(N)}{\sqrt{N}}\Big(\mathrm{KL}(\rho||\pi)+\ln\frac{1}{\delta}+$$
$$C_{u,1}+C_{u,2}+\frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}(\frac{1}{4},\bar{\alpha},G_e+2G_{e,1},1,\frac{1}{\|\alpha_g\|_{\ell_1}K})\Big)\Big)\geq 1-2\delta$$

and for $\epsilon=-1$:

$$\mathbf{P}\Big(\forall\rho\in\mathcal{M}_\pi:E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta)\leq E_{\theta\sim\rho}\mathcal{L}(\theta)+\frac{16L_\ell\ln(N)}{\sqrt{N}}\Big(\mathrm{KL}(\rho||\pi)+\ln\frac{1}{\delta}+$$
$$C_{u,1}+C_{u,2}+\frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}(\frac{1}{4},\bar{\alpha},G_e+2G_{e,1},1,\frac{1}{\|\alpha_g\|_{\ell_1}K})\Big)\Big)\geq 1-2\delta$$

and with the union bound

$$\mathbf{P}\Big(\forall\rho\in\mathcal{M}_\pi:\Big|E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta)-E_{\theta\sim\rho}\mathcal{L}(\theta)\Big|\leq\frac{16L_\ell\ln(N)}{\sqrt{N}}\Big(\mathrm{KL}(\rho||\pi)+\ln\frac{1}{\delta}+$$
$$C_{u,1}+C_{u,2}+\frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}(\frac{1}{4},\bar{\alpha},G_e+2G_{e,1},1,\frac{1}{\|\alpha_g\|_{\ell_1}K})\Big)\Big)\geq 1-4\delta.$$

I.e., the upper bound on the generalization gap is of the form $O\left(\frac{\ln(N)}{\sqrt{N}}\right)(KL(\rho\|\pi)+\ln(\frac{1}{\delta})+O(1))$.

The additional terms in the bound suggest that the convergence guarantees for system identification are looser than in the bounded case. It also suggest that the learning problem for bounded data might be easier (require less data to achieve a good performance).

Unfortunately, the bound above has the drawback that it holds only for Lipschitz loss functions. However, for system identification, the preferred loss function is the quadratic loss $\|\hat{y}-y\|_2^2$, which is not Lipschitz. This motivates the last case:

Unbounded data and quadratic loss function We can show that we high probability (at least $1-\frac{2\delta}{3}$) the prediction errors $\|\hat{\mathbf{y}}_\theta(t \mid 0) - \mathbf{y}(t)\|_2$ and $\|\hat{\mathbf{y}}_\theta(t) - \mathbf{y}(t)\|_2$ are smaller than the constant $C_N = 6G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{\ln\left(\frac{3N^2}{2\delta}\right)} = O((\ln(N))^{1/2})$. When the prediction errors are smaller than C_N, we can replace the quadratic loss function in the empirical error with the truncated quadratic loss function $\ell_{C_N}(y', y)$, which is defined by $\ell_{C_N}(y', y) = \|y' - y\|_2^2$ if $\|y' - y\|_2 \leq C_N$ and $\ell_{C_N}(y', y) = C_N$ otherwise. This truncated loss is Lipschitz with the Lipschitz constant $L_{qd} = 12C_N$. We can then reuse the bound for L_{qd}-Lipschitz loss functions, stated for $\frac{\delta}{3}$ instead of δ in order to get a bound for the true error of the loss function ℓ_{C_N}. By using the union bound with the event that $\|\hat{\mathbf{y}}_\theta(t \mid 0) - \mathbf{y}(t)\|_2 \leq C_N$ and $\|\hat{\mathbf{y}}_\theta(t) - \mathbf{y}(t)\|_2 \leq C_N$, we can obtain a high probability (at least $1-2\delta$) bound on the true error for the loss function ℓ_{C_N} in terms of the empirical error for the original quadratic loss function. Finally, we can show that the true error for the quadratic loss functions is smaller than the true error for the loss function ℓ_{C_N} plus an additional terms which is $O(1/N)$. The bound of Theorem 3.1 then becomes as follows.

Theorem B.5 (Theorem 3.1: unbounded data, quadratic loss case)**.** *Let Assumptions 2.2 and 2.1 hold, and assume that ℓ is the quadratic loss function. For any a prior density π over Θ, confidence level $\delta \in (0, 0.5)$, integer $\epsilon \in \{-1, 1\}$, and $N \geq 3$, and $0 < \lambda < \frac{N}{32L_{qd}}$, the inequality*

$$\mathbf{P}\Big(\forall\rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}\mathcal{L}(\theta) \leq \epsilon E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\Big(\mathrm{KL}(\rho\|\pi) + \ln\left(\frac{2}{3\delta}\right) + \frac{1}{2}\sum_{i=1}^{2}\ln E_{\theta\sim\pi}\exp\left(\mathcal{C}_i(2\lambda, \theta, N, \delta, \epsilon)\right)\Big)\Big) \geq 1 - 2\delta$$

with $\mathrm{KL}(\rho\|\pi) = E_{\theta\sim\rho}\ln\frac{\rho(\theta)}{\pi(\theta)}$ the Kullback–Leibler divergence, and $\mathcal{C}_i(\lambda, \theta, N, \delta, \epsilon)$ is defined as

$$\begin{aligned}
\mathcal{C}_i(\lambda, \theta, N, \delta, 1) &\triangleq \frac{\lambda L_{qd}^2}{2N} + \mathcal{C}_{ub,i}\left(\lambda, \theta, L_{qd}, N\right)\\
\mathcal{C}_i(\lambda, \theta, N, \delta, -1) &\triangleq \mathcal{C}_{ub,i}\left(\lambda, \theta, L_{qd}, N\right)\\
L_{qd} &= 12G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{\ln\left(\frac{3N^2}{2\delta}\right)}
\end{aligned}$$

where $\mathcal{C}_{ub,i}(\lambda, \theta, L_{qd}, N)$ is defined in Theorem B.3.

In particular, for $\epsilon = 1$ the theorem states:

$$\mathbf{P}\Big(\forall\rho \in \mathcal{M}_\pi : E_{\theta\sim\rho}\mathcal{L}(\theta) \leq E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\Big(\mathrm{KL}(\rho\|\pi) + \ln\left(\frac{2}{3\delta}\right) + \frac{1}{2}\sum_{i=1}^{2}\ln E_{\theta\sim\pi}\exp\left(\mathcal{C}_{ub,i}(2\lambda, \theta, L_{qd}, N)\right) + \frac{\lambda L_{qd}^2}{2N}\Big)\Big) \geq 1 - 2\delta$$

and for $\epsilon = -1$ it states:

$$\mathbf{P}\Big(\forall\rho \in \mathcal{M}_\pi : E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) \leq E_{\theta\sim\rho}\mathcal{L}(\theta) + \frac{1}{\lambda}\Big(\mathrm{KL}(\rho\|\pi) + \ln\left(\frac{2}{3\delta}\right) + \frac{1}{2}\sum_{i=1}^{2}\ln E_{\theta\sim\pi}\exp\left(\mathcal{C}_{ub,i}(2\lambda, \theta, L_{qd}, N)\right)\Big)\Big) \geq 1 - 2\delta$$

and by using the union bound we get the following bound on the absolute value of the generalization gap:

$$\mathbf{P}\Big(\forall\rho \in \mathcal{M}_\pi : \Big|E_{\theta\sim\rho}\mathcal{L}(\theta) - E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta)\Big| \leq \frac{1}{\lambda}\Big(\mathrm{KL}(\rho\|\pi) + \ln\left(\frac{2}{3\delta}\right) + \frac{1}{2}\sum_{i=1}^{2}\ln E_{\theta\sim\pi}\exp\left(\mathcal{C}_{ub,i}(2\lambda, \theta, L_{qd}, N)\right) + \frac{\lambda L_{qd}^2}{2N}$$

That is, in this case the term $\mathcal{C}_i(\lambda, \theta, N, \delta, \epsilon)$ of Theorem 3.1 depends on δ via $L_g d$ and on ϵ. By right-hand side of the PAC-Bayesian inequality then becomes

$$\frac{1}{\lambda}\Big[KL(\rho\|\pi) + \ln\left(\frac{2}{3\delta}\right) + \frac{1}{2}\ln E_{\theta\sim\rho}\exp(\frac{1}{2}C_{b,1}(4\lambda, \theta, c_N, N) + C_{ub}(2\lambda, L_{qd}, N) + \frac{\lambda L_{qd}^2}{2N}) + \frac{1}{2}\ln E_{\theta\sim\rho}\exp(\frac{1}{2}C_{b,2}(4\lambda, \theta, c_N, N) + C_{ub}(2\lambda, L_{qd}, N) + \frac{\lambda L_{qd}^2}{2N})\Big]$$

where $c_N = \frac{2\ln(N)}{\|\alpha_g\|_{\ell_1} K}$ as in the case of unbounded data and Lipschitz loss. If we choose $\lambda = \frac{\sqrt{N}}{16\ln(N)L_{qd}}) = \frac{\sqrt{N}}{384\sqrt{2}G_e\|\mathbf{e}_g\|_\psi\sqrt{(1+\frac{2}{3}\delta)}\ln(N))^{3/2}}$ we get the bound Corollary 3.2:

Corollary B.6 (Corollary 3.2: unbounded, quadratic loss case)**.** *For every prior* π, $\delta(0, 0.5)$, $\epsilon \in \{-1, 1\}$, $L = 192G_e\|\mathbf{e}_g\|_\psi\sqrt{\ln\left(1+\frac{2}{3\delta}\right)}$,

$$\mathbf{P}\Big(\forall\rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}\mathcal{L}(\theta) \leq \epsilon E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{L(\ln(N))^{3/2}}{\sqrt{N}}\Big(\mathrm{KL}(\rho||\pi) + \ln\left(\frac{2}{3\delta}\right) + \frac{\max\{0,\epsilon\}}{128} + C_{u,1} + C_{u,2} + \frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}(2, \bar{\alpha}, G_e + 2G_{e,1}, 1, \frac{1}{\|\alpha_g\|_{\ell_1}K})\Big)\Big) \geq 1 - 2\delta$$

That is, for $\epsilon = 1$, it states that:

$$\mathbf{P}\Big(\forall\rho \in \mathcal{M}_\pi : E_{\theta\sim\rho}\mathcal{L}(\theta) \leq E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{L(\ln(N))^{3/2}}{\sqrt{N}}\Big(\mathrm{KL}(\rho||\pi) + \ln\left(\frac{2}{3\delta}\right) + \frac{1}{128} + C_{u,1} + C_{u,2} + \frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}\left(\frac{1}{4}, \bar{\alpha}, G_e + 2G_{e,1}, 1, \frac{1}{\|\alpha_g\|_{\ell_1}K}\right)\Big)\Big) \geq 1 - 2\delta$$

and for $\epsilon = -1$:

$$\mathbf{P}\Big(\forall\rho \in \mathcal{M}_\pi : E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) \leq E_{\theta\sim\rho}\mathcal{L}(\theta) + \frac{L(\ln(N))^{3/2}}{\sqrt{N}}\Big(\mathrm{KL}(\rho||\pi) + \ln\left(\frac{2}{3\delta}\right) + C_{u,1} + C_{u,2} + \frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}(\frac{1}{4}, \bar{\alpha}, G_e + 2G_{e,1}, 1, \frac{1}{\|\alpha_g\|_{\ell_1}K})\Big)\Big) \geq 1 - 2\delta$$

and with the union bound

$$\mathbf{P}\Big(\forall\rho \in \mathcal{M}_\pi : \Big|E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) - E_{\theta\sim\rho}\mathcal{L}(\theta)\Big| \leq \frac{L(\ln(N))^{3/2}}{\sqrt{N}}\Big(\mathrm{KL}(\rho||\pi) + \ln\left(\frac{2}{3\delta}\right) + \frac{1}{128} + C_{u,1} + C_{u,2} + \frac{1}{2}\sum_{i=1}^{2}\mathcal{C}_{m,b,i}(\frac{1}{4}, \bar{\alpha}, G_e + 2G_{e,1}, 1, \frac{1}{\|\alpha_g\|_{\ell_1}K})\Big)\Big) \geq 1 - 4\delta.$$

That is the upper bound on the generalization gap is of the form

$$O\left(\sqrt{\ln\left(1+\frac{2}{3\delta}\right)}\right)O\left(\frac{\ln(N)^{3/2}}{\sqrt{N}}\right)\left(KL(\rho\|\pi) + \ln\left(\frac{2}{3\delta}\right) + O(1)\right)$$

The presence of addition terms indicate that the bound for quadratic loss and unbounded data is larger than for the case of Lipschitz loss or for bounded data.

To sum up, the bound for each case was derived from the previous case by adding an extra term. In all cases, the bound depends only on the prior and the distance of the posterior from the prior, and the bound of Corollary 3.2 is an upper bound of the bound of Theorem 3.1 for a certain choice of λ as an increasing function of N.

Additional intuition behind Gibbs posteriors and Corollary 3.4 We can find the posterior which minimizes the bound of Theorem 3.1 or Corollary 3.2, which is the so called Gibbs posterior, and we can think of learning as taking the average (MEP), taking the most likely model (MAP) or taking a random model from this Gibb posterior. For all these cases we can derive a bound on the generalization gap by exploiting a closed-form formula for the KL divergence if the prior is uniform or Gaussian, and the posterior is the Gibbs posterior. Moreover, using the fact that the bounds of Theorem 3.1/Corollary 3.2 hold for both the generalization gap and its multiple by -1, using the union bound, we can derive a bound on the difference between the true error of the learned model and the smallest achievable true error. For quadratic loss, under the assumption that the data generator is included among the predictors, the latter can be related to the parameter estimation error. These results are collected in Corollary 3.4.

C Proofs of Section 2

Remark C.1 (LTI in innovation form)**.** *Let us assume that $\mathbf{e}_g$ has zero mean. It follows that $\mathbf{e}_g(t)$ is the innovation process of $[\mathbf{y}^T(t)\ \mathbf{u}^T(t)]^T$ and* (4) *is in the so called (forward) innovation form, see [22, Section 8.5.2]. Intuitively, it means that $\mathbf{e}_g$ is the difference between $[\mathbf{y}^T(t)\ \mathbf{u}^T(t)]^T$ and its best linear prediction based on its own past values.*

Remark C.2 (Equivalence of (4) and (2))**.** *Let us assume that $\mathbf{e}_g$ and $\mathbf{e}_s$ are both zero mean. More precisely, if there is no feedback from $\mathbf{y}$ to $\mathbf{u}$ according to Definition 17.1.1. of [25], then* (2) *holds and by [10] the matrices A_0, B_0, C_0, K_0, D_0 can be computed from the matrices A_g, K_g, C_g.*

Conversely, if (2) *holds and $\mathbf{u}(t)$ is generated by an ARMA process, and there is no feedback from $\mathbf{y}(t)$ to $\mathbf{u}(t)$, then Assumption 2.2 holds and A_g, K_g, C_g can be computed from those of* (2) *and the ARMA representation of $\mathbf{u}(t)$, again by reversing the equations from [10]. As a passing remark, if there is a feedback from $\mathbf{y}(t)$ to $\mathbf{u}(t)$ and $\mathbf{u}(t)$ is generated by an LTI system driven by $\mathbf{y}(t)$, i.e.,*

$$\mathbf{s}_\mathbf{u}(t+1) = A_u \mathbf{s}_\mathbf{u}(t) + K_u \mathbf{y}(t), \quad \mathbf{u}(t) = C_u \mathbf{s}_\mathbf{u}(t) \tag{15}$$

then again the generator described in Assumption 2.2 can be obtained by adding (15) *to* (2)*.*

Remark C.3 (Methods for stable parametrisations)**.** *The construction of [33, Section 3.2.3] relies on canonical forms based on structure indices, to which the diffeomorphism from [18] between general and stable parametersations are applied. Intuitively, the latter transformation maps (A, B, C, D) to $(\frac{A}{\mu+\epsilon}, B, C, D)$, where μ is the maximum of moduli of the eigenvalues of A and $\epsilon > 0$. The construction of [28, Section 4.6] uses Ober's balanced canonical forms, and it relies on fixing the observability and controllability Grammians to being equal and diagonal. That is, stability is assured by using diagonal quadratic Lyapunov functions. In this case, by fixing upper bounds on the diagonals of the Grammians and on the norms of the system matrices occurring in the parameterization, we ensure existence of constants M, C, γ mentioned in Assumption 2.1.*

Proof of Lemma 2.4. We have $\mathbf{y}_\theta(t\,|\,t_0) = \sum_{k=t_0}^{t-1} H_\theta(t-1-k)\,\mathbf{w}(k) + D_\theta \mathbf{w}(t) = \sum_{k=0}^{t-1-t_0} H_\theta(k)\,\mathbf{w}(t-1-k) + D_\theta \mathbf{w}(t)$ with the impulse response $H_\theta(k) \triangleq C_\theta A_\theta^k B_\theta$ for $k \geq 0$. Because A_θ is Schur, there exist $M > 0$ and $\rho \in (0,1)$ such that $\|A_\theta^k\|_2 \leq M\rho^k$, hence $\sum_{k=s}^{\infty} \|H_\theta(k)\|_2 \leq \|C_\theta\|_2 \|B_\theta\|_2 M \frac{\rho^s}{1-\rho} \to 0$ as $s \to \infty$. For $t_1 < t_2 \leq t$ we get, using stationarity of $\mathbf{w}$ ($\sigma_\mathbf{w} = (\mathbf{E}\|\mathbf{w}(s)\|_2^2)^{1/2}$), that

$$(\mathbf{E}[\|\mathbf{y}_\theta(t\,|\,t_2) - \mathbf{y}_\theta(t\,|\,t_1)\|_2^2])^{1/2} \leq \sum_{m=t-t_2}^{t-1-t_1} (\mathbf{E}\left[\|H_\theta(m)\mathbf{w}(t-1-m)\|_2^2\right])^{1/2} \tag{16a}$$

$$\leq \sigma_\mathbf{w} \sum_{m=t-t_2}^{t-1-t_1} \|H_\theta(m)\|_2 \tag{16b}$$

$$\leq \sigma_\mathbf{w} \sum_{m=t-t_2}^{\infty} \|H_\theta(m)\|_2 \to 0 \quad \text{as } t_2 \to -\infty \tag{16c}$$

Therefore $\{\mathbf{y}_\theta(t\,|\,t_0)\}_{t_0 \leq t}$ is Cauchy (in L^2) and hence converges in mean square to $\mathbf{y}_\theta(t) = \sum_{k=-\infty}^{t-1} H_\theta(t-1-k)\,\mathbf{w}(k) + D_\theta \mathbf{w}(t) = \sum_{k=0}^{\infty} H_\theta(k)\,\mathbf{w}(t-1-k) + D_\theta \mathbf{w}(t)$. This representation shows that $\mathbf{y}_\theta$ is the output of a causal, time-invariant, stable LTI filter (with impulse response $\{H_\theta(k)\}$ and feed-through D_θ) driven by the stationary input $\mathbf{w}$. Hence $\mathbf{y}_\theta$ is stationary, since stable LTI filtering preserves stationarity. Moreover, it follows that

$$\mathbf{y}_\theta(t) - \mathbf{y}_\theta(t\mid t_0) = \sum_{k=t-t_0}^{\infty} H_\theta(k)\,\mathbf{w}(t-1-k). \tag{17}$$

so $\mathbf{y}_\theta(t) - \mathbf{y}_\theta(t\,|\,0) \to 0$ in mean square as $t \to \infty$ by arguments as in (16).

Finally, consider

$$\left|\mathbf{E}\big[\ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t)\big)\big] - \mathbf{E}\big[\ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t\,|\,t_0)\big)\big]\right| \leq \mathbf{E}\big[\,\big|\ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t)\big) - \ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t\,|\,t_0)\big)\big|\,\big] \tag{18}$$

If ℓ is L_ℓ-Lipschitz, then the right hand side of (18) is bounded by $L_\ell \mathbf{E}[\|\mathbf{y}_\theta(t) - \mathbf{y}_\theta(t \mid t_0)\|_2]$, and hence, as $\mathbf{y}_\theta(t) - \mathbf{y}_\theta(t \,|\, 0) \to 0$ in mean square as $t \to \infty$ or as $t_0 \to -\infty$, the right hand side of (18) also tends to 0, i.e.,

$$\mathbf{E}\big[\,\big|\ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t)\big) - \ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t \,|\, t_0)\big)\big|\,\big] \to 0.$$

as $t \to \infty$ or $t_0 \to -\infty$.

If ℓ is quadratic, then

$$\begin{aligned}
&|\ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t)\big) - \ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t \,|\, t_0)\big)| = \big|\|\mathbf{y}(t) - \mathbf{y}_\theta(t)\|_2^2 - \|\mathbf{y}(t) - \mathbf{y}_\theta(t \,|\, t_0)\|_2^2\big| = \\
&|2\mathbf{y}^T(t)(-\mathbf{y}_\theta(t) + \mathbf{y}_\theta(t \,|\, t_0)) + \big(\|\mathbf{y}_\theta(t)\|_2^2 - \|\mathbf{y}_\theta(t \,|\, t_0)\|_2^2\big)\,| \leq \\
&2\|\mathbf{y}^T(t)\|_2 \|\mathbf{y}_\theta(t) - \mathbf{y}_\theta(t \,|\, t_0)\|_2 + |\|\mathbf{y}_\theta(t)\|_2 - \|\mathbf{y}_\theta(t \,|\, t_0)\|_2|\,\big(\|\mathbf{y}_\theta(t)\|_2 + \|\mathbf{y}_\theta(t \,|\, t_0)\|_2\big)
\end{aligned}$$

Hence, using the Cauchy-Schwartz inequality

$$\begin{aligned}
&\mathbf{E}\left[|\ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t)\big) - \ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t \,|\, t_0)\big)|\right] \leq \\
&2\mathbf{E}\left[\|\mathbf{y}^T(t)\|_2 \|\mathbf{y}_\theta(t) - \mathbf{y}_\theta(t \,|\, t_0)\|_2\right] + \mathbf{E}\left[|\|\mathbf{y}_\theta(t)\|_2 - \|\mathbf{y}_\theta(t \,|\, t_0)\|_2|\,\big(\|\mathbf{y}_\theta(t)\|_2 + \|\mathbf{y}_\theta(t \,|\, t_0)\|_2\big)\right] \leq \\
&2\sqrt{\mathbf{E}\left[\|\mathbf{y}_\theta(t) - \mathbf{y}_\theta(t \mid t_0)\|_2^2\right]}\sqrt{\mathbf{E}\left[\|\mathbf{y}(t)\|_2^2\right]} + \\
&\sqrt{\mathbf{E}\left[\|\mathbf{y}_\theta(t) - \mathbf{y}_\theta(t \,|\, t_0)\|_2^2\right)]}\sqrt{\big(\mathbf{E}[\|\mathbf{y}_\theta(t)\|_2^2] + 2\mathbf{E}[\|\mathbf{y}_\theta(t)\|_2 \|\mathbf{y}_\theta(t \,|\, t_0)\|_2] + \mathbf{E}[\|\mathbf{y}_\theta(t \,|\, t_0)\|_2^2]\big)}
\end{aligned}$$

Note that $\mathbf{E}\left[\|\mathbf{y}_\theta(t) - \mathbf{y}_\theta(t \mid t_0)\|_2^2\right] \to 0$ as $t_0 \to -\infty$ or $t \to \infty$, since $\mathbf{y}_\theta(t \mid t_0) - \mathbf{y}_\theta(t \mid t_0) \to 0$ in the mean-square sense as $t_0 \to -\infty$ or $t \to \infty$. Moreover,

$$2\mathbf{E}[\|\mathbf{y}_\theta(t)\|_2 \|\mathbf{y}_\theta(t \,|\, t_0)\|_2] \leq \mathbf{E}[\|\mathbf{y}_\theta(t)\|_2^2] + \mathbf{E}[\|\mathbf{y}_\theta(t \,|\, t_0)\|_2^2],$$

and $\mathbf{E}[\|\mathbf{y}_\theta(t)\|_2^2]$ is constant due to stationarity of $\mathbf{y}_\theta(t)$. Now for sufficiently large $t > T_\epsilon$ or $-t_0 > T_\epsilon$, $\mathbf{E}\left[\|\mathbf{y}_\theta(t) - \mathbf{y}_\theta(t \mid t_0)\|_2^2\right] < \epsilon^2$, and it follows that

$$\begin{aligned}
\mathbf{E}[\|\mathbf{y}_\theta(t \mid t_0)\|_2^2] &\leq \mathbf{E}[(\|\mathbf{y}_\theta(t)\|_2 + \|\mathbf{y}_\theta(t \mid t_0) - \mathbf{y}_\theta(t)\|_2)^2] \\
&\leq 2(\mathbf{E}[\|\mathbf{y}_\theta(t)\|_2^2] + \mathbf{E}[\|\mathbf{y}_\theta(t) - \mathbf{y}_\theta(t \mid t_0)\|_2^2]) < 2\mathbf{E}[\|\mathbf{y}_\theta(t)\|_2^2] + 2\epsilon^2
\end{aligned}$$

Hence, it follows that for $t > T_\epsilon$ or $-t_0 > T_\epsilon$,

$$\sqrt{\big(\mathbf{E}[\|\mathbf{y}_\theta(t)\|_2^2] + 2\mathbf{E}[\|\mathbf{y}_\theta(t)\|_2]\|\mathbf{y}_\theta(t \,|\, t_0)\|_2] + \mathbf{E}[\|\mathbf{y}_\theta(t \,|\, t_0)\|_2^2]\big)} \leq \sqrt{6\mathbf{E}[\|\mathbf{y}_\theta(t)\|_2^2] + 4\epsilon^2}$$

and $\sqrt{6\mathbf{E}[\|\mathbf{y}_\theta(t)\|_2^2] + 4\epsilon^2} = C$ does not depend on t. Hence, if $t > T_\epsilon$ or $-t_0 > T_\epsilon$, then

$$\mathbf{E}\big[\,\big|\ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t)\big) - \ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t \,|\, t_0)\big)\big|\,\big] \leq \epsilon(2\mathbf{E}[\|\mathbf{y}(t)\|_2^2] + C)$$

from which it follows that $\mathbf{E}\big[\,\big|\ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t)\big) - \ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t \,|\, t_0)\big)\big|\,\big] \to 0$ as either $t_0 \to -\infty$ or $t \to \infty$.

That is, in all cases,

$$\lim_{t\to\infty} E\big[\ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t \,|\, 0)\big)\big] = E\big[\ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t)\big)\big] = \lim_{t_0\to-\infty} E\big[\ell\big(\mathbf{y}(t), \mathbf{y}_\theta(t \,|\, t_0)\big)\big] \tag{19}$$

proving the lemma since $\mathcal{L}(\theta)$ is the right hand side of (19) by definition. □

D Proofs of Section 2.3

Remark D.1 (α as Markov-parameters)**.** *If α is the sequence of Markov parameters of a stable deterministic LTI system determined by matrices (A, B, C, D), i.e., $\alpha(0) = D$, $\alpha(k) = CA^{k-1}B$, $k \geq 1$, then the ℓ_1 norm $\|\alpha\|_{\ell_1}$ and the WDMC $\theta_\infty(\alpha)$ are both finite. In this case, $\|\alpha\|_{\ell_1}$ is the classical ℓ_1 norm of the underlying LTI system. Moreover, if $M > 0$, $\gamma \in [0, 1)$ are such that $\|A^k\|_2 \leq M\gamma^k$, then $\|\alpha\|_{\ell_1} \leq \|D\|_2 + M\|B\|_2\|C\|_2/(1-\gamma)$ and $\theta_\infty(\alpha) \leq M\|B\|_2\|C\|_2/(1-\gamma)^2$. In turn, such M always exist if A is a Schur matrix, and γ can be chosen arbitrarily close to $\rho(A)$ from above, while remaining below 1. In particular, the upper bounds decrease as γ decreases, provided M, $\|B\|_2$, and $\|C\|_2$ remain fixed.*

Finally, as promised before Lemma D.3, we can state that all the constants from Definition 2.6 are well-defined.

Lemma D.2. *If Assumption 2.2 and 2.1 hold, then the quantities $\|\alpha_\theta\|_{\ell_1}$ and $\theta_\infty(\alpha_\theta)$, $\|\bar{\alpha}\|_{\ell_1}$, $\theta_\infty(\bar{\alpha})$, $G_e(\theta)$, $G_{e,1}(\theta)$, G_e, $G_{e,1}$ are well defined real numbers.*

Proof. From Remark D.1 and the assumption that A_g is Schur it follows that $\|\alpha_g\|_{\ell_1}$ and $\theta_\infty(\alpha_g)$ are well-defined. By Assumption 2.1 there exists constants M, γ, C such that for all $\theta \in \Theta$, $\|A_\theta^k\|_2 \leq M\gamma^k$, $\|B_\theta\|_2 \leq C$, and $\|C_\theta\|_2 \leq C$. Therefore, $\bar{\alpha}(k) = \sup_\theta \|\alpha_\theta(k)\|_2 \leq M\gamma^{k-1}C^2$ and hence, by using the argument of Remark D.1, $\|\alpha_\theta\|_{\ell_1} \leq \sum_{k=0}^\infty \|\alpha\|_{\ell_1} \leq \sum_{k=0}^\infty \bar{\alpha}(k) = \|\bar{\alpha}\|_{\ell_1} \leq C + \frac{MC^2}{1-\gamma} < \infty$ and $\theta_\infty(\alpha_\theta) = \sum_{k=0}^\infty k\|\alpha_\theta(k)\|_2 \leq \sum_{k=0}^\infty k\bar{\alpha}(k) = \theta_\infty(\bar{\alpha}) \leq \frac{MC^2}{(1-\gamma)^2} < \infty$. Therefore, the quantities $\|\alpha_\theta\|_{\ell_1}$ and $\theta_\infty(\alpha_\theta)$, $\|\bar{\alpha}\|_{\ell_1}$, $\theta_\infty(\bar{\alpha})$ defining $G_e(\theta)$, $G_{e,1}(\theta)$, G_e, $G_{e,1}$ are well-defined, and hence so are $G_e(\theta)$, $G_{e,1}(\theta)$, G_e, $G_{e,1}$ □

Lemma D.3. *Let $\alpha_{g,y}(k)$ and $\alpha_{g,w}(k)$ be the matrix formed by the first $n_{\mathbf{y}}$ and the last $n_{\mathbf{w}}$ rows of $\alpha_g(k)$ respectively, where $n_{\mathbf{w}}$ is the number of entries of $\mathbf{w}(t)$. Let $c_{\theta,s}(k) = \left(\sum_{l=0}^{\min\{k,s\}} \alpha_\theta(l)\alpha_{g,w}(k-l)\right)$ and $c_\theta(k) = \sum_{l=0}^k \alpha_\theta(l)\alpha_{g,w}(k-l)$. Then*

$$\begin{aligned}
&\begin{bmatrix}\mathbf{y}(t)\\ \hat{\mathbf{y}}_\theta(t)\end{bmatrix} = \left(\begin{bmatrix}\alpha_{g,y}\\ c_\theta\end{bmatrix} \star \mathbf{e}_g\right)(t), \quad \begin{bmatrix}\mathbf{y}(t)\\ \hat{\mathbf{y}}_\theta(t \mid t_0)\end{bmatrix} = \left(\begin{bmatrix}\alpha_{g,y}\\ c_{\theta,t-t_0}\end{bmatrix} \star \mathbf{e}_g\right)(t), \\
&\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t) = ((\alpha_{g,y} - c_\theta) \star \mathbf{e}_g)(t), \quad \mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t \mid t_0) = ((\alpha_{g,y} - c_{\theta,t-t_0}) \star \mathbf{e}_g)(t)
\end{aligned} \tag{20}$$

Moreover, the constants from Definition 2.6 satisfy:

$$\max\{\|c_\theta\|_{\ell_1}, \|c_{\theta,s}\|_{\ell_1}\} \leq \|\alpha_\theta\|_{\ell_1}\|\alpha_g\|_{\ell_1}, \tag{21}$$

$$\max\left\{\left\|\begin{bmatrix}\alpha_{g,y}\\ c_\theta\end{bmatrix}\right\|_{\ell_1}, \left\|\begin{bmatrix}\alpha_{g,y}\\ c_{\theta,t-t_0}\end{bmatrix}\right\|_{\ell_1}, \|\alpha_{g,y} - c_\theta\|_{\ell_1}, \|\alpha_{g,y} - c_{\theta,t-t_0}\|_{\ell_1}\right\} \leq G_e(\theta) \leq G_e \tag{22}$$

$$\max\left\{\theta_\infty\left(\begin{bmatrix}\alpha_{g,y}\\ c_\theta\end{bmatrix}\right), \theta_\infty\left(\begin{bmatrix}\alpha_{g,y}\\ c_{\theta,t-t_0}\end{bmatrix}\right), \theta_\infty(\alpha_{g,y} - c_\theta), \theta_\infty(\alpha_{g,y} - c_{\theta,t-t_0})\right\} \leq G_{e,1}(\theta) \tag{23}$$

$$\sum_{t=t_0}^{N+t_0} \|c_{\theta,t-t_0} - c_\theta\|_{\ell_1} \leq \theta_\infty(\alpha_\theta)\|\alpha_g\|_{\ell_1} \tag{24}$$

In particular, $[\mathbf{y}^T(t) \;\; \hat{\mathbf{y}}_\theta^T(t)]^T$ and $\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)$ from (20) *are stationary.*

Proof of Lemma D.3. Let C_w denote the matrix formed by the last n_w rows of C_g. From Assumption 2.2 and the proof of Lemma 2.4 we get that

$$\begin{aligned}
\mathbf{w}(t) &= \sum_{k=1}^\infty C_w A_g^{k-1} K_g \mathbf{e}_g(t-k) + \alpha_{g,w}(0)\mathbf{e}_g(t) = \sum_{k=0}^\infty \alpha_{g,w}(k)\mathbf{e}_g(t-k) = (\alpha_{g,w} \star \mathbf{e}_g)(t) \\
\mathbf{y}(t) &= \sum_{k=0}^\infty \alpha_{g,y}(k)\mathbf{e}_g(t-k) \\
\hat{\mathbf{y}}_\theta(t \mid t_0) &= \sum_{k=1}^{t-t_0} \hat{C}_\theta \hat{A}_\theta^{k-1} \hat{B}_\theta \mathbf{w}(t-k) + \hat{D}_\theta \mathbf{w}(t) = \sum_{k=0}^{t-t_0} \alpha_\theta(k)\mathbf{w}(t-k) \\
\hat{\mathbf{y}}_\theta(t) &= \lim_{t_0 \to -\infty} \hat{\mathbf{y}}_\theta(t \mid t_0) = \sum_{k=0}^\infty \alpha_\theta(k)\mathbf{w}(t-k)
\end{aligned}$$

It then follows using the new index $n = k + j \geq 0$ and the standard argument involving Cauchy products of series,

$$\begin{aligned}\hat{\mathbf{y}}_\theta(t \mid t_0) &= \sum_{k=0}^{t-t_0} \alpha_\theta(k) \sum_{j=0}^{\infty} \alpha_{g,w}(j)\mathbf{e}_g(t-k-j) \\ &= \sum_{n=0}^{\infty} \left(\sum_{k=0}^{\min(n,t-t_0)} \alpha_\theta(k)\alpha_{g,w}(n-k) \right) \mathbf{e}_g(t-n) \\ &= \sum_{n=0}^{\infty} c_{\theta,t-t_0}(n)\mathbf{e}_g(t-n) = (c_{\theta,t-t_0} \star \mathbf{e}_g)(t)\end{aligned}$$

Notice that since A_g and $\hat{A}_\theta$ are Schur, the series $\sum_{k=0}^{\infty} \alpha_\theta(k)$ and $\sum_{k=0}^{\infty} \alpha_{g,w}(k)$ are absolutely convergent. With a similar argument, using that absolutely convergent series can be rearranged, it follows that

$$\begin{aligned}\hat{\mathbf{y}}_\theta(t) &= \sum_{k=0}^{\infty} \alpha_\theta(k) \sum_{j=0}^{\infty} \alpha_{g,w}(j)\mathbf{e}_g(t-k-j) \\ &= \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} \alpha_\theta(k)\alpha_{g,w}(n-k) \right) \mathbf{e}_g(t-n) \\ &= \sum_{n=0}^{\infty} c_\theta(n)\mathbf{e}_g(t-n) = (c_\theta \star \mathbf{e}_g)(t)\end{aligned}$$

From this the first claim of the lemma, i.e., (20), follows. The first inequality of the second claim, i.e., (21), follows by first noticing that

$$\begin{aligned}&\|c_\theta(k)\|_2 \leq \sum_{i=0}^{k} \|\alpha_\theta(i)\|_2 \|\alpha_{g,w}(k-i)\|_2, \\ &\|c_{\theta,t-t_0}(k)\|_2 \leq \sum_{i=0}^{\min(k,t-t_0)} \|\alpha_\theta(i)\|_2 \|\alpha_{g,w}(k-i)\|_2 \leq \sum_{i=0}^{k} \|\alpha_\theta(i)\|_2 \|\alpha_{g,w}(k-i)\|_2\end{aligned}$$

Since $\sum_{i=0}^{k} \|\alpha_\theta(i)\|_2 \|\alpha_{g,w}(k-i)\|_2$ is the k'th coefficient of the Cauchy-product of $\sum_{k=0}^{\infty} \|\alpha_\theta(k)\|_2$ and $\sum_{k=0}^{\infty} \|\alpha_{g,w}(k)\|_2$, we obtain (21) by

$$\|c\|_{\ell_1} = \sum_{k=0}^{\infty} \|c(k)\|_2 \leq \sum_{i=0}^{\infty} \|\alpha_\theta(i)\|_2 \sum_{j=0}^{\infty} \|\alpha_{g,w}(j)\|_2 = \|\alpha_\theta\|_{\ell_1} \|\alpha_{g,w}\|_{\ell_1} \leq \|\alpha_\theta\|_{\ell_1} \|\alpha_g\|_{\ell_1} \quad (25)$$

with c denoting either c_θ or $c_{\theta,t-t_0}$. Concerning the second inequality (22), notice that

$$\left\| \begin{bmatrix} \alpha_{g,y}(k) \\ c_\theta(k) \end{bmatrix} \right\|_2 \leq \|\alpha_{g,y}(k)\|_2 + \|c_\theta(k)\|_2, \quad \|\alpha_{g,y}(k) - c_\theta(k)\|_2 \leq \|\alpha_{g,y}(k)\|_2 + \|c_\theta(k)\|_2 \quad (26a)$$

$$\left\| \begin{bmatrix} \alpha_{g,y}(k) \\ c_{\theta,t-t_0}(k) \end{bmatrix} \right\|_2 \leq \|\alpha_{g,y}(k)\|_2 + \|c_{\theta,t-t_0}(k)\|_2, \quad (26b)$$

$$\|\alpha_{g,y}(k) - c_{\theta,t-t_0}(k)\|_2 \leq \|\alpha_{g,y}(k)\|_2 + \|c_{\theta,t-t_0}(k)\|_2 \quad (26c)$$

and hence,

$$\max \left\{ \left\| \begin{bmatrix} \alpha_{g,y} \\ c_\theta \end{bmatrix} \right\|_{\ell_1}, \|\alpha_{g,y} - c_\theta\|_{\ell_1} \right\} \leq \|\alpha_{g,y}\|_{\ell_1} + \|c_\theta\|_{\ell_1} \quad (27a)$$

$$\max \left\{ \left\| \begin{bmatrix} \alpha_{g,y} \\ c_{\theta,t-t_0} \end{bmatrix} \right\|_{\ell_1}, \|\alpha_{g,y} - c_{\theta,t-t_0}\|_{\ell_1} \right\} \leq \|\alpha_{g,y}\|_{\ell_1} + \|c_{\theta,t-t_0}\|_{\ell_1} \quad (27b)$$

From (27) and (25), the inequality of (22) follows by using the definition of $G_e(\theta)$ and the fact that $\max\{\|\alpha_{g,y}\|_{\ell_1}, \|\alpha_{g,w}\|_{\ell_1}\} \leq \|\alpha_g\|_{\ell_1}$.

Concerning (23), notice that by defininig $d_1(i) = i\|\alpha_\theta(i)\|_2$ and $d_2(i) = i\|\alpha_{g,w}(i)\|_2$,

$$\theta_\infty(c_{\theta,t-t_0}) = \sum_{k=0}^{\infty} k\|c_{\theta,t-t_0}(k)\|_2 \leq \sum_{k=0}^{\infty} k \sum_{i=0}^{\min(k,t-t_0)} \|\alpha_\theta(i)\|_2\|\alpha_{g,w}(k-i)\|_2 =$$
$$\sum_{k=0}^{\infty} \sum_{i=0}^{\min(k,t-t_0)} i\|\alpha_\theta(i)\|_2\|\alpha_{g,w}(k-i)\|_2 + \sum_{k=0}^{\infty} \sum_{i=0}^{\min(k,t-t_0)} (k-i)\|\alpha_\theta(i)\|_2\|\alpha_{g,w}(k-i)\|_2 =$$
$$\sum_{k=0}^{\infty} \sum_{i=0}^{\min(k,t-t_0)} d_1(i)\|\alpha_{g,w}(k-i)\|_2 + \sum_{k=0}^{\infty} \sum_{i=0}^{\min(k,t-t_0)} d_2(k-i)\|\alpha_\theta(i)\|_2 \leq$$
$$\sum_{k=0}^{\infty}\sum_{i=0}^{k} d_1(i)\|\alpha_{g,w}(k-i)\|_2 + \sum_{k=0}^{\infty}\sum_{i=0}^{k} d_2(k-i)\|\alpha_\theta(i)\|_2$$

Similarly,

$$\theta_\infty(c_\theta) = \sum_{k=0}^{\infty} k\|c_\theta(k)\|_2 \leq \sum_{k=0}^{\infty} k \sum_{i=0}^{k} \|\alpha_\theta(i)\|_2\|\alpha_{g,w}(k-i)\|_2 =$$
$$\sum_{k=0}^{\infty}\sum_{i=0}^{k} i\|\alpha_\theta(i)\|_2\|\alpha_{g,w}(k-i)\|_2 + \sum_{k=0}^{\infty}\sum_{i=0}^{k} (k-i)\|\alpha_\theta(i)\|_2\|\alpha_{g,w}(k-i)\|_2 =$$
$$\sum_{k=0}^{\infty}\sum_{i=0}^{k} d_1(i)\|\alpha_{g,w}(k-i)\|_2 + \sum_{k=0}^{\infty}\sum_{i=0}^{k} d_2(k-i)\|\alpha_\theta(i)\|_2 \leq$$
$$\left(\sum_{k=0}^{\infty} d_1(k)\right)\left(\sum_{k=0}^{\infty}\|\alpha_{g,w}(k)\|_2\right) + \left(\sum_{k=0}^{\infty} d_2(k)\right)\left(\sum_{k=0}^{\infty}\|\alpha_\theta(k)\|_2\right) =$$
$$\theta_\infty(\alpha_\theta)\|\alpha_{g,w}\|_{\ell_1} + \theta_\infty(\alpha_{g,w})\|\alpha_\theta\|_{\ell_1} \leq \theta_\infty(\alpha_\theta)\|\alpha_g\|_{\ell_1} + \theta_\infty(\alpha_g)\|\alpha_\theta\|_{\ell_1}$$

In particular,

$$\max\{\theta_\infty(c_\theta), \theta_\infty(c_{\theta,t-t_0})\} \leq \theta_\infty(\alpha_\theta)\|\alpha_g\|_{\ell_1} + \theta_\infty(\alpha_g)\|\alpha_\theta\|_{\ell_1} \tag{28}$$

Moreover, using (26) it follows that

$$\max\left\{\theta_\infty\left(\begin{bmatrix}\alpha_{g,y}\\ c_\theta\end{bmatrix}\right), \theta_\infty\left(\begin{bmatrix}\alpha_{g,y}\\ c_{\theta,t-t_0}\end{bmatrix}\right), \theta_\infty(\alpha_{g,y}-c_\theta), \theta_\infty(\alpha_{g,y}-c_{\theta,t-t_0})\right\} \leq$$
$$\theta_\infty(\alpha_{g,y}) + \max\{\theta_\infty(c_\theta), \theta_\infty(c_{\theta,t-t_0})\}$$

from which inequality (23) follows by using (28). Finally, notice that

$$c_\theta(k) - c_{\theta,t-t_0}(k) = \begin{cases} 0 & k \leq t-t_0 \\ \sum_{i=t-t_0+1}^{k} \alpha_\theta(i)\alpha_{g,w}(k-i) & k > t-t_0 \end{cases}$$

Hence, by setting $d_3(k) = \|\alpha_\theta(k+t-t_0+1)\|_2$

$$\begin{aligned}
\|c_\theta - c_{\theta,t-t_0}\|_{\ell_1} &= \sum_{k=t-t_0+1}^{\infty} \Big\| \sum_{i=t-t_0+1}^{k} \alpha_\theta(i)\alpha_{g,w}(k-i)\Big\|_2 \\
&\leq \sum_{k=t-t_0+1}^{\infty} \sum_{i=t-t_0+1}^{k} \|\alpha_\theta(i)\|_2\|\alpha_{g,w}(k-i)\|_2 \\
&= \sum_{k=t-t_0+1}^{\infty} \sum_{i=0}^{k-(t-t_0+1)} d_3(i)\|\alpha_{g,w}(k-i-(t-t_0+1))\|_2 \\
&= \sum_{k=0}^{\infty}\sum_{i=0}^{k} d_3(i)\|\alpha_{g,w}(k-i)\|_2 = \left(\sum_{k=t-t_0+1}^{\infty}\|\alpha_\theta(k)\|_2\right)\cdot\|\alpha_{g,w}\|_{\ell_1}
\end{aligned}$$

By noticing that

$$\sum_{t=t_0}^{N+t_0}\sum_{k=t-t_0+1}^{\infty}\|\alpha_\theta(k)\|_2 \le \sum_{k=1}^{\infty}\min\{k,N+1\}\|\alpha_\theta(k)\|_2 \le \sum_{k=1}^{\infty}k\|\alpha_\theta(k)\|_2 = \theta_\infty(\alpha_\theta)$$

the last inequality (24) follows. Finally, the stationarity is a consequence of the standard result that stable LTI filters maps stationary process to stationary process, see e.g., [16, Section 5.6].

□

E Proof of Theorem 3.1

The proof of Theorem 3.1 is organized as follows. First, we define an auxiliary quantity called *the infinite past empirical error*

$$V_N(\theta) = \frac{1}{N}\sum_{t=0}^{N-1}\ell(\mathbf{y}(t),\hat{\mathbf{y}}_\theta(t))$$

It follows that $V_N(\theta)$ is an unbiased estimate of $\mathcal{L}(\theta)$, i.e., $\mathbf{E}[V_N(\theta)] = \mathcal{L}(\theta)$.

As the first step, we show the following general theorem.

Theorem E.1. *Given a parameter set Θ, a prior distribution π over Θ, $\delta \in (0, 0.5)$, $\lambda > 0$ and $\epsilon \in \{-1, 1\}$. Then with probability at least $1 - 2\delta$, the following holds*

$$\forall\rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}\mathcal{L}(\theta) \le \epsilon E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \tilde{\mathbb{B}}_N(\rho,\pi,\lambda,\epsilon,\delta) \tag{29}$$

$$\tilde{\mathbb{B}}_N(\rho,\pi,\lambda,\epsilon,\delta) \triangleq \frac{1}{\lambda}\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \Psi(\lambda,\pi,N,\epsilon)\right) \tag{30}$$

$$\Psi(\lambda,\pi,N,\epsilon) \triangleq \frac{1}{2}\Big(\ln E_{\theta\sim\pi}\mathbf{E}[e^{2\epsilon\lambda(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}] + \ln E_{\theta\sim\pi}\mathbf{E}[e^{2\epsilon\lambda(\mathcal{L}(\theta)-V_N(\theta))}]\Big), \tag{31}$$

with $\mathrm{KL}(\rho\|\pi) = E_{\theta\sim\rho}\ln\frac{\rho(\theta)}{\pi(\theta)}$ *the Kullback–Leibler divergence.*

In the theorem above, the moment generating functions $\mathbf{E}[e^{2\epsilon\lambda(\mathcal{L}(\theta)-V_N(\theta))}]$ and $\mathbf{E}[e^{2\epsilon\lambda(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}]$ in the right-hand side are interpreted as ∞ whenever $2\epsilon\lambda$ falls outside of their domain of definition. In this case, $\Psi(\lambda,\pi,N,\epsilon)$ is taken to mean ∞ and the bounds are trivially true.

Proof of Theorem E.1. By [15, Theorem 3], for any measurable functions $X(\theta,\omega), Y(\theta,\omega)$, $\tilde{\lambda} > 0$, and $\epsilon \in \{-1, 1\}$, with probability at least $1 - \delta$ we have:

$$\forall\rho \in \mathcal{M}_\pi : E_{\theta\sim\rho}X(\theta,\omega) \le E_{\theta\sim\rho}Y(\theta,\omega) + \frac{1}{\tilde{\lambda}}\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \Psi_{X,Y}(\tilde{\lambda})\right), \tag{32}$$

with $\Psi_{X,Y}(\tilde{\lambda}) = \ln E_{\theta\sim\pi}\mathbf{E}[e^{\tilde{\lambda}(X(\theta,\omega)-Y(\theta,\omega))}]$. Now apply the above theorem for $X = \epsilon\mathcal{L}(\theta), Y = \epsilon V_N(\theta)$, and $X = \epsilon V_N(\theta), Y = \epsilon\hat{\mathcal{L}}_N(\theta)$. Then by applying a union bound on the two obtained probabilistic inequalities, we obtain with probability at least $1 - 2\delta$

$$\forall\rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}\mathcal{L}(\theta) \le \epsilon E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\tilde{\lambda}}\left(2\mathrm{KL}(\rho\|\pi) + 2\ln\frac{1}{\delta} + \Psi_{\epsilon V_N,\epsilon\hat{\mathcal{L}}_N}(\tilde{\lambda}) + \Psi_{\epsilon\mathcal{L},\epsilon V_N}(\tilde{\lambda})\right)$$

With some algebraic manipulation and $\tilde{\lambda} = 2\lambda$, we obtain the statement of the theorem. □

For the case of bounded innovation noise or Lipschitz loss function, we then show that $\mathbf{E}[e^{2\epsilon\lambda(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}]$ and $\mathbf{E}[e^{2\epsilon\lambda(\mathcal{L}(\theta)-V_N(\theta))}]$ can be both upper bounded by $e^{\mathcal{C}_i(\theta,2\lambda,N,\delta,\epsilon)}$, from which the statement of the theorem follows. For the case of quadratic loss function, we will show that it can be reduced to the case of Lipschitz loss function with probability at least $1 - 2\delta/3$ and then we apply the result for Lipschitz loss with $2\delta/3$ instead of δ.

E.1 Proof of Theorem 3.1: bounded case

For the sake of readability, we repeat below the basic assumptions for the case of bounded data

Assumption E.2 (Bounded noise and Lipschitz loss)**.** *The innovation noise* $\mathbf{e}_g(t)$ *from* (4) *is essentially bounded , i.e.,*

$$\|\mathbf{e}_g(t)\|_\infty \leq c_{\mathbf{e}},\ \forall t \in \mathbb{Z} \tag{33}$$

and the loss function ℓ *is either* L_ℓ*-Lipschitz or quadratic.*

The following Lemma is obvious in the L_ℓ-Lipschits case, and follow by applying Cauchy–Schwarz in the quadratic case.

Lemma E.3. *Let* $B = \{(z, z') \in \mathbb{R}^{n_{\mathbf{y}}} \times \mathbb{R}^{n_{\mathbf{y}}} \mid \|z - z'\|_2 \leq G_e(\theta)\sqrt{n_{\mathbf{e}}}c_{\mathbf{e}}\}$*. Then* $(\mathbf{y}(t), \mathbf{y}_\theta(t)), (\mathbf{y}(t), \mathbf{y}_\theta(t \mid 0)) \in B$ *for all* t*,* $t_0 \leq t$*, and for all* $(z_1, z_2), (z_3, z_4) \in B$*,*

$$\ell(z_1, z_2) - \ell(z_3, z_4) \leq L(\theta)\left(\|z_1 - z_3\|_2 + \|z_2 - z_4\|_2\right)$$

where $L(\theta)$ *is as in Theorem 3.1, i.e.,* $L(\theta) = 2G_e(\theta)\sqrt{n_{\mathbf{e}}}c_{\mathbf{e}}$ *if* ℓ*-quadratic, and* $L(\theta) = L_\ell$ *if* ℓ *is* L_ℓ*-Lipschitz.*

Remark E.4 (Norm loss L_ℓ-Lipschitz)**.** *Note that if the loss* ℓ *is the* p *norm of the difference, i.e.,* $\ell(y, \hat{y}) = \|y - \hat{y}\|_p$*, then* $\ell(y, \hat{y}) \leq C_p\|y - \hat{y}\|_2$ *with* $C_p = \sqrt{n_{\mathbf{y}}}$ *if* $p = 1$*, and* $C_p = 1$ *if* $p = \infty$ *or* $p = 2$*.*

Lemma E.5. *Let assumptions 2.2, 2.1 and E.2 hold. Then*

$$\mathbf{E}[e^{\left(\epsilon\lambda|V_N(\theta) - \hat{\mathcal{L}}_N(\theta)|\right)}] \leq e^{\left(\frac{L(\theta)\lambda}{N}\theta_\infty(\alpha_\theta)\|\alpha_g\|_{\ell_1}(c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}})\right)} = e^{C_{b,1}(\lambda,\theta,L(\theta),c_{\mathbf{e}},N)}$$

Proof E.5 . Using the Lipschitz property of the loss function it follows that

$$\begin{aligned}\epsilon(V_N(\theta) - \hat{\mathcal{L}}_N(\theta)) \leq |V_N(\theta) - \hat{\mathcal{L}}_N(\theta)| \leq \frac{1}{N}\sum_{t=0}^{N-1}|\ell(\mathbf{y}(t), \hat{\mathbf{y}}_\theta(t)) - \ell(\mathbf{y}(t), \hat{\mathbf{y}}_\theta(t \mid 0))| \leq \\ \frac{L(\theta)}{N}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t) - \hat{\mathbf{y}}_\theta(t \mid 0)\|_2\end{aligned} \tag{34}$$

From Lemma D.3 it follows that

$$\hat{\mathbf{y}}_\theta(t) - \hat{\mathbf{y}}_\theta(t \mid 0) = ((c_\theta - c_{\theta,t}) \star \mathbf{e}_g)(t) = \sum_{k=0}^{\infty}(c_\theta - c_{\theta,t})(k)\mathbf{e}_g(t - k)$$

and hence

$$\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t) - \hat{\mathbf{y}}_\theta(t \mid 0)\|_2 \leq \sum_{t=0}^{N-1}\sum_{k=0}^{\infty}\|(c_\theta - c_{\theta,t})(k)\|_2\|\mathbf{e}_g(t - k)\|_2 \leq \sum_{t=0}^{N-1}\|c_\theta - c_{\theta,t}\|_{\ell_1}\sqrt{n_{\mathbf{e}}}c_{\mathbf{e}}$$

From (24) it follows that

$$\sum_{t=0}^{N-1}\|c_\theta - c_{\theta,t}\|_{\ell_1} \leq \theta_\infty(\alpha_\theta)\|\alpha_g\|_{\ell_1}$$

and hence,

$$\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t) - \hat{\mathbf{y}}_\theta(t \mid 0)\|_2 \leq \theta_\infty(\alpha_\theta)\|\alpha_g\|_{\ell_1}\sqrt{n_{\mathbf{e}}}c_{\mathbf{e}}$$

from which the statement of the lemma follows using the definition of $\mathcal{C}_{b,1}(\lambda, \theta, L(\theta), c_{\mathbf{e}}, N)$. □

Lemma E.6. *Let assumptions 2.2, 2.1 and E.2 hold. Then*

$$\mathbf{E}[e^{\epsilon\lambda(\mathcal{L}(\theta) - V_N(\theta))}] \leq e^{\frac{(L(\theta))^2\lambda^2}{2N}(G_e(\theta) + 2G_{e,1}(\theta))(c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}})^2} = e^{\mathcal{C}_{b,2}(\lambda,\theta,L(\theta),c_{\mathbf{e}},N)} \tag{35}$$

Proof of Lemma E.6 . For each $\Sigma(\theta) \in \mathcal{F}$, consider $\mathbf{z}_t = \left[\mathbf{y}^T(t)\ \hat{\mathbf{y}}_\theta^T(t)\right]^T$. Then by Lemma D.3, $\mathbf{z}_t = \left(\begin{bmatrix}\alpha_{g,y}\\ c_\theta\end{bmatrix} \star \mathbf{e}_g\right)(t)$ and $\mathbf{z}_t$ is stationary. Moreover, as $\left\|\begin{bmatrix}\alpha_{g,y}\\ c_\theta\end{bmatrix}\right\|_{\ell_1} \leq G_e(\theta) < \infty$ and $\theta_\infty\left(\begin{bmatrix}\alpha_{g,y}\\ c_\theta\end{bmatrix}\right) \leq G_{e,1}(\theta) < \infty$, it follows that $\mathbf{z}_t$ is a Bernoulli-shift. In particular, $\|\mathbf{z}_t\|_\infty \leq G_e(\theta)c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}}$, and by [1, Proposition 4.2] $\mathbf{z}_t$ is a weakly dependent process in the terminology of [1], with $\|\mathbf{z}_0\|_\infty \leq G_e(\theta)c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}}$, and the mixing coefficient $\theta_{\infty,N}(1)$ of $\mathbf{z}_t$ satisfies $\theta_{\infty,N}(1) < 2G_{e,1}(\theta)c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}}$.

If ℓ is L_ℓ-Lipschitz, then let $B = \mathbb{R}^{n_{\mathbf{y}}} \times \mathbb{R}^{n_{\mathbf{y}}}$. If ℓ is quadratic, then consider the set B from Lemma E.3. Define the function $h(z_0, z_0', \ldots, z_{N-1}, z_{N-1}') = \epsilon\frac{1}{L(\theta)}\sum_{i=0}^{N-1}\ell(z_i, z_i')$, $z_0, \ldots, z_{N-1}, z_0', \ldots, z_{N-1}' \in \mathbb{R}^{n_{\mathbf{y}}}$ on B^N. We claim that h has Lipschitz constant 1 in the following sense:

$$|h(z_0, z_0', \ldots, z_{N-1}, z_{N-1}') - h(\tilde{z}_0, \tilde{z}_0', \ldots, \tilde{z}_{N-1}, \tilde{z}_{N-1}')| \leq \sum_{i=0}^{N-1} \|z_i - \tilde{z}_i\|_2 + \|z_i' - \tilde{z}_i'\|_2$$

Indeed, using Lemma E.3,

$$|h(z_0, z_0', \ldots, z_{N-1}, z_{N-1}') - h(\tilde{z}_0, \tilde{z}_0', \ldots, \tilde{z}_{N-1}, \tilde{z}_{N-1}')| = \left|\frac{\epsilon}{L(\theta)}\sum_{i=0}^{N-1}\ell(z_i, z_i') - \ell(\tilde{z}_i, \tilde{z}_i')\right|$$
$$\leq \frac{1}{L(\theta)}\sum_{i=0}^{N-1} L(\theta)(\|z_i - \tilde{z}_i\|_2 + \|z_i' - \tilde{z}_i'\|_2) = \sum_{i=0}^{N-1}\|z_i - \tilde{z}_i\|_2 + \|z_i' - \tilde{z}_i'\|_2$$

If ℓ is L_ℓ-Lipschitz, then h is already defined on $(\mathbb{R}^{n_{\mathbf{y}}} \times \mathbb{R}^{n_{\mathbf{y}}})^N$. If ℓ is quadratic, we can use Kirszbraun's theorem for scalar valued functions [13, 2.10.44,page 202], from which it then follows that h can be extended to a 1-Lipschitz map on $(\mathbb{R}^{n_{\mathbf{y}}} \times \mathbb{R}^{n_{\mathbf{y}}})^N$ without changing its values on B^N. In fact, such an extension can be defined as $h_{ext}(z) = \inf_{x\in B^N}\{h(x) + \|z - x\|_2\}$. By abuse of notation this extension will be denote by h too. Notice that by Lemma E.3, $(\mathbf{y}(t), \mathbf{y}_\theta(t)) \in B$ for all t, hence $h(\mathbf{y}(0), \mathbf{y}_\theta(0), \ldots, \mathbf{y}(N-1), \mathbf{y}_\theta(N-1)) = \frac{\epsilon}{L(\theta)}\sum_{i=0}^{N-1}\ell(\mathbf{y}(i), \mathbf{y}_\theta(i)) = \frac{\epsilon}{L(\theta)}NV_N(\theta)$.

Notice that,

$$\epsilon\lambda(\mathbf{E}[V_N(\theta)] - V_N(\theta)) = \frac{\lambda L(\theta)}{N}\left(\mathbf{E}[h] - h\right).$$

Now we can apply [1, Theorem 6.6], from which it follows that for any $q \geq 0$

$$\mathbf{E}[e^{q(\mathbf{E}[h]-h)}] \leq e^{\frac{q^2 N}{2}(\|\mathbf{z}_t\|_\infty + \theta_{\infty,N}(1))^2} \tag{36}$$

By choosing $q = \frac{\lambda L(\theta)}{N}$ and using the definition of $\mathcal{C}_{b,2}(\lambda, \theta, L(\theta), c_{\mathbf{e}}, N)$, we obtain the statement of the lemma . □

Proof of Theorem 3.1: bounded processes. The statement of the theorem follows from Theorem E.1 by replacing $\mathbf{E}[e^{2\lambda\epsilon(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}]$ and $\mathbf{E}[e^{2\lambda\epsilon(\mathcal{L}(\theta)-V_N(\theta))}]$ with their upper bound from Lemma E.5 and Lemma E.6 respectively. □

E.2 Proof of Theorem 3.1: unbounded processes and Lipschitz loss functions

Recall that according to our assumptions, the 2-norm of the noise process $\mathbf{e}_g$ is subgaussian. Bounded processes have this property, however, in this section we will not assume that the process $\mathbf{e}_g$ is bounded.

Recall that we also consider Lipschitz losses.

Assumption E.7 (Lipschitz loss)**.** *We assume that the loss function ℓ is L_ℓ-Lipschitz, i.e., for all $y, \hat{y}, y', \hat{y}' \in \mathbb{R}^{n_{\mathbf{y}}}$, $|\ell(y, \hat{y}) - \ell(y', \hat{y}')| \leq L_\ell(\|y - y'\|_2 + \|\hat{y} - \hat{y}'\|_2)$.*

The main idea is to use Theorem E.1 and to derive bounds for the moment generating functions $\mathbf{E}[e^{2\lambda\epsilon(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}]$ and $\mathbf{E}[e^{2\lambda\epsilon(\mathcal{L}(\theta)-V_N(\theta))}]$. In order to achieve this, we will reduce the problem of bounding $\mathbf{E}[e^{2\lambda\epsilon(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}]$ and $\mathbf{E}[e^{2\lambda\epsilon(\mathcal{L}(\theta)-V_N(\theta))}]$ to the same problem for bounded noises, by truncating the unbounded noise $\mathbf{e}_g$ which drives the data generator. This problem was solved in Lemma E.5 and Lemma E.6, which thus can be re-used.

To this end, let $C > 0$ be any constant and let $\bar{\mathbf{e}}_g(t)$ be the truncated version of $\mathbf{e}_g(t)$, i.e.,

$$\bar{\mathbf{e}}_g(t) = \mathbf{e}_g(t)\chi_{\{\|\mathbf{e}_g(t)\|\leq C\}} := \mathbf{e}_g(t)\chi(\|\mathbf{e}_g(t)\| \leq C). \tag{37}$$

Let $\bar{\mathbf{w}}(t) = \begin{bmatrix}\bar{\mathbf{y}}^T(t) & \bar{\mathbf{u}}^T(t)\end{bmatrix}^T$ be the output of the data generator, if $\mathbf{e}_g(t)$ is replaced by $\bar{\mathbf{e}}_g(t)$, and let $\hat{\bar{\mathbf{y}}}_\theta(t \mid t_0)$ be the output of the predictor (1) parametrized by θ with $\mathbf{w}$ replaced by $\bar{\mathbf{w}}$, and let $\hat{\bar{\mathbf{y}}}_\theta(t)$ be the limit $\lim_{t_0\to-\infty}\hat{\bar{\mathbf{y}}}_\theta(t \mid t_0)$. Note that Lemma 2.4 applies, as $\bar{\mathbf{e}}_g(t)$ satisfies all the assumptions of Assumption 2.2. Let $\bar{V}_N(\theta) = \frac{1}{N}\sum_{t=0}^{N-1}\ell(\bar{\mathbf{y}}(t), \hat{\bar{\mathbf{y}}}_\theta(t))$ be the corresponding infinite past empirical error, $\hat{\bar{\mathcal{L}}}_N(\theta) = \frac{1}{N}\sum_{t=0}^{N-1}\ell(\bar{\mathbf{y}}(t), \hat{\bar{\mathbf{y}}}_\theta(t \mid 0))$ the finite past empirical error, and $\bar{\mathcal{L}}(\theta) = \mathbf{E}[\ell(\bar{\mathbf{y}}(t), \hat{\bar{\mathbf{y}}}_\theta(t))]$ the true error for the model corresponding to $\theta \in \Theta$, if it is fed with the truncated signal $\bar{\mathbf{w}}$ instead of $\mathbf{w}$.

It turns out that we can bound the difference between these quantities and their non-truncated counterparts. The basic steps are as follows: First, using Lipschitz properties of the loss function, we bound the moment generating functions of the differences $\sup_{\theta\in\Theta}|\hat{\mathcal{L}}_N(\theta) - \hat{\bar{\mathcal{L}}}_N(\theta)|$, $\sup_{\theta\in\Theta}|V_N(\theta) - \bar{V}_N(\theta)|$ by the corresponding quantity for the ℓ_2 loss, i.e.; by the moment generating functions of $\sup_{\theta\in\Theta}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t \mid 0) - \hat{\bar{\mathbf{y}}}_\theta(t \mid 0)\|_2$ and $\sum_{t=0}^{N-1}\|\mathbf{y}(t) - \bar{\mathbf{y}}(t)\|_2$.

Lemma E.8.

$$\mathbf{E}[e^{\lambda\sup_{\theta\in\Theta}|\hat{\mathcal{L}}_N(\theta)-\hat{\bar{\mathcal{L}}}_N(\theta)|}] \leq \sqrt{\mathbf{E}[e^{\frac{2\lambda L_\ell}{N}\sum_{t=0}^{N-1}\|\mathbf{y}(t)-\bar{\mathbf{y}}(t)\|_2}]}\sqrt{\mathbf{E}[e^{\frac{2\lambda L_\ell}{N}\sup_{\theta\in\Theta}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t|0)-\hat{\bar{\mathbf{y}}}_\theta(t|0)\|_2}]} \tag{38}$$

$$\mathbf{E}[e^{\lambda\sup_{\theta\in\Theta}|V_N(\theta)-\bar{V}_N(\theta)|}] \leq \sqrt{\mathbf{E}[e^{\frac{2\lambda L_\ell}{N}\sum_{t=0}^{N-1}\|\mathbf{y}(t)-\bar{\mathbf{y}}(t)\|_2}]}\sqrt{\mathbf{E}[e^{\frac{2\lambda L_\ell}{N}\sup_{\theta\in\Theta}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t)-\hat{\bar{\mathbf{y}}}_\theta(t)\|_2}]} \tag{39}$$

Then, using the technique of [3] we bound the moment generating functions of the truncation errors $\sup_{\theta\in\Theta}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t \mid 0)-\hat{\bar{\mathbf{y}}}_\theta(t \mid 0)\|_2$ and $\sum_{t=0}^{N-1}\|\mathbf{y}(t)-\bar{\mathbf{y}}(t)\|_2$. To this end, we use the following bound on the moment generating function of $\|\mathbf{e}_g(t)\|_2$.

Lemma E.9 (Subgaussian noise)**.** *If $\|\mathbf{e}_g(t)\|_2$ is sub-Gaussian then $\Psi_{\mathbf{e}_g}(\lambda) = \mathbf{E}[e^{\lambda\|\mathbf{e}_g(0)\|_2}]$ is well-defined for all $\lambda > 0$. In fact,*

$$\Psi_{\mathbf{e}_g}(\lambda) \leq \bar{\Psi}_{\mathbf{e}_g}(\lambda) = e^{(\lambda^2 14^2\|\mathbf{e}_g\|_{\psi_2}^2 + 3\lambda\|\mathbf{e}_g\|_{\psi_2})}$$

Using this bound, the moment generating function of the truncation error can be bounded as follows:

Lemma E.10. *For any $\lambda < 0.5N$*

$$\mathbf{E}[e^{\frac{\lambda}{N}\sum_{t=0}^{N-1}\|\mathbf{y}(t)-\bar{\mathbf{y}}(t)\|_2}] \leq e^{\mathcal{C}(\|\alpha_g\|_{\ell_1},\lambda,N,C)}$$
$$\mathbf{E}[e^{\frac{\lambda}{N}\sup_{\theta\in\Theta}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}(t)-\hat{\bar{\mathbf{y}}}_\theta(t)\|_2}] \leq e^{\mathcal{C}(\|\alpha_g\|_{\ell_1}\|\bar{\alpha}\|_{\ell_1},\lambda,N,C)}$$
$$\mathbf{E}[e^{\frac{\lambda}{N}\sup_{\theta\in\Theta}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}(t|0)-\hat{\bar{\mathbf{y}}}_\theta(t|0)\|_2}] \leq e^{\mathcal{C}(\|\alpha_g\|_{\ell_1}\|\bar{\alpha}\|_{\ell_1},\lambda,N,C)}$$

where

$$\mathcal{C}(s,\lambda,N,C) = \lambda\frac{\bar{\Psi}_{\mathbf{e}_g}(s)sC}{(e^{sC}-1)} + \frac{2\lambda^2}{N}\bar{\Psi}_{\mathbf{e}_g}(s) \tag{40}$$

From this we derive

Lemma E.11. *For any $\lambda < N/4L_\ell$*

$$\max\{\mathbf{E}[e^{\lambda\sup_{\theta\in\Theta}|\hat{\mathcal{L}}_N(\theta)-\hat{\bar{\mathcal{L}}}_N(\theta)|}], \mathbf{E}[e^{\lambda\sup_{\theta\in\Theta}|V_N(\theta)-\bar{V}_N(\theta)|}], e^{\lambda\sup_{\theta\in\Theta}|\mathcal{L}(\theta)-\bar{\mathcal{L}}(\theta)|}\} \leq e^{\frac{1}{2}\left(\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N,C)+\mathcal{C}(\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N,C)\right)}$$

In order to go back to bounding the moment generating functions of $\epsilon(V_N(\theta) - \hat{\mathcal{L}}_N(\theta))$ and $\epsilon(\mathcal{L}(\theta) - V_N(\theta))$, we use the previous lemma, and the following inequality which we derive using the Cauchy-Schwarz inequality:

Lemma E.12. *For* $\epsilon \in \{-1, 1\}$

$$\mathbf{E}[e^{\lambda\epsilon(\mathcal{L}(\theta)-V_N(\theta))}] \leq$$

$$\sqrt{\mathbf{E}[e^{2\lambda\epsilon(\bar{\mathcal{L}}(\theta)-\bar{V}_N(\theta))}]}\sqrt{\sqrt{\mathbf{E}[e^{4\lambda\sup_{\theta\in\Theta}|\bar{V}_N(\theta)-V_N(\theta)|}]}\sqrt{e^{4\lambda\sup_{\theta\in\Theta}|\bar{\mathcal{L}}(\theta)-\mathcal{L}(\theta)|}}} \tag{41}$$

$$\mathbf{E}[e^{\lambda\epsilon(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}] \leq$$

$$\sqrt{\mathbf{E}[e^{2\lambda\epsilon(\bar{V}_N(\theta)-\hat{\bar{\mathcal{L}}}_N(\theta))}]}\sqrt{\sqrt{\mathbf{E}[e^{4\lambda\sup_{\theta\in\Theta}|\bar{V}_N(\theta)-V_N(\theta)|}]}\sqrt{\mathbf{E}[e^{4\lambda\sup_{\theta\in\Theta}|\hat{\bar{\mathcal{L}}}_N(\theta)-\hat{\mathcal{L}}_N(\theta)|}]}} \tag{42}$$

If we apply Lemma E.11 to the right-hand sides of the latter inequalities, then we obtain.

Lemma E.13. *For* $\epsilon \in \{1, -1\}$, *for* $\lambda < \frac{N}{16L_\ell}$,

$$\mathbf{E}[e^{\lambda\epsilon(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}] \leq$$

$$e^{\frac{1}{4}(\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},8L_\ell\lambda,N,C)+\mathcal{C}(\|\alpha_g\|_{\ell_1},8L_\ell\lambda,N,C))}\sqrt{\mathbf{E}[e^{2\lambda\epsilon(\bar{V}_N(\theta)-\hat{\bar{\mathcal{L}}}_N(\theta))}]}$$

$$\mathbf{E}[e^{\lambda\epsilon(\mathcal{L}(\theta)-V_N(\theta))}] \leq e^{\frac{1}{4}(\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},8L_\ell\lambda,N,C)+\mathcal{C}(\|\alpha_g\|_{\ell_1},8L_\ell\lambda,N,C))}\sqrt{\mathbf{E}[e^{2\lambda\epsilon(\bar{\mathcal{L}}(\theta)-\bar{V}_N(\theta))}]}$$

Finally, Lemma E.5 and Lemma E.6 derived for the bounded case can be applied to the moment generating functions of $\epsilon(\bar{\mathcal{L}}(\theta) - \bar{V}_N(\theta))$ and $\epsilon(\bar{V}_N(\theta) - \hat{\bar{\mathcal{L}}}_N(\theta))$, for a suitable choice of C. The detailed proof is presented below.

Proof of Theorem 3.1: unbounded noise and Lipschitz loss. Let $C = \frac{2\ln(N)}{K\|\alpha_g\|_{\ell_1}}$ where $K = \min\{1, \|\bar{\alpha}\|_{\ell_1}\}$. Then

$$C\|\alpha_g\|_{\ell_1} \geq 2\ln(N), \quad C\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1} \geq 2\ln(N)$$

and hence for $N \geq 2$,

$$e^{C\|\alpha_g\|_{\ell_1}} - 1 \geq e^{2\ln(N)} - 1 = N^2 - 1 \geq N, \quad e^{C\|\alpha_g\|_{\ell_1}\|\bar{\alpha}\|_{\ell_1}} - 1 \geq e^{2\ln(N)} - 1 = N^2 - 1 \geq N$$

and therefore

$$\frac{1}{e^{C\|\alpha_g\|_{\ell_1}} - 1} \leq \frac{1}{N}, \quad \frac{1}{e^{C\|\alpha_g\|_{\ell_1}\|\bar{\alpha}\|_{\ell_1}} - 1} \leq \frac{1}{N}$$

Therefore, as $C\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1} = 2\ln(N)\max\{1, \|\bar{\alpha}\|_{\ell_1}\}$,

$$\mathcal{C}(\|\alpha_g\|_{\ell_1}, 8L_\ell\lambda, N, C) \leq \frac{8L_\ell\lambda}{N}\bar{\Psi}_{\mathbf{e}_g}(\|\alpha_g\|_{\ell_1})\|\alpha_g\|_{\ell_1}\frac{2\ln(N)}{K\|\alpha_g\|_{\ell_1}} + \frac{(8L_\ell\lambda)^2}{N}2\bar{\Psi}_{\mathbf{e}_g}(\|\alpha_g\|_{\ell_1}) \leq$$

$$\frac{8L_\ell\lambda}{N}\bar{\Psi}_{\mathbf{e}_g}(\|\alpha_g\|_{\ell_1})\frac{1}{\min\{1, \|\bar{\alpha}\|_{\ell_1}\}}2\ln(N) + \frac{128L_\ell^2\lambda^2}{N}\bar{\Psi}_{\mathbf{e}_g}(\|\alpha_g\|_{\ell_1})$$

$$\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1}, 8L_\ell\lambda, N, C) \leq$$

$$\frac{8L_\ell\lambda}{N}\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1})(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1})C + \frac{(8L_\ell\lambda)^2}{N}2\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1}) =$$

$$\frac{8L_\ell\lambda}{N}\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1})2\ln(N)\max\{1, \|\bar{\alpha}\|_{\ell_1}\} + \frac{128L_\ell^2\lambda^2}{N}\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1})$$

Hence,

$$\frac{1}{4}(\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1}, 8L_\ell\lambda, N, C) + \mathcal{C}(\|\alpha_g\|_{\ell_1}, 8L_\ell\lambda, N, C)) \leq$$

$$\frac{4\lambda\ln(N)L_\ell}{N}\left(\underbrace{\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1})\max\{1, \|\bar{\alpha}\|_{\ell_1}\} + \bar{\Psi}_{\mathbf{e}_g}(\|\alpha_g\|_{\ell_1})\frac{1}{\min\{1, \|\bar{\alpha}\|_{\ell_1}\}}}_{C_{u,2}}\right) \tag{43}$$

$$+ \frac{32\lambda^2L_\ell^2}{N}\underbrace{\left(\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1}) + \bar{\Psi}_{\mathbf{e}_g}(\|\alpha_g\|_{\ell_1})\right)}_{C_{u,1}} = \bar{\mathcal{C}}_{ub}(\lambda, L_\ell, N)$$

From Lemma E.5 and Lemma E.6 applied to $\bar{V}_N(\theta)$, $\bar{\mathcal{L}}(\theta)$, $\hat{\bar{\mathcal{L}}}_N(\theta)$, with $c_{\mathbf{e}} = C = \frac{2\ln(N)}{K\|\alpha_g\|_{\ell_1}}$ and $L(\theta) = L_\ell$, it follows that

$$\sqrt{\mathbf{E}[e^{2\lambda\epsilon(\bar{V}_N(\theta)-\hat{\bar{\mathcal{L}}}_N(\theta))}]} \leq \sqrt{e^{\mathcal{C}_{b,1}(2\lambda,\theta,L_\ell,\frac{2\ln(N)}{K\|\alpha_g\|_{\ell_1}},N)}} = e^{\frac{1}{2}\mathcal{C}_{b,1}(2\lambda,\theta,L_\ell,\frac{2\ln(N)}{K\|\alpha_g\|_{\ell_1}},N)}$$
$$\sqrt{\mathbf{E}[e^{\epsilon 2\lambda(\bar{\mathcal{L}}(\theta)-\bar{V}_N(\theta))}]} \leq \sqrt{e^{\mathcal{C}_{b,2}(2\lambda,\theta,L_\ell,\frac{2\ln(N)}{K\|\alpha_g\|_{\ell_1}},N)}} = e^{\frac{1}{2}\mathcal{C}_{b,2}(2\lambda,\theta,L_\ell,\frac{2\ln(N)}{K\|\alpha_g\|_{\ell_1}},N)}$$

Using Lemma E.13 and the inequalities above for any $\lambda < \frac{N}{32L_\ell}$,

$$\begin{aligned}\mathbf{E}[e^{2\lambda\epsilon(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}] &\leq e^{\bar{C}_{ub}(2\lambda,L_\ell,N)+\frac{1}{2}\mathcal{C}_{b,1}(4\lambda,\theta,L_\ell,\frac{2\ln(N)}{K\|\alpha_g\|_{\ell_1}},N)} = e^{\mathcal{C}_1(2\lambda,\theta,N,\delta,\epsilon)} \\ \mathbf{E}[e^{2\lambda\epsilon(\mathcal{L}(\theta)-V_N(\theta))}] &\leq e^{\bar{C}_{ub}(2\lambda,L_\ell,N)+\frac{1}{2}\mathcal{C}_{b,2}(4\lambda,\theta,L_\ell,\frac{2\ln(N)}{K\|\alpha_g\|_{\ell_1}},N)} = e^{\mathcal{C}_2(2\lambda,\theta,N,\delta,\epsilon)}\end{aligned} \tag{44}$$

The statement of the theorem follows from Theorem E.1 by replacing $\mathbf{E}[e^{2\lambda\epsilon(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}]$ and $\mathbf{E}[e^{2\lambda\epsilon(\mathcal{L}(\theta)-V_N(\theta))}]$ with their upper bound from (44).

□

The rest of the section is devoted to the proof of the lemmas used above.

First, we prove Lemma E.9

Proof of Lemma E.9. First note that $\mathbf{z}(t) := \|\mathbf{e}_g(t)\|_2 - \mathbf{E}[\|\mathbf{e}_g(t)\|_2]$ is zero mean sub-Gaussian, and using sub-additivity of the sub-Gaussian norm we get $\|\mathbf{z}(t)\|_{\psi_2} \leq \|\|\mathbf{e}_g(t)\|_2\|_{\psi_2} + \|\mathbf{E}[\|\mathbf{e}_g(t)\|_2]\|_{\psi_2} = \|\mathbf{e}_g\|_{\psi_2} + \|\mathbf{E}[\|\mathbf{e}_g(t)\|_2]\|_{\psi_2}$. Now from the proof of [43, Proposition 2.6.1] it follows that $\mathbf{E}[\|\mathbf{e}_g(t)\|_2] \leq 3\|\mathbf{e}_g\|_{\psi_2}$, and by [43, Exercise 2.24], $\|\mathbf{E}[\|\mathbf{e}_g(t)\|_2]\|_{\psi_2} \leq \mathbf{E}[\|\mathbf{e}_g(t)\|_2]/\sqrt{\ln 2} \leq 2\mathbf{E}[\|\mathbf{e}_g(t)\|_2]$. Hence $\|\mathbf{z}(t)\|_{\psi_2} \leq 7\|\mathbf{e}_g\|_{\psi_2}$.

Then from the proof of [43, Proposition 2.6.1] it follows that $\mathbf{E}[e^{\lambda(\|\mathbf{e}_g(t)\|_2-\mathbf{E}[\|\mathbf{e}_g(t)\|_2])}] \leq e^{\lambda^2 K_z^2}$ and $K_z := \sqrt{\frac{3}{2}}\|\mathbf{z}(t)\|_{\psi_2} \leq 14\|\mathbf{e}_g\|_{\psi_2}$. It then follows that $\Psi_{\mathbf{e}_g}(\lambda) \leq e^{(\lambda^2 K_z^2+\lambda\mathbf{E}[\|\mathbf{e}_g(t)\|_2])}$ and the statement of the lemma follows. □

In order to present the proof of Lemma E.8-E.13, we need the following sequence of results.

Lemma E.14. *Let* $\mathbf{z}$ *be nonnegative valued random variable such that for all* $0 < s \leq \bar{\lambda}$ *its moment generating function is well-defined. Then for any* $C > 0$*,*

$$\mathbf{E}[\mathbf{z}\chi(\mathbf{z} > C)] \leq \frac{C\mathbf{E}[e^{\bar{\lambda}\mathbf{z}}]}{e^{\bar{\lambda}C} - 1}$$

Note that, as in (37), we use $\chi(\mathbf{z} > C)$ as a shorthand for the indicator function $\chi_{\{\mathbf{z}>C\}}(\mathbf{z})$. This practices will be used in the sequel.

Proof of Lemma E.14. Let us consider $\phi(x) = \frac{e^x-1}{x}$ with the continuous extension $\phi(0) = 1$. It then follows that ϕ is a monotonically increasing function of x for $x > 0$. Hence,

$$\frac{\phi(\bar{\lambda}\mathbf{z})}{\phi(\bar{\lambda}C)} \geq \chi(\mathbf{z} > C).$$

Indeed, if $\mathbf{z} > C$, then $\frac{\phi(\bar{\lambda}\mathbf{z})}{\phi(\bar{\lambda}C)} \geq 1 = \chi(\mathbf{z} > C)$. If $\mathbf{z} \leq C$, then $1 \geq \frac{\phi(\bar{\lambda}\mathbf{z})}{\phi(\bar{\lambda}C)} \geq 0 = \chi(\mathbf{z} > C)$. It then follows that

$$\mathbf{z}\frac{\phi(\bar{\lambda}\mathbf{z})}{\phi(\bar{\lambda}C)} \geq \mathbf{z}\chi(\mathbf{z} > C)$$

and by taking expectations it follows that

$$\mathbf{E}[\mathbf{z}\chi(\mathbf{z} > C)] \leq \frac{1}{\phi(\bar{\lambda}C)}\mathbf{E}[\mathbf{z}\phi(\bar{\lambda}\mathbf{z})]$$

Note that

$$\mathbf{z}\phi(\bar{\lambda}\mathbf{z}) = \frac{e^{\bar{\lambda}\mathbf{z}} - 1}{\bar{\lambda}}$$

and hence by using the inequality $e^x - 1 \leq e^x$ for $x \geq 0$ it follows that

$$\mathbf{E}[\mathbf{z}\phi(\bar{\lambda}\mathbf{z})] = \frac{\mathbf{E}[e^{\bar{\lambda}\mathbf{z}}] - 1}{\bar{\lambda}} \leq \frac{\mathbf{E}[e^{\bar{\lambda}\mathbf{z}}]}{\bar{\lambda}}$$

By noticing that

$$\bar{\lambda}\phi(\bar{\lambda}C) = \frac{e^{\bar{\lambda}C} - 1}{C}$$

the statement of the lemma follows. □

Lemma E.15.

$$\mathbf{E}[e^{s\|e_g(t)\|_2\chi(\|\mathbf{e}_g(t)\|_2 > C)}] \leq \left(1 + s\frac{\bar{\Psi}_{\mathbf{e}_g}(\bar{\lambda})C}{e^{\bar{\lambda}C} - 1} + s^2\frac{2\bar{\Psi}_{\mathbf{e}_g}(\bar{\lambda})}{\bar{\lambda}^2}\right) := \bar{\Psi}_{\mathbf{e}_g}(s, \bar{\lambda}, C)$$

for all $s \in [0, 0.5\bar{\lambda})$.

Proof Lemma E.15. First note that

$$\begin{aligned}\mathbf{E}[e^{s\|\mathbf{e}_g(t)\|_2\chi(\|\mathbf{e}_g(t)\|_2 > C)}] = 1 + s\mathbf{E}[\|\mathbf{e}_g(t)\|_2\chi(\|\mathbf{e}_g(t)\|_2 > C)] + \\ \sum_{k=2}^{\infty}\frac{\mathbf{E}[\|\mathbf{e}_g(t)\|_2^k\chi(\|\mathbf{e}_g(t)\|_2 > C)]s^k]}{k!}.\end{aligned} \tag{45}$$

and $\mathbf{E}[\|\mathbf{e}_g(t)\|_2^k\chi(\|\mathbf{e}_g(t)\|_2 > C)] \leq \mathbf{E}[\|\mathbf{e}_g(t)\|_2^k]$.

As in Lemma E.9 let $\Psi_{\mathbf{e}_g}(\bar{\lambda}) = \mathbf{E}[e^{\bar{\lambda}\|\mathbf{e}_g(0)\|_2}] = \mathbf{E}[e^{\bar{\lambda}\|\mathbf{e}_g(t)\|_2}]$ be the moment generating function for $\|\mathbf{e}_g(t)\|_2$. Then

$$\Psi_{\mathbf{e}_g}(\bar{\lambda}) = \sum_{k=0}^{\infty}\frac{\mathbf{E}[\|\mathbf{e}_g(t)\|_2^k]}{k!}\bar{\lambda}^k$$

and we get

$$\frac{\mathbf{E}[\|\mathbf{e}_g(t)\|_2^k]}{k!} \leq \frac{\Psi_{\mathbf{e}_g}(\bar{\lambda})}{\bar{\lambda}^k}.$$

It follows that for all $k \geq 2$,

$$\frac{\mathbf{E}[\|\mathbf{e}_g(t)\|_2^k\chi(\|\mathbf{e}_g(t)\|_2 > C)]s^k}{k!} \leq \frac{\mathbf{E}[\|\mathbf{e}_g(t)\|_2^k]s^k}{k!} \leq \Psi_{\mathbf{e}_g}(\bar{\lambda})\frac{s^k}{\bar{\lambda}^k}.$$

Therefore, for all $s \in [0, 0.5\bar{\lambda})$.

$$\begin{aligned}\sum_{k=2}^{\infty}\frac{\mathbf{E}[\|e_g(t)\|_2^k\chi(\|\mathbf{e}_g(t)\|_2 > C)]s^k]}{k!} \leq \sum_{k=2}^{\infty}\Psi_{\mathbf{e}_g}(\bar{\lambda})\frac{s^k}{\bar{\lambda}^k} = \Psi_{\mathbf{e}_g}(\bar{\lambda})\frac{s^2}{\bar{\lambda}^2}\frac{1}{1 - \frac{s}{\bar{\lambda}}} = \\ \Psi_{\mathbf{e}_g}(\bar{\lambda})\frac{s^2}{\bar{\lambda}(\bar{\lambda} - s)} \leq \Psi_{\mathbf{e}_g}(\bar{\lambda})\frac{s^2}{0.5\bar{\lambda}^2} = 2\Psi_{\mathbf{e}_g}(\bar{\lambda})\frac{s^2}{\bar{\lambda}^2}\end{aligned} \tag{46}$$

where in the last inequality we used that as $s < 0.5\bar{\lambda}$, $\bar{\lambda} - s \geq 0.5\bar{\lambda}$. Finally, use Lemma E.14 with $\mathbf{z} = \|\mathbf{e}_g(t)\|_2$ to obtain

$$\mathbf{E}[\|\mathbf{e}_g(t)\|_2\chi(\|\mathbf{e}_g(t)\|_2 > C)] \leq \frac{\Psi_{\mathbf{e}_g}(\bar{\lambda})C}{e^{\bar{\lambda}C} - 1} \tag{47}$$

The result now follows by combining (45), (46) and (47), and then use Lemma E.9. □

We will also need the following technical lemma

Lemma E.16. *Let* $a_k \geq 0$ *be a sequence of nonnegative numbers such that the series* $\sum_{k=0}^{\infty} a_k$ *is convergent. For any* N*, define* $c_{k,N} = \sum_{l=\max\{0,k-N+1\}}^{k} a_l$*. Then for every* $L \geq 0$*,* $\sum_{k=0}^{L} c_{k,N} \leq N\sum_{k=0}^{L} a_k$*, the series* $\sum_{k=0}^{\infty} c_{k,N}$ *is convergent, and* $\sum_{k=0}^{\infty} c_{k,N} \leq N\sum_{k=0}^{\infty} a_k$*.*

Proof of Lemma E.16. We will prove by induction on $L \geq N-1$ that

$$\sum_{k=0}^{L} c_{k,N} = \sum_{k=0}^{L-N+1} N a_k + \sum_{k=L-N+2}^{L} (L-k+1) a_k \tag{48}$$

For $L = N-1$, by induction on N, it can be shown that

$$\sum_{k=0}^{N-1} c_{k,N} = \sum_{k=0}^{N-1} \sum_{l=0}^{k} a_l = \sum_{k=0}^{N-1} (N-k) a_k.$$

which yield (48) for $L = N-1$. Assume (48) holds for $N-1 \leq L' \leq L$. Notice that $c_{L+1,N} = \sum_{l=L-N+2}^{L+1} a_l$ and hence

$$\sum_{k=0}^{L+1} c_{k,N} = \sum_{k=0}^{L} c_{k,N} + c_{L+1,N} = \sum_{k=0}^{L-N+1} N a_k + \sum_{k=L-N+2}^{L} (L-k+1) a_k + \sum_{l=L-N+2}^{L} a_l =$$

$$\sum_{k=0}^{L-N+1} N a_k + \sum_{k=L-N+2}^{L} (L-k+2) a_k = \sum_{k=0}^{L-N+1} N a_k + (L-(L-N+2)+2) a_{L-N+2} +$$

$$\sum_{k=L-N+3}^{L} ((L+1)-k+1) a_k = \sum_{k=0}^{L-N+2} N a_k + \sum_{k=L-N+3}^{L} ((L+1)-k+1) a_k$$

The last expression is the right-hand side of (48) for L being replaced by $L+1$, thus proving (48). Note that for $k = L-N+2, \ldots, L$ we have $L-k+1 \leq N$, and hence (48) implies that $\sum_{k=0}^{L} c_{k,N} \leq N \sum_{k=0}^{L} a_k$. Since a_k and c_k are nonnegative, and $\sum_{k=0}^{\infty} a_k$ is convergent, the rest of the lemma follows. □

Lemma E.17. *Let* $\mathbf{I}$ *be an index set, and* $\{\beta_{i,t}(k)\}_{k=0}^{\infty}$, $i \in \mathbf{I}$, *be a sequence of matrices such that there exists a sequence* $\{\bar{\beta}(k)\}_{k=0}^{\infty}$ *with* $\|\bar{\beta}\|_{\ell_1} < \infty$ *and* $\bar{\beta}(k) \geq \sup_{i \in \mathbf{I}, t \geq 0} \|\beta_{i,t}(k)\|_2$. *Let* $\mathbf{z}_i(t) = (\beta_{i,t} \star \mathbf{e}_g)(t)$ *and* $\bar{\mathbf{z}}_i(t) = (\beta_{i,t} \star \bar{\mathbf{e}}_g)(t)$. *Then for* $\lambda < 0.5N$

$$\mathbf{E}[e^{\frac{\lambda}{N} \sup_{i \in \mathbf{I}} \sum_{t=0}^{N-1} \|\mathbf{z}_i(t) - \bar{\mathbf{z}}_i(t)\|_2}] \leq \prod_{k=0}^{\infty} \bar{\Psi}_{\mathbf{e}_g} \left(\frac{c(k)\lambda}{N}, \|\bar{\beta}\|_{\ell_1}, C \right) \leq e^{\mathcal{C}(\|\bar{\beta}\|_{\ell_1}, \lambda, N, C)}$$

$$\mathcal{C}(s, \lambda, N, C) = \lambda \frac{\bar{\Psi}_{\mathbf{e}_g}(s) s C}{(e^{sC} - 1)} + \frac{\lambda^2}{N} 2 \bar{\Psi}_{\mathbf{e}_g}(s)$$

where $c(k) = \sum_{l=\max\{0, k-N+1\}}^{k} \|\bar{\beta}(l)\|_2$ *and the infinite product on the right-hand side is well-defined.*

Proof of Lemma E.17. Note that

$$c(k) \leq \|\bar{\beta}\|_{\ell_1} \tag{49}$$

and by Lemma E.16

$$\sum_{k=0}^{\infty} c(k) \leq N \sum_{k=0}^{\infty} \|\bar{\beta}(k)\|_2 = N \|\bar{\beta}\|_{\ell_1} \tag{50}$$

Note now that

$$\begin{aligned}
&\sum_{t=0}^{N-1} \|\mathbf{z}_i(t) - \bar{\mathbf{z}}_i(t)\|_2 \\
&\leq \sum_{t=0}^{N-1} \sum_{k=0}^{\infty} \|\beta_{i,t}(k)\|_2 \|\mathbf{e}_g(t-k) - \bar{\mathbf{e}}_g(t-k)\|_2 \\
&\leq \sum_{t=0}^{N-1} \sum_{k=0}^{\infty} \|\bar{\beta}(k)\|_2 \|\mathbf{e}_g(t-k) \chi(\|\mathbf{e}_g(t-k)\|_2 > C)\|_2 \\
&= \sum_{k=0}^{\infty} c(k) \|\mathbf{e}_g(N-k-1) \chi(\|\mathbf{e}_g(N-k-1)\|_2 > C)\|_2
\end{aligned}$$

Hence,

$$\mathbf{E}[e^{\frac{\lambda}{N}\sup_{i\in\mathbf{I}}\sum_{t=1}^{N}\|\mathbf{z}_i(t)-\bar{\mathbf{z}}_i(t)\|_2}] \leq$$
$$\leq \mathbf{E}[e^{\frac{\lambda}{N}\sum_{k=0}^{\infty}c(k)\|\mathbf{e}_g(N-k-1)\chi(\|\mathbf{e}_g(N-k-1)\|_2>C)\|_2}]$$
$$= \lim_{L\to\infty}\prod_{k=0}^{L}\bar{\Psi}_{\mathbf{e}_g}\left(\frac{c(k)\lambda}{N},\|\bar{\beta}\|_{\ell_1},C\right)$$

where we applied Lemma E.15 to $\mathbf{E}[e^{\frac{\lambda}{N}c(k)\|e_g(N-k-1)\chi(\|\mathbf{e}_g(N-k-1)\|_2>C)\|_2}]$ with $\bar{\lambda}=\|\bar{\beta}\|_{\ell_1}$ and $s=\frac{c(k)\lambda}{N}<0.5\|\bar{\beta}\|_{\ell_1}$ since $\lambda<0.5N$ by assumption. If $\|\bar{\beta}\|_{\ell^1}=0$, the result is trivial. Hence assume $\|\bar{\beta}\|_{\ell^1}>0$.

Note that

$$d_L=\ln\left(\prod_{k=0}^{L}\bar{\Psi}_{\mathbf{e}_g}\left(\frac{c(k)\lambda}{N},\|\bar{\beta}\|_{\ell_1},C\right)\right)=\sum_{k=0}^{L}r_k$$
$$r_k:=\ln\left(1+\frac{c(k)\lambda}{N}\frac{\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})C}{e^{\|\bar{\beta}\|_{\ell_1}C}-1}+\left(\frac{c(k)\lambda}{N}\right)^2\frac{2\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})}{\|\bar{\beta}\|_{\ell_1}^2}\right)\geq 0$$

We show that the series $\sum_{k=0}^{\infty}r_k$ is bounded, more precisely,

$$\sum_{k=0}^{L}r_k\leq\|\bar{\beta}\|_{\ell_1}\lambda\frac{\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})\|C}{e^{\|\bar{\beta}\|_{\ell_1}C}-1}+\frac{\lambda^2}{N}(\|\bar{\beta}\|_{\ell_1})^2\frac{2\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})}{\|\bar{\beta}\|_{\ell_1}^2}=\mathcal{C}(\|\bar{\beta}\|_{\ell_1},\lambda,N,C)$$

From this it follows that d_L is convergent, and from

$$\prod_{k=0}^{L}\bar{\Psi}_{\mathbf{e}_g}\left(\frac{c(k)\lambda}{N},\|\bar{\beta}\|_{\ell_1},C\right)=e^{d_L}$$

it follows that

$$\lim_{L\to\infty}\prod_{k=0}^{L}\bar{\Psi}_{\mathbf{e}_g}\left(\frac{c(k)\lambda}{N},\|\bar{\beta}\|_{\ell_1},C\right)\leq e^{\mathcal{C}(\|\bar{\beta}\|_{\ell_1},\lambda,N,C)} \tag{51}$$

i.e., the infinite product is well-defined, and the statement of the lemma holds.

In order to show (51), note that using $\ln(1+x)\leq x$, it follows that

$$r_k\leq\frac{c(k)\lambda}{N}\frac{\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})\|C}{e^{\|\bar{\beta}\|_{\ell_1}C}-1}+\left(\frac{c(k)\lambda}{N}\right)^2\frac{2\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})}{\|\bar{\beta}\|_{\ell_1}^2}$$

Hence,

$$\sum_{k=0}^{L}r_k\leq\frac{\lambda\sum_{k=0}^{L}c(k)}{N}\frac{\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})\|C}{e^{\|\bar{\beta}\|_{\ell_1}C}-1}+\lambda^2\left(\frac{\sum_{k=0}^{L}(c(k))^2}{N^2}\right)\frac{2\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})}{\|\bar{\beta}\|_{\ell_1}^2}$$

From (49) it follows that $c(k)\leq\|\bar{\beta}\|_{\ell_1}$. Hence

$$\left(\frac{1}{N}c(k)\right)^2\leq\frac{1}{N^2}c(k)\|\bar{\beta}\|_{\ell_1}$$

and as from (50) it follows that $\sum_{k=0}^{L}c(k)\leq N\|\bar{\beta}\|_{\ell_1}$, we get that

$$\sum_{k=0}^{L}\left(\frac{1}{N}c(k)\right)^2\leq\frac{1}{N^2}\sum_{k=0}^{L}c(k)\|\bar{\beta}\|_{\ell_1}\leq\frac{1}{N}\|\bar{\beta}\|_{\ell_1}^2$$

Hence,

$$\sum_{k=0}^{L}r_k\leq\frac{\lambda\sum_{k=0}^{L}c(k)}{N}\frac{\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})\|C}{e^{\|\bar{\beta}\|_{\ell_1}C}-1}+\lambda^2\left(\frac{\sum_{k=0}^{L}(c(k))^2}{N^2}\right)\frac{2\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})}{\|\bar{\beta}\|_{\ell_1}^2}\leq$$
$$\frac{\lambda N\|\bar{\beta}\|_{\ell_1}}{N}\frac{\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})\|C}{e^{\|\bar{\beta}\|_{\ell_1}C}-1}+\lambda^2\frac{1}{N}\|\bar{\beta}\|_{\ell_1}^2\frac{2\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})}{\|\bar{\beta}\|_{\ell_1}^2}=\lambda\|\bar{\beta}\|_{\ell_1}\frac{\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})\|C}{e^{\|\bar{\beta}\|_{\ell_1}C}-1}+\lambda^2\frac{2\bar{\Psi}_{\mathbf{e}_g}(\|\bar{\beta}\|_{\ell_1})}{N}$$

This completes the proof. □

Lemma E.17 provides the basis for several technical results.

Proof of Lemma E.10. From Lemma D.3 it follows that $\mathbf{y}(t) = (\alpha_{g,y} \star \mathbf{e}_g)(t)$ and $\bar{\mathbf{y}}(t) = (\alpha_{g,y} \star \bar{\mathbf{e}}_g)(t)$. By setting $I = \{g\}$, $\beta_{i,t} = \alpha_{g,y}$, $i = g$ and taking $\bar{\beta}(k) = \|\alpha_g(k)\|_2 \geq \|\alpha_{g,y}(k)\|_2$, and using Lemma E.17, the first inequality follows.

Similarly, from Lemma D.3 it follows that $\hat{\mathbf{y}}(t) = (c_\theta \star \mathbf{e}_g)(t)$ and $\hat{\bar{\mathbf{y}}}(t) = (c_\theta \star \bar{\mathbf{e}}_g)(t)$. Note that by definition $c_\theta(k) = \sum_{l=0}^{k} \alpha_\theta(l)\alpha_{g,w}(k-l)$ and hence

$$\|c_\theta(k)\|_2 \leq \sum_{l=0}^{k} \|\alpha_\theta(l)\|_2 \|\alpha_{g,w}(k-l)\|_2 \leq \sum_{l=0}^{k} \bar{\alpha}(l)\|\alpha_g(k-l)\|_2$$

Set

$$\bar{\beta}(k) = \sum_{l=0}^{k} \bar{\alpha}(l)\|\alpha_g(k-l)\|_2$$

from which by the well-known property of Cauchy products of series (the infinite sum of the Cauchy product of two absolutely convergent series is equal to the product of their infinite sums) it follows that

$$\|\bar{\beta}\|_{\ell_1} = \|\bar{\alpha}\|_{\ell_1} \|\alpha_g\|_{\ell_1}.$$

Moreover, $\bar{\beta}(k) \geq \sup_{\theta\in\Theta} \|c_\theta(k)\|_2$. By setting $I = \Theta$, $\beta_{i,t} = c_\theta$, $i = \theta \in \Theta$, and using $\bar{\beta}(k) = \sum_{l=0}^{k} \bar{\alpha}(l)\|\alpha_g(k-l)\|_2$, from Lemma E.17 it then follows that

$$\mathbf{E}[e^{\frac{\lambda}{N} \sup_{\theta\in\Theta} \sum_{t=0}^{N-1} \|\hat{\mathbf{y}}_\theta(t) - \hat{\bar{\mathbf{y}}}_\theta(t)\|_2}] \leq e^{\mathcal{C}(\|\bar{\beta}\|_{\ell_1}, \lambda, N, C)} = e^{\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1}, \lambda, N, C)}$$

Finally, the proof of the third inequality is similar to the second one. By Lemma D.3 it follows that $\hat{\mathbf{y}}_\theta(t \mid 0) = (c_{\theta,t} \star \mathbf{e}_g)(t)$ and $\hat{\bar{\mathbf{y}}}_\theta(t \mid 0) = (c_{\theta,t} \star \bar{\mathbf{e}}_g)(t)$. Notice that

$$\|c_{\theta,t}(k)\|_2 = \| \sum_{l=0}^{\min\{k,t\}} \alpha_\theta(l)\alpha_{g,w}(k-l)\|_2 \leq \sum_{l=0}^{\min\{k,t\}} \|\alpha_\theta(l)\|_2 \|\alpha_{g,w}(k-l)\|_2$$

$$\leq \sum_{l=0}^{k} \bar{\alpha}(l)\|\alpha_{g,w}(k-l)\|_2 \leq \sum_{l=0}^{k} \bar{\alpha}(l)\|\alpha_g(k-l)\|_2 = \bar{\beta}(k)$$

Notice that

$$\bar{\beta}(k) \geq \sup_{\theta\in\Theta, t\geq 0} \|c_{\theta,t}(k)\|_2,$$

and $\|\bar{\beta}\|_{\ell_1} = \|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1}$. By applying Lemma E.17 to $I = \Theta$, $\beta_{i,t} = c_{\theta,t}$ and choosing $\bar{\beta}$ as above, it follows that

$$\mathbf{E}[e^{\frac{\lambda}{N} \sup_{\theta\in\Theta} \sum_{t=0}^{N-1} \|\hat{\mathbf{y}}_\theta(t|0) - \hat{\bar{\mathbf{y}}}_\theta(t|0)\|_2}] \leq e^{\mathcal{C}(\|\bar{\beta}\|_{\ell_1}, \lambda, N, C)} = e^{\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1}, \lambda, N, C)}$$

□

Proof of Lemma E.8. Notice that

$$|\hat{\mathcal{L}}_N(\theta) - \hat{\bar{\mathcal{L}}}_N(\theta)| \leq \frac{1}{N} \sum_{t=0}^{N-1} |\ell(\mathbf{y}(t), \hat{\mathbf{y}}_\theta(t \mid 0)) - \ell(\bar{\mathbf{y}}(t), \hat{\bar{\mathbf{y}}}_\theta(t \mid 0))| \leq$$

$$\frac{L_\ell}{N} \sum_{t=0}^{N-1} \|\mathbf{y}(t) - \bar{\mathbf{y}}(t)\|_2 + \frac{L_\ell}{N} \sum_{t=0}^{N-1} \|\hat{\mathbf{y}}_\theta(t \mid 0) - \hat{\bar{\mathbf{y}}}_\theta(t \mid 0)\|_2$$

and hence

$$\sup_{\theta\in\Theta} |\hat{\mathcal{L}}_N(\theta) - \hat{\bar{\mathcal{L}}}_N(\theta)| \leq \frac{L_\ell}{N} \sum_{t=0}^{N-1} (\|\mathbf{y}(t) - \bar{\mathbf{y}}(t)\|_2 + \frac{L_\ell}{N} \sup_{\theta\in\Theta} \sum_{t=0}^{N-1} \|\hat{\mathbf{y}}_\theta(t \mid 0) - \hat{\bar{\mathbf{y}}}_\theta(t \mid 0)\|_2)$$

Then

$$\mathbf{E}[e^{\lambda \sup_{\theta\in\Theta}|\hat{\mathcal{L}}_N(\theta)-\hat{\bar{\mathcal{L}}}_N(\theta)|}] \le \mathbf{E}[e^{\frac{\lambda L_\ell}{N}\sum_{t=0}^{N-1}\|\mathbf{y}(t)-\bar{\mathbf{y}}(t)\|_2} e^{\frac{\lambda L_\ell}{N}\sup_{\theta\in\Theta}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t|0)-\hat{\bar{\mathbf{y}}}_\theta(t|0)\|_2}] \le$$
$$\sqrt{\mathbf{E}[e^{\frac{2\lambda L_\ell}{N}\sum_{t=0}^{N-1}(\|\mathbf{y}(t)-\bar{\mathbf{y}}(t)\|_2}]}\sqrt{\mathbf{E}[e^{\frac{2\lambda L_\ell}{N}\sup_{\theta\in\Theta}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t|0)-\hat{\bar{\mathbf{y}}}_\theta(t|0)\|_2}]}$$

from which (38) follows. Similarly,

$$|V_N(\theta)-\bar{V}_N(\theta)| \le \frac{1}{N}\sum_{t=0}^{N-1}|\ell(\mathbf{y}(t),\hat{\mathbf{y}}_\theta(t))-\ell(\bar{\mathbf{y}}(t),\hat{\bar{\mathbf{y}}}_\theta(t))| \le$$
$$\frac{L_\ell}{N}\sum_{t=0}^{N-1}\|\mathbf{y}(t)-\bar{\mathbf{y}}(t)\|_2 + \frac{L_\ell}{N}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t)-\hat{\bar{\mathbf{y}}}_\theta(t)\|_2$$

and hence

$$\sup_{\theta\in\Theta}|V_N(\theta)-\bar{V}_N(\theta)| \le \frac{L_\ell}{N}\sum_{t=0}^{N-1}(\|\mathbf{y}(t)-\bar{\mathbf{y}}(t)\|_2 + \frac{L_\ell}{N}\sup_{\theta\in\Theta}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t)-\hat{\bar{\mathbf{y}}}_\theta(t)\|_2)$$

so that

$$\mathbf{E}[e^{\lambda \sup_{\theta\in\Theta}|V_N(\theta)-\bar{V}_N(\theta)|}] \le \mathbf{E}[e^{\frac{\lambda L_\ell}{N}\sum_{t=0}^{N-1}\|\mathbf{y}(t)-\bar{\mathbf{y}}(t)\|_2} e^{\frac{\lambda L_\ell}{N}\sup_{\theta\in\Theta}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t)-\hat{\bar{\mathbf{y}}}_\theta(t)\|_2}] \le$$
$$\sqrt{\mathbf{E}[e^{\frac{2\lambda L_\ell}{N}\sum_{t=0}^{N-1}(\|\mathbf{y}(t)-\bar{\mathbf{y}}(t)\|_2}]}\sqrt{\mathbf{E}[e^{\frac{2\lambda L_\ell}{N}\sup_{\theta\in\Theta}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t)-\hat{\bar{\mathbf{y}}}_\theta(t)\|_2}]}$$

from which (39) follows. □

Proof of Lemma E.11. From Lemma E.10 it follows that

$$\sqrt{\mathbf{E}[e^{\frac{2\lambda L_\ell}{N}\sum_{t=0}^{N-1}\|\mathbf{y}(t)-\bar{\mathbf{y}}(t)\|_2}]} \le e^{\frac{1}{2}\mathcal{C}(\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N,C)} \tag{52}$$

and

$$\sqrt{\mathbf{E}[e^{\frac{2\lambda L_\ell}{N}\sup_{\theta\in\Theta}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t|0)-\hat{\bar{\mathbf{y}}}_\theta(t|t_0)\|}]} \le e^{\frac{1}{2}\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N,C)} \tag{53}$$

from which by using (38) it follows that

$$\mathbf{E}[e^{\lambda \sup_{\theta\in\Theta}|\hat{\mathcal{L}}_N(\theta)-\hat{\bar{\mathcal{L}}}_N(\theta)|}] \le e^{\frac{1}{2}(\mathcal{C}(\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N,C)+\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N,C))}$$

Likewise,from Lemma E.10 it follows that

$$\sqrt{\mathbf{E}[e^{\frac{2\lambda L_\ell}{N}\sup_{\theta\in\Theta}\sum_{t=0}^{N-1}\|\hat{\mathbf{y}}_\theta(t)-\hat{\bar{\mathbf{y}}}_\theta(t)\|_2}]} \le e^{\frac{1}{2}\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N,C)}$$

By combining this with (52), it follows by (39) that

$$\mathbf{E}[e^{\lambda \sup_{\theta\in\Theta}|V_N(\theta)-\bar{V}_N(\theta)|}] \le e^{\frac{1}{2}(\mathcal{C}(\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N,C)+\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N,C))} \tag{54}$$

Finally, as $\mathcal{L}(\theta)=\mathbf{E}[V_N(\theta)]$ and $\bar{\mathcal{L}}(\theta)=\mathbf{E}[\bar{V}_N(\theta)]$, it follows that

$$\sup_{\theta\in\Theta}|\mathcal{L}(\theta)-\bar{\mathcal{L}}(\theta)| = \sup_{\theta\in\Theta}|\mathbf{E}[V_N(\theta)-\bar{V}_N(\theta)]|$$
$$\le \sup_{\theta\in\Theta}\mathbf{E}[|V_N(\theta)-\bar{V}_N(\theta)|] \le \mathbf{E}[\sup_{\theta\in\Theta}|V_N(\theta)-\bar{V}_N(\theta)|]$$

and hence

$$e^{\lambda \sup_{\theta\in\Theta}|\mathcal{L}(\theta)-\bar{\mathcal{L}}(\theta)|} \le e^{\lambda\mathbf{E}[\sup_{\theta\in\Theta}|V_N(\theta)-\bar{V}_N(\theta)|]} \tag{55}$$

As the function $\phi: x\mapsto e^{\lambda x}$ is convex, by Jensen's inequality it follows that

$$e^{\lambda\mathbf{E}[\sup_{\theta\in\Theta}|V_N(\theta)-\bar{V}_N(\theta)|]} = \phi(\mathbf{E}[\sup_{\theta\in\Theta}|V_N(\theta)-\bar{V}_N(\theta)|]) \le$$
$$\mathbf{E}[\phi(\sup_{\theta\in\Theta}|V_N(\theta)-\bar{V}_N(\theta)|)] = \mathbf{E}[e^{\lambda \sup_{\theta\in\Theta}|V_N(\theta)-\bar{V}_N(\theta)|}]$$

and then the statement of the lemma follows from (54) and (55). □

Proof of Lemma E.12. Note that

$$\epsilon(V_N(\theta)-\hat{\mathcal{L}}_N(\theta)) = \epsilon(\hat{\bar{\mathcal{L}}}_N(\theta)-\hat{\mathcal{L}}_N(\theta)) + \epsilon(\bar{V}_N(\theta)-\hat{\bar{\mathcal{L}}}_N(\theta)) + \epsilon(V_N(\theta)-\bar{V}_N(\theta))$$
$$\leq \sup_{\theta\in\Theta}|\hat{\mathcal{L}}_N(\theta)-\hat{\bar{\mathcal{L}}}_N(\theta)| + \epsilon(\bar{V}_N(\theta)-\hat{\bar{\mathcal{L}}}_N(\theta)) + \sup_{\theta\in\Theta}|\bar{V}_N(\theta)-V_N(\theta)|$$

Hence, by using monotonicity of $x\mapsto e^x$ and Cauchy-Schwarz inequality for expectations,

$$\mathbf{E}[e^{\lambda\epsilon(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}] \leq \mathbf{E}[e^{\lambda\sup_{\theta\in\Theta}|\bar{V}_N(\theta)-V_N(\theta)|}e^{\lambda\sup_{\theta\in\Theta}|\hat{\bar{\mathcal{L}}}_N(\theta)-\hat{\mathcal{L}}_N(\theta)|}e^{\lambda\epsilon(\bar{V}_N(\theta)-\hat{\bar{\mathcal{L}}}_N(\theta))}]$$
$$\leq \sqrt{\mathbf{E}[e^{2\lambda\epsilon(\bar{V}_N(\theta)-\hat{\bar{\mathcal{L}}}_N(\theta))}]}\sqrt{\mathbf{E}[e^{2\lambda\sup_{\theta\in\Theta}|\bar{V}_N(\theta)-V_N(\theta)|}e^{2\lambda\sup_{\theta\in\Theta}|\hat{\bar{\mathcal{L}}}_N(\theta)-\hat{\mathcal{L}}_N(\theta)|}]}$$
$$\leq \sqrt{\mathbf{E}[e^{2\lambda\epsilon(\bar{V}_N(\theta)-\hat{\bar{\mathcal{L}}}_N(\theta))}]}\sqrt{\sqrt{\mathbf{E}[e^{4\lambda\sup_{\theta\in\Theta}|\bar{V}_N(\theta)-V_N(\theta)|}]}\sqrt{\mathbf{E}[e^{4\lambda\sup_{\theta\in\Theta}|\hat{\bar{\mathcal{L}}}_N(\theta)-\hat{\mathcal{L}}_N(\theta)|}]}}$$

proving (42). A similar argument proves (41). □

Proof of Lemma E.13. From Lemma E.11 it follows that

$$\sqrt{\mathbf{E}[e^{4\lambda\sup_{\theta\in\Theta}|\bar{V}_N(\theta)-V_N(\theta)|}]} \leq e^{\frac{1}{4}(\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},8L_\ell\lambda,N,C)+\mathcal{C}(\|\alpha_g\|_{\ell_1},8L_\ell\lambda,N,C))}$$
$$\sqrt{\mathbf{E}[e^{4\lambda\sup_{\theta\in\Theta}|\hat{\bar{\mathcal{L}}}_N(\theta)-\hat{\mathcal{L}}_N(\theta)|}]} \leq e^{\frac{1}{4}(\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},8L_\ell\lambda,N,C)+\mathcal{C}(\|\alpha_g\|_{\ell_1},8L_\ell\lambda,N,C))}$$

and hence

$$\sqrt{\sqrt{\mathbf{E}[e^{4\lambda\sup_{\theta\in\Theta}|\bar{V}_N(\theta)-V_N(\theta)|}]}\sqrt{\mathbf{E}[e^{4\lambda\sup_{\theta\in\Theta}|\hat{\bar{\mathcal{L}}}_N(\theta)-\hat{\mathcal{L}}_N(\theta)|}]}}$$
$$\leq e^{\frac{1}{4}(\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},8L_\ell\lambda,N,C)+\mathcal{C}(\|\alpha_g\|_{\ell_1},8L_\ell\lambda,N,C))}$$

By using (42) the first statement of the lemma follows. The proof of the second claim is similar. □

E.3 Proof of Theorem 3.1: quadratic loss

The core of the proof is to show that the prediction errors $\|\mathbf{y}(t)-\hat{\mathbf{y}}_\theta(t)\|_2$ and $\|\mathbf{y}(t)-\hat{\mathbf{y}}_\theta(t\mid 0)\|_2$ are sub-Gaussian and hence they are bounded with high probability as follows.

Lemma E.18. *For $N\geq 2$ and $\delta\in(0,0.5)$*

$$\mathbf{P}\left(\forall\theta\in\Theta,\ \forall t=0,\dots,N-1:\right.$$
$$\left.\max\{\|\mathbf{y}(t)-\hat{\mathbf{y}}_\theta(t)\|_2, \|\mathbf{y}(t)-\hat{\mathbf{y}}_\theta(t\mid 0)\|_2\} \leq 6G_e(\|\mathbf{e}_g\|_{\psi_2})\sqrt{\ln\left(\frac{3N^2}{2\delta}\right)}\right) \geq 1-\frac{2\delta}{3} \tag{56}$$

and

$$\mathbf{E}[\|\mathbf{y}(t)-\hat{\mathbf{y}}_\theta(t)\|_2^2\chi(\|\mathbf{y}(t)-\hat{\mathbf{y}}_\theta(t)\|_2^2 > (6G_e\|\mathbf{e}_g\|_{\psi_2})^2\ln\left(\frac{3N^2}{2\delta}\right))] \leq \frac{2\cdot 6^2G_e^2\|\mathbf{e}_g\|_{\psi_2}^2\ln(\frac{3N^2}{2\delta})}{N} \tag{57}$$

We replace the quadratic loss by the globally Lipschitz truncated loss ℓ_C, which coincides with the quadratic loss whenever the prediction error is bounded by C. More precisely, let us define the loss function ℓ_C as follows:

$$\ell_C(y,\hat{y}) = \min\{\|y-\hat{y}\|_2^2, C^2\}$$

Lemma E.19. *For any $C>0$, ℓ_C is $2C$-Lipschitz.*

The proof then reduces the quadratic case to the case of Lipschitz loss ℓ_C for a suitable C as follows.

Proof of Theorem 3.1:quadratic case. Let us set $C = 6G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{\ln(\frac{3N^2}{2\delta})}$ and let us consider the true and empirical errores for the loss function ℓ_C:

$${}^C\hat{\mathcal{L}}_N(\theta) = \frac{1}{N}\sum_{t=0}^{N-1}\ell_C(y(t),\hat{y}_\theta(t\mid 0)), \quad {}^C\mathcal{L}(\theta) = \mathbf{E}[\ell_C(y(t),\hat{y}_\theta(t))] \tag{58}$$

By applying Theorem 3.1 for the case of Lipschitz loss function ℓ_C and with $\frac{2\delta}{3}$ instead of δ, it follows that for any fixed $\epsilon \in \{-1, 1\}$, for all $0 < \lambda < \frac{N}{32 \cdot 2C}$

$$\begin{aligned}\forall \rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}{}^C\mathcal{L}(\theta) \leq \epsilon E_{\theta\sim\rho}{}^C\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\left(\mathrm{KL}(\rho\|\pi) + \ln\left(\frac{3}{2\delta}\right) + \right.\\ \left.\frac{1}{2}\left(\ln E_{\theta\sim\pi}\exp(\bar{\mathcal{C}}_1(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon)) + \ln E_{\theta\sim\pi}\exp(\bar{\mathcal{C}}_2(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon))\right)\right)\end{aligned} \tag{59}$$

holds with probability at least $1 - \frac{4\delta}{3}$, where for $i = 1, 2$.

$$\bar{C}_i(\lambda,\theta,N,\frac{2\delta}{3},\epsilon) = \mathcal{C}_{ub,i}(\lambda,\theta,12G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{\ln\left(\frac{3N^2}{2\delta}\right)},N) \tag{60}$$

By using Lemma E.18, (59) and the union bound, it follows that

$$\begin{aligned}\forall \rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}{}^C\mathcal{L}(\theta) \leq \epsilon E_{\theta\sim\rho}{}^C\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\left(\mathrm{KL}(\rho\|\pi) + \ln\left(\frac{3}{2\delta}\right) + \right.\\ \left.\frac{1}{2}\left(\ln E_{\theta\sim\pi}\exp(\bar{\mathcal{C}}_1(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon)) + \ln E_{\theta\sim\pi}\exp(\bar{\mathcal{C}}_2(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon))\right)\right)\end{aligned} \tag{61}$$

and

$$\forall\theta\in\Theta,\ \forall t = 0,\ldots,N-1: \quad \|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2 \leq C, \|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t \mid 0)\|_2 \leq C$$

holds with probability at least $1 - (\frac{4\delta}{3} + \frac{2\delta}{3}) = 1 - 2\delta$ over data.

Notice that if $\omega \in \Omega$ is such that

$$\forall\theta\in\Theta, t = 0,\ldots,N-1: \quad \|\mathbf{y}(t)(\omega) - \hat{\mathbf{y}}_\theta(t)(\omega)\|_2 \leq C, \|\mathbf{y}(t)(\omega) - \hat{\mathbf{y}}_\theta(t \mid 0)(\omega)\|_2 \leq C$$

then $\ell_C(\mathbf{y}(t)(\omega), \hat{\mathbf{y}}_\theta(t \mid 0)(\omega)) = \|\mathbf{y}(t)(\omega) - \hat{\mathbf{y}}_\theta(t \mid 0)(\omega)\|_2^2$ and hence

$$\hat{\mathcal{L}}_N(\theta)(\omega) = {}^C\hat{\mathcal{L}}_N(\theta)(\omega)$$

Therefore, the event defined by (61) equals the event defined by

$$\begin{aligned}\forall \rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}{}^C\mathcal{L}(\theta) \leq \epsilon E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{3}{2\delta} + \right.\\ \left.\frac{1}{2}\left(\ln E_{\theta\sim\pi}\exp(\bar{\mathcal{C}}_1(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon)) + \ln E_{\theta\sim\pi}\exp(\bar{\mathcal{C}}_2(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon))\right)\right)\end{aligned} \tag{62}$$

and

$$\forall\theta\in\Theta,\ \forall t = 0,\ldots,N-1: \quad \|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2 \leq C, \|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t \mid 0)\|_2 \leq C$$

For $\epsilon = -1$ we will show (62) implies

$$\begin{aligned}\forall \rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}\mathcal{L}(\theta) \leq \epsilon E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\left(\mathrm{KL}(\rho\|\pi) + \ln\left(\frac{3}{2\delta}\right) + \right.\\ \left.\frac{1}{2}\left(\ln E_{\theta\sim\pi}\exp(\bar{\mathcal{C}}_1(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon)) + \ln E_{\theta\sim\pi}\exp(\bar{\mathcal{C}}_2(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon))\right)\right)\end{aligned} \tag{63}$$

and as (62) holds with probability at least $1 - 2\delta$, it follows that (63) also holds with probability at least $1 - 2\delta$. To this end, notice that

$$\begin{aligned}\mathcal{L}(\theta) = \mathbf{E}[\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2^2] = \mathbf{E}[\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2^2\chi(\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2 \leq C)] + \\ \mathbf{E}[\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2^2\chi(\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2 > C)]\end{aligned} \tag{64}$$

and that $\mathbf{E}[\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2^2\chi(\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2 > C)] \geq C^2\mathbf{P}(\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2 > C)$. Hence

$$\begin{aligned}\mathcal{L}(\theta) &\geq \mathbf{E}[\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2^2\chi(\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2 \leq C)] + C^2\mathbf{P}(\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2 > C) \\ &= \mathbf{E}[\ell_C(\mathbf{y}(t), \hat{\mathbf{y}}_\theta(t))] = {}^C\mathcal{L}(\theta)\end{aligned} \tag{65}$$

i.e., $\epsilon\mathcal{L}(\theta) \leq \epsilon^C\mathcal{L}(\theta)$ for $\epsilon = -1$, from which it follows that (62) implies (63).

For $\epsilon = 1$, we will show that (62) implies

$$\forall \rho \in \mathcal{M}_\pi : \epsilon E_{\theta\sim\rho}\mathcal{L}(\theta) \leq \epsilon E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{72G_e^2\|\mathbf{e}_g\|_{\psi_2}^2 \ln\left(\frac{3N^2}{2\delta}\right)}{N} + \frac{1}{\lambda}\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{3}{2\delta} + \frac{1}{2}\left(\ln E_{\theta\sim\pi}\exp(\bar{\mathcal{C}}_1(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon)) + \ln E_{\theta\sim\pi}\exp(\bar{\mathcal{C}}_2(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon))\right)\right) \tag{66}$$

First note that

$$\frac{72G_e^2\|\mathbf{e}_g\|_{\psi_2}^2 \ln\left(\frac{3N^2}{2\delta}\right)}{N} = \frac{1}{\lambda}\ln\left(\exp\left(\lambda\frac{72G_e^2\|\mathbf{e}_g\|_{\psi_2}^2 \ln\left(\frac{3N^2}{2\delta}\right)}{N}\right)\right)$$

and hence after algebraic manipulations

$$\frac{72G_e^2\|\mathbf{e}_g\|_{\psi_2}^2 \ln\left(\frac{3N^2}{2\delta}\right)}{N} + \frac{1}{\lambda}\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{3}{2\delta} + \frac{1}{2}\left(\ln E_{\theta\sim\pi}\exp(\bar{\mathcal{C}}_1(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon)) + \ln E_{\theta\sim\pi}\exp(\bar{\mathcal{C}}_2(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon))\right)\right) \leq \frac{1}{\lambda}\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{3}{2\delta} + \frac{1}{2}\left(\ln E_{\theta\sim\pi}\exp(\mathcal{C}_1(2\lambda,\theta,N,\delta,\epsilon)) + \ln E_{\theta\sim\pi}\exp(\mathcal{C}_2(2\lambda,\theta,N,\delta,\epsilon))\right)\right)$$

where for $i = 1, 2$,

$$\mathcal{C}_i(\lambda,\theta,N,\delta,\epsilon) = \lambda\frac{72G_e^2\|\mathbf{e}_g\|_{\psi_2}^2 \ln\left(\frac{3N^2}{2\delta}\right)}{N} + \mathcal{C}_{ub,i}(\lambda,\theta,12G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{\ln\left(\frac{3N^2}{2\delta}\right)},N)$$

To show that (62) implies (66), notice that

$$\begin{aligned}
&\mathcal{L}(\theta) = \mathbf{E}[\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2^2] = \\
&\mathbf{E}[\ell_C(\mathbf{y}(t),\hat{\mathbf{y}}_\theta(t))\chi(\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2 \leq C)] + \mathbf{E}[\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2^2\chi(\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2 > C)] \leq \\
&\mathbf{E}[\ell_C(\mathbf{y}(t),\hat{\mathbf{y}}_\theta(t))] + \mathbf{E}[\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2^2\chi(\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2 > C)] \\
&= {}^C\mathcal{L}(\theta) + \mathbf{E}[\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2^2\chi(\|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2 > C)]
\end{aligned} \tag{67}$$

Using (57) it follows that

$$\mathcal{L}(\theta) \leq {}^C\mathcal{L}(\theta) + 72G_e^2\|\mathbf{e}_g\|_{\psi_2}^2\frac{\ln\left(\frac{3N^2}{2\delta}\right)}{N} \tag{68}$$

Hence, by adding $72G_e^2\|\mathbf{e}_g\|_{\psi_2}^2\frac{\ln\left(\frac{3N^2}{2\delta}\right)}{N}$ to both sides of the inequality (62), it follows that (66) is implied by (62). Combining the two case it follows that for quadratic loss and $\epsilon \in \{-1, 1\}$, the inequality (8) also holds with probability at least $1 - 2\delta$ over data. □

We will conclude by presenting the proof of Lemma E.18-E.19. To this end, we need the following technical result.

Lemma E.20. *Let* $\mathbf{I}$ *be an index set,* $\{\beta_{i,t}(k)\}_{k=0}^\infty$*,* $i \in \mathbf{I}$*, be a sequence of matrices such that* $\|\bar{\beta}\|_{\ell_1} < \infty$*, where* $\bar{\beta}(k) = \sup_{i\in\mathbf{I},t\geq 0}\|\beta_{i,t}(k)\|_2$*, and let* $\mathbf{z}_i(t) = (\beta_{i,t} \star \mathbf{e}_g)(t)$*. Then* $\|\mathbf{z}_i(t)\|_2$*,* $i \in \mathbb{I}$ *and* $\mathbf{z}(t) = \sup_{i\in\mathbf{I}}\|\mathbf{z}_i(t)\|_2$ *are sub-Gaussian with sub-Gaussian norm*

$$\|\|\mathbf{z}_i(t)\|_2\|_{\psi_2} \leq \|\mathbf{z}(t)\|_{\psi_2} \leq 6\|\bar{\beta}\|_{\ell_1}\|\mathbf{e}_g\|_{\psi_2} \tag{69}$$

and the moments and the moment generating function of $\mathbf{z}^2(t) = \sup_{i\in\mathbf{I}}\|\mathbf{z}_i(t)\|_2^2$ *satisfy*

$$\mathbf{E}[\mathbf{z}^{2r}(t)] \leq \|\bar{\beta}\|_{\ell_1}^{2r}(3\|\mathbf{e}_g\|_{\psi_2})^{2r}2r! \tag{70}$$

$$\mathbf{E}[e^{\lambda\mathbf{z}^2(t)}] \leq \frac{1 + \lambda\|\bar{\beta}\|_{\ell_1}^2(3\|\mathbf{e}_g\|_{\psi_2})^2}{1 - \lambda\|\bar{\beta}\|_{\ell_1}^2(3\|\mathbf{e}_g\|_{\psi_2})^2}, \quad |\lambda| < \frac{1}{\|\bar{\beta}\|_{\ell_1}^2(3\|\mathbf{e}_g\|_{\psi_2})^2} \tag{71}$$

Moreover, for any $C > 0$,

$$\mathbf{E}[\|\mathbf{z}_i(t)\|_2^2 \chi(\|\mathbf{z}_i(t)\|_2^2 > C^2)] \leq \frac{2C^2}{e^{\frac{C^2}{(6\|\bar{\beta}\|_{\ell_1}\|\mathbf{e}_g\|_{\psi_2})^2}} - 1} \tag{72}$$

Proof Lemma E.20. Notice that

$$\begin{aligned}\mathbf{z}(t) &= \sup_{i\in\mathbf{I}} \|\sum_{k=0}^{\infty} \beta_{i,t}(k)\mathbf{e}_g(t-k)\|_2 \\ &\leq \sup_{i\in\mathbf{I}} \sum_{k=0}^{\infty} \|\beta_{i,t}(k)\|_2 \|\mathbf{e}_g(t-k)\|_2 \leq \sum_{k=0}^{\infty} \|\bar{\beta}(k)\|_2 \|\mathbf{e}_g(t-k)\|_2\end{aligned}$$

Hence, for any $r \geq 1$, by using the arithmetic-geometric mean inequality we get $a_1 \cdots a_k \leq \frac{1}{k}\sum_{i=1}^{k} a_i^k$, $a_1, \ldots, a_k \geq 0$,

$$\mathbf{z}^{2r}(t) \leq \left(\sum_{k=0}^{\infty} \|\bar{\beta}(k)\|_2 \|\mathbf{e}_g(t-k)\|_2\right)^{2r} = \sum_{k_1,\ldots,k_{2r}=0}^{\infty} \prod_{j=1}^{2r} \|\bar{\beta}(k_j)\|_2 \|\mathbf{e}_g(t-k_j)\|_2 \leq$$
$$\sum_{k_1,\ldots,k_{2r}=0}^{\infty} \prod_{j=1}^{2r} \|\bar{\beta}(k_j)\|_2 \frac{1}{2r} \sum_{j=1}^{2r} \|\mathbf{e}_g(t-k_j)\|_2^{2r}$$

By taking expectations, using the fact that as $\|\mathbf{e}_g(t)\|_2$ is sub-Gaussian, and the proof of [43, Proposition 2.6.1] we obtain

$$\mathbf{E}[\|\mathbf{e}_g(t-k_j)\|_2^{2r}] = \mathbf{E}[\|\mathbf{e}_g(t)\|_2^{2r}] \leq 2r(r-1)!(3\|\mathbf{e}_g\|_{\psi_2})^{2r} = 2r!(3\|\mathbf{e}_g\|_{\psi_2})^{2r} \tag{73}$$

it follows that

$$\mathbf{E}[\mathbf{z}^{2r}(t)] \leq \left(\sum_{k=0}^{\infty} \|\bar{\beta}(k)\|_2\right)^{2r} 2r!(3\|\mathbf{e}_g(t)\|_{\psi_2})^{2r} = \|\bar{\beta}\|_{\ell_1}^{2r} 2r!(3\|\mathbf{e}_g\|_{\psi_2})^{2r}$$

i.e., (70) holds. By noticing that

$$\mathbf{E}[e^{\lambda \mathbf{z}^2(t)}] = 1 + \sum_{r=1}^{\infty} \frac{\lambda^r \mathbf{E}[\mathbf{z}^{2r}(t)]}{r!} \leq 1 + \sum_{r=1}^{\infty} \frac{\lambda^r \|\bar{\beta}\|_{\ell_1}^{2r} (3\|\mathbf{e}_g\|_{\psi_2})^{2r} 2r!}{r!} =$$
$$1 + 2\lambda\|\bar{\beta}\|_{\ell_1}^2 (3\|\mathbf{e}_g\|_{\psi_2})^2 \sum_{r=0}^{\infty} (\lambda\|\bar{\beta}\|_{\ell_1}^2 (3\|\mathbf{e}_g\|_{\psi_2})^2)^r = 1 + \frac{2\lambda\|\bar{\beta}\|_{\ell_1}^2 (3\|\mathbf{e}_g\|_{\psi_2})^2}{1 - \lambda\|\bar{\beta}\|_{\ell_1}^2 (3\|\mathbf{e}_g\|_{\psi_2})^2} =$$
$$\frac{1 + \lambda\|\bar{\beta}\|_{\ell_1}^2 (3\|\mathbf{e}_g\|_{\psi_2})^2}{1 - \lambda\|\bar{\beta}\|_{\ell_1}^2 (3\|\mathbf{e}_g\|_{\psi_2})^2}$$

it follows that (71) holds. Finally, for[3] $\lambda = \frac{1}{4\|\bar{\beta}\|_{\ell_1}^2 (3\|\mathbf{e}_g\|_{\psi_2})^2}$, the bound becomes

$$\mathbf{E}[e^{\lambda \mathbf{z}^2(t)}] \leq \frac{1 + \frac{1}{4}}{1 - \frac{1}{4}} = \frac{5/4}{3/4} = \frac{5}{3} \leq 2$$

from which by the definition of sub-Gaussian norm of $\mathbf{z}(t)$ it follows

$$\|\mathbf{z}(t)\|_{\psi_2} \leq \sqrt{\frac{1}{\lambda}} = 6\|\bar{\beta}\|_{\ell_1}\|\mathbf{e}_g\|_{\psi_2}$$

i.e., (69) holds. In particular, the sub-Gaussian norm of $\mathbf{z}(t)$ is bounded and hence by [43, Proposition 2.6.1] $\mathbf{z}(t)$ is sub-Gaussian. Moreover, as $\mathbf{z}^2(t) \geq \|\mathbf{z}_i(t)\|_2^2$, it follows that the sub-Gaussian norm of $\|\mathbf{z}_i(t)\|_2$ is also bounded and it is also sub-Gaussian.

[3] If $\|\bar{\beta}\|_{\ell^1} = 0$, the result is trivial; hence assume $\|\bar{\beta}\|_{\ell^1} > 0$.

Finally, from Lemma E.14 applied to $\|\mathbf{z}_i(t)\|_2^2$ it follows that

$$\mathbf{E}[\|\mathbf{z}_i(t)\|_2^2 \chi(\|\mathbf{z}_i(t)\|_2^2 > C^2)] \leq \frac{\mathbf{E}[e^{\bar{\lambda}\|\mathbf{z}_i(t)\|_2^2}]C^2}{e^{C^2\bar{\lambda}} - 1} \tag{74}$$

if $\bar{\lambda}$ is such that the moment generating function $\mathbf{E}[e^{s\|\mathbf{z}_i(t)\|_2^2}]$ is well-defined on $[0, \bar{\lambda}]$. Note that $\|\mathbf{z}_i(t)\|_2^2 \leq \mathbf{z}^2(t)$ and hence $\mathbf{E}[e^{s\|\mathbf{z}_i(t)\|^2}] \leq \mathbf{E}[e^{s\mathbf{z}^2(t)}]$. Now $\bar{\lambda}$ can be chosen as $\bar{\lambda} = \frac{1}{4\|\bar{\beta}\|_{\ell_1}^2 (3\|\mathbf{e}_g\|_{\psi_2})^2}$ and hence $\mathbf{E}[e^{\bar{\lambda}\|\mathbf{z}_i(t)\|^2}] \leq \mathbf{E}[e^{\bar{\lambda}\mathbf{z}^2(t)}] \leq 2$. Combining (74) and the bound on $\mathbf{E}[e^{\bar{\lambda}\|\mathbf{z}_i(t)\|^2}]$, we obtain (72). □

Proof of Lemma E.18. Let $\mathbb{I} = \Theta \times \{1, 2\}$, $\mathbf{z}_{(\theta,1)}(t) = \mathbf{y}(t) - \mathbf{y}_\theta(t)$, $\mathbf{z}_{(\theta,2)}(t) = \mathbf{y}(t) - \mathbf{y}_\theta(t \mid 0)$, $\beta_{(\theta,1),t} = \alpha_{g,y} - c_\theta$, and $\beta_{(\theta,2),t} = \alpha_{g,y} - c_{\theta,t}$. Note that from the proof of Lemma D.3 it follows that

$$\begin{aligned}
\|(\alpha_{g,y} - c_\theta)(k)\|_2 &\leq \|\alpha_{g,y}(k)\|_2 + \sum_{l=0}^{k} \|\alpha_\theta(l)\|_2 \|\alpha_{g,w}(k-l)\|_2 \\
&\leq \|\alpha_{g,y}(k)\|_2 + \sum_{l=0}^{k} \bar{\alpha}(l) \|\alpha_{g,w}(k-l)\|_2 \\
\|(\alpha_{g,y} - c_{\theta,t})(k)\|_2 &\leq \|\alpha_{g,y}(k)\|_2 + \sum_{l=0}^{\min\{k,t\}} \|\alpha_\theta(l)\|_2 \|\alpha_{g,w}(k-l)\|_2 \\
&\leq \|\alpha_{g,y}(k)\|_2 + \sum_{l=0}^{k} \bar{\alpha}(l) \|\alpha_{g,w}(k-l)\|_2.
\end{aligned}$$

Hence $\bar{\beta}(k) = \sup_{i\in\mathbb{I}, t\geq 0} \|\beta_{i,t}(k)\|_2 \leq \|\alpha_{g,y}(k)\|_2 + \sum_{l=0}^{k} \bar{\alpha}(l)\|\alpha_{g,w}(k-l)\|_2$ and

$$\|\bar{\beta}\|_{\ell_1} \leq \sum_{k=0}^{\infty} \|\alpha_{g,y}(k)\|_2 + \sum_{k=0}^{\infty}\sum_{l=0}^{k} \bar{\alpha}(l)\|\alpha_{g,w}(k-l)\|_2 \leq \|\alpha_g\|_{\ell_1} + \|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1} = G_e$$

From this and Lemma E.20 and Lemma D.3 it follows that for any $t = 0, \ldots, N-1$ $\sup_{\theta\in\Theta} \max\{\|\mathbf{y}(t) - \mathbf{y}_\theta(t)\|_2, \|\mathbf{y}(t) - \mathbf{y}_\theta(t \mid 0)\|_2\}$ are sub-Gaussian random variable whose sub-Gaussian norms are upper bounded by $6G_e\|\mathbf{e}_g\|_{\psi_2}$. Hence, using the sub-Gaussian tail bound (proof of (iii) $\Longrightarrow$ (i) in [43, Proposition 2.6.1]), we obtain the desired result

$$\begin{aligned}
\mathbf{P}\left(\sup_{\theta\in\Theta} \max\{\|\mathbf{y}(t) - \mathbf{y}_\theta(t)\|_2, \|\mathbf{y}(t) - \mathbf{y}_\theta(t \mid 0)\|_2\} \geq 6G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{\ln(\frac{3N^2}{2\delta})}\right) \leq \\
= 2e^{-\ln(\frac{3N^2}{2\delta})} = \frac{2}{N}\frac{2\delta}{3N} \leq \frac{2\delta}{3N}
\end{aligned}$$

Therefore, by applying a union bound over $t = 0, \ldots, N-1$, the first inequality (56) of the lemma follows.

In order to obtain the second inequality, we use (72), and notice that

$$\begin{aligned}
\frac{2C^2}{e^{\frac{C^2}{(6G_e\|\mathbf{e}_g\|_{\psi_2})^2}} - 1} &= \frac{2(6G_e\|\mathbf{e}_g\|_{\psi_2})^2 \ln(\frac{3N^2}{2\delta})}{e^{\ln(\frac{3N^2}{2\delta})} - 1} = \frac{2(6G_e\|\mathbf{e}_g\|_{\psi_2})^2 \ln(\frac{3N^2}{2\delta}) 2\delta}{3N^2 - 2\delta} \\
&\leq 72G_e^2\|\mathbf{e}_g\|_{\psi_2}^2 \frac{\ln(\frac{3N^2}{2\delta})}{N}
\end{aligned}$$

where we used that $2\delta \leq 1$, $3N^2 - 2\delta \geq N^2 \geq N$, i.e., (57). □

Proof of Lemma E.19. Notice that

$$|\ell_C(y_1, \hat{y}_1) - \ell_C(y_2, \hat{y}_2)| = \begin{cases}
|\|y_1 - \hat{y}_1\|_2^2 - \|y_2 - \hat{y}_2\|_2^2| & \|y_1 - \hat{y}_1\|_2^2 \leq C^2, \quad \|y_2 - \hat{y}_2\|_2^2 \leq C^2 \\
|\|y_2 - \hat{y}_2\|_2^2 - C^2| & \|y_1 - \hat{y}_1\|_2^2 > C^2, \quad \|y_2 - \hat{y}_2\|_2^2 \leq C^2 \\
|\|y_1 - \hat{y}_1\|_2^2 - C^2| & \|y_2 - \hat{y}_2\|_2^2 > C^2, \quad \|y_1 - \hat{y}_1\|_2^2 \leq C^2 \\
0 & \|y_1 - \hat{y}_1\|_2^2 > C^2, \quad \|y_2 - \hat{y}_2\|_2^2 > C^2
\end{cases} \tag{75}$$

For each of the cases above we will show that the righ-hand side of (75) is bounded by $2C(\|y_1 - y_2\|_2 + \|\hat{y}_1 - \hat{y}_2\|_2)$.

First, consider the case when $\|y_1 - \hat{y}_1\|_2^2 \leq C^2$ and $\|y_2 - \hat{y}_2\|_2^2 \leq C^2$. Then by repeated use of the triangle inequality,

$$|\|y_1 - \hat{y}_1\|_2^2 - \|y_2 - \hat{y}_2\|_2^2| = |\|y_1 - \hat{y}_1\|_2 - \|y_2 - \hat{y}_2\|_2|(\|y_1 - \hat{y}_1\|_2 + \|y_2 - \hat{y}_2\|_2) \leq \|(y_1 - y_2) - (\hat{y}_1 - \hat{y}_2)\|_2(2C) \leq 2C(\|y_1 - y_2\|_2 + \|\hat{y}_1 - \hat{y}_2\|_2)$$

Consider the case when $\|y_1 - \hat{y}_1\|_2^2 > C^2$, $\|y_2 - \hat{y}_2\|_2^2 \leq C^2$. Then by the definition of ℓ_C,

$$|\ell_C(y_1, \hat{y}_1) - \ell_C(y_2, \hat{y}_2)| = |C^2 - \|y_2 - \hat{y}_2\|_2^2| = (C - \|y_2 - \hat{y}_2\|_2)(\|y_2 - \hat{y}_2\|_2 + C) \leq (C - \|y_2 - \hat{y}_2\|_2)2C \leq 2C(\|y_1 - \hat{y}_1\|_2 - \|y_2 - \hat{y}_2\|_2) \leq 2C\|(y_2 - y_1) - (\hat{y}_2 - \hat{y}_1)\|_2 \leq 2C(\|y_1 - y_2\|_2 + \|\hat{y}_1 - \hat{y}_2\|_2)$$

The case when $\|y_2 - \hat{y}_2\|_2^2 > C^2$, $\|y_1 - \hat{y}_1\|_2^2 \leq C^2$ is symmetric.

Finally, consider the case when $\|y_1 - \hat{y}_1\|_2^2 > C^2$, $\|y_2 - \hat{y}_2\|_2^2 > C^2$. Then by the definition of ℓ_C,

$$|\ell_C(y_1, \hat{y}_1) - \ell_C(y_2, \hat{y}_2)| = |C^2 - C^2| = 0 \leq 2C(\|y_1 - y_2\|_2 + \|\hat{y}_1 - \hat{y}_2\|_2)$$

This completes the proof. □

E.4 Proof of Corollary 3.2

The corollary follows from Theorem 3.1 by choosing λ as follows:

$$\lambda = \frac{\sqrt{N}}{L(\ln(N))^r} = \begin{cases} \sqrt{N} & \mathbf{e}_g \text{ is bounded} \\ \frac{\sqrt{N}}{16L_\ell \ln(N)} & \ell \text{ is } L_\ell\text{-Lipschitz} \\ \frac{\sqrt{N}}{192G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{1+\ln\left(\frac{3}{2\delta}\right)}(\ln(N))^{3/2}} & \ell \text{is quadratic} \end{cases} \tag{76}$$

and upper bounding the right-hand side of (8), i.e., by showing that

$$\frac{1}{2}\left(\ln E_{\theta\sim\pi} \exp(\mathcal{C}_1(2\lambda, \theta, N, \delta, \epsilon)) + \ln E_{\theta\sim\pi} \exp(\mathcal{C}_2(2\lambda, \theta, N, \delta, \epsilon))\right) \leq \mathcal{C}(\delta, \epsilon) \tag{77}$$

Bounded case In order to prove the bounded case, note that from Lemma D.3

$$\|\alpha_\theta\|_{\ell_1} \leq \|\bar{\alpha}\|_{\ell_1}, \quad \theta_\infty(\alpha_\theta) \leq \theta_\infty(\bar{\alpha}), \quad G_e(\theta) \leq G_e, \quad G_{e,1}(\theta) \leq G_{e,1} \tag{78}$$

In addition, notice that the functions $\mathcal{C}_{m,b,i}(\lambda, \alpha, X, L, c)$ have the following property; if $\lambda \leq \bar{\lambda}\sqrt{N}$, $L_1 \leq L_2$, $X_1 \leq X_2$, $c_1 \leq c_2$, and $\|\alpha(k)\|_2 \leq \|\beta(k)\|_2$ for all $k \in \mathbb{N}$, then

$$\mathcal{C}_{m,b,i}(\lambda, \alpha, X_1, L_1, c_1) \leq N\mathcal{C}_{m,b,i}(\bar{\lambda}, \beta, X_2, L_2, c_2) \tag{79}$$

Note that $L(\theta) \leq L_b$, and by using (78), it follows that $\|\alpha_\theta(k)\|_2 \leq \bar{\alpha}(k)$, $G_e(\theta) \leq G_e$, $G_{e,1}(\theta) \leq G_{e,1}$, and hence, by using (79), it follows that for all $i = 1, 2$

$$\mathcal{C}_i(2\lambda, \theta, N, \delta, \epsilon) = \frac{\mathcal{C}_{m,b,i}(2\sqrt{N}, \alpha_\theta, G_e(\theta) + 2G_{e,1}(\theta), L(\theta), c_\mathbf{e})}{N} \leq \mathcal{C}_{m,b,i}(2, \bar{\alpha}, G_e + 2G_{e,1}, L_b, c_\mathbf{e})$$

Hence for $i = 1, 2$,

$$\begin{aligned} \ln E_{\theta\sim\pi} \exp(\mathcal{C}_i(2\lambda, \theta, N, \delta, \epsilon)) &\leq \ln E_{\theta\sim\pi} \exp(\mathcal{C}_{m,b,i}(2, \bar{\alpha}, G_e + 2G_{e,1}, L_b, c_\mathbf{e})) \\ &= \mathcal{C}_{b,m,i}(2, \bar{\alpha}, G_e + 2G_{e,1}, L_b, c_\mathbf{e}) \end{aligned}$$

and hence (77).

Lipschitz loss case: First, notice that for $N \geq 4$, $\ln(N) \geq 1$ and hence

$$\begin{aligned}\bar{\mathcal{C}}_{ub}(2\lambda, L_\ell, N) =& \frac{1}{N}\left(\frac{2\sqrt{N}}{16L_\ell \ln(N)}\right)^2 \left(32L_\ell^2 C_{u,1}\right) + 4\frac{1}{N}\frac{2\sqrt{N}}{16L_\ell \ln(N)} L_\ell C_{u,2}\ln(N) \\ =& \frac{1}{2\ln(N)^2}C_{u,1} + \frac{1}{2}C_{u,2}\frac{1}{\sqrt{N}} \leq C_{u,1} + C_{u,2}\end{aligned} \tag{80}$$

Moreover, using (79),

$$\begin{aligned}\mathcal{C}_{b,i}\Big(4\lambda,&\theta, L_\ell, \frac{2\ln(N)}{\|\alpha_g\|_{\ell_1}K}, N\Big) = \mathcal{C}_{b,i}\left(\frac{4\sqrt{N}}{16L_\ell \ln(N)}, \theta, L_\ell, \frac{2\ln(N)}{\|\alpha_g\|_{\ell_1}K}, N\right) \\ &= \frac{\mathcal{C}_{m,b,i}\left(\frac{4}{16L_\ell}\frac{\sqrt{N}}{\ln(N)}, \alpha_\theta, G_e(\theta) + 2G_{e,1}(\theta), L_\ell, \frac{2\ln(N)}{\|\alpha_g\|_{\ell_1}K}\right)}{N} \\ &\leq \mathcal{C}_{m,b,i}\left(\frac{1}{4L_\ell \ln(N)}, \bar{\alpha}, G_e + 2G_{e,1}, L_\ell, \frac{2\ln(N)}{\|\alpha_g\|_{\ell_1}K}\right) \\ &= \begin{cases} \frac{1}{4L_\ell \ln(N)} L_\ell \frac{2\ln(N)}{\|\alpha_g\|_{\ell_1}K}\sqrt{n_{\mathbf{e}}}\theta_\infty(\bar{\alpha})\|\alpha_g\|_{\ell_1} & i = 1 \\ \frac{1}{(4L_\ell \ln(N))^2}\frac{L_\ell^2}{2}(G_e + 2G_{e,1})\left(\frac{2\ln(N)}{\|\alpha_g\|_{\ell_1}K}\right)^2 n_{\mathbf{e}} & i = 2 \end{cases} \\ &= \begin{cases} \frac{1}{2K}\sqrt{n_{\mathbf{e}}}\theta_\infty(\bar{\alpha}) & i = 1 \\ \frac{1}{8}\frac{(G_e+2G_{e,1})}{2}\left(\frac{1}{\|\alpha_g\|_2 K}\right)^2 n_{\mathbf{e}} & i = 2 \end{cases} \\ &\leq \mathcal{C}_{m,b,i}\left(\frac{1}{2}, \bar{\alpha}, G_e + 2G_{e,1}, 1, \frac{1}{\|\alpha_g\|_{\ell_1}K}\right)\end{aligned} \tag{81}$$

From this it then follows that

$$\mathcal{C}_{ub,i}\left(\frac{2}{16L_\ell}\frac{\sqrt{N}}{\ln(N)}, \theta, L_\ell, N\right) \leq C_{u,1} + C_{u,2} + \frac{1}{2}\mathcal{C}_{m,b,i}\left(\frac{1}{2}, \bar{\alpha}, G_e + 2G_{e,1}, 1, \frac{1}{\|\alpha_g\|_{\ell_1}K}\right) \tag{82}$$

and hence

$$\begin{aligned}&\frac{1}{2}\left(\ln E_{\theta\sim\pi}\exp(\mathcal{C}_1(2\lambda, \theta, N, \delta, \epsilon)) + \ln E_{\theta\sim\pi}\exp(\mathcal{C}_2(2\lambda, \theta, N, \delta, \epsilon))\right) \\ &\leq \frac{1}{2}\sum_{i=1}^{2}\left(C_{u,1} + C_{u,2} + \frac{1}{2}\mathcal{C}_{m,b,i}\left(\frac{1}{2}, \bar{\alpha}, G_e + 2G_{e,1}, 1, \frac{1}{\|\alpha_g\|_{\ell_1}K}\right)\right) = \mathcal{C}(\delta, \epsilon)\end{aligned}$$

so (77) holds.

Quadratic loss case: Notice that as $N^2 \geq 4$, it follows that $\ln(N^2) \geq 1$, and hence

$$\ln\left(\frac{3N^2}{2\delta}\right) = \left(\ln(N^2) + \ln\left(\frac{3}{2\delta}\right)\right) \leq \ln(N^2)\left(1 + \ln\left(\frac{3}{2\delta}\right)\right) \tag{83}$$

By using (83) and applying (82) with $L_\ell = \frac{L(\ln(N))^{1/2}}{16}$ and noticing that $\lambda = \frac{\sqrt{N}}{16L_\ell \ln(N)}$ it follows that

$$\begin{aligned}\mathcal{C}_{ub,i}(2\lambda,&\theta, 12G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{\ln\left(\frac{3N^2}{2\delta}\right)}, N) \leq \mathcal{C}_{ub,i}(2\lambda, \theta, \frac{L(\ln(N))^{1/2}}{16}, N) \\ &\leq C_{u,1} + C_{u,2} + \frac{1}{2}\mathcal{C}_{m,b,i}\left(1/2, \bar{\alpha}, G_e + 2G_{e,1}, 1, \frac{1}{\|\alpha_g\|_{\ell_1}K}\right)\end{aligned} \tag{84}$$

Moreover, from (83) and the fact that $\ln(N) \geq 1$ and $\ln\left(\frac{3N^2}{2\delta}\right) \geq 1$, it follows that

$$
\begin{aligned}
&2\lambda \frac{\ln\left(\frac{3N^2}{2\delta}\right)}{N} 72 G_e^2 \|\mathbf{e}_g\|_{\psi_2}^2 \max\{0, \epsilon\} \\
&= \frac{2\sqrt{N}}{192 G_e \|\mathbf{e}_g\|_{\psi_2} \sqrt{2} \ln(N) \sqrt{1 + \ln\left(\frac{3}{2\delta}\right)} \sqrt{\ln(N)}} (72 G_e^2 \|\mathbf{e}_g\|_{\psi_2}^2 \max\{0, \epsilon\}) \frac{\ln\left(\frac{3N^2}{2\delta}\right)}{N} \\
&\leq \frac{2\sqrt{N}}{192 G_e \|\mathbf{e}_g\|_{\psi_2} \ln(N) \sqrt{1 + \ln\left(\frac{3}{2\delta}\right)} \sqrt{\ln(N^2)}} (72 G_e^2 \|\mathbf{e}_g\|_{\psi_2}^2 \max\{0, \epsilon\}) \frac{(1 + \ln\left(\frac{3}{2\delta}\right)) \ln(N^2)}{N} \\
&= \frac{2 \cdot 72 G_e^2 \|\mathbf{e}_g\|_{\psi_2}^2 \max\{0, \epsilon\} \sqrt{1 + \ln\left(\frac{3}{2\delta}\right)} \sqrt{\ln(N^2)}}{192 G_e \|\mathbf{e}_g\|_{\psi_2} \ln(N) \sqrt{N}} \\
&= \frac{6}{16} \frac{2 G_e \|\mathbf{e}_g\|_{\psi_2} \sqrt{2} (\ln(N))^{1/2} \sqrt{1 + \ln\left(\frac{3}{2\delta}\right)} \max\{0, \epsilon\}}{\ln(N) \sqrt{N}} \\
&= \frac{6}{8 \cdot 192} \frac{L (\ln(N))^{1/2}}{\sqrt{N} \ln(N)} \max\{0, \epsilon\} \\
&= \frac{1}{128} \frac{L}{\sqrt{N} (\ln(N))^{1/2}} \max\{0, \epsilon\}
\end{aligned}
\tag{85}
$$

Hence, for $i = 1, 2$,

$$
\begin{aligned}
\mathcal{C}_i(2\lambda, &\theta, N, \delta, \epsilon) \leq \frac{L}{128} \max\{0, \epsilon\} + \\
&C_{u,1} + C_{u,2} + \frac{1}{2} \mathcal{C}_{b,m,i} \left(1/2, \bar{\alpha}, G_e + 2G_{e,1}, 1, \frac{1}{\|\alpha_g\|_2 K}\right)
\end{aligned}
$$

and thus

$$
\begin{aligned}
&\frac{1}{2} \left(\ln E_{\theta \sim \pi} \exp(\mathcal{C}_1(2\lambda, \theta, N, \delta, \epsilon)) + \ln E_{\theta \sim \pi} \exp(\mathcal{C}_2(2\lambda, \theta, N, \delta, \epsilon))\right) \\
&\leq \frac{L}{128} \max\{0, \epsilon\} + \frac{1}{2} \sum_{i=1}^{2} \left(C_{u,1} + C_{u,2} + \frac{1}{2} \mathcal{C}_{m,b,i} \left(1/2, \bar{\alpha}, G_e + 2G_{e,1}, 1, \frac{1}{\|\alpha_g\|_2 K}\right)\right) \\
&= \mathcal{C}(\delta, \epsilon)
\end{aligned}
$$

which completes the proof.

E.5 Uniform bounds derived from Corollary 3.2

In order to present the uniform bounds, we need the following assumptions from Assumption 3.3: **(LipModel)** and either **(Gauss)** or **(Uni)**.

Corollary E.21 (Uniform [PAC bounds, Example 2.1,2.2 [2])**.** *With the Assumption **(LipModel)** and either **(Gauss)** or **(Uni)** from Assumption 3.3 the following holds. Let $L_\Theta(\delta)$, $\mathcal{C}_{L,emp}$, $\mathcal{C}_{L,true}$, $C(\theta)$, $r_{Lip,1}$, $r_{Lip,2}$ be defined as in Table 1, and let L, C_δ, $\mathcal{C}(\delta, \epsilon)$, r be defined as in Corollary 3.2 For any $\delta \in (0, 0.5)$, $\sigma > 0$, $N \geq 4$, $\epsilon \in \{-1, 1\}$,*

$$
\begin{aligned}
&\forall \theta \in \Theta: \ \epsilon \mathcal{L}(\theta) \leq \epsilon \hat{\mathcal{L}}_N(\theta) + \\
&\frac{L (\ln(N))^{r+1}}{\sqrt{N}} \left(\mathcal{C}_{L,emp} + \mathcal{C}_{L,true} + C(\theta) + \ln\left(\frac{C_\delta}{\delta}\right) + 2 r_{Lip,1} \ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta, \epsilon)\right)
\end{aligned}
\tag{86}
$$

holds with probability at least $1 - (2 + r_{Lip,1})\delta$ over data.

The proof of Corollary E.21 is based on a number of technical lemmas, which will also be used for the proof of Corollary 3.4.

Lemma E.22 (KL divergence)**.** *Assume that either **(Gauss)** or **(Uni)** holds and $\sigma > 0$ is arbitrary if **(Gauss)** holds, and $\sigma < 2R$ if **(Uni)** holds. For every $\theta \in \Theta$, define the density ρ_θ as follows*

$$\rho_\theta = \begin{cases} \mathcal{N}(\theta, \frac{\sigma^2}{N} I) & \text{if } \textbf{(Gauss)} \text{ holds} \\ \mathcal{U}(\theta, \frac{\sigma}{\sqrt{N}}) & \text{if } \textbf{(Uni)} \text{ holds} \end{cases}$$

where $\mathcal{N}(m, \sigma^2 I)$ denotes the Gaussian distribution with mean m and variance $\sigma^2 I$, and $\mathcal{U}(m, r)$ denotes the uniform distribution on the ball with center in m and radius r intersected with Θ. Then for $N \geq 4$,

$$\mathrm{KL}(\rho_\theta \| \pi) \leq C(\theta) \ln(N)$$

Proof of Lemma E.22. If **(Gauss)** holds, then using KL divergence for Gaussian densities ([25, Lemma 11.3.4]) it follows that

$$\begin{aligned} &\mathrm{KL}(\rho_\theta \| \pi) = \\ &\frac{1}{2}\left(\mathrm{trace}\left(\frac{\sigma^2}{N} P^{-1}\right) + (\theta_m - \theta)^T P^{-1}(\theta_m - \theta) - n_\theta + \ln \frac{\det(P) N^{n_\theta}}{\sigma^{2n_\theta}}\right) \\ &= \frac{1}{2}\left(\mathrm{trace}\left(\frac{\sigma^2}{N} P^{-1}\right) + (\theta_m - \theta)^T P^{-1}(\theta_m - \theta) - n_\theta + \ln \frac{\det(P)}{\sigma^{2n_\theta}} + n_\theta \ln(N)\right) \\ &\leq \ln(N) C(\theta) \end{aligned}$$

where we used the fact that $\ln(N) \geq 1$ and $\frac{1}{N} \leq 1$ if $N \geq 4$. Likewise, if **(Uni)** holds, denote by $B(\theta)$ the intersection of the ball of radius $\frac{\sigma}{\sqrt{N}}$ centered at θ with Θ. We claim that $B(\theta)$ contains a ball of radius $\frac{\sigma}{2\sqrt{N}}$. Indeed, since $\sigma < 2R$, it follows that $\frac{\sigma}{2\sqrt{N}R} < 1$. Consider

$$\theta_r = \frac{\sigma}{2\sqrt{N}R}\theta_m + \left(1 - \frac{\sigma}{2\sqrt{N}R}\right)\theta.$$

It then follows that

$$\|\theta_r - \theta_m\|_2 = \left(1 - \frac{\sigma}{2\sqrt{N}R}\right)\|\theta - \theta_m\|_2 \leq \left(1 - \frac{\sigma}{2\sqrt{N}R}\right) R \leq R,$$

hence θ_r belongs to Θ. Likewise,

$$\|\theta_r - \theta\|_2 = \frac{\sigma}{2\sqrt{N}R}\|\theta_m - \theta\|_2 \leq \frac{\sigma}{2\sqrt{N}R} R \leq \frac{\sigma}{2\sqrt{N}} \leq \frac{\sigma}{\sqrt{N}},$$

i.e., θ_r belongs to the ball of radius $\frac{\sigma}{\sqrt{N}}$ centered at θ. Finally, we argue that the ball $B_{\frac{\sigma}{2\sqrt{N}}}(\theta_r)$ of radius $\frac{\sigma}{2\sqrt{N}}$ centered at θ_r is contained in $B(\theta)$ and Θ. Let $\tilde{\theta} \in B_{\frac{\sigma}{2\sqrt{N}}}(\theta_r)$. Then

$$\begin{aligned} &\|\tilde{\theta} - \theta_m\|_2 = \|(\tilde{\theta} - \theta_r) + (\theta_r - \theta_m)\|_2 \leq \|\tilde{\theta} - \theta_r\|_2 + \|\theta_r - \theta_m\|_2 \\ &\leq \frac{\sigma}{2\sqrt{N}} + \left(1 - \frac{\sigma}{2\sqrt{N}R}\right)\|\theta - \theta_m\|_2 \leq \frac{\sigma}{2\sqrt{N}} + \left(1 - \frac{\sigma}{2\sqrt{N}R}\right) R \leq R \end{aligned}$$

i.e., $\tilde{\theta} \in \Theta$. Similarly,

$$\|\tilde{\theta} - \theta\|_2 = \|(\tilde{\theta} - \theta_r) + (\theta_r - \theta)\|_2 \leq \|\tilde{\theta} - \theta_r\|_2 + \|\theta_r - \theta\|_2 \leq \frac{\sigma}{2\sqrt{N}} + \frac{\sigma}{2\sqrt{N}} = \frac{\sigma}{\sqrt{N}},$$

i.e., $\tilde{\theta} \in B(\theta)$. Then $C_2 = \mathrm{vol}(B(\theta))$ is larger than the volume of $B_{\frac{\sigma}{2\sqrt{N}}}(\theta_r)$, i.e., from the formula for the volume of a high-dimensional ball [40] it follows that

$$C_2 \geq \mathrm{vol}(B_{\frac{\sigma}{2\sqrt{N}}}(\theta_r)) = \frac{\pi^{n_\theta/2}}{\Gamma(n_\theta/2 + 1)}\left(\frac{\sigma}{2\sqrt{N}}\right)^{n_\theta}$$

Then

$$\begin{aligned} KL(\rho_\theta \| \pi) &= \int_{\|\tilde{\theta} - \theta\|_2 \leq \frac{\sigma}{\sqrt{N}}} \ln\left(\frac{\rho_\theta(\tilde{\theta})}{\pi(\tilde{\theta})}\right) \frac{1}{C_2} dm(\tilde{\theta}) \\ &= \int_{\|\tilde{\theta} - \theta\|_2 \leq \frac{\sigma}{\sqrt{N}}} \frac{1}{C_2} \ln\left(\frac{\mathrm{vol}(\Theta)}{C_2}\right) dm(\tilde{\theta}) = \ln\left(\frac{\mathrm{vol}(\Theta)}{C_2}\right) \end{aligned}$$

and hence

$$\begin{aligned}\mathrm{KL}(\rho_\theta\|\pi) &= \ln\left(\frac{\mathrm{vol}(\Theta)}{C_2}\right) \leq \ln\left(\frac{\mathrm{vol}(\Theta)}{\frac{\pi^{n_\theta/2}}{\Gamma(n_\theta/2+1)}\left(\frac{\sigma}{2\sqrt{N}}\right)^{n_\theta}}\right)\\ &= \ln\left(\frac{\mathrm{vol}(\Theta)\Gamma(n_\theta/2+1)}{\pi^{n_\theta/2}(\sigma/2)^{n_\theta}}\right) + \ln\left(\left(\sqrt{N}\right)^{n_\theta}\right)\\ &= \ln\left(\frac{\mathrm{vol}(\Theta)\Gamma(n_\theta/2+1)}{\sqrt{\pi(\sigma^2/4)}^{n_\theta}}\right) + \frac{n_\theta}{2}\ln(N) \leq C(\theta)\ln(N)\end{aligned}$$

□

Another ingredient is the following:

Lemma E.23 (Lipschitz continuity of the true and empirical errors)**.** *Using the constants defined in Corollary 3.2-E.21, the following inequality holds*

$$\begin{aligned}&\forall\theta_1,\theta_2\in\Theta:\\ &|\hat{\mathcal{L}}_N(\theta_1)-\hat{\mathcal{L}}_N(\theta_2)| \leq L_\Theta(\delta)L(\ln(N))^r\|\theta_1-\theta_2\|_2 +\\ &\qquad 2r_{Lip,I}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1}+C_{u,2}+\ln\left(\frac{C_\delta}{\delta}\right)\right),\end{aligned} \tag{87}$$

with probability $1-r_{Lip,1}\delta$ *over data, where*

$$r_{Lip,1} = \begin{cases} 0 & \mathbf{e}_g \textit{ is bounded}\\ 1 & \textit{otherwise}\end{cases}$$

and

$$\begin{aligned}&\forall\,\theta_1,\theta_2\in\Theta:\\ &|\mathcal{L}(\theta_1)-\mathcal{L}(\theta_2)| \leq L_\Theta(\delta)L(\ln(N))^r\|\theta_1-\theta_2\|_2 + 2r_{Lip,1}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1}+C_{u,2}\right)+\\ &\qquad 2r_{Lip,2}\frac{L}{512}\frac{L(\ln(N))^r}{\sqrt{N}}\end{aligned} \tag{88}$$

where

$$r_{Lip,2} = \begin{cases} 2 & \ell \textit{ is quadratic and } \mathbf{e}_g \textit{ is unbounded}\\ 0 & \textit{otherwise}\end{cases}.$$

Moreover, the inequalities below

(10) *from Corollary 3.2 holds, and*

$$\begin{aligned}&\forall\theta_1,\theta_2\in\Theta:\\ &|\hat{\mathcal{L}}_N(\theta_1)-\hat{\mathcal{L}}_N(\theta_2)| \leq L_\Theta(\delta)L(\ln(N))^r\|\theta_1-\theta_2\|_2 +\\ &\qquad 2r_{Lip,1}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1}+C_{u,2}+\ln\left(\frac{C_\delta}{\delta}\right)\right),\end{aligned}$$

and (89)

$$\begin{aligned}&\forall\,\theta_1,\theta_2\in\Theta:\\ &|\mathcal{L}(\theta_1)-\mathcal{L}(\theta_2)| \leq L_\Theta(\delta)L(\ln(N))^r\|\theta_1-\theta_2\|_2 + 2r_{Lip,1}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1}+C_{u,2}\right)+\\ &\qquad 2r_{Lip,2}\frac{L}{512}\frac{L(\ln(N))^r}{\sqrt{N}}\end{aligned}$$

hold simultaneously with probability at least $1-(2+r_{Lip,1})\delta$.

Proof of Lemma E.23. **Bounded noise case:** If ℓ is L_ℓ-Lipschitz and $\mathbf{e}_g$ is bounded, then using (**LipModel**) and the argument of the proof from Lemma D.3, it follows that

$$|\ell(\mathbf{y}(t),\hat{\mathbf{y}}_{\theta_1}(t))-\ell(\mathbf{y}(t),\hat{\mathbf{y}}_{\theta_2}(t))|\leq L_\ell\|\hat{\mathbf{y}}_{\theta_1}(t)-\hat{\mathbf{y}}_{\theta_2}(t)\|_2$$
$$\leq L_\ell\sum_{k=0}^{\infty}\|c_{\theta_1}(k)-c_{\theta_2}(k)\|_2\|\mathbf{e}_g(t-k)\|_2\leq L_\ell\sum_{k=0}^{\infty}\sum_{l=0}^{k}\|\alpha_{\theta_1}(l)-\alpha_{\theta_2}(l)\|_2\|\alpha_{g,w}(k-l)\|_2c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}}$$
$$\leq L_\ell\|\alpha_{\theta_1}-\alpha_{\theta_2}\|_{\ell_1}\|\alpha_g\|_{\ell_1}c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}}=\underbrace{L_\alpha L_\ell\|\alpha_g\|_{\ell_1}c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}}}_{L_\Theta(\delta)}\|\theta_1-\theta_2\|_2$$

and

$$|\ell(\mathbf{y}(t),\hat{\mathbf{y}}_{\theta_1}(t\mid 0))-\ell(\mathbf{y}(t),\hat{\mathbf{y}}_{\theta_2}(t\mid 0))|\leq L_\ell\|\hat{\mathbf{y}}_{\theta_1}(t\mid 0)-\hat{\mathbf{y}}_{\theta_2}(t\mid 0)\|_2$$
$$\leq L_\ell\sum_{k=0}^{\infty}\|c_{\theta_1,t}(k)-c_{\theta_2,t}(k)\|_2\|\mathbf{e}_g(t-k)\|_2$$
$$\leq L_\ell\sum_{k=0}^{\infty}\sum_{l=0}^{\min\{k,t\}}\|\alpha_{\theta_1}(l)-\alpha_{\theta_2}(l)\|_2\|\alpha_{g,w}(k-l)\|_2c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}}$$
$$\leq L_\ell\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\|\alpha_{\theta_1}(l)-\alpha_{\theta_2}(l)\|_2\|\alpha_{g,w}(k-l)\|_2c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}}$$
$$\leq L_\ell\|\alpha_{\theta_1}-\alpha_{\theta_2}\|_{\ell_1}\|\alpha_g\|_{\ell_1}c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}}=\underbrace{L_\alpha L_\ell\|\alpha_g\|_{\ell_1}c_{\mathbf{e}}\sqrt{n_{\mathbf{e}}}}_{L_\Theta(\delta)}\|\theta_1-\theta_2\|_2$$

From this it then follows that for all $\theta_1,\theta_2\in\Theta$:

$$|\mathcal{L}(\theta_1)-\mathcal{L}(\theta_2)|=|\mathbf{E}[\ell(\mathbf{y}(t),\hat{\mathbf{y}}_{\theta_1}(t))-\ell(\mathbf{y}(t),\hat{\mathbf{y}}_{\theta_2}(t))]|$$
$$\leq\mathbf{E}[|\ell(\mathbf{y}(t),\hat{\mathbf{y}}_{\theta_1}(t))-\ell(\mathbf{y}(t),\hat{\mathbf{y}}_{\theta_2}(t)|]\leq L_\Theta(\delta)\|\theta_1-\theta_2\| \tag{90}$$
$$|\hat{\mathcal{L}}_N(\theta_1)-\hat{\mathcal{L}}_N(\theta_2)|=\frac{1}{N}\left|\sum_{t=0}^{N-1}\ell(\mathbf{y}(t),\hat{\mathbf{y}}_{\theta_1}(t\mid 0))-\ell(\mathbf{y}(t),\hat{\mathbf{y}}_{\theta_2}(t\mid 0))\right|\leq$$
$$\frac{1}{N}\sum_{t=0}^{N-1}|\ell(\mathbf{y}(t),\hat{\mathbf{y}}_{\theta_1}(t\mid 0))-\ell(\mathbf{y}(t),\hat{\mathbf{y}}_{\theta_2}(t\mid 0))|\leq L_\Theta(\delta)\|\theta_1-\theta_2\| \tag{91}$$

By noticing that $r_{\text{Lip},1}=r=0$ and $L=1$ for the bounded case, the claim of the lemma follows.

Unbounded noise case: Assume that the noise is not necessarily bounded and ℓ is L_ℓ-Lipschitz. From Subsection E.2, recall the definition of truncated empirical and true errors $\bar{\mathcal{L}}_N(\theta)$ and $\bar{\mathcal{L}}(\theta)$, corresponding to the output $\bar{\mathbf{y}}(t)$ and $\bar{\mathbf{y}}_\theta(t)$ and $\bar{\mathbf{y}}_\theta(t\mid 0)$ generated for the truncated noise $\bar{\mathbf{e}}_g(t)=\mathbf{e}_g(t)\chi(\mathbf{e}_g(t)<C)$ for $C=\frac{\ln(N)}{K\|\alpha_g\|_{\ell_1}}$. Using Chernoff's inequality and Lemma E.11, it follows that

$$\mathbf{P}(\forall\theta\in\Theta:\left|\hat{\mathcal{L}}_N(\theta)-\bar{\mathcal{L}}_N(\theta)\right|>c)\leq\frac{e^{\lambda\frac{1}{2}\left(\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N)+\mathcal{C}(\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N)\right)}}{e^{\lambda c}} \tag{92}$$

for all $0<\lambda<\frac{N}{4L_\ell}$ and hence

$$\mathbf{P}\left(\forall\theta\in\Theta:\left|\hat{\mathcal{L}}_N(\theta)-\bar{\mathcal{L}}_N(\theta)\right|\leq\frac{1}{2\lambda}\left(\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N)+\right.\right.$$
$$\left.\left.\mathcal{C}(\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N)+\ln\left(\frac{1}{\delta}\right)\right)\right)\geq 1-\delta \tag{93}$$

Similarly, from Lemma E.11 it follows that

$$\forall\theta\in\Theta:\quad|\mathcal{L}(\theta)-\bar{\mathcal{L}}(\theta)|\leq\frac{1}{2\lambda}\left(\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N)+\mathcal{C}(\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N)\right) \tag{94}$$

From (43) in Subsection E.2 for $C=\frac{\ln(N)}{K\|\alpha_g\|_{\ell_1}}$ it follows that

$$\frac{1}{2}\left(\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N)+\mathcal{C}(\|\alpha_g\|_{\ell_1},2L_\ell\lambda,N)\right)=2\bar{\mathcal{C}}_{ub}(\frac{\lambda}{8},L_\ell,N)$$

and hence by setting $\lambda = \frac{\sqrt{N}}{16L_\ell \ln(N)}$ it follows that

$$\begin{aligned}&\frac{1}{\lambda}(\frac{1}{2}\left(\mathcal{C}(\|\bar{\alpha}\|_{\ell_1}\|\alpha_g\|_{\ell_1}, 2L_\ell\lambda, N) + \mathcal{C}(\|\alpha_g\|_{\ell_1}, 2L_\ell\lambda, N)\right)\\ &\leq (\frac{1}{2}C_{u,1} + \frac{1}{8}C_{u,2})\frac{16L_\ell\ln(N)}{\sqrt{N}} \leq (C_{u,1}+C_{u,2})\frac{16L_\ell\ln(N)}{\sqrt{N}}\end{aligned} \tag{95}$$

and hence, appplying it to (93)-(94) and recalling that $L = 16L_\ell$ results in

$$\begin{aligned}&\mathbf{P}\left(\forall\theta\in\Theta : \left|\hat{\mathcal{L}}_N(\theta) - \bar{\mathcal{L}}_N(\theta)\right| \leq (C_{u,1}+C_{u,2}+\ln(\frac{1}{\delta}))\frac{16L_\ell\ln(N)}{\sqrt{N}}\right) \geq 1-\delta\\ &\forall\theta_1,\theta_2\in\Theta : |\mathcal{L}(\theta_1)-\mathcal{L}(\theta_2)| \leq (C_{u,1}+C_{u,2})\frac{L\ln(N)}{\sqrt{N}}\end{aligned} \tag{96}$$

Note that (90)-(91) holds for $\bar{\mathcal{L}}_N(\theta)$ and $\bar{\hat{\mathcal{L}}}_N(\theta)$ with $L_\Theta(\delta) = L_\alpha L_\ell\sqrt{n_{\mathbf{e}}}\frac{\ln(N)}{K\|\alpha_g\|_{\ell_1}}\|\alpha_g\|_{\ell_1} = L_\alpha L_\ell\sqrt{n_{\mathbf{e}}}\frac{\ln(N)}{K}$ and hence by using (96) and the observation

$$\begin{aligned}&|\hat{\mathcal{L}}_N(\theta_1)-\hat{\mathcal{L}}_N(\theta_2)| \leq |\hat{\mathcal{L}}_N(\theta_1)-\bar{\hat{\mathcal{L}}}_N(\theta_1)| + |\bar{\hat{\mathcal{L}}}_N(\theta_1)-\bar{\hat{\mathcal{L}}}_N(\theta_2)| + |\bar{\hat{\mathcal{L}}}_N(\theta_2)-\hat{\mathcal{L}}_N(\theta_2)|\\ &|\mathcal{L}(\theta_1)-\mathcal{L}(\theta_2)| \leq |\mathcal{L}(\theta_1)-\bar{\mathcal{L}}(\theta_1)| + |\bar{\mathcal{L}}(\theta_1)-\bar{\mathcal{L}}(\theta_2)| + |\bar{\mathcal{L}}(\theta_2)-\mathcal{L}(\theta_2)|\end{aligned}$$

it follows

$$\begin{aligned}&\mathbf{P}\left(\forall\theta_1,\theta_2\in\Theta : \left|\hat{\mathcal{L}}_N(\theta_1)-\hat{\mathcal{L}}_N(\theta_1)\right| \leq 2(C_{u,1}+C_{u,2}+\ln(\frac{1}{\delta}))\frac{16L_\ell\ln(N)}{\sqrt{N}})+\right.\\ &\left.L_\alpha L_\ell\sqrt{n_{\mathbf{e}}}\frac{\ln(N)}{K}\|\theta_1-\theta_2\|_2\right) \geq 1-\delta\\ &\forall\theta_1,\theta_2\in\Theta : |\mathcal{L}(\theta_1)-\mathcal{L}(\theta_2)|\\ &\leq 2(C_{u,1}+C_{u,2}+\ln(\frac{1}{\delta}))\frac{L\ln(N)}{\sqrt{N}}) + L_\alpha L_\ell\sqrt{n_{\mathbf{e}}}\frac{\ln(N)}{K}\|\theta_1-\theta_2\|_2\end{aligned} \tag{97}$$

Since $L_\alpha L_\ell\sqrt{n_{\mathbf{e}}}\frac{\ln(N)}{K} = L_\Theta L\ln(N)$ with the definition of L_Θ as in Corollary E.21 for the unbounded Lipschitz case (97) can be rewritten as

$$\begin{aligned}&\mathbf{P}\big(\forall\theta_1,\theta_2\in\Theta : \left|\hat{\mathcal{L}}_N(\theta_1)-\hat{\mathcal{L}}_N(\theta_1)\right| \leq 2(C_{u,1}+C_{u,2}+\ln(\frac{1}{\delta}))\frac{L\ln(N)}{\sqrt{N}})+\\ &L_\Theta(\delta)L\ln(N)\|\theta_1-\theta_2\|_2\big) \geq 1-\delta\\ &\forall\theta_1,\theta_2\in\Theta : |\mathcal{L}(\theta_1)-\mathcal{L}(\theta_2)|\\ &\leq 2(C_{u,1}+C_{u,2}+\ln(\frac{1}{\delta}))\frac{L\ln(N)}{\sqrt{N}}) + L_\Theta(\delta)L\ln(N)\|\theta_1-\theta_2\|_2\end{aligned} \tag{98}$$

hence (87)-(88). By using the union bound, and that (10) holds with probability $1-2\delta$, (89) follows.

Quadratic case: Recall from (58), Subsection E.3 the definition of the loss function ℓ_C and the true and empirical errors ${}^C\mathcal{L}(\theta)$ and ${}^C\mathcal{L}_N(\theta)$. Recall that for $C = 6G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{\ln(\frac{3N^2}{2\delta})}$.

Recall that by Lemma E.18 that

$$\sup_{t=0,\ldots,N-1,\theta\in\Theta}\max\{\|\mathbf{y}(t)-\mathbf{y}_\theta(t)\|_2, \|\mathbf{y}(t)-\mathbf{y}_\theta(t[0])\|_\infty\} \leq C \tag{99}$$

holds with probability at least $1-\frac{\delta}{3}$. Also recall that ℓ_C is L_ℓ-Lipschitz with $L_\ell = 12G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{\ln\left(\frac{3N^2}{2\delta}\right)}$.

By applying (98) to ${}^C\mathcal{L}_N(\theta)$ and ${}^C\mathcal{L}(\theta)$ and replacing δ with $\frac{2\delta}{3}$ and using the union bound, it follows that

$$\begin{aligned}&\sup_{t=0,\ldots,N-1,\theta\in\Theta}\max\{\|\mathbf{y}(t)-\mathbf{y}_\theta(t)\|_2, \|\mathbf{y}(t)-\mathbf{y}_\theta(t[0])\|_\infty\} \leq 6G_e(\|\mathbf{e}_g\|_{\psi_2})\sqrt{\ln\left(\frac{3N^2}{2\delta}\right)}\\ &\sup_{\theta_1,\theta_2\in\Theta}\left|{}^C\mathcal{L}_N(\theta_1) - {}^C\mathcal{L}_N(\theta_2)\right|\\ &\leq L_\alpha L_\ell\frac{\sqrt{n_{\mathbf{e}}}}{K}\ln(N)\|\theta_1-\theta_2\| + 2\frac{16L_\ell\ln(N)}{\sqrt{N}}\left(C_{u,1}+C_{u,2}+\ln\left(\frac{3}{2\delta}\right)\right),\end{aligned}$$

holds with probability at least $1-\delta$, and

$$\begin{aligned}&\sup_{\theta_1,\theta_2\in\Theta}\left|{}^{C}\mathcal{L}(\theta_1)-{}^{C}\mathcal{L}(\theta_2)\right|\\&\leq L_\alpha L_\ell\frac{\sqrt{n_{\mathbf{e}}}}{K}\ln(N)\|\theta_1-\theta_2\|+2\frac{16L_\ell\ln(N)}{\sqrt{N}}\left(C_{u,1}+C_{u,2}\right)\end{aligned}\tag{100}$$

where $L_\ell=12G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{\ln\left(\frac{3N^2}{2\delta}\right)}$. Recall from Subsection E.3, proof of Theorem 3.1 that if

$$\forall:\quad t=0,1,\ldots,N-1:\quad \max\{\|\mathbf{y}(t)-\mathbf{y}_\theta(t)\|_2,\|\mathbf{y}(t)-\mathbf{y}_\theta(t\mid 0)\|_\infty\}\leq C$$

holds, then ${}^{C}\hat{\mathcal{L}}_N(\theta)=\hat{\mathcal{L}}_N(\theta)$, and hence

$$\begin{aligned}&\sup_{t=0,\ldots,N-1,\theta\in\Theta}\max\{\|\mathbf{y}(t)-\mathbf{y}_\theta(t)\|_2,\|\mathbf{y}(t)-\mathbf{y}_\theta(t[0])\|_\infty\}\leq 6G_e(\|\mathbf{e}_g\|_{\psi_2})\sqrt{\ln\left(\frac{3N^2}{2\delta}\right)}\\&\sup_{\theta_1,\theta_2\in\Theta}|\mathcal{L}_N(\theta_1)-\mathcal{L}_N(\theta_2)|\\&\leq L_\alpha L_\ell\frac{\sqrt{n_{\mathbf{e}}}}{K}\ln(N)\|\theta_1-\theta_2\|+2\frac{16L_\ell\ln(N)}{\sqrt{N}}\left(C_{u,1}+C_{u,2}+\ln\left(\frac{3}{2\delta}\right)\right),\end{aligned}\tag{101}$$

holds with probability at least $1-\delta$. By using

$$\sqrt{\ln\left(\frac{3N^2}{2\delta}\right)}\leq\sqrt{2}(\ln(N))^{1/2}\sqrt{1+\ln\left(\frac{3}{2\delta}\right)}$$

it follows that

$$L_\alpha L_\ell\frac{\sqrt{n_{\mathbf{e}}}}{K}\ln(N)\leq 12L_\alpha G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{1+\ln\left(\frac{3}{2\delta}\right)}\sqrt{2}\frac{\sqrt{n_{\mathbf{e}}}}{K}(\ln(N))^{3/2}$$

Moreover, recall (67)-(68) holds, and thus using (83), it follows that

$$\begin{aligned}&\left|\mathcal{L}(\theta)-{}^{C}\mathcal{L}(\theta)\right|\leq 72G_e^2\|\mathbf{e}_g\|_{\psi_2}^2\frac{\ln\left(\frac{3N^2}{2\delta}\right)}{N}\leq 72\frac{L^2}{192^2}\frac{\ln(N)}{N}=\\&\frac{6}{16\cdot 192}L^2\frac{\ln(N)}{N}\leq\frac{L}{512}\frac{L(\ln(N))^{\frac{3}{2}}}{\sqrt{N}}\end{aligned}\tag{102}$$

In then follows that

$$\begin{aligned}&|\mathcal{L}(\theta_1)-\mathcal{L}(\theta_2)|\leq\left|\mathcal{L}(\theta_1)-{}^{C}\mathcal{L}(\theta_1)\right|+\left|{}^{C}\mathcal{L}(\theta_1)-{}^{C}\mathcal{L}(\theta_2)\right|+\left|{}^{C}\mathcal{L}(\theta_2)-\mathcal{L}(\theta_2)\right|\\&\leq\frac{\sqrt{n_{\mathbf{e}}}}{K}L\ln(N)\|\theta_1-\theta_2\|+2\frac{16L_\ell\ln(N)}{\sqrt{N}}\left(C_{u,1}+C_{u,2}\right)+\frac{2L}{512}\frac{L(\ln(N))^{\frac{3}{2}}}{\sqrt{N}}\end{aligned}\tag{103}$$

By repeating the argument of the proof of Theorem 3.1 for the quadratic case, and using (103) and

$$\begin{aligned}&\frac{1}{L}12L_\alpha G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{1+\ln\left(\frac{3}{2\delta}\right)}\sqrt{2}\frac{\sqrt{n_{\mathbf{e}}}}{K}=\\&\frac{12L_\alpha G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{1+\ln\left(\frac{3}{2\delta}\right)}\sqrt{2}\frac{\sqrt{n_{\mathbf{e}}}}{K}}{192G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{2}\sqrt{1+\ln\left(\frac{3}{2\delta}\right)}}=L_\alpha\frac{\sqrt{n_{\mathbf{e}}}}{16K}\end{aligned}$$

it then follows that (101) implies that

$$\text{(99) holds}\tag{104}$$

$$\sup_{\theta_1,\theta_2\in\Theta}|\mathcal{L}_N(\theta_1)-\mathcal{L}_N(\theta_2)|\tag{105}$$

$$\leq L_\Theta(\delta)L(\ln(N))^{3/2}\|\theta_1-\theta_2\|+\frac{2L\ln(N)^r}{\sqrt{N}}\left(C_{u,1}+C_{u,2}+\ln\left(\frac{3}{2\delta}\right)\right)\tag{106}$$

which is implies (87), and

$$\sup_{\theta_1,\theta_2\in\Theta} |\mathcal{L}(\theta_1) - \mathcal{L}(\theta_2)| \tag{107}$$

$$\leq L_\Theta L(\ln(N))^{3/2}\|\theta_1 - \theta_2\| + \frac{2L\ln(N)^{3/2}}{\sqrt{N}}(C_{u,1} + C_{u,2}) + \tag{108}$$

$$+ 2\frac{L}{512}\frac{L(\ln(N))^{3/2}}{\sqrt{N}} \tag{109}$$

which is equivalent to (88). Since (101) holds with probability at least $1-\delta$, it then follows that (87) also holds with probability at least $1-\delta$.

Finally, from Corollary 3.2 it follows that (10) holds with probability at least $1-\delta$, and hence by the union bound (89) follows.

□

Proof of Corollary E.21. Consider the inequality (89) which holds with probability at least $1-(2+r_{\text{Lip},1})\delta$. If (89) holds, then for all $\theta\in\Theta$,

$$\epsilon(\mathcal{L}(\theta)) - (\mathcal{L}(\tilde{\theta})) \leq |\epsilon(\mathcal{L}(\theta) - \mathcal{L}(\tilde{\theta}))| = |\mathcal{L}(\theta) - \mathcal{L}(\tilde{\theta})|$$

and hence

$$\epsilon\mathcal{L}(\theta) \leq \epsilon\mathcal{L}(\tilde{\theta}) + L_\Theta(\delta)L(\ln(N))^r\|\theta-\tilde{\theta}\|_2 + 2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}(C_{u,1}+C_{u,2}) + 2r_{\text{Lip},2}\frac{L}{512}\frac{L(\ln(N))^r}{\sqrt{N}}$$

Then $E_{\tilde{\theta}\sim\rho_\theta}\|\tilde{\theta}-\theta\|_2^2 \leq \frac{\sigma^2}{N}$ and hence

$$E_{\tilde{\theta}\sim\rho_\theta}\|\tilde{\theta}-\theta\|_2 \leq \sqrt{E_{\tilde{\theta}\sim\rho_\theta}\|\tilde{\theta}-\theta\|_2^2} \leq \frac{\sigma}{\sqrt{N}}$$

Therefore

$$\begin{aligned}\mathcal{L}(\theta) &\leq E_{\tilde{\theta}\sim\rho_\theta}\mathcal{L}(\tilde{\theta}) + L_\Theta(\delta)L(\ln(N))^r\|\theta-\tilde{\theta}\|_2 \\ &+ 2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}(C_{u,1}+C_{u,2}) + 2r_{\text{Lip},2}\frac{L}{512}\frac{L(\ln(N))^r}{\sqrt{N}} \\ &\leq \frac{\sigma L_\Theta(\delta)}{\sqrt{N}} + 2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}(C_{u,1}+C_{u,2}) + 2r_{\text{Lip},2}\frac{L}{512}\frac{L(\ln(N))^r}{\sqrt{N}}\end{aligned} \tag{111}$$

Similary,

$$\epsilon\hat{\mathcal{L}}_N(\tilde{\theta}) \leq \epsilon\hat{\mathcal{L}}_N(\theta) + L_\Theta(\delta)L(\ln(N))^r\|\theta-\tilde{\theta}\|_2 + 2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1}+C_{u,2}+\ln\left(\frac{C_\delta}{\delta}\right)\right)$$

and hence

$$\begin{aligned}\epsilon E_{\tilde{\theta}\sim\rho_\theta}\hat{\mathcal{L}}_N(\tilde{\theta}) &\leq \epsilon\hat{\mathcal{L}}_N(\theta) + L_\Theta(\delta)L(\ln(N))^r E_{\tilde{\theta}\sim\rho_\theta}\|\theta-\tilde{\theta}\|_2 \\ &+ 2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1}+C_{u,2}+\ln\left(\frac{C_\delta}{\delta}\right)\right) \\ &\leq \frac{\sigma L_\Theta(\delta)}{\sqrt{N}} + 2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1}+C_{u,2}\ln\left(\frac{C_\delta}{\delta}\right)\right)\end{aligned} \tag{112}$$

In addition, since (10) also holds, it follows that

$$\epsilon E_{\tilde{\theta}\sim\rho_\theta}\mathcal{L}(\tilde{\theta}) \leq \epsilon E_{\tilde{\theta}\sim\rho_\theta}\hat{\mathcal{L}}_N(\tilde{\theta}) + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\text{KL}(\rho_\theta\|\pi) + \ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,\epsilon)\right) \tag{113}$$

By combining (111), (112), and (113), it follows that

$$
\begin{aligned}
\epsilon\mathcal{L}(\theta) \leq & \epsilon E_{\tilde{\theta}\sim\rho_\theta}\mathcal{L}(\tilde{\theta}) + \frac{\sigma L_\Theta(\delta)}{\sqrt{N}} + 2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1}+C_{u,2}\right) \\
& + 2r_{\text{Lip},2}\frac{L}{512}\frac{L(\ln(N))^r}{\sqrt{N}} \\
& \leq \epsilon E_{\tilde{\theta}\sim\rho_\theta}\hat{\mathcal{L}}_N(\tilde{\theta}) + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\mathrm{KL}(\rho_\theta\|\pi) + \ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,\epsilon)\right) + \\
& + \frac{\sigma L_\Theta(\delta)}{\sqrt{N}} + 2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1}+C_{u,2}\right) + 2r_{\text{Lip},2}\frac{L}{512}\frac{L(\ln(N))^r}{\sqrt{N}} \\
& \leq \epsilon\hat{\mathcal{L}}_N(\theta) + \frac{\sigma L_\Theta(\delta)}{\sqrt{N}} + 2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1}+C_{u,2}+\ln\left(\frac{C_\delta}{\delta}\right)\right) \\
& + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\mathrm{KL}(\rho_\theta\|\pi) + \ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,\epsilon)\right) + \\
& + \frac{\sigma L_\Theta(\delta)}{\sqrt{N}} + 2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1}+C_{u,2}\right) + 2r_{\text{Lip},2}\frac{L}{512}\frac{L(\ln(N))^r}{\sqrt{N}} \\
& = \epsilon\hat{\mathcal{L}}_N(\theta) + \frac{2\sigma L_\Theta(\delta)}{\sqrt{N}} + 4r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1}+C_{u,2}\right) + 2r_{\text{Lip},2}\frac{L}{512}\frac{L(\ln(N))^r}{\sqrt{N}} \\
& + \frac{L(\ln(N))^r}{\sqrt{N}}\left(C(\theta)\ln(N) + \ln\left(\frac{C_\delta}{\delta}\right) + 2r_{\text{Lip},1}\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\epsilon,\delta)\right) \\
& \leq \epsilon\hat{\mathcal{L}}_N(\theta) + \frac{L(\ln(N))^{r+1}}{\sqrt{N}}\left(2\sigma L_\Theta(\delta) + 4r_{\text{Lip},1}\left(C_{u,1}+C_{u,2}\right)\right. \\
& \left. + 2r_{\text{Lip},2}\frac{L}{512} + \left(C(\theta) + \ln\left(\frac{C_\delta}{\delta}\right) + 2r_{\text{Lip},1}\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\epsilon,\delta)\right)\right)
\end{aligned}
$$

where in the last step we used that $r \geq 0$ and $\ln(N) \geq 1$ for $N \geq 4$. □

F Proof of Corollary 3.4

We start with the following remark explaining the notion of random models drawn from a data-dependent posterior.

Remark F.1 (Joint probability over data and parameters)**.** *Let $\rho^{\mathcal{D}}$ be a function on $\Omega \times \Theta$ which is measurable w.r.t. the direct product $\mathcal{F} \times \mathcal{B}_\Theta$ of σ-algebras, and such that for any $\omega \in \Omega$, $\rho^{\mathcal{D}}(\omega) : \Theta \ni \theta \mapsto \rho^{\mathcal{D}}(\omega,\theta)$ is a probability density on Θ. We refer to $\rho^{\mathcal{D}}$ as a* random density on Θ*. Let B be a subset of $\mathcal{F}\times\mathcal{B}_\Theta$ and denote by $P_{\theta\sim\rho^{\mathcal{D}}}(B)$ the random variable $\omega \mapsto P_{\theta\sim\rho^{\mathcal{D}}(\omega)}(\{\theta \mid (\omega,\theta)\in B\})$, where $P_{\theta\sim\rho^{\mathcal{D}}(\omega)}$ is the probability measure induced by the density $\rho^{\mathcal{D}}(\omega)$. Define the probability measure $\mathbf{P}\times P_{\theta\sim\rho^{\mathcal{D}}}$ on $\mathcal{F}\times\mathcal{B}_\Theta$ as follows:*

$$\left(\mathbf{P}\times P_{\theta\sim\rho^{\mathcal{D}}}\right)(B) \triangleq \mathbf{E}[P_{\theta\sim\rho^{\mathcal{D}}}(B)]$$

Similarly to the previous definition of relationship holding with probability over data, if $\diamond$ is a binary operator and $\{X_i, Y_i\}_{i\in I}$ is family of functions taking values in $\mathbb{R}$ and which are measurable w.r.t. $(\Omega\times\Theta, \mathcal{F}\times\mathcal{B}_\Theta)$, then we say that $\forall i \in I : X_i \diamond Y_i$, holds with probability at least $1-\delta$ over data and over samples from $\rho^{\mathcal{D}}$, if the set $B = \{(\omega,\theta)\in\Omega\times\Theta \mid \forall i\in I : X_i(\omega,\theta)\diamond Y_i(\omega,\theta)\}$ is measurable in $(\Omega\times\Theta,\mathcal{F}\times\mathcal{B}_\Theta)$ and $\left(\mathbf{P}\times P_{\theta\sim\rho^{\mathcal{D}}}\right)(B) \geq 1-\delta$.

Note that if a relationship holds with a probability at least $1-\delta$ over data, then it also holds with a probability at least $1-\delta$ over data and sampled from $\rho^{\mathcal{D}}$: if F is an event from $\mathcal{F}$, then $\left(\mathbf{P}\times P_{\theta\sim\rho^{\mathcal{D}}}\right)(F) = \mathbf{P}(F) \geq 1-\delta$.

F.1 Proof of the (11) on empirical error

First, we present a useful property of Gibbs posteriors.

Lemma F.2 (Property of Gibbs distribution)**.**

$$\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\ln\left(\frac{\rho^{Gibbs}(\theta)}{\pi(\theta)}\right) = -\frac{1}{\lambda}\ln E_{\theta\sim\pi}e^{-\lambda\hat{\mathcal{L}}_N(\theta)}$$
$$= \inf_{\rho\in\mathcal{M}_\pi}\left(\mathbf{E}_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\mathrm{KL}(\rho\|\pi)\right) \tag{114}$$

$$E_{\theta\sim\rho^{Gibbs}}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\mathrm{KL}(\rho^{Gibbs}\|\pi) = -\frac{1}{\lambda}\ln\mathbf{E}_{\theta\sim\pi}e^{-\lambda\hat{\mathcal{L}}_N(\theta)} =$$
$$= \inf_{\rho\in\mathcal{M}_\pi}\left(\mathbf{E}_{\theta\sim\rho}(\hat{\mathcal{L}}_N(\theta)) + \frac{1}{\lambda}\mathrm{KL}(\rho\|\pi)\right) \tag{115}$$

Proof of Lemma F.2. Using the definition of the Gibbs distribution $\rho^{\mathrm{Gibbs}}(\theta) = \frac{\pi(\theta)e^{-\lambda\hat{\mathcal{L}}_N(\theta)}}{Z}$, $Z = E_{\theta\sim\pi}e^{-\lambda\hat{\mathcal{L}}_N(\theta)}$, it follows that

$$\ln\left(\frac{\rho^{\mathrm{Gibbs}}(\theta)}{\pi(\theta)}\right) = -\lambda\hat{\mathcal{L}}_N(\theta) - \ln Z$$

and hence

$$\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\ln\left(\frac{\rho^{\mathrm{Gibbs}}(\theta)}{\pi(\theta)}\right) = \hat{\mathcal{L}}_N(\theta) - \hat{\mathcal{L}}_N(\theta) - \frac{1}{\lambda}\ln Z = -\frac{1}{\lambda}\ln Z$$

This gives the first equality of (114).

As to the first equality of (115), it follows directly from the first equality of (114) by taking the expectation with respect to $\theta\sim\rho^{\mathrm{Gibbs}}$ and recalling that $\mathrm{KL}(\rho^{\mathrm{Gibbs}}\|\pi) = \mathbf{E}_{\theta\sim\rho^{\mathrm{Gibbs}}}\left[\ln\frac{\rho^{\mathrm{Gibbs}}(\theta)}{\pi(\theta)}\right]$.

The second equalities of (114)-(115) are Donsker-Varadhan's variational formula [2, Lemma 2.2]

□

Next, we present a version of Theorem 3.1 where the inequality holds not only over data but over data and any model randomly drawn from a data-dependent posterior.

Lemma F.3 (Single draw bound)**.** *Let $\rho^{\mathcal{D}}$ be a a random density on Θ. For any prior π, any $\delta\in(0,0.5)$ with a probability at least $(1-2\delta)$ over the data and over all samples $\theta_\star$ drawn from $\rho_{\mathcal{D}}$, for any λ satisfying the condition of Theorem 3.1*

$$\epsilon\mathcal{L}(\theta_\star) \le \epsilon\hat{\mathcal{L}}_N(\theta_\star) + \frac{1}{\lambda}\left(\ln\left(\frac{\rho^{\mathcal{D}}(\theta_\star)}{\pi(\theta_\star)}\right) + \ln\left(\frac{C_\delta}{\delta}\right) + \frac{1}{2}\sum_{i=1}^{2}\ln\mathbf{E}_{\theta\sim\pi}\exp(\mathcal{C}_i(2\lambda,\theta_\star,N,\delta,\epsilon))\right) \tag{116}$$

In particular, with probability at least $(1-2\delta)$ over data and over all samples $\theta_\star$ drawn from $\rho_{\mathcal{D}}$,

$$\epsilon\mathcal{L}(\theta_\star) \le \hat{\mathcal{L}}_N(\theta_\star) + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\left(\frac{\rho^{\mathcal{D}}(\theta_\star)}{\pi(\theta_\star)}\right) + \ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,\epsilon)\right) \tag{117}$$

for all $N\ge K_\delta$, where the constants K_δ, C_δ, L, r are defined as in Corollary 3.2. Moreover, if $\epsilon=1$ and $\rho^{\mathcal{D}} = \rho_{Gibbs}$, then with probability at least $(1-2\delta)$ over the data and over $\theta_\star$ sampled from ρ_{Gibbs},

$$\mathcal{L}(\theta_\star) \le -\frac{1}{\lambda}\ln\left(E_{\theta\sim\pi}e^{-\lambda\hat{\mathcal{L}}_N(\theta)}\right) + \frac{1}{\lambda}\left(\ln\left(\frac{C_\delta}{\delta}\right) + \frac{1}{2}\sum_{i=1}^{2}\ln\mathbf{E}_{\theta\sim\pi}\exp(\mathcal{C}_i(2\lambda,\theta,N,\delta,\epsilon))\right) \tag{118}$$

In particular, if the Gibbs posterior ρ^{Gibbs} is defined for $\lambda=\lambda_N = \frac{\sqrt{N}}{L(\ln(N))^r}$ then for $N\ge K_\delta$

$$\mathcal{L}(\theta_\star) \le -\frac{1}{\lambda}\ln E_{\theta\sim\pi}e^{-\lambda\hat{\mathcal{L}}_N(\theta)} + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right) \tag{119}$$

Proof of Lemma F.3. Following the argument of the proof of [2, Theorem 2.7], it follows that for any measurable function X on $\Omega \times \Theta$, $E_{\theta\sim\pi}e^{2\lambda X(\omega,\theta)} = E_{\theta\sim\rho^{\mathcal{D}}}e^{2\lambda X(\omega,\theta)-\ln\frac{\rho^{\mathcal{D}}(\omega,\theta)}{\pi(\theta)}}$ and hence $\mathbf{E}E_{\theta\sim\pi}e^{2\lambda X(\theta)} = \mathbf{E}E_{\theta\sim\rho^{\mathcal{D}}}e^{2\lambda X(\theta)-\ln\frac{\rho^{\mathcal{D}}(\theta)}{\pi(\theta)}}$ with the random variables $X(\theta) : \omega \ni \Omega \mapsto X(\omega,\theta)$. Notice that $\mathbf{E}E_{\theta\sim\rho^{\mathcal{D}}}$ is the expectation operator corresponding to the probability measure $\mathbf{P}\times P_{\theta\sim\rho^{\mathcal{D}}}$, and that $\mathbf{E}[E_{\theta\sim\pi}X] = E_{\theta\sim\pi}\mathbf{E}[X]$, and hence, by Chernoff's bound:

$$\mathbf{P}\times P_{\theta\sim\rho^{\mathcal{D}}}\big(X < \frac{1}{2\lambda}\big(\ln\left(\frac{1}{\delta}\right) + \ln\left(\frac{\rho^{\mathcal{D}}(\theta)}{\pi(\theta)}\right) + \ln\mathbf{E}E_{\theta\sim\pi}e^{2\lambda X}\big)\big)$$

$$= \mathbf{P}\times P_{\theta\sim\rho^{\mathcal{D}}}\Big(\left(X - \frac{1}{2\lambda}\ln\left(\frac{\rho^{\mathcal{D}}(\theta)}{\pi(\theta)}\right)\right) < \frac{1}{2\lambda}\left(\ln\left(\frac{1}{\delta}\right) + \ln\mathbf{E}E_{\theta\sim\pi}e^{2\lambda X}\right)\Big) \leq \frac{\mathbf{E}E_{\theta\sim\rho^{\mathcal{D}}}e^{2\lambda X-\ln\frac{\rho^{\mathcal{D}}(\theta)}{\pi(\theta)}}}{e^{\ln\frac{1}{\delta}+\ln\mathbf{E}E_{\theta\sim\pi}e^{2\lambda X}}}$$

$$= \frac{\mathbf{E}E_{\theta\sim\pi}e^{2\lambda X}}{\frac{1}{\delta}\mathbf{E}E_{\theta\sim\pi}e^{2\lambda X}} = \delta$$

Hence,

$$X \leq \frac{1}{2\lambda}\big(\ln\left(\frac{1}{\delta}\right) + \ln\left(\frac{\rho^{\mathcal{D}}(\theta)}{\pi(\theta)}\right) + \ln E_{\theta\sim\pi}\mathbf{E}e^{2\lambda X(\theta)}\big) \tag{120}$$

with probability at least $1-\delta$ over data and samples from $\rho^{\mathcal{D}}$. Hence, by taking $X = \epsilon(\mathcal{L}(\theta) - V_N(\theta))$ and then $X = \epsilon(V_N(\theta)) - \hat{\mathcal{L}}_N(\theta))$, and by taking the union bound as in the proof of Theorem E.1, it follows that

$$\begin{aligned}\epsilon\mathcal{L}(\theta) \leq& \epsilon\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\Big(\ln\left(\frac{1}{\delta}\right) + \ln\left(\frac{\rho^{\mathcal{D}}(\theta)}{\pi(\theta)}\right)\\ &+ \frac{1}{2}\Big(\ln E_{\theta\sim\pi}\mathbf{E}[e^{2\epsilon\lambda(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}] + \ln E_{\theta\sim\pi}\mathbf{E}[e^{2\epsilon\lambda(\mathcal{L}(\theta)-V_N(\theta))}]\Big)\Big)\end{aligned} \tag{121}$$

holds with probability at least $1-2\delta$ over data and samples from $\rho^{\mathcal{D}}$.

Proof of (116) **for bounded noise** Let us apply the inequality (121) and use the bound on $\mathbf{E}[e^{2\epsilon\lambda(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}]$ and $\mathbf{E}[e^{2\epsilon\lambda(\mathcal{L}(\theta)-V_N(\theta))}]$ from Lemma E.6-E.5, from which (116) follows.

Proof of (116) **for Lipschitz loss** Let us apply (44) from Subsection E.2 to bound the bound on $\mathbf{E}[e^{2\epsilon\lambda(V_N(\theta)-\hat{\mathcal{L}}_N(\theta))}]$ and $\mathbf{E}[e^{2\epsilon\lambda(\mathcal{L}(\theta)-V_N(\theta))}]$, from which (116) follows.

Proof of (116) **quadratic loss** Recall from Subsection E.3 the definition of the losses ${}^C\mathcal{L}(\theta)$, ${}^C\hat{\mathcal{L}}_N(\theta)$. corresponding to the loss ℓ_C. Note that by Lemma E.19 ℓ_C is $2C$-Lipschitz. Notice that from the statement of the present lemma proven for the case of Lipschitz losses for $\lambda \leq \frac{\sqrt{N}}{32\cdot 2C\ln(N)}$, with $C = 6G_e\|\mathbf{e}_g\|_{\psi_2}\sqrt{\ln(\frac{3N^2}{2\delta})}$ probability $1-\frac{4}{3\delta}$ over data and parameters sampled from $\rho^{\mathcal{D}}$, the following holds:

$$\begin{aligned}\epsilon\,{}^C\mathcal{L}(\theta) \leq& \epsilon\,{}^C\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\left(\ln\left(\frac{\rho(\theta)}{\pi(\theta)}\right) + \ln\left(\frac{3}{2\delta}\right) + \right.\\ &\left.\frac{1}{2}\sum_{i=1}^{2}\ln\mathbf{E}_{\theta\sim\pi}\exp\left(\bar{C}_i(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon)\right)\right)\end{aligned} \tag{122}$$

where $\bar{C}_i(2\lambda,\theta,N,\frac{2\delta}{3},\epsilon)$ satisfies (60). From Lemma E.18 it follows that (56) holds with probability at least $1-\frac{2\delta}{3}$ over data and hence (56) holds with probability at least $1-\frac{2\delta}{3}$ over data and parameters sampled from ρ^{θ}. Hence, by using the union bound, it follows that (122) and (56) hold simultaneously with probability at least $1-2\delta$ over data and parameters sampled from $\rho^{\mathcal{D}}$.

Moreover, if (57) holds, then it holds that ${}^C\hat{\mathcal{L}}_N(\theta) = \hat{\mathcal{L}}_N(\theta)$ and by and (68) – (67)

$${}^C\mathcal{L}(\theta) \leq \mathcal{L}(\theta) \leq {}^C\mathcal{L}(\theta) + 72G_e^2\|\mathbf{e}_g\|_{\psi_2}^2\frac{\ln\left(\frac{3N^2}{2\delta}\right)}{N}$$

Hence,

$$\epsilon\mathcal{L}(\theta) \leq \epsilon^{C}\mathcal{L}(\theta) + 72G_e^2\|\mathbf{e}_g\|_{\psi_2}^2 \frac{\ln\left(\frac{3N^2}{2\delta}\right)}{N}\max\{\epsilon, 0\} \tag{123}$$

Hence, by using (122) and ${}^{C}\hat{\mathcal{L}}_N(\theta) = \hat{\mathcal{L}}_N(\theta)$, it follows that if (123) and (56) both hold, then (116) holds. Since (123) and (56) holds with probability at least $1-2\delta$ over data and samples from $\rho^{\mathcal{D}}$, so does (116).

Proof of (116) **implies** (118) It follows by applying Lemma F.2.

Proof of (117) From the proof of Corollary 3.2 it follows that for $\lambda = \frac{\sqrt{N}}{L(\ln(N))^r}$ where r is as in Corollary 3.2,

$$\frac{1}{2\lambda}\left(\ln E_{\theta\sim\pi}\exp(\mathcal{C}_1(2\lambda,\theta,N,\delta,\epsilon)) + \ln E_{\theta\sim\pi}\exp(\mathcal{C}_2(2\lambda,\theta,N,\delta,\epsilon))\right) \leq \mathcal{C}(\delta,\epsilon)$$

Using this inequality, it follows that (116) implies (117). □

Lemma F.4. *If $\mathcal{L}(\theta)$ is convex, ρ is a probability density over Θ, and $\theta_\star = E_{\theta\sim\rho}\theta$ is the mean of the density ρ, then*

$$\mathcal{L}(\theta_\star) \leq E_{\theta\sim\rho}[\mathcal{L}(\theta)]. \tag{124}$$

Proof of Lemma F.4. If $\mathcal{L}(\theta)$ is convex in θ, then $\mathcal{L}(\theta_\star) = \mathcal{L}(E_{\theta\sim\rho^{\mathcal{D}}}\theta) \leq E_{\theta\sim\rho^{\mathcal{D}}}\mathcal{L}(\theta)$ by Jensen's inequality. □

Proof of (11). For the single draw case, from (118) of Lemma F.3, it follows that with probability at least $1-2\delta$ over data and samples from ρ^{Gibbs},

$$\mathcal{L}(\theta_\star) \leq -\ln E_{\theta\sim\pi}e^{-\lambda\hat{\mathcal{L}}_N(\theta)} + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right) \tag{125}$$

for $\lambda = \sqrt{N}/(L(\ln(N))^r)$ For the MEP case, using Lemma F.4 and Lemma F.2 and Corollary 3.2it follows that with probability at least $1-2\delta$ over data, it holds that

$$\begin{aligned}
\mathcal{L}(\theta_\star) &\leq E_{\theta\sim\rho^{\text{Gibbs}}}\mathcal{L}(\theta) \leq E_{\theta\sim\rho^{\text{Gibbs}}}\hat{\mathcal{L}}_N(\theta) + \underbrace{\frac{L(\ln(N))^r}{\sqrt{N}}}_{1/\lambda}\left(\text{KL}(\rho^{\text{Gibbs}} \mid \pi) + \ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right) \\
&= \left(E_{\theta\sim\rho^{\text{Gibbs}}}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\text{KL}(\rho^{\text{Gibbs}} \mid \pi)\right) + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right) \\
&= -\ln E_{\theta\sim\pi}e^{-\lambda\hat{\mathcal{L}}_N(\theta)} + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right)
\end{aligned} \tag{126}$$

with $\lambda = \sqrt{N}/(L(\ln(N))^r)$. Using the second equality of (115), it follows that for any posterior $\rho \in \mathcal{M}_\pi$,

$$-\frac{1}{\lambda}\ln E_{\theta\sim\pi}e^{-\lambda\hat{\mathcal{L}}_N(\theta)} \leq E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\text{KL}(\rho\|\pi) \tag{127}$$

Choose $\rho_{\theta_\star}$ as in Lemma E.22, i.e.,

$$\rho_{\theta_\star} = \begin{cases} \mathcal{N}(\theta_\star, \frac{\sigma^2}{N}) & \text{if } \textbf{(Gauss)} \text{ holds} \\ \mathcal{U}(\theta_\star, \frac{\sigma}{\sqrt{N}}) & \text{if } \textbf{(Uni)} \text{ holds} \end{cases}$$

From Lemma E.23 it follows that (87) holds with probability at least $1 - r_{\text{Lip},1}\delta$ over data, and hence over data and samples from ρ^{Gibbs}. In particular, as (87) implies

$$\begin{aligned}
E_{\theta\sim\rho_{\theta_\star}}\hat{\mathcal{L}}_N(\theta) \leq \hat{\mathcal{L}}_N(\theta_\star) + L_\Theta(\delta)L(\ln(N))^{r+1}\underbrace{E_{\theta\sim\rho_{\theta_\star}}\|\theta-\theta_\star\|_2}_{\leq\frac{\sigma}{\sqrt{N}}} + \\
2r_{\text{Lip},1}\frac{L(\ln(N))^r}{\sqrt{N}}\left(C_{u,1} + C_{u,2} + \ln\left(\frac{C_\delta}{\delta}\right)\right),
\end{aligned} \tag{128}$$

It then follows that (128) holds with probability at least $1 - r_{\text{Lip},1}\delta$ over data and samples from ρ^{Gibbs}. Then using the union bound on (125),(128), and (127) and Lemma E.22 it follows that

$$\begin{aligned}
&\mathcal{L}(\theta_\star) \leq E_{\theta\sim\rho^{\text{Gibbs}}}\hat{\mathcal{L}}_N(\theta) \leq -\ln E_{\theta\sim\pi}e^{-\lambda\hat{\mathcal{L}}_N(\theta)} + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\frac{C_\delta}{\delta} + \mathcal{C}(\delta,1)\right)\\
&\leq E_{\theta\sim\rho_{\theta_\star}}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\,KL(\rho_{\theta_\star}\|\pi) + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right)\\
&\leq \hat{\mathcal{L}}_N(\theta_\star) + L_\Theta(\delta)L(\ln(N))^{r+1}\frac{\sigma}{\sqrt{N}} + 2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1} + C_{u,2} + \ln\left(\frac{C_\delta}{\delta}\right)\right)\\
&+ \frac{L(\ln(N))^{r+1}}{\sqrt{N}}C(\theta_\star) + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right)\\
&= \hat{\mathcal{L}}_N(\theta_\star) + \frac{L(\ln(N))^{r+1}\sigma}{\sqrt{N}}\Big(\underbrace{L_\Theta(\delta)\sigma + 2r_{\text{Lip},1}(C_{u,1}+C_{u,2})}_{\mathcal{C}_{L,emp}} + 2r_{\text{Lin},1}\ln\left(\frac{C_\delta}{\sigma}\right)\\
&+ \underbrace{\left(C(\theta_\star) + \ln\left(\frac{C_\delta}{\delta}\right) + C(\delta,1)\right)}_{\mathcal{C}_L(\sigma,1,\theta_\star)}\Big) \leq \hat{\mathcal{L}}_N(\theta_\star) + \frac{L(\ln(N))^{r+1}\sigma}{\sqrt{N}}\mathcal{C}_{em}(\delta,\sigma)
\end{aligned} \tag{129}$$

with probability at least $1 - (2 + r_{\text{Lip},1})\delta$ over data and $\theta_\star$ sampled from ρ^{Gibbs}. Similarly, by using that (126) holds with probability at least $1 - 2\delta$, the fact that (87) holds with probability as least $1 - r_{\text{Lip},1}\delta$ over data, and that (87) implies (127), it follows, using the union bound that for MEP,

$$\begin{aligned}
\mathcal{L}(\theta_\star) \leq& E_{\theta\sim\rho^{\text{Gibbs}}}\hat{\mathcal{L}}_N(\theta) \leq -\ln e^{-\lambda E_{\theta\sim\pi}\hat{\mathcal{L}}_N(\theta)} + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right)\\
&\leq E_{\theta\sim\rho_{\theta_\star}}\hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\,KL(\rho_{\theta_\star}\|\pi) + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right)\\
&\leq \hat{\mathcal{L}}_N(\theta_\star) + L_\Theta(\delta)L(\ln(N))^{r+1}\frac{\sigma}{\sqrt{N}} + \frac{L(\ln(N))^{r+1}}{\sqrt{N}}\mathcal{C}(\theta_\star)\\
&+ 2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}\left(C_{u,1} + C_{u,2} + \ln\left(\frac{C_\delta}{\delta}\right)\right)\\
&+ \frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right)\\
&= \hat{\mathcal{L}}_N(\theta_\star) + \frac{L(\ln(N))^r}{\sqrt{N}}\Big(\underbrace{(L_\Theta(\delta)\sigma + 2r_{\text{Lip},1}\,(C_{u,1}+C_{u,2}))}_{\mathcal{C}_{L,emp}} + 2r_{\text{Lip},1}\ln\left(\frac{C_\delta}{\delta}\right) +\\
&+ \underbrace{\left(\mathcal{C}(\theta_\star) + \ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right)}_{\mathcal{C}_L(\delta,1,\theta_\star)}\Big) \leq \hat{\mathcal{L}}_N(\theta_\star) + \frac{L(\ln(N))^{r+1}\sigma}{\sqrt{N}}\mathcal{C}_{em}(\delta,\sigma)
\end{aligned} \tag{130}$$

holds with probability at least $1 - (2 + r_{\text{Lip},1})\delta$ over data.

Finally, the case of MAP follows from Lemma E.21 by applying it to the random model $\theta_\star$: as

$$\begin{aligned}
&\forall\theta\in\Theta: \mathcal{L}(\theta) \leq \hat{\mathcal{L}}_N(\theta) +\\
&\quad\frac{L(\ln(N))^{r+1}}{\sqrt{N}}\left(\mathcal{C}_{L,true} + \mathcal{C}_{L,emp} + C(\theta) + \ln\left(\frac{C_\delta}{\delta}\right) + 2r_{\text{Lip},1}\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right)
\end{aligned}$$

holds with probability at least $1-(2+r_{\text{Lip},1})\delta$ over data, it then follows that for $\theta=\theta_\star$,

$$\begin{aligned}\mathcal{L}(\theta_\star) &\leq \hat{\mathcal{L}}_N(\theta_\star)+\frac{L(\ln(N))^{r+1}}{\sqrt{N}}\Bigg(\mathcal{C}_{L,true}+\\ &+\underbrace{\left(\mathcal{C}_{L,emp}+C(\theta_\star)+\ln\left(\frac{C_\delta}{\delta}\right)+\mathcal{C}(\delta,1)\right)}_{\mathcal{C}_L(\sigma,1,\theta_\star)}+2r_{\text{Lip},1}\ln\left(\frac{C_\delta}{\delta}\right)\Bigg)\\ &=\hat{\mathcal{L}}_N(\theta_\star)+\frac{L(\ln(N))^{r+1}}{\sqrt{N}}\mathcal{C}_{em}(\delta,\sigma)\end{aligned}$$

□

F.2 Proof of (12)

From (10) it follows that each of the inequalities below hold with probability at least $1-2\delta$ over data

$$\forall\rho\in\mathcal{M}_\pi: E_{\theta\sim\rho}\mathcal{L}(\theta)\leq E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta)+\frac{L(\ln(N))^r}{\sqrt{N}}(\text{KL}(\rho||\pi)+\ln\left(\frac{C_\delta}{\delta}\right)+\mathcal{C}(\delta,1)) \quad (131)$$

$$\forall\rho\in\mathcal{M}_\pi: -E_{\theta\sim\rho}\mathcal{L}(\theta)\leq -E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta)+\frac{L(\ln(N))^r}{\sqrt{N}}(\text{KL}(\rho||\pi)+\ln\left(\frac{C_\delta}{\delta}\right)+\mathcal{C}(\delta,-1)) \quad (132)$$

In particular, by using the union bound it follows that (131) and (132) hold simultaneously with probability at least $1-4\delta$. Using the fact that for any posterior $\rho\in\mathcal{M}_\pi$,

$$E_{\theta\sim\rho^{\text{Gibbs}}}\hat{\mathcal{L}}_N(\theta)+\frac{L(\ln(N))^r}{\sqrt{N}}\text{KL}(\rho^{\text{Gibbs}}||\pi)\leq E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta)+\frac{L(\ln(N))^r}{\sqrt{N}}(\text{KL}(\rho||\pi)$$

it then follows that with probability at least $1-4\delta$ over data

$$\begin{aligned}\forall\rho\in\mathcal{M}_\pi: &E_{\theta\sim\rho^{\text{Gibbs}}}\mathcal{L}(\theta)\leq E_{\theta\sim\rho^{\text{Gibbs}}}\hat{\mathcal{L}}_N(\theta)+\frac{L(\ln(N))^r}{\sqrt{N}}(\text{KL}(\rho^{\text{Gibbs}}||\pi)+\ln\left(\frac{C_\delta}{\delta}\right)+\mathcal{C}(\delta,1))\\ &\leq E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta)+\frac{L(\ln(N))^r}{\sqrt{N}}\text{KL}(\rho||\pi)+\frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\left(\frac{C_\delta}{\delta}\right)+\mathcal{C}(\delta,1)\right)\\ &\leq E_{\theta\sim\rho}\mathcal{L}(\theta)+\frac{L(\ln(N))^r}{\sqrt{N}}\left(\text{KL}(\rho||\pi)+\ln\left(\frac{C_\delta}{\delta}\right)+\mathcal{C}(\delta,-1)\right)+\\ &+\frac{L(\ln(N))^r}{\sqrt{N}}\left(\text{KL}(\rho||\pi)+\ln\left(\frac{C_\delta}{\delta}\right)+\mathcal{C}(\delta,1)\right)\\ &=E_{\theta\sim\rho}\mathcal{L}(\theta)+\frac{L(\ln(N))^r}{\sqrt{N}}\left(2\text{KL}(\rho||\pi)+2\ln\left(\frac{C_\delta}{\delta}\right)+\mathcal{C}(\delta,-1)+\mathcal{C}(\delta,1)\right) \quad (133)\end{aligned}$$

For the posterior $\rho_{\theta_{true}}$ from Lemma E.22, using (88) from Lemma E.23, it follows that

$$\begin{aligned}&E_{\theta\sim\rho_{\theta_{true}}}\mathcal{L}(\theta)\leq\mathcal{L}(\theta_{true})\\ &+L_\Theta(\delta)L(\ln(N))^rE_{\theta\sim\rho_{\theta_{true}}}\|\theta-\theta_{true}\|_2\\ &+2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}(C_{u,1}+C_{u,2})+2r_{\text{Lip},2}\frac{L}{512}\frac{L(\ln(N))^r}{\sqrt{N}}\\ &\leq\mathcal{L}(\theta_{true})+L_\Theta(\delta)\sigma\frac{L(\ln(N))^r}{\sqrt{N}}+2r_{\text{Lip},1}\frac{L\ln(N)^r}{\sqrt{N}}(C_{u,1}+C_{u,2})+2r_{\text{Lip},2}\frac{L}{512}\frac{L(\ln(N))^r}{\sqrt{N}}\\ &\leq\mathcal{L}(\theta_{true})+\frac{L\ln(N)^r}{\sqrt{N}}\mathcal{C}_{L,true} \quad (134)\end{aligned}$$

By applying (133) with $\rho=\rho_{\theta_{true}}$ and using that by Lemma E.22

$$\text{KL}(\rho_{\theta_{true}}||\pi)\leq\mathcal{C}(\theta_{true})\ln(N)$$

and the fact that $\ln(N) \geq 1$ it follows that

$$E_{\theta\sim\rho^{\text{Gibbs}}}\mathcal{L}(\theta) \leq \mathcal{L}(\theta_{true}) + \frac{L\ln(N)^{r+1}}{\sqrt{N}}\left(\mathcal{C}_{L,true} + \underbrace{2\mathcal{C}(\theta_{true}) + 2\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,-1) + \mathcal{C}(\delta,1)}_{\mathcal{C}_L(\delta,1,\theta_{true})+\mathcal{C}_L(\delta,-1,\theta_{true})-2\mathcal{C}_{L,emp}}\right) \tag{135}$$

with probability at least $1-4\delta$. If $\theta_\star$ is MEP, then by Lemma F.4 $\mathcal{L}(\theta_\star) \leq E_{\theta\sim\rho^{\text{Gibbs}}}\mathcal{L}(\theta)$. Combining this with (135), (12) for MEP follows.

For single draw, from Lemma F.3, (119), it follows that with probability at least $1-2\delta$ overa data and samples $\theta_\star$ from ρ^{Gibbs} it holds that

$$\mathcal{L}(\theta_\star) \leq -\frac{1}{\lambda}\ln E_{\theta\sim\pi}e^{-\lambda\hat{\mathcal{L}}_N(\theta)} + \frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right) \tag{136}$$

with $\lambda = \lambda_N = \frac{\sqrt{N}}{L(\ln N)^r}$. By using(114) from Lemma F.2, it follows that for $\rho_{\theta_{true}}$ from Lemma E.22

$$-\frac{1}{\lambda}\ln E_{\theta\sim\pi}e^{-\lambda\hat{\mathcal{L}}_N(\theta)} \leq E_{\theta\sim\rho_{\theta_{true}}}\mathcal{L}_N(\theta) + \frac{L(\ln(N))^r}{\sqrt{N}}\text{KL}(\rho_{\theta_{true}}||\pi) \tag{137}$$

Using that 132 holds with probability at least $1-2\delta$ over data, and using (134), and $\ln(N) \geq 1$, it follows that

$$\begin{aligned}
-\frac{1}{\lambda}\ln E_{\theta\sim\pi}e^{-\lambda\hat{\mathcal{L}}_N(\theta)} &\leq E_{\theta\sim\rho_{\theta_{true}}}\mathcal{L}_N(\theta) + \frac{L(\ln(N))^r}{\sqrt{N}}\text{KL}(\rho_{\theta_{true}}||\pi) \\
&\leq E_{\theta\sim\rho_{\theta_{true}}}\mathcal{L}(\theta) + \frac{L(\ln(N))^r}{\sqrt{N}}\,KL(\rho_{\theta_{true}}||\pi) \\
&+ \frac{L(\ln(N))^r}{\sqrt{N}}\left(\text{KL}(\rho_{\theta_{true}}||\pi) + \ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,-1)\right) \\
&\leq \mathcal{L}(\theta_{true}) + \frac{L(\ln(N))^{r+1}}{\sqrt{N}}\left(\mathcal{C}_{L,true} + 2\mathcal{C}(\theta_{true}) + \ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,-1)\right)
\end{aligned} \tag{138}$$

with probability at least $1-2\delta$ over data and hence over data and samples from ρ^{Gibbs}. By combining (136) and (138) using the union bounds, it follows that with probability $1-4\delta$ over data and samples $\theta_\star$ from ρ^{Gibbs},

$$\begin{aligned}
\mathcal{L}(\theta_\star) \leq& \mathcal{L}(\theta_{true}) + \frac{L(\ln(N))^{r+1}}{\sqrt{N}}\left(\mathcal{C}_{L,true} + 2\mathcal{C}(\theta_{true}) + \ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,-1))\right) + \\
&+ \frac{L(\ln(N))^r}{\sqrt{N}}\left(\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1)\right) \\
&\leq \mathcal{L}(\theta_{true}) + \frac{L(\ln(N))^{r+1}}{\sqrt{N}}\left(\mathcal{C}_{L,true} + \underbrace{2\mathcal{C}(\theta_{true}) + 2\ln\left(\frac{C_\delta}{\delta}\right) + \mathcal{C}(\delta,1) + \mathcal{C}(\delta,-1)}_{\mathcal{C}_L(\delta,1,\theta_{true})+\mathcal{C}_L(\delta,-1,\theta_{true})-2\mathcal{C}_{L,emp}}\right)
\end{aligned}$$

from which the statement of (12) for single draw follows, by replacing δ by $\frac{\delta}{2}$.

Finally, we consider the case of MAP. Since (11), i.e.,

$$\begin{aligned}
\mathcal{L}(\theta_\star) \leq& \hat{\mathcal{L}}_N(\theta_\star) + \\
&\frac{L(\ln(N))^{r+1}}{\sqrt{N}}\Big(\mathcal{C}_{L,true} + \mathcal{C}_L(\delta,1,\theta_\star) + 2r_{\text{Lip},1}\ln\left(\frac{C_\delta}{\delta}\right)\Big)
\end{aligned} \tag{139}$$

holds with probability at least $1-(2+r_{\text{Lip},1})\delta$. From (86) applied to $\theta = \theta_{true}$ and $\epsilon = -1$ it also follows that

$$\begin{aligned}
\hat{\mathcal{L}}_N(\theta_{true}) \leq& \mathcal{L}(\theta_{true}) + \\
&\frac{L(\ln(N))^{r+1}}{\sqrt{N}}\Big(\mathcal{C}_{L,true} + \mathcal{C}_L(\delta,-1,\theta_{true}) + 2r_{\text{Lip},1}\ln\left(\frac{C_\delta}{\delta}\right)\Big)
\end{aligned} \tag{140}$$

holds with probability at least $1-(2+\mathrm{r}_{\mathrm{Lip},1})\delta$ over data. If $\theta_\star$ is MAP, then $\theta_\star$ minimizes $\hat{\mathcal{L}}_N(\theta) - \frac{1}{\lambda}\ln\pi(\theta)$ and hence $\hat{\mathcal{L}}_N(\theta_\star) - \frac{1}{\lambda}\ln\pi(\theta_\star) \leq \hat{\mathcal{L}}_N(\theta_{true}) - \frac{1}{\lambda}\ln\pi(\theta_{true})$. Hence,

$$\hat{\mathcal{L}}_N(\theta_\star) \leq \hat{\mathcal{L}}_N(\theta_{true}) + \frac{1}{\lambda}\ln\left(\frac{\pi(\theta_\star)}{\pi(\theta_{true})}\right)$$

and then by using (139)-(140) and the fact that $\lambda = \frac{\sqrt{N}}{L(\ln(N))^r}$ and hence $\frac{1}{\lambda} \leq \frac{L(\ln(N))^{r+1}}{\sqrt{N}}$, it follows that with probability $1-(4+2r_{\mathrm{Lip},1})\delta$ over data,

$$\begin{aligned}
&\hat{\mathcal{L}}_N(\theta_\star) \leq \mathcal{L}(\theta_{true}) \\
&+ \frac{L(\ln(N))^{r+1}}{\sqrt{N}}\Big(2\mathcal{C}_{L,true} + \mathcal{C}_L(\delta,1,\theta_\star) + \mathcal{C}_L(\delta,-1,\theta_{true}) + 4r_{\mathrm{Lip},1}\ln\left(\frac{C_\delta}{\delta}\right)\Big) \\
&+ \frac{L(\ln(N))^{r+1}}{\sqrt{N}}\ln\left(\frac{\pi(\theta_\star)}{\pi(\theta_{true})}\right)
\end{aligned} \tag{141}$$

holds, hence, by replacing δ with $\delta/2$, it follows that (12) holds with probability at least $1-(2+r_{\mathrm{Lip},1})\delta$.

F.3 Proof of (13)

If **(SConv)** holds, then by [30, Theorem 2.1.11, Theorem 2.1.8]

$$m_\Theta\|\theta_\star - \theta_{true}\|^2 \leq \mathcal{L}(\theta_\star) - \mathcal{L}(\theta_{true})$$

from which using (12) the inequality (13) follows.

F.4 Proof of (14)

From [26, Chapter 4, page 87] it follows that $\sigma^2_{\mathbf{e}^s} = \mathcal{L}(\theta_{\mathrm{true}}) = \inf_{\theta\in\Theta}\mathcal{L}(\theta)$, where $\sigma^2_{\mathbf{e}^s} = \mathbf{E}[\|\mathbf{e}^s(t)\|_2^2]$, if ℓ is the quadratic loss function, i.e. the true system yields the best possible prediction in mean square sense.

$$\forall\theta\in\Theta : \sigma^2_{\mathbf{e}^s} = \mathcal{L}(\theta_{true}) \leq \mathcal{L}(\theta) \tag{142}$$

where $\sigma^2_{\mathbf{e}^s}$ is the variance of $\mathbf{e}^s(t)$. That is, the parameter θ_{true} corresponding to the data generator results in the smallest possible generalisation loss for quadratic loss function. Let H_θ be the transfer function of the predictor parametrized by θ

It can be shown (see [12, Appendix, (82)]) and [26, Chapter 4, page 87]) that

$$\|H_{\theta_{\mathrm{true}}} - H_\theta\|^2_{H_2} \leq \frac{1}{m_{\mathbf{w}}}(\mathcal{L}(\theta) - \mathcal{L}(\theta_{true})), \tag{143}$$

Then (14) follows from (12) and (143).

F.5 Remark on relationship with system identification and learning

Remark F.5 (Identification of FIR models)**.** *Consider the SISO case ($n_y = n_u = 1$), fix an integer T_0 and let $\Theta = \mathbb{R}^{T_0}$ and for any $\theta = (\theta_1,\ldots,\theta_{T_0})^T \in \Theta$, let $D_\theta = \theta_1$,*

$$A = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix},\ B = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},\ C_\theta = \begin{bmatrix} \theta_2 \\ \vdots \\ \theta_{T_0} \\ 1 \end{bmatrix}^T$$

Then $\hat{\mathbf{y}}_\theta(t) = u(t-T_0) + \sum_{i=0}^{T_0-1}\theta_{i+1}\mathbf{u}(t-i)$. Assume that there exists $\theta_{true}\in\Theta$, such that $\mathbf{y}(t) = \hat{\mathbf{y}}_{\theta_{true}}(t) + \mathbf{e}(t)$, where $\mathbf{e}(t)$ is i.i.d. with zero mean and variance σ^2, and independent of $\mathbf{u}(s)$, $s\in\mathbb{Z}$. Let $\hat{\mathbf{u}}(t) = \mathbf{u}(t)$ for $t>0$ and let $\hat{\mathbf{u}}(t) = 0$ for $t<0$. Let Φ be the regressor matrix constructed from $\tilde{\mathbf{u}}$ as in [34, (3.45), page 48] and let $Y = [\mathbf{y}(0) \quad \cdots \quad \mathbf{y}(N-1)]^T$. Then $\hat{\mathcal{L}}_N(\theta) = \frac{1}{N}\|Y-\Phi\theta\|_2^2$. Assume that the prior π is the Gaussian density $\mathcal{N}(\tilde{\theta},P)$, $P=P^T>0$. Then a lengthy but straightforward calculation reveal that the Gibbs posterior is the Gaussian

density $\mathcal{N}(\tilde{\theta} + P\Phi^T(\Phi P\Phi^T + \frac{N}{2\lambda}I)^{-1}(Y - \Phi\tilde{\theta}), P - P\Phi^T(\Phi P\Phi^T + \frac{N}{2\lambda}I)^{-1}\Phi P)$*, which is the conditional density of* θ *w.r.t.* Y, Φ *presented in [34, (4.6), page 101], if* $Y = \Phi\theta + E$ *is assumed with* $E \sim \mathcal{N}(0, \frac{N}{2\lambda}I_{T_0})$*. Then the MAP and MEP estimates are both* $\theta_\star = \tilde{\theta} + P\Phi^T(\Phi P\Phi^T + \frac{N}{2\lambda}I)^{-1}(Y - \Phi\theta_{true})$ *and they solve the regularized optimization problem* $\arg\min_{\theta\in\Theta} \lambda\hat{\mathcal{L}}_N(\theta) + (\theta - \tilde{\theta})P^{-1}(\theta - \tilde{\theta})$*. Moreover, they are also the maximum likelihood of the posterior* $P(\theta \mid Y, \Phi)$*.*

Then Corollary 3.4 applies to $\theta_\star$*, yielding a* $O(\frac{(\ln(N))^r}{\sqrt{N}})$ *bound on the generalization gap of* $\theta_\star$ *and a* $O(\frac{(\ln(N))^{r+1}}{\sqrt{N}})$ *bound on the square parameter estimation error* $\|\theta_\star - \theta_{true}\|_2^2$*.*

Remark F.6. *(*Maximum likelihood estimates of state-space models*). Consider the parametrization of LTI systems*

$$\begin{aligned}\hat{\mathbf{x}}(t+1) &= (\hat{A}_\theta + \hat{L}_\theta\hat{C}_\theta)\hat{\mathbf{x}}(t) + (\hat{B}_\theta + L_\theta D_\theta)\mathbf{u}(t) + \hat{L}_\theta\mathbf{e}_\theta(t),\\ \mathbf{y}_\theta(t) &= \hat{C}_\theta\hat{\mathbf{x}}(t) + \hat{D}_\theta\mathbf{u}(t) + \mathbf{e}_\theta(t)\end{aligned} \tag{144}$$

where A_θ *and* $A_\theta + L_\theta C_\theta$ *are Schur,* $\mathbf{e}_\theta(t)$ *is i.i.d., zero mean Gaussian with variance* $\frac{1}{\lambda}$*,* $(\mathbf{u}, \mathbf{e}_\theta)$ *is jointly Gaussian,* $\mathbf{e}_\theta(t)$ *is independent of* $\mathbf{u}(s)$ *for* $s < t$*. We can associate with* (144) *the predictor*

$$\begin{aligned}\hat{\mathbf{x}}(t+1) &= \hat{A}_\theta\hat{\mathbf{x}}(t) + \hat{B}_\theta\mathbf{u}(t) + \hat{L}_\theta\mathbf{y}(t),\\ \hat{\mathbf{y}}_\theta(t \mid 0) &= \hat{C}_\theta\hat{\mathbf{x}}(t) + \hat{D}_\theta\mathbf{u}(t),\end{aligned} \tag{145}$$

Assume that $\mathbf{y}$ *is generated by an LTI system of the form* (144) *which satisfies [4, (1), INP B, page 427].*

In this case, by [4, (2.7), page 422, Chapter 7] the conditional likelihood function $p_\theta(\{y(t)\}_{t=-1}^{N-1} \mid \{u(t)\}_{t=0}^{N-1})$ $\{\mathbf{y}_\theta(t)\}_{t=0}^{N-1}$ *w.r.t* $\{\mathbf{u}(t)\}_{t=0}^{N-1}$*, evaluated at* $y(t) = \mathbf{y}(t)$ *and* $u(t) = \mathbf{u}(t)$*,* $t = -1, \ldots, N-1$ *equals* $Ze^{-\frac{N\lambda}{2}\hat{\mathcal{L}}_N(\theta)}$ *for a suitable constant* Z*, where the empirical error* $\hat{\mathcal{L}}_N(\theta)$ *is defined for the quadratic loss. That is, i.e., up to an additive constant, its logarithm equals* $\ln\rho^{Gibbs}(\theta)$*. In particular, the maximum likelihood estimate* $\operatorname{argmax}_{\theta\in\Theta} p_\theta(\{y(t)\}_{t=0}^{N-1} \mid \{u(t)\}_{t=0}^{N-1})$*, see [4, Chapter 7,page 428] equals the MAP* $\operatorname{argmax}_{\theta\in\Theta}\rho^{Gibbs}(\theta)$ *of the Gibbs posterior corresponding to the uniform distribution* $\pi = \frac{1}{vol(\Theta)}$ *as a prior. Here we assume that* $vol(\Theta)$ *is finite.*

We can then use the bounds (11)*,* (12)*,* (13)*,* (14)*, for the generalization gap, the gap between the true error of the learned model and the best true error and the parameter estimation error respectively.*

Remark F.7 (Bayesian methods for estimating LTI systems)**.** *The Bayesian method for system identification of LTI systems [31] corresponds to taking a single draw, MEP or MAP from the conditional distribution of the parameter given the past observations. If the parametrization is in innovation form, i.e., of the form* (144) *and all the signals are jointly Gaussian, then by Remark F.6 the joint density of* $\{\mathbf{y}(t)\}_{t=-1}^{N-1}$ *given* $\{\mathbf{u}(t)\}_{t=0}^{N-1}$ *and* $\{\mathbf{y}(s), \mathbf{u}(s)\}_{s<0}$ *and a parameter* θ *is* $Ze^{-N\lambda\hat{\mathcal{L}}_N(\theta)}$ *for a suitable constant, and hence by [31, Section 3, eq. (12)] the conditional density of* θ *with respect to* $\{\mathbf{y}(t), \mathbf{u}(t)\}_{t=0}^{N-1}$ *is the Gibbs posterior* $\frac{e^{-N\lambda\hat{\mathcal{L}}_N(\theta)}p_\theta(\mathbf{y}(-1))\pi(\theta)}{E_{\theta\in\pi}e^{-\lambda\hat{\mathcal{L}}_N(\theta)}}$ *for the quadratic loss and a suitable prior, and hence Gibbs posterior for the predictors* (145)*, and hence taking a random draw, MEP or MAP from this conditional density, as suggested by [31], corresponds to taking a random draw, MEP or MAP from the Gibbs posterior. In particular, the bounds* (11)*,* (12)*,* (13)*,* (14) *apply.*

Remark F.8 (PEM methods)**.** *To begin with if a PEM algorithm can be reduced to optimizing a cost function* $\hat{\mathcal{L}}_N(\theta) + g(\theta)$*, then the algorithm is equivalent to taking the MAP estimate of the Gibbs posterior with a prior* π *such that* $-\frac{1}{\lambda}\ln\pi(\theta) = \frac{g(\theta)}{\lambda}$*. In particular,* (11)-(14) *from Corollary 3.4 applies in this case. For example, the quadratic regularization term* $g(\theta) = (\theta - \theta_m)^T P^{-1}(\theta - \theta_m)$ *corresponds to* $\pi \sim \mathcal{N}(\theta_m, P/\lambda)$*. If* π *can be interpreted as a Gaussian density, then MAP is MEP and the corresponding sharper bounds can be used.*

Furthermore, many PEM algorithms rely on choosing randomly the initial parameter estimate. The outcomes of such algorithms can be viewed as drawing a sample from a suitable posterior distribution. More precisely, consoider an algorithm A_N*. For every initial value* $\theta_0 \in \Theta$ *and data* $\mathcal{D}$*, the algorithm generates an estimate* $A_N(\theta_0, \mathcal{D})$*. Most optimization algorithms used for PEM methods (Gradient, Gauss-Newton), etc. are of this form, if we fix the number of steps. Assume that the initial*

parameter values θ_{init} are i.i.d., samples from a prior density π. Then the resulting parameter values $A_N(\theta_{init}, \mathcal{D})$ are i.i.d. samples from a distribution $P_{A_N}(B) = \int_{A_N^{-1}(B,\mathcal{D})} \pi(\theta) dm(\theta)$ for any $B \in B_\theta$. Assume that the latter distribution has a density $\hat{\rho}$ which is in $\mathcal{M}_\pi$. Then Lemma F.3 provides an error bound for the true error of such an algorithm, when applied to an initial parameter value randomly sampled from π.

Remark F.9 (Strong convexity of the true error)**.** *Strong convexity implies that there is a unique minimizer of the true error, which is the true system. It can be seen as an* identifiability and persistence of excitation assumption*, as it implies that there is no parameter value which gives the same input-output behavior as the true system (otherwise their true errores would be the same), and* $\mathbf{w}$ *is rich enough (the average response to* $\mathbf{w}$ *allows us to find the true system). In fact, strong convexity of the true error (at least locally), is a standard assumption for studying the asympotic distribution of the parameter estimation error [26, Theorem 9.1].*

Remark F.10 (Bounds on the difference between system matrices)**.** *Assuming that the LTI systems $\Sigma(\theta)$ are all minimal, we can use* (14) *and [32, Theorem V.2] to derive an error bound (in high-probability) on the difference between the matrices of the learned and true system. Alternatively, if we consider parametrizations based on structure indices [33, Chapter 3] we could use the fact that H_2 norms can be used to approximate locally the Riemannian distance between parameter values [33, Chapter 4, page 129].*

Remark F.11 (Relationship with [11, 8])**.** *For the case of bounded noise in the data generator, Theorem 3.1 provides a PAC-Bayesian bound on the generalization gap which is similar to [11, Theorem 2, Corollary 1] and [8, Theorem 5.2]. However, [11] considers RNNs, and the corresponding constants $L_{\mathbf{v}}(\theta), L_{g,\mathbf{s}}(\theta), L_{g,\mathbf{v}}$ correspond to $\|\hat{B}_\theta\|_2$, $\|\hat{C}_\theta\|_2$, and $\|\hat{D}_\theta\|_2$, respectively, and $\tau(\theta)$ corresponds to an upper bound on the spectrum of $\hat{A}_\theta$ and $C(\theta)$ is such that $\|\hat{A}_\theta^k\|_2 \leq \tau(\theta)^k C(\theta)$. In our setting, $L_{\mathbf{v}}(\theta), L_{g,\mathbf{s}}(\theta), L_{g,\mathbf{v}}$ can be taken to be C from Assumption 2.1, $C(\theta)$ can be taken to be M and $\tau(\theta)$ can be taken to be γ from Assumption 2.1. Then the constants G_θ, H_θ from [11] are upper bounds on $\|\alpha_\theta\|_{\ell_1}$ and $\theta_\infty(\alpha_\theta)$ respectively. Hence the term $\hat{\Psi}_\pi(\lambda, N)$ of [11, Theorem 2, Corollary 1] is an upper bound on the term $\frac{1}{2}\sum_{i=1}^{2} \ln E_{\theta\sim\pi} \exp\left(\mathcal{C}_i(2\lambda, \theta, N, \delta, \epsilon)\right)$ from* (8)*, i.e., the bound of [11, Theorem 2, Corollary 1] is more conservative than the one of this paper.*

As to [8, Theorem 5.2], it uses $\|\alpha_{g,y} - c_\theta\|_{\ell_1} \leq G_e(\theta)$ and $\theta_\infty(\alpha_{g,y} - c_\theta) \leq G_{e,1}(\theta)$ in the bounds, where c_θ and $\alpha_{g,y}$ are defined in Lemma D.3. While this may appear less conservative, the use of c_θ makes the bound difficult to evaluate in practice, as c_θ depends on the parameters of the unknown data generator. In contrast, our bound uses only $\|\alpha_g\|_{\ell_1}$ and $\theta_\infty(\alpha_g)$, which, although they also depend on the data generator, are easier to estimate or upper bound in practice.

Note that [11, 8] do not address the case of unbounded signals or parameter estimation errors. They also do not provide generalization bounds for single draw, MEP, or MAP estimators.

Remark F.12 (Relationship with finite-sample bounds [32, 24, 39, 42, 20])**.** *Recent works such as [32, 24, 39, 42, 20] provide finite-sample parameter estimation bounds for specific algorithms (e.g., least-squares for ARX models, subspace methods), typically assuming the data generator starts from a deterministic initial state and often neglecting the stochastic component. However, assuming a deterministic initial state can be problematic, since process noise generally drives the system to a random initial state determined by past disturbances.*

In contrast, our results: ***(1)*** *apply to a broad class of learning algorithms (including PEM, maximum likelihood, Bayesian, and regularized FIR methods),* ***(2)*** *assume the data generator is in a stationary regime,* ***(3)*** *explicitly account for both deterministic and stochastic components, and* ***(4)*** *provide generalization gap bounds. The trade-off is that our bound on the square of the parameter estimation error is $O\left(\frac{(\ln N)^{r+1}}{\sqrt{N}}\right)$, which is more conservative than the $O\left(\frac{\ln N}{N}\right)$ rates in the cited works.*

G Numerical Experiments

The code for the numerical example can be found at `https://github.com/MOHAMMADZAHD93/PACneurips2026`.

We evaluate the right-hand sides of (8) (Theorem 3.1) and (10) (Corollary 3.2), compare them with the absolute generalization gap $\left|\mathbf{E}_{\theta\sim\rho_D}\mathcal{L}(\theta) - \mathbf{E}_{\theta\sim\rho_D}\hat{\mathcal{L}}_N(\theta)\right|$, and report how its median and

0.9 quantile ($q_{.90}$) change with N, using $\mathcal{L}$ from (5) and $\hat{\mathcal{L}}_N$ from Section 2. At $N = 1000$ and $N = 64000$, we compare the bounds directly with the observed gap. We also test non-vacuity in the following sense: whether there exists a sample size N_1 for which a bound is below the 90th-percentile error observed at $N = 4$. For the Gibbs posterior, the error compared with the right-hand side of (13) is the squared parameter error $\|\theta_\star - \theta_{true}\|_2^2$.

We use the data generator in Assumption 2.2 and the predictor class in Assumption 2.1.

G.1 Data generator and predictors

The generator has three states, two inputs, and two outputs. In the notation of (4), it is

$$\mathbf{x}(t+1) = A_g\mathbf{x}(t) + K_g\mathbf{e}_g(t), \qquad \begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix} = C_g\mathbf{x}(t) + \mathbf{e}_g(t), \tag{146}$$

where $A_g = \begin{bmatrix}0.25 & 0 & 0\\ 0 & 0.20 & 0\\ 0 & 0 & 0.125\end{bmatrix}$, $K_g = \begin{bmatrix}0 & 0 & 0.02 & 0\\ 0 & 0 & 0 & 0.02\\ 0 & 0 & 0.70 & 0.63\end{bmatrix}$, and $C_g = \begin{bmatrix}0.70 & 0 & 0\\ 0 & 0.70 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$.

The innovation is $\mathbf{e}_g(t) = [\mathbf{v}^\mathsf{T}(t), \mathbf{u}^\mathsf{T}(t)]^\mathsf{T}$, where the coordinates of $\mathbf{v}(t)$ and $\mathbf{u}(t)$ are independent across coordinates and time. The first two rows of the second equation in (146) generate $\mathbf{y}(t)$, and its last two rows generate $\mathbf{u}(t)$. The input to the predictors is $\mathbf{w}(t) = [\mathbf{y}^\mathsf{T}(t), \mathbf{u}^\mathsf{T}(t)]^\mathsf{T}$.

The predictor matrices in (1) are $\hat{A}_\theta = A_g + (\theta_1 - 1)D_1 + (\theta_2 - 1)D_2$, $D_1 = \begin{bmatrix}0 & 0 & 3.700\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$, $D_2 = \begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 3.404\\ 0 & 0 & 0\end{bmatrix}$, $\hat{B}_\theta = K_g$, $\hat{C}_\theta = \begin{bmatrix}0.70 & 0 & 0\\ 0 & 0.70 & 0\end{bmatrix}$, and $\hat{D}_\theta = 0_{2\times 4}$. Thus, $\hat{\mathbf{x}}(t+1) = \hat{A}_\theta\hat{\mathbf{x}}(t) + K_g\mathbf{w}(t)$, $\hat{\mathbf{y}}_\theta(t \mid 0) = \hat{C}_\theta\hat{\mathbf{x}}(t)$, and $\hat{\mathbf{x}}(0) = 0$, with $\hat{A}_\theta = \begin{bmatrix}0.25 & 0 & 3.700(\theta_1 - 1)\\ 0 & 0.20 & 3.404(\theta_2 - 1)\\ 0 & 0 & 0.125\end{bmatrix}$.

The parameter set is $\Theta = [0.6, 1.4]^2$ and $\theta_{true} = (1, 1)$. At this parameter, $\hat{A}_{\theta_{true}} = A_g$. For every $\theta \in \Theta$, the eigenvalues of $\hat{A}_\theta$ are 0.25, 0.20, and 0.125, so the predictors satisfy the Schur condition in Assumption 2.1.

In the bounded case, $u_i(t) \sim U[-0.05, 0.05]$ and $v_i(t) \sim U[-0.005, 0.005]$ for $i \in \{1, 2\}$. In the unbounded case, $u_i(t) = 0.0125R_{u,i}(t) + 0.0000625Z_{u,i}(t)$ and $v_i(t) = 0.25\big(0.0125R_{v,i}(t) + 0.0000625Z_{v,i}(t)\big)$ for $i \in \{1, 2\}$, where the R variables are Rademacher, the Z variables are standard Gaussian, and all these variables are independent across coordinates and time.

For each data realization, we initialize the generator state at zero, simulate the generator for 160 initial time steps, and discard those input-output pairs so that the generator state approaches its stationary regime. We retain the next N input-output pairs, initialize the predictor state at zero, and update it on those pairs as in (1). These N pairs are used to compute $\hat{\mathcal{L}}_N$.

G.2 Data-dependent uniform posterior

We first use the prior $\pi = U([0.6, 1.4]^2)$. The posterior ρ_D is uniform on a square of side length 0.1 centred at a constrained minimizer of $\hat{\mathcal{L}}_N$. Its centre is constrained so that the square is contained in Θ. Hence, $\mathrm{KL}(\rho_D\|\pi) = \log(0.8^2/0.1^2) = \log 64$.

We consider the ℓ_2 loss, $\ell(\mathbf{y}, \hat{\mathbf{y}}) = \|\mathbf{y} - \hat{\mathbf{y}}\|_2$, and the quadratic loss, $\ell(\mathbf{y}, \hat{\mathbf{y}}) = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$, from Assumption 2.3. For each realization and sample size N, we compute $\Big|\mathbf{E}_{\theta\sim\rho_D}\mathcal{L}(\theta) - \mathbf{E}_{\theta\sim\rho_D}\hat{\mathcal{L}}_N(\theta)\Big|$. The sample sizes are $N \in \{4, 500, 1000, 2000, 4000, 8000, 16000, 32000, 64000\}$. Table 4 reports the median and $q_{.90}$ of this absolute gap over 40 independent data realizations.

The posterior integrals are evaluated deterministically. For the quadratic loss, $\mathcal{L}(\theta)$ is computed from stationary second moments. For the ℓ_2 loss, it is estimated using a fixed stationary Sobol reference.

Table 4: Posterior-averaged absolute generalization gap and the right-hand sides of (8) and (10).

Innovations	Loss	N	Median	$q_{.90}$	(8) / (10)
Bounded	ℓ_2	1,000	3.53×10^{-5}	7.27×10^{-5}	0.0564 / 0.2678
Bounded	ℓ_2	64,000	3.62×10^{-6}	6.64×10^{-6}	0.00703 / 0.0335
Bounded	Quadratic	1,000	3.44×10^{-7}	9.04×10^{-7}	0.0321 / 0.2591
Bounded	Quadratic	64,000	3.64×10^{-8}	7.64×10^{-8}	0.00400 / 0.0324
Unbounded	ℓ_2	1,000	7.52×10^{-6}	2.00×10^{-5}	22.30 / 72.06
Unbounded	ℓ_2	64,000	3.56×10^{-7}	7.59×10^{-7}	4.35 / 14.43
Unbounded	Quadratic	1,000	6.71×10^{-8}	1.39×10^{-7}	73.20 / 480.68
Unbounded	Quadratic	64,000	5.34×10^{-9}	1.43×10^{-8}	17.26 / 121.84

Table 5: Fitted exponents from least-squares fits of the logarithms of the median and $q_{.90}$ of the posterior-averaged absolute generalization gap against $\ln N$, over $N \in \{500, 1000, 2000, 4000, 8000, 16000, 32000, 64000\}$.

Innovation law	Loss	Median slope	$q_{.90}$ slope
Bounded	ℓ_2	-0.567	-0.505
Bounded	Quadratic	-0.583	-0.534
Unbounded	ℓ_2	-0.789	-0.810
Unbounded	Quadratic	-0.648	-0.578

We use $\delta = 0.025$ in (8) and (10). Applying the one-sided inequalities to both signs $\epsilon \in \{-1, 1\}$ gives a 90% probability statement for the absolute gap. We evaluate the prior expectations in (8) by 61×61 tensor Gauss–Legendre quadrature. For each sample size, innovation law, and loss, we minimize the right-hand side of (8) over the admissible λ range in Theorem 3.1. We evaluate 161 equally spaced points in $\log \lambda$, then apply bounded scalar minimization between the adjacent grid points around the smallest grid value, with absolute tolerance 10^{-10}. For the unbounded absolute-gap bound, we minimize separately for both signs in (8) and report the larger value. The last column of Table 4 reports the right-hand sides of (8) and (10), in that order.

At both displayed sample sizes, the median and $q_{.90}$ of the gap are below both right-hand sides in every row. Relative to the corresponding $q_{.90}$, the two bounds range from 7.8×10^2 to 4.2×10^5 for bounded innovations and from 1.1×10^6 to 8.5×10^9 for unbounded innovations. In each row, the right-hand side of (8) is smaller than that of (10). These comparisons show that the bounds are loose at the displayed sample sizes and that (10) is looser than (8) in this experiment.

To assess whether the observed sample-size dependence of the gap is consistent with the upper-bound dependence in (8) and (10), we compute its median and $q_{.90}$ from 40 independent data realizations at each $N \in \{500, 1000, 2000, 4000, 8000, 16000, 32000, 64000\}$. For each innovation law and loss, we fit separate least-squares lines to the logarithms of these two summaries against $\ln N$. Each entry in Table 5 is the fitted exponent for one summary of the posterior-averaged absolute generalization gap.

Across all four innovation-law and loss combinations, the median and $q_{.90}$ of the posterior-averaged absolute generalization gap decrease over the sample sizes considered. The bounded-innovation slopes, from -0.583 to -0.505, are close to the $-1/2$ exponent in the upper bounds. The unbounded-innovation slopes, from -0.810 to -0.578, are more negative over this finite range. Thus, the observed gaps approximately follow the stated dependence in the bounded cases and decrease faster over this range in the unbounded cases. This is consistent with (8) and (10), whose rates are upper bounds.

To test non-vacuity in the sense stated above, we estimate the $q_{.90}$ of the posterior-averaged absolute generalization gap at $N = 4$ from 400 independent data realizations. For each innovation law and loss, we ask whether the right-hand side of (8) or (10), evaluated at some N_1, is below this fixed 90th-percentile gap for $N = 4$. We locate a crossing by doubling and integer bisection in the final bracket, multiply the resulting integer by 1.01, round up to obtain N_1, and verify directly that the right-hand side at N_1 is below the fixed $N = 4$ quantile. The reported N_1 values are verified witnesses with a 1% margin, not claims that they are the smallest possible sample sizes. Table 6 reports the $N = 4$ quantiles and these witnesses.

Table 6: The $q_{.90}$ values at $N = 4$ and verified sample-size witnesses N_1 for (8) and (10).

Innovations	Loss	$q_{.90}$ at $N = 4$	N_1 for (8)	N_1 for (10)
Bounded	ℓ_2	5.9934×10^{-3}	8.8835×10^{4}	2.0171×10^{6}
Bounded	Quadratic	8.7284×10^{-5}	1.3551×10^{8}	8.9030×10^{9}
Unbounded	ℓ_2	9.4409×10^{-3}	6.8737×10^{10}	9.3633×10^{11}
Unbounded	Quadratic	2.4156×10^{-4}	1.0744×10^{16}	8.5340×10^{17}

Table 7: Constants used in (13).

Innovations	L	r	m_Θ	$\mathcal{C}_{or}(\delta, \sigma)$
Bounded	1	0	0.00933654	13.38869
Unbounded	37.90410	1.5	0.00175065	59.65784

For every innovation-law and loss combination, Table 6 gives an N_1 at which each bound is below the 90th-percentile generalization gap observed at $N = 4$. Thus, the bounds are not vacuous under the stated criterion. The N_1 values for Theorem 3.1 range from 8.8835×10^4 to 1.0744×10^{16}. The corresponding values for Corollary 3.2 are larger by factors of approximately 14 to 79. These large sample sizes show that non-vacuity in this sense does not mean the bounds are tight at the displayed sample sizes. The comparison establishes that each bound can fall below an observed generalization gap, while Table 4 shows that both bounds remain much larger than the observed gaps at $N = 1{,}000$ and $N = 64{,}000$.

G.3 Gibbs posterior

We next evaluate (13) for the Gibbs posterior defined in Corollary 3.4, using the quadratic loss. We retain the generator in (146), the predictors in (1), $\Theta = [0.6, 1.4]^2$, $\pi = U([0.6, 1.4]^2)$, and the innovation laws. We change only the posterior from ρ_D to ρ_{Gibbs}. The quantity compared with the right-hand side of (13) is $\|\theta_\star - \theta_{true}\|_2^2$, where $\theta_\star = \mathbf{E}_{\theta \sim \rho_{\mathrm{Gibbs}}} \theta$.

We compute $\|\theta_\star - \theta_{true}\|_2^2$ by evaluating the normalizing constant and posterior mean with 1089 fixed quadrature nodes. Refining the integration to 2601 nodes changes the posterior mean by less than 2.9×10^{-15}. We use $\delta = 0.05$ in (13), which gives a 90% probability statement for the bound. We use 40 independent data trajectories for $N \in \{500, 1000, 2000, 4000, 8000, 16000, 32000, 64000\}$ and 1000 additional independent trajectories at $N = 4$ to estimate the $q_{.90}$ used in Table 9.

At $N = 4$, the $q_{.90}$ of $\|\theta_\star - \theta_{true}\|_2^2$ over 1000 independent data realizations is $3.0956784 \times 10^{-10}$ for bounded innovations and $1.5926441 \times 10^{-14}$ for unbounded innovations. We apply the same non-vacuity criterion: we find the smallest integer $N_1 > 4$ for which the right-hand side of (13), evaluated at N_1, falls below the corresponding 90th-percentile squared parameter error observed at $N = 4$. Table 9 reports the fixed $N = 4$ quantiles and $\log_{10} N_1$, avoiding the long decimal representations of the exact integers.

At $N = 1000$ and $N = 64000$, the right-hand side of (13) exceeds the median and $q_{.90}$ of $\|\theta_\star - \theta_{true}\|_2^2$ in Table 8 by about 11-12 orders for bounded innovations and 22-23 orders for unbounded innovations. Nevertheless, the bound is not vacuous in the sense above: it falls below the 90th-percentile squared parameter error observed at $N = 4$ when $\log_{10} N_1 = 28.9802$ for bounded innovations and $\log_{10} N_1 = 50.1296$ for unbounded innovations. These very large N_1 values show that non-vacuity under this criterion coexists with substantial looseness at the sample sizes in Table 8. The common prior is symmetric about θ_{true}, so the very small posterior-mean errors, especially for unbounded innovations, must not be interpreted as evidence of strong learning from the data.

Table 8: Squared parameter error $\|\theta_\star - \theta_{true}\|_2^2$ and the right-hand side of (13).

Innovations	N	Median	$q_{.90}$	(13)
Bounded	1,000	1.20×10^{-10}	3.97×10^{-10}	313.25
Bounded	64,000	1.11×10^{-10}	5.28×10^{-10}	62.73
Unbounded	1,000	6.32×10^{-17}	2.41×10^{-16}	5.12×10^{6}
Unbounded	64,000	1.23×10^{-17}	5.87×10^{-17}	2.08×10^{6}

Table 9: The $q_{.90}$ values at $N = 4$ and the logarithms of the first integer sample sizes N_1 at which the right-hand side of (13) falls below them.

Innovations	$q_{.90}$ at $N = 4$	$\log_{10} N_1$
Bounded	$3.0956784 \times 10^{-10}$	28.9802
Unbounded	$1.5926441 \times 10^{-14}$	50.1296

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: the main contribution of the paper is a generalization bound for a class of dynamical systems, this contribution is clearly stated in the abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: the paper discusses the limitations of the proposed approach, including assumptions made and potential impacts on generalization, both in the main text and in the remarks at the end of the paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: all the assumptions are formally stated, all the detailed proofs are in the supplementary material.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: the paper is a theoretical one, the experiments serve only to illustrate the theoretical results, and they are presented the appendix. The description of the experiments contains all the necessary details and a link to the github repository with the code is provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: the paper provides open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: the paper provides all the necessary details for the experimental setup, including data splits, hyperparameters, and optimizer settings, a link to the github repository with the code is also provided.

Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: the main contribution of the paper is a mathematical result, proof of which is included in the supplementary material. The experiments serve as an illustration, not as the justification of the main claim. Moreover, the nature of the experiments does not easily lend itself to computation of statistical significance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: the experiment was performed on a desktop computer, no special hardware or large-scale resources were required.

   Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: the research conducted in the paper adheres to the NeurIPS Code of Ethics.

Guidelines:

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

10. **Broader impacts**

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

Answer: [No]

Justification: the contribution of the paper is a theoretical result, no societal impacts are directly associated with the work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A]

Justification: the paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: no assets are used.

    Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: the paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: the paper does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

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