ENERGY-GUIDED LEARNING FOR HEAT EQUATION WITH NOISY SOURCE: OPTIMAL ESTIMATION AND GENERALIZATION BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Data and the governing partial differential equations (PDEs) of continuous-time dynamical systems provide rich structure for statistical learning. A central challenge is how to integrate data with PDE constraints in a principled way to improve generalization. In this paper, we characterize the statistical estimation and generalization properties of learning solutions to the heat equation with a noisy source function from finite, randomly sampled data. We first derive a novel variational formulation, which we characterize as the minimizer of loss function for an energy regularized and ellipticity debiased Sobolev regression. By formulating the learning task as a supervised problem over space–time domains under physically meaningful norms, our approach goes beyond the traditional supervised learning framework by effectively leveraging *auxiliary data*—samples of the source function of the heat equation—to design a *physics-informed energy regularization*. Our analysis reveals that learning in linear parabolic PDEs exhibits intrinsic anisotropy between spatial and temporal regularity, in sharp contrast to the static elliptic problems, leading to fundamentally different estimation and generalization behaviors. Specifically, we derive non-asymptotically rate-optimal estimation errors, as well as generalization error, which explicitly capture the interplay among temporal sampling, spatial resolution, and solution smoothness. Our numerical experiments demonstrate substantial orders of magnitude of reduction in testing errors against various physics-informed neural network methods on various parabolic and hyperbolic PDEs.

1 INTRODUCTION

Continuous-time dynamical systems governed by partial differential equations (PDEs) arise across scientific and engineering fields involving the study of time evolution behaviors of physical phenomena ranging from fluid dynamics to quantum mechanics (Batchelor, 2000; Gasiorowicz, 2007). Due to their instability and nonlinearity, learning solutions of time-dependent PDEs from noisy data is a central yet challenging problem in scientific machine learning (Brunton & Kutz, 2024; Couairon & Chomaz, 1996). Classical numerical methods such as finite element and Galerkin methods provide highly accurate solvers but depend critically on the availability of a suitable computational mesh in high dimensional domain (Persson, 2005; Ceniceros & Hou, 2001; Bern & Plassmann, 2000). This limitation motivates data-driven approaches that aim to bypass explicit discretization.

A recent popular mesh-free paradigm is the physics-informed neural networks (PINNs), which enforce PDE constraints via *pointwise residual minimization* to extrapolate (Karniadakis et al., 2021). Despite empirical successes in several applications (Sharma et al., 2023; Raissi et al., 2020; Hu et al., 2024; Haghighat et al., 2021; Mo et al., 2022; Norambuena et al., 2024; Kurth et al., 2023; Yang et al., 2024), this formulation evaluates differential operators at collocation points and therefore implicitly assumes sufficient smoothness of the solution. As a result, PINNs are fundamentally tied to strong solutions and can fail in regimes involving nonsmooth dynamics or weak solutions, with documented instability and training pathologies (Cuomo et al., 2022; Krishnapriyan et al., 2021; Wang et al., 2022a).

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

Variational approaches, exemplified by deep Ritz method (E & Yu, 2018), instead minimize an energy functional whose minimizer characterized by the Euler-Lagrange equation coincides with the PDE solution. This provides global control in Sobolev spaces and enables learning of *weak solutions* with provable guarantees (Lu et al., 2021a; Jiao et al., 2022; Duan et al., 2022). However, such formulations crucially rely on the existence of an exact variational principle and are therefore limited to *static* elliptic PDEs. For time-dependent systems governed by parabolic/hyperbolic PDEs, no corresponding *exact* space-time energy functional exists, preventing a direct extension of deep Ritz method.

In this work, we focus on the statistical learning problem for weak solution of the heat equation

$$\begin{cases} \partial_t u(x,t) - \Delta u(x,t) = f(x,t), & (x,t) \in \Omega \times (0,T], \\ u(x,t) = 0, & (x,t) \in \partial\Omega \times [0,T], \\ u(x,0) = u_0(x), & x \in \Omega, \end{cases} \tag{1.1}$$

where f is an ***unknown*** source function in a closed, bounded spatial domain $\Omega \subset \mathbb{R}^d$ over a finite time horizon $[0,T]$, u_0 is a known initial condition, and the above equation satisfies the zero Dirichlet boundary condition. When we study the impact of the heat source f on the solution u, solving the PDE (1.1) in an appropriate sense is often a well-posed forward problem with mild boundary and initial conditions (Evans, 2010). Thus without loss of generality, we assume that the true PDE solution of (1.1) is unique and identifiable, which is denoted by $u^* := u^*(x,t)$ throughout the paper.

In reality, neither the source function f nor the true solution u^* is directly observed over the domain $\Omega \times [0,T]$. Instead, we are given observations of these functions at discrete time points $t_1, \ldots, t_n \in [0,T]$. At each time t_j, we observe m spatial points $X_{1j}, \ldots, X_{mj} \in \Omega$. In this paper, we work with noisy data of the source function and its corresponding solution discretized at the same spatio-temporal points generated by the following sampling model

$$S = \{(X_{ij}, t_j, f_{ij}, u^*_{ij})\}_{i=1,j=1}^{m,n}, \tag{1.2}$$

where

$$f_{ij} = f(X_{ij}, t_j) + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad u^*_{ij} = u^*(X_{ij}, t_j). \tag{1.3}$$

Interestingly, we may conceptually partition the resulting training dataset S into two components: (1) **Labeled Data** (i.e., standard regression data for the target u^*): $S^{labeled} = \{(x_{ij}, t_j, u^*_{ij})\}_{i=1,j=1}^{m,n}$, and (2) **Auxiliary Data** (consisting of additional (noisy) observations of the source term): $S^{aux} = \{(x_{ij}, t_j, f_{ij})\}_{i=1,j=1}^{m,n}$. The observational model (1.2) and (1.3) we consider here aligns with many real-world scenarios. In physical systems governed by PDEs, $f(x,t)$ is often unknown or only indirectly observed through sensors (Cannon & Esteva, 1986; Arridge et al., 2019; Hasanoğlu & Romanov, 2021; Iglesias et al., 2013; Nickl, 2024), which are typically noisy, while quantities represented by the solution $u^*(x,t)$, such as temperature in the reaction-diffusion equation, are more readily measurable. Furthermore, the governing PDE imposes additional structure on the learning problem. Characterizing how to exploit the structural richness of the linear parabolic PDEs in learning their dynamics is a fundamental problem in this area.

1.1 Our Contribution

In this paper, we address the fundamental challenge of incorporating PDE constraints into data-driven learning and demonstrate improved generalization and learning efficiency. We develop an extended supervised learning framework for learning the dynamics of the heat equation, a classical linear parabolic PDEs. Moving beyond the traditional regression approach based solely on labeled solution data $\{x_{ij}, t_j, u^*(x_{ij}, t_j)\}$, which aims to learn a mapping $(x,t) \mapsto u^*(x,t)$, our framework incorporates *energy-guided learning with auxiliary data* S^{aux}.

Our key idea is to use noisy observations of the source function f to construct an approximate energy functional via an *elliptic regularization* of the dynamical system (1.1), resulting in a PDE learner with improved regularity properties. By coupling these auxiliary observations with labeled solution data, we obtain a penalized estimator that leverages both statistical supervision and the underlying physical structure. Conceptually, this approach corresponds to regression in Sobolev spaces under energy-based regularization.

On the theoretical front, we show that this combination is necessary: purely supervised estimator (with labeled solution data) and the approximate Ritz method (with auxiliary data) are both suboptimal in sample complexity when the degree of spatial smoothness of the solution is below a threshold, whereas our energy-guided method with a suitable balance between data-fitting and physics constraint simultaneously achieves sharp estimation convergence rate, excess risk and generalization error bounds. Next, we turn to a more detailed summary of these contributions.

A new variational formulation. We introduce a new variational formulation for the heat equation. The idea is to characterize its solution as the minimizer of an energy-regularized, ellipticity-debiased Sobolev regression objective. Moreover, the derived total loss function with data fidelity and physics prior enjoys a stability estimate. This physics-informed energy-regularized formulation naturally integrates labeled and auxiliary data, which together provide global control in Sobolev spaces and enable efficient learning of weak solutions that least-PDE-residual methods (e.g., PINNs) fail to handle. Even though our theoretical analysis is limited to the heat equation with unknown source, we expect that the proposed energy-guided approach with elliptic regularization technique can be adapted to more general linear parabolic and hyperbolic PDEs (e.g., with techniques in Serra & Tilli (2012)). Towards this goal, we show some strong empirical performance of the proposed method against various PINNs when applied to those equations, cf. Section G.1 and Appendix G.2.

Optimal estimation and generalization error bounds. We establish a sharp statistical estimation and generalization error analysis for learning solution u^* to the heat equation. Our statistical estimation error exhibits an interesting smoothness-aware adaptation of the proposed energy-guided approach. On one hand, if the solution of the heat equation is spatially smooth (i.e., the subcritical regime $q > d/2$ in Theorem 3.1), then labeled data alone can achieve the optimal rate of convergence (cf. Theorem 3.3). On the other hand, if the spatial smoothness of the heat equation is below the threshold $q \leqslant d/2$, then convergence rates depend on the balance of data fidelity and energy regularization. When the regularization parameters is chosen optimally, the rate attains the minimax lower bound. In addition, we also derive a fast generalization error bound via a delicate localization technique.

Performance Improvement. To corroborate our theoretical findings, we provide substantial experimental results to see the improvements on generalizationG.3. Our approach reduces the test error of 50d Hamilton Jacobi equation by up to **79.5**$\% - $ **89.4**$\%$ relative to strong baselines in physical learning literature.

Overall, this work contributes to the growing body of literature in addressing the limitations of PINNs (Krishnapriyan et al., 2021; Wang et al., 2022b;a). Our energy-guided approach is particularly strong at capturing transient physical phenomena when initial data u_0^* exhibit singularities, while PINNs fail to model them. By characterizing these failure modes, we offer the perspective that integrating labeled and auxiliary data is a useful way to improve stability and convergence of neural solvers in dynamical systems.

1.2 Relevant Work

Physics-Informed Machine Learning. Despite the empirical advances of numerous PINN variants across diverse applications (McClenny & Braga-Neto, 2023; Shukla et al., 2020; Patel et al., 2022; Pfau et al., 2020; De Ryck et al., 2024; Kharazmi et al., 2021; Wang et al., 2022a), a rigorous theoretical understanding of machine learning based PDEs solvers remains largely vacant. In recent works, Lu et al. (2021a); Duan et al. (2022); Jiao et al. (2022) have established minimax optimal convergence rates for elliptic PDEs comparing the performance of the Deep Ritz Method (DRM) and PINNs. Another line of work focuses on the green function hidden in differential operator mapping u to f for elliptic and parabolic PDEs (Boullé et al., 2023; Negi et al., 2024; Boullé et al., 2022).

Variational Methods and Energy Regularization. In calculus of variation literature, variational methods to solve elliptic PDEs can be extended to solve not only elliptic PDEs but also parabolic and even hyperbolic ones (Serra & Tilli, 2012). There are extensive works (Mielke & Stefanelli, 2011; Stefanelli, 2025; Audrito & Sanz-Perela, 2024; Akagi et al., 2025b; Akagi & Stefanelli, 2011; Rossi et al., 2019) in this direction to establish approximate solution from regularized PDEs with various boundary conditions, which are relatively easier to tackle with compared to original PDEs problem. In our case, the penalized PDE, which we are concerned with, is called the telegraph heat equation, a hyperbolic modification of heat equation (Dehghan & Shokri, 2008; Orsingher & Beghin, 2004; Goldstein, 1951).

Semi-Supervised Learning in Physical Systems. In many real-world scenarios, the source function $f(x,t)$ is either unknown or indirectly observed through noisy sensors, whereas the solution $u^*(x,t)$ is more readily observable. We formulate this learning task as a semi-supervised problem where the solution labels u^* provide rich empirical information while the auxiliary information $\{f(x,t)\}$ enables an energy-based regularizer that enforces the underlying physical constraints. Similar to our approach, Cui et al. (2024); Lin et al. (2015) employ partially observed labels to learn solutions to complex systems such as minimal surface equations and Hamilton-Jacobi equations.

2 METHOD

2.1 DESIGN OF UNBIASED AND STABLE LOSS FUNCTION

There are two key ingredients to build a loss function whose minimization can lead to *efficient* learning of the solution function to the heat equation (1.1) from the data sample (1.2). First, viewing the heat equation as an input-output system $f \mapsto u$, we can determine the solution u from the source function f, which we have the auxiliary data access to. It is well-known that solution to the time-independent Poisson equation $\Delta u = f$ is the minimizer of the Dirichlet energy functional $I(u) = \int_\Omega [\frac{1}{2}|\nabla u(x)|^2 - f(x)u(x)]dx$. For the heat equation, such variational formulation does not exist and we adopt the following weighted energy functional (WEF) as a proxy, defined for $\epsilon > 0$,

$$I_\epsilon(u) := \int_0^T e^{-\frac{t}{\epsilon}} \int_\Omega \left[\frac{1}{2}|u_t|^2(x,t) + \frac{1}{2\epsilon}|\nabla u(x,t)|^2 - \frac{1}{\epsilon} f(x,t)u(x,t)\right] dxdt. \tag{2.1}$$

Our motivation of WEF comes from observing that $I_\epsilon(u)$ is *strongly convex* in u and we can identify its first-order optimality condition via the Euler-Lagrange equation given by the hyperbolic PDE

$$-\epsilon\, u_{tt} + u_t - \Delta u = f \text{ for } (x,t) \in \Omega \times (0,T), \tag{2.2}$$

cf. Appendix C.2 for derivation. In literature, (2.2) is called the *elliptic regularization* of the heat equation (Lions & Magenes, 1972), which originated in the study of geometric evolution equations (Ilmanen, 1994). When ϵ is small, the relaxed solution u_ϵ^* of (2.2) enjoys the uniform-in-time approximation for the heat equation solution u^* (Mielke & Stefanelli, 2011; Rossi et al., 2019) in the sense that: for all $t > 0$, $u_{\epsilon_n}^* \to u^*$ in $L^2(\Omega)$ for a subsequence $\epsilon_n \to 0$.

Second, from a statistical learning perspective, we can construct a data-fitting term using the features (x,t) to predict the solution $u^*(x,t)$. It turns out the appropriate loss metric to compensate the WEF curvature is the Sobolev geometry in the solution or label space (cf. Proposition 2.2 below for precise statement in terms of stability). Specifically, we choose

$$D(u,u^*) := ||u - u^*||^2_{H_t^1 H_x^1}, \tag{2.3}$$

where we recall $H_t^1 H_x^1 = H_t^1(0,T; H_x^1(\Omega))$. Note that this $H_t^1 H_x^1$ geometry has close connection to the Sobolev training in literature (Czarnecki et al., 2017; Lu et al., 2022). Combining the two pieces of information together, we propose the population total loss defined as

$$L_{\epsilon,\beta}(u) := D(u,u^*) + \beta\big[I_\epsilon(u) + M_\epsilon(u,u^*)\big], \tag{2.4}$$

where $\beta > 0$ is a tuning parameter to balance data fidelity and physics constraint, and

$$M_\epsilon(u,u^*) := \int_0^T \int_\Omega e^{-t/\epsilon} u(x,t) u_{tt}^*(x,t)\, dx\, dt - e^{-T/\epsilon} \int_\Omega u(x,T) u_t^*(x,T)\, dx. \tag{2.5}$$

is an extra bias correction term due to the ellipticity offset from (2.2). With debiasing, we can show that the (exact) heat equation is the minimizer of the total loss function $L_{\epsilon,\beta}(u)$.

Theorem 2.1 (Minimizer of the total loss function $L_{\epsilon,\beta}(u)$)**.** *Let $s \geqslant 1$ and $p \in [0,s]$. Then, the solution u^* to heat equation in (1.1) is the minimizer of the total loss function $L_{\epsilon,\beta}(u)$ over $\mathcal{U}_0 := H_t^p(0,T; H_x^{2s+2-2p}(\Omega)) \cap L_t^2(0,T; H_{0,x}^1(\Omega)) \cap \{u : \Omega \times [0,T] \to \mathbb{R} | u(x,0) = u_0^*(x)\}$, i.e., $u^* = \arg\min_{u \in \mathcal{U}_0} L_{\epsilon,\beta}(u)$ for any $\epsilon, \beta > 0$.*

We remark that p is the temporal regularity parameter and s is the total space-time smoothness parameter. To summarize, our learning objective has three components: (i) a data-fitting term $D(u,u^*)$ that provides the distance between a candidate solution function u and the true label u^*,

(ii) a WEF term $I_\epsilon(u)$ that navigates the underlying PDE dynamics during training, and (iii) a model debiasing term $M_\epsilon(u, u^*)$ that has a negligible impact in the sample version. Detailed derivations and discussions on our loss function can be found in Appendix C. Moreover, we can prove that the total loss function $L_{\epsilon,\beta}(u)$ has a good lower and upper curvature around its minimizer u^*.

Proposition 2.2 (Stability estimate of the total loss function $L_{\epsilon,\beta}(u)$)**.** *Let u^* be the true solution to the heat equation defined in (1.1). Then for any $u \in H_t^1 H_x^1$, we have*

$$M_1(\beta,\epsilon)||u-u^*||^2_{H_t^1 H_x^1} \leqslant L_{\epsilon,\beta}(u) - L_{\epsilon,\beta}(u^*) \leqslant M_2(\epsilon,\beta)||u-u^*||^2_{H_t^1 H_x^1},$$

where

$$M_1(\beta,\epsilon) := \frac{1}{3} + \min(\frac{1}{3}, \frac{\beta e^{-\frac{\epsilon}{T}}}{C(d,\epsilon,\Omega)}, \frac{\beta e^{-\frac{\epsilon}{T}}}{4}), \qquad M_2(\beta,\epsilon) := 1 + \frac{\beta}{2} + \frac{\beta}{2\epsilon}, \tag{2.6}$$

and $C(d,\epsilon,\Omega) > 0$ is a constant depending on d, ϵ, and Ω.

We highlight that stability is a central structure to establish error estimation in PDE solutions in terms of the input data (e.g., initial and boundary conditions, source function) and it is also a cornerstone of statistical learning theory to prove sharp rate of convergences in estimation (cf. Theorem 3.1) and generalization error bounds (cf. Theorem 3.5).

2.2 Empirical Loss Minimization

Given the data $\{(X_{ij}, t_j, f_{ij}, u^*_{ij})\}_{i=1,j=1}^{m,n}$ and their computed derivatives $\{\nabla u^*, \partial_t u^*, \nabla\partial_t u^*, u^*_{tt}\}$, it is natural to minimize the empirical version of the total loss in (2.4) as

$$\hat{u}_\theta = \underset{u\in\mathcal{U}_J}{\arg\min} \left\{ \hat{L}_{\epsilon,\beta}(u) := \frac{1}{nm}\left[\hat{D}(u,u^*) + \beta\big(\hat{I}_\epsilon(u) + \hat{M}_\epsilon(u,u^*)\big)\right]\right\}, \tag{2.7}$$

where $\mathcal{U}_J$ is an appropriately chosen function space such as parameterized neural networks or an orthogonal subspace spanned by the truncated local basis in (3.1). For the explicit formula of the empirical loss function terms, please refer to Appendices C.

Note that the empirical loss minimization procedure (2.7) involves data from the *noiseless* solution and its derivative through numerical computation (e.g., calculated by the Python `findiff` library with high-order accuracy). This is a consequence of Sobolev norm underlying (2.7). The way to compute these derivatives follow previous works in physical informed learning literature Long et al. (2019); Rudy et al. (2017); Baydin et al. (2018). In our theory (Section 3), the solution data and its numerically computed derivatives are treated as noiseless and consistent with the statistical frameworks for learning elliptic equations under Sobolev-type losses (Lu et al., 2022). Since derivative computation is known to amplify the noise, we complement our experiments with ablation studies that evaluate the sensitivity of the proposed estimator to noise in both the solution and derivative data. In practice, our experiments also show using only solution labels can achieve significant performance improvement.

3 Theoretical Results

For convenience of our theoretical analysis, we use rescaled regularization parameter $\lambda = \beta/(nm)$ in the empirical total loss function in (2.7) and consider a subspace in the span of truncated local basis in space and time

$$\mathcal{U}_J := \left\{ u \in \mathcal{U}_0 \,\middle|\, \left\|\frac{d^p u}{dt^p}\right\|^2_{L_t^2(0,T;H_x^{2s+2-2p}(\Omega))} \leqslant R \right\} \cap V_{J,J'}, \tag{3.1}$$

where

$$V_{J,J'} := \left\{ u : \Omega\times[0,T] \to \mathbb{R} \,\middle|\, u(x,t) = \sum_{|\xi|_\infty \leqslant J, \tau \leqslant J'} a^2_{\tau,\xi}\phi^2_{\tau,\xi}(x,t) \right\}. \tag{3.2}$$

Note that one can view local basis as decoupled basis acting on space and time dimension, respectively $\phi_{\tau,\xi}(x,t) := \phi_\xi(x)\phi_\tau(t)$ to satisfy certain properties. In our analysis, we choose the local basis

to be wavelet expansions that capture the local information in the dynamics. For more details about these function space and construction of estimator $\hat{u}_J$ and local basis $\phi(x,t)$, please refer to Appendices A.2 and D. In our numerical experiments in Section 4, we use simple multi-layer perceptrons (MLPs) to compare with other standard neural solvers such as PINNs.

To derive theoretical results, we work with the standard statistical sampling model in which the space-time point locations $\{(X_{ij}, t_j)\}_{i=1,j=1}^{m,n}$ are sampled independently and uniformly from distribution P_0, where P_0 is the product Lebesgue measure on $\Omega \times [0,T]$.

3.1 Estimation Error Bounds

We denote $N := nm$ as the total sample size. Our first main result is an upper bound for the empirical loss minimization procedure of the energy-guided estimator in (2.7).

Theorem 3.1 (Convergence rates of energy-guided estimator)**.** *Under Assumptions 1-3 in Appendix A.2 and the sampling model* (1.2) *and* (1.3)*. We have the following bounds on the excess risk.*

Case I: supercritical regime $q \leqslant \frac{d}{2}$**.** *The error is further divided into three sub-regimes:*

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H_t^1 H_x^1} \lesssim_{\Omega_T,R} \begin{cases} \left(\frac{\log(1+C_2N)}{N}\right)^{\frac{4s-2}{4s+2d+2}}, & \text{if } 0 \leqslant \lambda \leqslant \frac{e^{T/\epsilon}}{N} \\ \left(\frac{\log(1+C_2N)e^{T/\epsilon}}{N^2\lambda}\right)^{\frac{4s-2}{4s+2d+2}}, & \text{if } \frac{e^{T/\epsilon}}{N} < \lambda \leqslant \lambda^* \\ \frac{\lambda}{e^{T/\epsilon}}, & \text{if } \lambda^* \leqslant \lambda < \infty \end{cases}, \tag{3.3}$$

where

$$\lambda^* := (\sqrt{\log(1+C_2N)}/N)^{\frac{4s-2}{4s+d}} e^{T/\epsilon}. \tag{3.4}$$

Case II: subcritical regime $q > \frac{d}{2}$**.** *For all* $\beta \geqslant 0$*, we have the following error bound*

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H_t^1 H_x^1} \lesssim_{\Omega_T,R} \left(\frac{\log(1+C_2N)}{N}\right)^{\frac{4s-2}{4s+d}}. \tag{3.5}$$

Remark 3.2. The relationship between n and m captures the parabolic scaling in the heat equation, where one temporal derivative is balanced against two spatial derivatives. This relation is also implicitly reflected in our proof using parabolic multi-level truncation (J, J^2) for space-time representation in the wavelet space. More precisely, we need $n \approx m^{2/d}$ in Proposition E.5 for Theorem 3.1 to hold.

Depending on the regularity of the ground-truth solution u^*, it is interesting to see from Theorem 3.1 that there are two regimes for the statistical convergence rate. In the subcritical regime, the prescribed regularity is sufficiently strong relative to the dimension for global Sobolev control to yield suitable metric entropy bounds. Consequently, the estimation problem behaves essentially as a standard nonparametric regression problem over the corresponding Sobolev class, and the energy-guidance term does not improve the leading-order statistical rate. In contrast, in the supercritical regime, the same prescribed regularity is insufficient, relative to the dimension, for such global Sobolev arguments to provide a sharp characterization of the statistical complexity. A more localized analysis, based on wavelet representations and the PDE energy structure, is then required. In this regime, energy guidance can exploit the local dynamical information encoded by the equation and lead to a sharper rate of convergence. In addition, it turns out that by properly choosing the tuning parameter λ^*, the rate of convergence for energy-guided learning with auxiliary data is minimax optimal up to a logarithmic factor, justified by Theorem 3.3 below.

Theorem 3.3 (Lower bound on estimation error)**.** *Fix* $\epsilon > 0$*,* $s \geqslant 1$ *and* $2p + q = 2s + 2$ *for* $0 \leqslant p \leqslant s$*. Under Assumptions 1-4 in Appendix A.2, we have the following result holds*

$$\inf_{\hat{u}(x,t)} \sup_{u^*(f)\in\mathcal{U}} \mathbb{E}\left[||\hat{u}(x,t) - u^*(f)||^2_{H_t^1 H_x^1}\right] \gtrsim N^{-\frac{4s-2}{4s+d}}. \tag{3.6}$$

Our minimax lower bound under the space-time Sobolev norm is a non-standard rate. However, (3.6) relates to the standard nonparametric rate $n^{-\frac{2s}{2s+d-2}}$ for estimating spatially Hölder smooth functions

under the L_2 evaluation metric (Tsybakov, 2009). To interpret the difference, our obtained rate suffers regularity loss by degree of three: one from spatial derivative and other two from temporal derivative, witnessing one versus two space-time regularity trade-off by using the $H_t^1 H_x^1$ metric.

3.2 Excess Risk and Generalization Error Bounds

There are two commonly used performance measures to evaluate the statistical and generalization capability of a learned model: (1) the *excess risk* $\Delta\mathcal{E}_S := L_{\epsilon,\beta}(\hat{u}_J) - L_{\epsilon,\beta}(u^*)$, reflecting the difference between the expected losses between a learned model and the best possible model (i.e., oracle); (2) the *generalization error* $\Delta\mathcal{E}_{S,gen} := \hat{L}_{\epsilon,\beta}(\hat{u}_J) - L_{\epsilon,\beta}(\hat{u}_J)$, quantifying the gap between the training and testing loss for the *learned model* on the training data. Bounds on excess risk and generalization error both measure the quality of prediction on unseen testing data. These two quantities are related and we remark that the excess risk bound is also a key step to establish the convergence rates of energy-guided estimator in Theorem 3.1.

Theorem 3.4 (Excess risk bound)**.** *Suppose the assumptions in Theorem 3.1 hold. For the empirical minimizer of the total loss function $\hat{u}_J$, we choose $J = J_1^*$ in* (E.166) *if $p \leqslant \frac{d}{2}$ and $J = J_2^*$ in* (E.176) *if $q > \frac{d}{2}$. Then the excess risk satisfies the following expectation bound over the training data distribution*

$$\mathbb{E}\left[\Delta\mathcal{E}_S\right] \lesssim \frac{e^{T/\epsilon}}{\epsilon}\Big(\frac{\sqrt{\log(1+C_2N)}}{N}\Big)^{\frac{4s-2}{4s+d}}, \tag{3.7}$$

where the constant of $\lesssim$ depends only on Ω_T and R.

Theorem 3.5 (Generalization error bound)**.** *Suppose the assumptions in Theorem 3.1 hold, the ground truth solution u^* to the PDE in (1.1), and $\hat{u}_J$ be empirical minimizer of the total loss function in* (2.7) *on the training data. We further assume $q > \frac{d}{2} + 1$ and $p \geqslant 2$, then the following result holds*

$$\mathbb{E}[\Delta\mathcal{E}_{S,gen}] \lesssim J^{\frac{d+2}{2}} \cdot \log J \cdot \frac{\log(N)}{N}, \tag{3.8}$$

where the constant of $\lesssim$ depends only on $\epsilon, d, q, R, T, \Omega$, and F.

Remark 3.6. Lu et al. (2021b) proved the slow generalization error using a one-layer neural network with m neurons for elliptic PDEs, and Lu et al. (2021a) improved this bound to fast generalization error with J-truncated Fourier series instead. Their generalization bounds also suffer from polynomial growth of model parameters in terms of m or J. Our Theorem 3.5 achieves the fast rate of generalization error in terms of the effective sample size nm. Note that the upper bound on excessive risk in Theorem 3.4 incorporates approximation error, while Theorem 3.5 only focus on statistical errors. We also impose a slightly stronger assumption $q > \frac{d}{2} + 1$ to enforce higher spatial regularity so the Sobolev embedding theorem is applied to enhance local integrability.

4 Numerical Experiments

In this section, we present some interesting examples using our energy-regularized approach for readers to understand how our approach works. The comprehensive experiment report can be found in G.1. We report using our energy-regularized approach to solve four parabolic linear PDEs and one nonlinear parabolic PDEs in various dimensions up to $d = 50$. In appendix G.2, we show our methods achieve dominant accuracy advantage in L_2 relative in comparison with various PINNs baseline. Furtheremore, we show our energy regularization effects beats purely sobolev regression using similar sample-adaptive regularization tuning in appendix G.3. Since we ideally assume solutions label to be noiseless and compute derivatives from solution labels, we also provide noise ablation on how additional noise injection will degrade our prediction; our ablation results show our method is robust against mild noisy solution labels and computed derivatives as well. To see the extension of our method, we provide straightforward extension to nonlinear hyperbolic PDEs in appendix I. Lastly, we also present parameter sensitivity analysis (see appendix H) for readers to understand how PDEs regularization parameter ϵ, β, and sample complexity N dynamically entangles.

4.1 Heat Equation

We first consider the heat equation in the domain $\Omega \times [-1,1]^d$ with a zero forcing term for $d = 1, 3$. Here we choose the initial condition $u_0(x) = \log\log(1 + |x|^{-1}) \in H^1(\Omega)$ which has a severe

singularity around $|x| = 0$. In this setting, strong solution does not exist and one only has weak solution in an appropriate Sobolev space. In practice, since u_0 cannot be numerically computed due to singularity, we perturb around $|x| = 0$ by adding a small perturbation 10^{-6}, leading to a modified initial condition $u_0(x) = \log\log(1 + (|x|^2 + 10^{-6})^{-1/2})$. For a fair comparison, we make the total sample size roughly the same between a vanilla PINN and our method, because they have distinct sampling schemes. For the 1D heat equation, Figure 1 shows that our solution is very accurate in the entire phase of space and time, while the PINN solution fails. Moreover, Figure 2 shows that PINN prioritizes learning snapshots around the singularity $|x| = 0$ at initial time and does not keep track of the overall dynamics; that is, PINN does not capture physical dynamics when t is away from zero. Our explanation is that the penalizing mechanism in the PINN loss design leads to the imbalance of initial condition and PDE residuals, causing the failure in capturing dynamics at larger $t > 0$, since one needs to allocate different penalty weights for learning interior solutions, boundary solutions, or initial solutions. As a singularity is present at $t = 0$, PINNs heavily enforce residuals to initial time points instead of overall dynamics, as shown in Figure 2 (top row).

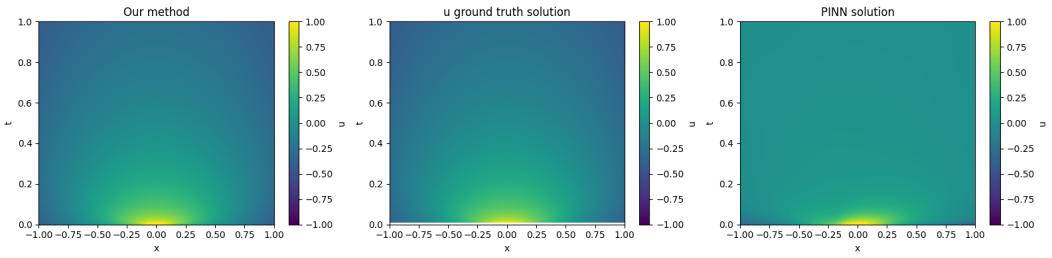

Figure 1: Comparison of solutions from our method and PINN for a 1-D heat equation.

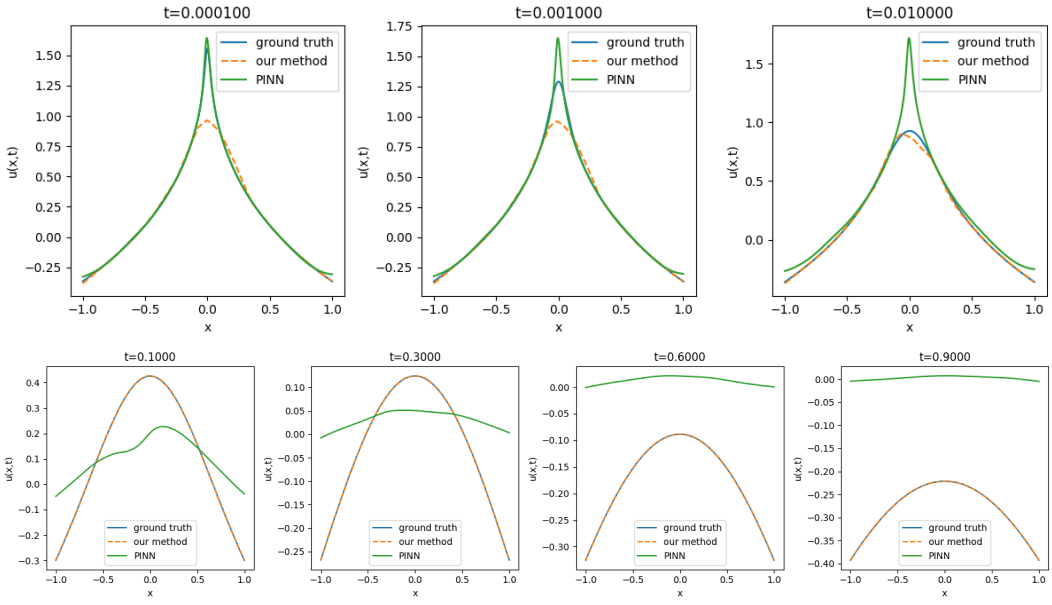

Figure 2: Top row: Near $t \approx 0$ snapshot comparison of solutions from our method and a PINN for a 1-D solution. Bottom row: zoom-in snapshot comparison of solutions from our method and a PINN different times.

We observe similar phenomena in higher-dimensional experimental settings. Figure 3 shows a 3D scenario in which PINN fails to solve the heat equation. In the literature, several approaches have been proposed to improve PINNs by rebalancing loss penalty weights to handle singularities formed in shock waves and to enhance estimation accuracy for weak solutions. Although these alternative methods improve upon the baseline PINN benchmark (De Ryck et al., 2024; Kharazmi et al., 2021; McClenny & Braga-Neto, 2023; Wang et al., 2022a), they remain underperformed relative to our approach by several orders of magnitudes in testing errors in Appendix G. Given that strong solutions

are generally easier to approximate, we anticipate that those competing methods will correspondingly fall short of the performance of our model in more complex scenarios where only weak solutions exist, e.g., with singularity initial conditions.

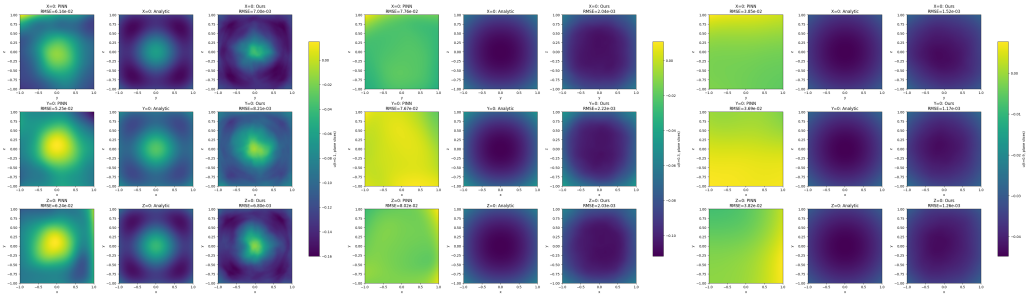

Figure 3: 3D heat equation example. Left plot is 3D solution projection onto two coordinates at $t = 0$. Middle plot is 3D solution projection onto two coordinates at $t = 0.3$. Right plot is 3D solution projection onto two coordinates at $t = 0.9$.

4.2 3D Fokker-Planck Equation

In this experiment, we show the applicability of our energy-guided method for more general linear parabolic PDEs. Specifically, we consider the Fokker-Planck equation $u_t + \nabla \cdot (b\, u) - \nabla \cdot (D\, \nabla u) = 0$ with standard Gaussian initial condition $u(x, 0) = (2\pi)^{-1/2} e^{-|x|^2}$. In this setting, we use a constant drift coefficient $b = -1$ and diffusion coefficient $D = 0.25$. Throughout the time evolution, our solution roughly preserves mass conversation inherited by physics law in the Fokker-Planck equation. The learned solutions of our method compared with the ground truth solutions are shown in Figure 4.

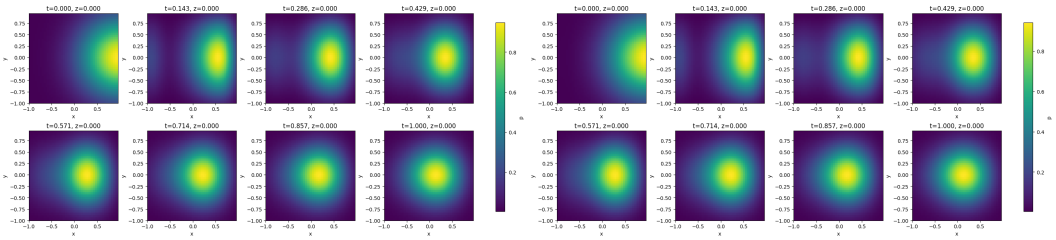

Figure 4: 3D Fokker-Planck equation example. Left plot is ground-truth solution projection onto x,y coordinates over time evolution. Right plot is our solution projection onto x,y coordinates over time evolution.

5 Discussion

Our work studies variational energy regularization for data-driven approximation of parabolic PDEs. Though our theoretical sheds some light in the guidance of energy regularization in physical learning, we admit our framework could be better implemented in suitable architecture including optimizer and neural network.Current popular architecture in time-dependent physical dynamic prediction only uses MLPs where temporal information is not well-represented in architecture realization. We believe networks with separate spatial and temporal representations can better capture parabolic regularity than standard MLPs. Current research on foundation models on dynamics system McCabe et al. (2025) is worth investigation to understand how these architectural choices interact with energy regularization to affect approximation error, optimization, and generalization remains an important direction.

AI USE STATEMENT

In this work, we used generative AI tools to edit portions of the manuscript for clarity, readability, grammar, and style. The tools were not used to generate research ideas, formulate mathematical claims or proofs, design the methodology or experiments, interpret results, or otherwise determine the scientific content of the paper. All AI-assisted edits were reviewed and revised by the authors. The authors take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Robert A. Adams. *Sobolev Spaces*, volume 65 of *Pure and Applied Mathematics*. Academic Press, New York–London, 1975.

Goro Akagi and Ulisse Stefanelli. Weighted energy-dissipation functionals for doubly nonlinear evolution. *Journal of Functional Analysis*, 260(9):2541–2578, 2011. doi: 10.1016/j.jfa.2010.12.027.

Goro Akagi, Verena Bögelein, Alice Marveggio, and Ulisse Stefanelli. Weighted inertia-dissipation-energy approach to doubly nonlinear wave equations. *Journal of Functional Analysis*, 289(8):111067, 2025a. doi: 10.1016/j.jfa.2025.111067.

Goro Akagi, Ulisse Stefanelli, and Riccardo Voso. Weighted energy-dissipation approach to semilinear gradient flows with state-dependent dissipation. *Journal of Differential Equations*, 440:113431, 2025b. doi: 10.1016/j.jde.2025.113431.

Herbert Amann. *Linear and Quasilinear Parabolic Problems. Volume II: Function Spaces*, volume 106 of *Monographs in Mathematics*. Springer, 2019. doi: 10.1007/978-3-030-11763-4.

Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019. doi: 10.1017/S0962492919000059.

Alessandro Audrito and Tomás Sanz-Perela. Elliptic regularization of some semilinear parabolic free boundary problems. *Interfaces and Free Boundaries*, 26(1):135–159, 2024.

G. K. Batchelor. *An Introduction to Fluid Dynamics*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000.

Atılım Güneş Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018.

Marshall W. Bern and Paul E. Plassmann. Mesh generation. In *Handbook of Computational Geometry*, pp. 291–332. North-Holland, 2000.

Nicolas Boullé, Seick Kim, Tianyi Shi, and Alex Townsend. Learning Green's functions associated with time-dependent partial differential equations. *Journal of Machine Learning Research*, 23(218):1–34, 2022.

Nicolas Boullé, Diana Halikias, and Alex Townsend. Elliptic PDE learning is provably data-efficient. *Proceedings of the National Academy of Sciences*, 120(39):e2303904120, 2023. doi: 10.1073/pnas.2303904120.

Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathématique*, 334(6):495–500, 2002. doi: 10.1016/S1631-073X(02)02292-6.

Steven L. Brunton and J. Nathan Kutz. Promising directions of machine learning for partial differential equations. *Nature Computational Science*, 4(7):483–494, 2024.

J. R. Cannon and S. P. Esteva. An inverse problem for the heat equation. *Inverse Problems*, 2(4):395, 1986. doi: 10.1088/0266-5611/2/4/007.

Hector D. Ceniceros and Thomas Y. Hou. An efficient dynamically adaptive mesh for potentially singular solutions. *Journal of Computational Physics*, 172(2):609–639, 2001.

A. Couairon and J.-M. Chomaz. Global instability in fully nonlinear systems. *Physical Review Letters*, 77(19):4015, 1996.

Jianbo Cui, Shu Liu, and Haomin Zhou. A supervised learning scheme for computing Hamilton–Jacobi equation via density coupling, 2024.

Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics-informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022. doi: 10.1007/s10915-022-01939-z.

Wojciech Marian Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4281–4290, 2017.

Tim De Ryck, Siddhartha Mishra, and Roberto Molinaro. wPINNs: Weak physics weak physics informed neural networks for approximating entropy solutions of hyperbolic conservation laws. *SIAM Journal on Numerical Analysis*, 62(2):811–841, 2024. doi: 10.1137/22M1522504.

Mehdi Dehghan and Ali Shokri. A numerical method for solving the hyperbolic telegraph equation. *Numerical Methods for Partial Differential Equations*, 24(4):1080–1093, 2008. doi: 10.1002/num.20306.

Jacques Deny and Jacques-Louis Lions. Les espaces du type de Beppo Levi. *Annales de l'Institut Fourier*, 5:305–370, 1954. doi: 10.5802/aif.55.

Hongjie Dong and Yanze Liu. Sobolev estimates for fractional parabolic equations with space-time non-local operators. *Calculus of Variations and Partial Differential Equations*, 62(3):96, 2023.

Hongjie Dong, Pilgyu Jung, and Doyoon Kim. L_p-estimates of the conormal derivative problem for parabolic equations with time measurable coefficients and A_p-weights, 2025.

Chenguang Duan, Yuling Jiao, Yanming Lai, Dingwei Li, Xiliang Lu, and Jerry Zhijian Yang. Convergence rate analysis for deep Ritz method. *Communications in Computational Physics*, 31(4):1020–1048, 2022. doi: 10.4208/cicp.OA-2021-0195.

Richard M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1:290–330, 1967.

Weinan E and Bing Yu. The deep Ritz method: A deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1):1–12, 2018. doi: 10.1007/s40304-018-0127-z.

Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2nd edition, 2010. Chapter 7, pp. 388–390.

Stephen Gasiorowicz. *Quantum Physics*. Wiley India Pvt. Limited, 3rd edition, 2007. ISBN 9788126511174.

Sydney Goldstein. On diffusion by discontinuous movements, and on the telegraph equation. *The Quarterly Journal of Mechanics and Applied Mathematics*, 4(2):129–156, 1951. doi: 10.1093/qjmam/4.2.129.

Ehsan Haghighat, Maziar Raissi, Adrian Moure, Hector Gomez, and Ruben Juanes. A physics-informed deep learning framework for inversion and surrogate modeling in solid mechanics. *Computer Methods in Applied Mechanics and Engineering*, 379:113741, 2021. doi: 10.1016/j.cma.2021.113741.

Alemdar Hasanov Hasanoğlu and Vladimir G. Romanov. *Introduction to Inverse Problems for Differential Equations*. Springer, 2nd edition, 2021.

Haoteng Hu, Lehua Qi, and Xujiang Chao. Physics-informed neural networks (PINN) for computational solid mechanics: Numerical frameworks and applications. *Thin-Walled Structures*, 205: 112495, 2024. doi: 10.1016/j.tws.2024.112495.

Marco A. Iglesias, Kody J. H. Law, and Andrew M. Stuart. Ensemble Kalman methods for inverse problems. *Inverse Problems*, 29(4):045001, 2013.

Tom Ilmanen. *Elliptic regularization and partial regularity for motion by mean curvature*, volume 108 of *Memoirs of the American Mathematical Society*. American Mathematical Society, Providence, RI, 1994. No. 520.

Yuling Jiao, Yanming Lai, Dingwei Li, Xiliang Lu, Fengru Wang, Yang Wang, and Jerry Zhijian Yang. A rate of convergence of physics informed neural networks for the linear second order elliptic PDEs. *Communications in Computational Physics*, 31(4):1272–1295, 2022. doi: 10.4208/cicp.OA-2021-0186.

Fritz John. *Partial Differential Equations*, volume 1 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 4th edition, 1982.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. doi: 10.1038/s42254-021-00314-5.

Ehsan Kharazmi, Zhongqiang Zhang, and George Em Karniadakis. hp-VPINNs: Variational physics-informed neural networks with domain decomposition. *Computer Methods in Applied Mechanics and Engineering*, 374:113547, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert M. Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26548–26560, 2021.

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. FourCastNet: Accelerating global high-resolution weather forecasting using adaptive Fourier neural operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference (PASC '23)*, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3592979.3593412.

Tong Lin, Hanlin Xue, Ling Wang, Bo Huang, and Hongbin Zha. Supervised learning via Euler's elastica models. *Journal of Machine Learning Research*, 16(111):3637–3686, 2015.

Jacques-Louis Lions and Enrico Magenes. *Non-Homogeneous Boundary Value Problems and Applications, Vol. I*, volume 181 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 1972. doi: 10.1007/978-3-642-65161-8.

Zichao Long, Yiping Lu, and Bin Dong. PDE-Net 2.0: Learning PDEs from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics*, 399:108925, 2019. doi: 10.1016/j.jcp.2019.108925.

Yiping Lu, Haoxuan Chen, Jianfeng Lu, Lexing Ying, and Jose Blanchet. Statistical numerical PDE: Fast rate, neural scaling law and when it's optimal. In *NeurIPS 2021 Workshop: The Symbiosis of Deep Learning and Differential Equations*, 2021a.

Yiping Lu, Jose Blanchet, and Lexing Ying. Sobolev acceleration and statistical optimality for learning elliptic equations via gradient descent. *Advances in Neural Information Processing Systems*, 35:33233–33247, 2022.

Yulong Lu, Jianfeng Lu, and Min Wang. A priori generalization analysis of the deep Ritz method for solving high dimensional elliptic partial differential equations. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3196–3241. PMLR, 2021b.

Tengyu Ma. Lecture notes for machine learning theory (CS229M/STATS214). Stanford University, 2022. URL https://github.com/tengyuma/cs229m_notes/blob/main/master.pdf. Course page: https://web.stanford.edu/class/stats214/.

Michael McCabe, Payel Mukhopadhyay, Tanya Marwah, Bruno Régaldo-Saint Blancard, François Rozet, Cristiana Diaconu, Lucas Meyer, Kaze W. K. Wong, Hadi Sotoudeh, Alberto Bietti, Irina Espejo, Rio Fear, Siavash Golkar, Tom Hehir, Keiya Hirashima, Geraud Krawezik, François Lanusse, Rudy Morel, Ruben Ohana, Liam Parker, Mariel Pettee, Jeff Shen, Kyunghyun Cho, Miles Cranmer, and Shirley Ho. Walrus: A cross-domain foundation model for continuum dynamics. *arXiv preprint arXiv:2511.15684*, 2025.

Levi D. McClenny and Ulisses M. Braga-Neto. Self-adaptive physics-informed neural networks. *Journal of Computational Physics*, 474:111722, 2023. doi: 10.1016/j.jcp.2022.111722.

Alexander Mielke and Michael Ortiz. A class of minimum principles for characterizing the trajectories and the relaxation of dissipative systems. *ESAIM: Control, Optimisation and Calculus of Variations*, 14(3):494–516, 2008. doi: 10.1051/cocv:2007064.

Alexander Mielke and Ulisse Stefanelli. Weighted energy-dissipation functionals for gradient flows. *ESAIM: Control, Optimisation and Calculus of Variations*, 17(1):52–85, 2011. doi: 10.1051/cocv/2009043.

Yifan Mo, Liming Ling, and Delu Zeng. Data-driven vector soliton solutions of coupled nonlinear Schrödinger equation using a deep learning algorithm. *Physics Letters A*, 421:127739, 2022. doi: 10.1016/j.physleta.2021.127739.

Youssef Mroueh, Tom Sercu, and Anant Raj. Sobolev descent. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2976–2985. PMLR, 2019.

Pawan Negi, Maggie Cheng, Mahesh Krishnamurthy, Wenjun Ying, and Shuwang Li. Learning domain-independent Green's function for elliptic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 421:116779, 2024. doi: 10.1016/j.cma.2024.116779.

Richard Nickl. Bernstein–von Mises theorems for time evolution equations, 2024.

Ariel Norambuena, Marios Mattheakis, Francisco J. González, and Raúl Coto. Physics-informed neural networks for quantum control. *Physical Review Letters*, 132(1):010801, 2024. doi: 10.1103/PhysRevLett.132.010801.

Enzo Orsingher and Luisa Beghin. Time-fractional telegraph equations and telegraph processes with Brownian time. *Probability Theory and Related Fields*, 128(1):141–160, 2004.

Ravi G. Patel, Indu Manickam, Nathaniel A. Trask, Mitchell A. Wood, Myoungkyu Lee, Ignacio Tomas, and Eric C. Cyr. Thermodynamically consistent physics-informed neural networks for hyperbolic systems. *Journal of Computational Physics*, 449:110754, 2022.

Per-Olof Persson. *Mesh generation for implicit geometries*. PhD thesis, Massachusetts Institute of Technology, 2005.

David Pfau, James S. Spencer, Alexander G. D. G. Matthews, and W. M. C. Foulkes. Ab initio solution of the many-electron Schrödinger equation with deep neural networks. *Physical Review Research*, 2:033429, 2020.

Maziar Raissi, Alireza Yazdani, and George Em Karniadakis. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 367(6481):1026–1030, 2020. doi: 10.1126/science.aaw4741.

Martin Rathmair. On how Poincaré inequalities imply weighted ones. *Monatshefte für Mathematik*, 188(4):753–763, 2019. doi: 10.1007/s00605-019-01266-w.

Riccarda Rossi, Giuseppe Savaré, Antonio Segatti, and Ulisse Stefanelli. Weighted energy-dissipation principle for gradient flows in metric spaces. *Journal de Mathématiques Pures et Appliquées*, 127:1–66, 2019. doi: 10.1016/j.matpur.2018.06.022.

Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017. doi: 10.1126/sciadv.1602614.

Enrico Serra and Paolo Tilli. Nonlinear wave equations as limits of convex minimization problems: Proof of a conjecture by De Giorgi. *Annals of Mathematics*, 175(3):1551–1574, 2012. doi: 10.4007/annals.2012.175.3.11.

Pushan Sharma, Wai Tong Chung, Bassem Akoush, and Matthias Ihme. A review of physics-informed machine learning in fluid mechanics. *Energies*, 16(5):2343, 2023. doi: 10.3390/en16052343.

Khemraj Shukla, Patricio Clark Di Leoni, James Blackshire, Daniel Sparkman, and George Em Karniadakis. Physics-informed neural network for ultrasound nondestructive quantification of surface breaking cracks. *Journal of Nondestructive Evaluation*, 39(3):61, 2020.

Hwijae Son, Jin Woo Jang, Woo Jin Han, and Hyung Ju Hwang. Sobolev training for physics informed neural networks, 2021.

Ulisse Stefanelli. The weighted inertia-energy-dissipation principle. *Mathematical Models and Methods in Applied Sciences*, 35(02):223–282, 2025. doi: 10.1142/S0218202525400019.

Elias M. Stein and Rami Shakarchi. *Functional Analysis: Introduction to Further Topics in Analysis*. Princeton University Press, 2011.

Roger Temam. *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, volume 68 of *Applied Mathematical Sciences*. Springer, New York, 2nd edition, 1997. doi: 10.1007/978-1-4612-0645-3.

Hans Triebel. *Theory of Function Spaces III*, volume 100 of *Monographs in Mathematics*. Birkhäuser Basel, 2006.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. doi: 10.1007/b13794.

Bengt Ove Turesson. *Nonlinear Potential Theory and Weighted Sobolev Spaces*, volume 1736 of *Lecture Notes in Mathematics*. Springer, Berlin, Heidelberg, 2000.

Sabir Umarov. On fractional Duhamel principle and its applications. *Journal of Differential Equations*, 252(10):5217–5234, 2012. doi: 10.1016/j.jde.2012.01.029.

Sara van de Geer. A new approach to least-squares estimation, with applications. *The Annals of Statistics*, 15(2):587–602, 1987. doi: 10.1214/aos/1176350362.

Sara van de Geer. M-estimation using penalties or sieves. *Journal of Statistical Planning and Inference*, 108:55–69, 2002. doi: 10.1016/S0378-3758(02)00270-7.

A. W. van der Vaart. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

Aad W. van der Vaart and Jon A. Wellner. Empirical processes indexed by estimated functions. In *Asymptotics: Particles, Processes and Inverse Problems*, volume 55 of *IMS Lecture Notes Monograph Series*, pp. 234–252. Institute of Mathematical Statistics, 2007.

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

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Chuwei Wang, Shanda Li, Di He, and Liwei Wang. Is L^2 physics-informed loss always suitable for training physics-informed neural network? In *Advances in Neural Information Processing Systems*, volume 35, 2022a.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021. doi: 10.1137/20M1318043.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022b. doi: 10.1016/j.jcp.2021.110768.

Qidong Yang, Alex Hernandez-Garcia, Paula Harder, Venkatesh Ramesh, Prasanna Sattigeri, Daniela Szwarcman, Campbell D. Watson, and David Rolnick. Fourier neural operators for arbitrary resolution climate data downscaling. *Journal of Machine Learning Research*, 25(420):1–30, 2024. URL http://jmlr.org/papers/v25/23-0597.html.

Appendix

A Notations and Assumptions

It this section, we collect the notations and the main assumptions that are used in this paper.

A.1 Notations

Throughout the paper, we let the domain be closed bounded set $\Omega \in \mathbb{R}^d$ with Lipschitz boundary $\partial\Omega$ over finite time horizon $[0, T]$, and spatial-temporal domain be $\Omega_T := \Omega \times [0, T]$. we denote implicit regularization parameter $0 < \epsilon < 1$, total smoothness parameter s, temporal regularity parameter p and spatial regularity parameter q such that $s \geqslant 1$ and $2p + q = 2s + 2$ and $0 \leqslant p \leqslant s$, and the size of Sobolev norm of u $R > 1$, and denote the functional spaces $\mathcal{U} := \mathcal{U}(R)$, $\mathcal{F} := \mathcal{F}(R)$, and $H_t^1 H_x^1 := H_t^1(0, T; H_x^1(\Omega))$. Unless we make additional specifications, we follow the following notation:

d : spatial dimension,
Ω : closed bounded domain with Lipschitz boundary $\partial\Omega$
ϵ : implicit regularization parameter for the regularized PDEs,
ϵ_1 : radius of net for construction of covering number
J : the truncation level of the local basis representation,
n : the number of temporal samples, m : the number of spatial samples
N : the number of total sample size
s : the total smoothness, q : the spatial smoothness, p : the temporal smoothness
β : explicit regularization parameter for the energy regularizer $I_\epsilon(u)$,
ρ : the radius of localized balls,
$L_{\epsilon,\beta}(u)$: our loss function ϵ and β,
$\hat{L}_{\epsilon,\beta}(u)$: the empirical counterpart of $L_{\epsilon,\beta}(u)$,
$I_\epsilon(u)$: weighted energy functional,
$\hat{I}_\epsilon(u)$: the empirical counterpart of $I_\epsilon(u)$,
$H_t^p H_x^q : H^p(0, T; H^q(\Omega))$, $H_t^1 H_x^1 : H^1(0, T; H^1(\Omega))$,
u^* : the ground truth solution satisfying the PDE problem,
u : estimator,
$\hat{u}$: empirical estimator of u^*,
$\mathcal{U}$: the search space of models/estimators,
$\mathcal{U}_J$: the search space of models/estimators with local basis representation at truncation level J,
$\hat{u}_J := \arg\min_{u \in \mathcal{U}_J} \widehat{L}_{\epsilon,\beta}(u)$,
$u_J := \arg\min_{u \in \mathcal{U}_J} L_{\epsilon,\beta}(u)$.

A.2 Main Assumptions

To follow up discussions on assumptions of this paper, we make a few comments on parameters used in our paper. ϵ is an implicit regularization parameter that serves to regularize PDEs model. The relation between p, q, and s is well-known as linear parabolic equations always gain more one more derivative as time evolves. The bounded Sobolev norm R is used to control the growth of our estimator $\hat{u}_J$ in (3.1).

A1 (Regularity Condition) Let $\mathcal{F}(R)$ be the functional space of forcing function $f : \Omega_T \to \mathbb{R}$ and $u_0^* := u^*(x, 0)$ be initial condition such that

$$(1)\ u_0^* \in H^{2s+1}(\Omega), \tag{A.1}$$

$$(2)\ ||\frac{d^p f}{dt^p}||^2_{L_t^2(0,T;H_x^{2s-2p}(\Omega))} \leqslant R,\ 0 \leqslant p \leqslant s. \tag{A.2}$$

A2 (Compatibility Condition) Let u_0^* be initial value of linear parabolic PDEs and f be right hand side of the PDEs, then u_0^* and f need to satisfy the following conditions

$$(3) = \begin{cases} g_0 := u_0^* \in H_0^1(\Omega),\ g_1 := f(0,x) - Lg_0 \in H_0^1(\Omega), \\ ,...., g_{p-1} := \dfrac{d^{p-1}}{dt^{p-1}} f(0,x) - Lg_{p-1} \in H_0^1(\Omega). \end{cases} \tag{A.3}$$

A3 (Search Space) Let $\mathcal{U}(R)$ be the set of all functions $u : \Omega_T \to \mathbb{R}$ such that

$$(4)\ u(x,t) \in H_t^p(0,T;H_x^{2s+2-2p}(\Omega)) \cap L_t^2(0,T;H_{0,x}^1(\Omega)),\ u(x,0) = u_0^*(x)\ , 0 \leqslant p \leqslant s, \tag{A.4}$$

$$(5)\ ||\frac{d^p u}{dt^p}||^2_{L_t^2(0,T;H_x^{2s+2-2p}(\Omega))} \leqslant R,\ 0 \leqslant p \leqslant s, \tag{A.5}$$

$$(6)\ u^* \in \mathcal{U}(R). \tag{A.6}$$

A4 (Invertibility Condition) Moreover, we require u^* to satisfy

$$(7)\ u^* \in \mathcal{U}(R) \cap \{u \mid u_t + Lu \in \mathcal{F}(R)\}. \tag{A.7}$$

Remark A.1. (1)-(3) are essential assumptions needed in the the parabolic regularity theory 3.1 in (Evans, 2010, Chapter 7, Section 7.1, Theorem 6) to establish higher order regularity result.; (5) is equivalent to $||f||^2_{H_t^p(0,T;H_x^{2s+2-2p}(\Omega))} \lesssim R$, which guarantees that the approximation errors used in theorem 3.1 are controllable, and (6) guarantees learning framework is realizable; (7) is sufficient condition to guarantee for $u^* \in \mathcal{U}(R)$, there always exists data f such that we can construct an estimator $\hat{u}$ in minimax lower bound.

A.3 Functional Space and Norms

In this section, we summarize the key functional spaces used in our analysis, with particular attention to their relationship with Fourier series representations. A more detailed treatment of these topics can be found in (Evans, 2010, Chapter 5) and Adams (1975).

Let $\Omega \subseteq \mathbb{R}^n$ be an open set with nonempty interior and integers $\alpha, \beta \geqslant 0$, we denote $W^{\alpha,\beta}$ as the set of the Sobolev function f with α degree of smoothness and β degree of integrability. We define the norms as

$$||f||_{W^{\alpha,\beta}(\Omega)} := \sum_{|k| \leqslant \alpha} ||\partial^k f||_{L^\beta(\Omega)}, \tag{A.8}$$

where the derivative ∂^k is defined as weak derivative in the distributional sense and detailed discussion of weak derivative is in (Stein & Shakarchi, 2011, Chapter 3, Section 1).

Sometimes, an alternative but equivalent definition is used:

$$\|f\|_{W^{\alpha,\beta}(\Omega)} := \|f\|_{L^\beta(\Omega)} + \sum_{|k|=\alpha} \|\partial^k f\|_{L^\beta(\Omega)}. \tag{A.9}$$

In this work, we adopt the full norm in (A.8). In particular, we denote

$$H^\alpha(\Omega) := W^{\alpha,2}(\Omega). \tag{A.10}$$

We also define the negative-order Sobolev space $H^{-1}(\Omega)$ as the dual of $H_0^1(\Omega)$, which satisfies the following Gelfand triple:

$$H_0^1(\Omega) \subseteq L^2(\Omega) \subseteq H^{-1}(\Omega). \tag{A.11}$$

Next, we define functional spaces involving time. Here, we consider function f as a mapping from $t \in [0, T]$ to Banach Space $\mathcal{X}$.

A.3.1 SPACES INVOLVING TIME

Let $\mathcal{X}$ denote a real Banach space, with the norm $||\cdot||$.

Definition 1. *The space*

$$L^\beta(0,T;\mathcal{X}) \tag{A.12}$$

consists of all measurable functions $f:[0,T]\to\mathcal{X}$ with

$$||f||_{L^\beta(0,T;\mathcal{X})} := \begin{cases} \left(\int_0^T ||f(t)||^\beta\right)^{\frac{1}{\beta}} < \infty, \quad 1\leqslant\beta<\infty, \\ \\ ess\sup_{0\leqslant t\leqslant T}||f(t)|| < \infty, \qquad \beta=\infty \end{cases} \tag{A.13}$$

Once we have defined $L^\beta(0,T;\mathcal{X})$, we can define the time-dependent Sobolev space

Definition 2. *The Sobolev space*

$$W^{\alpha,\beta}(0,T;\mathcal{X}) \tag{A.14}$$

consists of all functions $f\in L^\beta(0,T;\mathcal{X})$ such that weak derivative $\frac{\partial^k u}{\partial t^k}$ exists in the distributional sense and belongs to $L^\beta(0,T;\mathcal{X})$. Furthermore,

$$||u||_{W^{\alpha,\beta}(0,T;\mathcal{X})} := \begin{cases} \left(\int_0^T \sum_{k=0}^{\alpha} ||\frac{\partial^k u(t)}{\partial t^k}(t,\cdot)||^\beta dt\right)^{1/\beta}, \quad 1\leqslant\beta<\infty \\ \\ ess\sup_{0\leqslant t\leqslant T}\sum_{k=0}^{\alpha}||\frac{\partial^k u(t)}{\partial t^k}(t,\cdot)||^\beta dt, \quad \beta=\infty \end{cases} . \tag{A.15}$$

We write $H^\alpha(0,T;\mathcal{X}) = W^{\alpha,2}(0,T;\mathcal{X})$. In particular, let $\mathcal{X}=H^{\alpha'}(\Omega)$, $\alpha'>0$, we abbreviate

$$H_t^\alpha H_x^{\alpha'} = H^\alpha(0,T;H^{\alpha'}(\Omega)). \tag{A.16}$$

A.3.2 SOBOLEV EXTENSION AND FOURIER CHARACTERIZATION

A very useful tool to handle function space defined on compact domains is using extension operators that preserve the norms. The following theorem ensures the existence of such an extension operator that is the same among all Sobolev spaces. This allows us to characterize Sobolev functions on compact domains via the Fourier coefficients of their extensions, a crucial step to derive approximation errors in our proof.

Theorem A.2 (Lipschitz Domain Extension Theorem)**.** *(Evans, 2010, Theorem 5.4, when α=1)(Triebel, 2006, Theorem 1.105) Let $\Omega\subseteq\mathbb{R}^d$ be a bounded Lipschitz domain and the smoothness parameter $\alpha>0$. Then, there exists a linear extension operator*

$$Ext: H^\alpha(\Omega)\to H^\alpha(\mathbb{R}^d) \tag{A.17}$$

such that for any $f\in H^\alpha(\Omega)$,

$$(i) Ext\, f\,|_\Omega = f, \ \text{a.s in } \Omega, \tag{A.18}$$

$$(ii) ||Ext\, f||_{W^{\alpha,2}(\mathbb{R}^d)} \leqslant C(\Omega)||f||_{W^{\alpha,2}(\Omega)}, \tag{A.19}$$

where the constant C depending only on Ω.

If $f(x)$ admits representation in finitely many orthornormal basis $\phi_\xi(x)$,

$$f := \sum_{\xi=0}^{\infty} a_\xi\phi_\xi, \quad \{\phi_\xi\} \text{ orthonormal basis in } L^2(\mathbb{R}^d). \tag{A.20}$$

By orthogonality of L_2 basis ϕ_ξ, f in Sobolev norm can be written

$$||f||^2_{H^\alpha(\mathbb{R}^d)} \lesssim \sum_{\xi=0}^{\infty}(1+|\xi|^\alpha)^2 a_\xi^2. \tag{A.21}$$

Moreover, we need Theorem A.2 to establish space-time extension lemma so that we can extend $u \in H_t^1([0,T]; H_x^1(\Omega))$ to $ext\, u \in H_t^1(\mathbb{R}; H_x^1(\mathbb{R}^d))$. Similar ideas that leverage approximation analysis on larger or full space-time domain and then restrict analytical results to space-time cylinder $[0,T] \times \Omega$ are common in PDEs and numerical analysis literature. see Dong et al. (2025); Dong & Liu (2023) (Amann, 2019, Section 1.3)

Lemma A.3 (Space–time extension)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $T > 0$, and smoothness parameters $\alpha_1, \alpha_2 \geqslant 0$. There exists a bounded linear operator*

$$Ext : H_t^{\alpha_2}\big(0,T; H_x^{\alpha_1}(\Omega)\big) \longrightarrow H_t^{\alpha_2}\big(\mathbb{R}; H_x^{\alpha_1}(\mathbb{R}^d)\big) \tag{A.22}$$

such that for every $u \in H_t^1(0,T; H_x^1(\Omega))$:

(i) $Ext\, u(t,x) = u(t,x)$ for a.e. $(t,x) \in (0,T) \times \Omega$,

(ii) There exists a constant $C = C(\Omega, T) > 0$ with

$$||Ext\ u||_{H_t^{\alpha_2}(\mathbb{R}; H_x^{\alpha_1}(\mathbb{R}^d))} \leq C \, \|u\|_{H_t^{\alpha_2}(0,T; H_x^{\alpha_1}(\Omega))}. \tag{A.23}$$

Proof of Lemma A.3. We first extend spatial domain Ω to $\mathbb{R}^d$. Since Ω is a bounded Lipschitz domain, the extension Theorem A.2 provides a bounded linear operator

$$E_x : H_x^{\alpha_1}(\Omega) \longrightarrow H_x^{\alpha_1}(\mathbb{R}^d) \tag{A.24}$$

such that for any $t \in [0,T]$, (i) and (ii) of Theorem A.2 hold for all $u(\cdot, t) \in H^{\alpha_1}(\Omega)$.

For $u \in H_t^{\alpha_2}(0,T; H_x^{\alpha_1}(\Omega))$, we can define

$$(\hat{E}_x u)(\cdot, t) := E_x\big(u(\cdot, t)\big), \quad \text{for a.e. } t \in (0,T). \tag{A.25}$$

Since E_x is bounded operator along with Fubini's theorem, we obtain for every temporal derivative index $\alpha_2 \geqslant 0$

$$\int_0^T \|\partial_t^{\alpha_2}(\hat{E}_x u)(t)\|^2_{H^{\alpha_1}(\mathbb{R}^d)}\, dt \leq C_\Omega^2 \int_0^T \|\partial_t^{\alpha_2} u(t)\|^2_{H^{\alpha_1}(\Omega)}\, dt. \tag{A.26}$$

This implies $\hat{E}_x u \in H^{\alpha_2}(0,T; H^{\alpha_1}(\mathbb{R}^d))$. Thus, the operator $\hat{E}_x : H_t^{\alpha_2}(0,T; H_x^{\alpha_1}(\Omega)) \to H_t^{\alpha_2}(0,T; H_x^{\alpha_1}(\mathbb{R}^d))$ is bounded with

$$\|\hat{E}_x u\|_{H_t^{\alpha_2}(0,T; H_x^{\alpha_1}(\mathbb{R}^d))} \leq C_\Omega \, \|u\|_{H_t^{\alpha_2}(0,T; H_x^{\alpha_1}(\Omega))}. \tag{A.27}$$

Moreover, $(\widetilde{E}_x u)(t,x) = u(t,x)$, for a.e. $(t,x) \in [0,T] \times \Omega$.

Next we apply Sobolev extension theorem again to the interval $(0,T) \subset \mathbb{R}$, with a bounded linear $E_t : H^{\alpha_2}(0,T) \to H^{\alpha_2}(\mathbb{R})$ satisfying (i) and (ii) of Theorem A.2 as well.

Let X be a separable Hilbert space and let $\{e_k\}_{k \geq 1}$ be an orthonormal basis of X. For $u \in H^{\alpha_2}(0,T; X)$, define scalar functions $u_k(t) := \langle u(t), e_k \rangle_X \in H^{\alpha_2}([0,T])$.

Define $\hat{E}_t u : \mathbb{R} \to X$ by

$$(\hat{E}_t u)(t) := \sum_{k=1}^{\infty} (E_t u_k)(t)\, e_k, \tag{A.28}$$

and similarly, for every temporal derivative index $\alpha_2 \geqslant 0$, we obtain

$$\|\hat{E}_t \partial_t^{\alpha_2} u\|^2_{L^2(\mathbb{R}; X)} = \sum_{k=1}^{\infty} \|E_t \partial_t^{\alpha_2} u_k\|^2_{L^2(\mathbb{R})} \leq C_T^2 \sum_{k=1}^{\infty} \|\partial_t^{\alpha_2} u_k\|^2_{L^2(0,T)} = C_T^2 \|\partial_t^{\alpha_2} u\|^2_{L^2(0,T;X)}. \tag{A.29}$$

Moreover, for a.e. $t \in [0,T]$, we have

$$(\hat{E}_t u)(t) = \sum_{k=1}^{\infty} (E_t u_k)(t)\, e_k = \sum_{k=1}^{\infty} u_k(t)\, e_k = u(t), \tag{A.30}$$

When $X = H_x^1(\mathbb{R}^d)$, $\hat{E}_t : H_t^{\alpha_2}\big(0,T;H_x^{\alpha_1}(\mathbb{R}^d)\big) \to H_t^{\alpha_2}\big(\mathbb{R};H_x^{\alpha_1}(\mathbb{R}^d)\big)$ is bounded and extends functions in time. Finally, we define

$$Ext := \hat{E}_t \circ \hat{E}_x. \tag{A.31}$$

For $u \in H_t^1(0,T;H_x^1(\Omega))$, we have

$$Ext\ u = \hat{E}_t\big(\hat{E}_x u\big) \in H_t^{\alpha_2}\big(\mathbb{R};H_x^{\alpha_1}(\mathbb{R}^d)\big), \tag{A.32}$$

and for a.e. $(t,x) \in [0,T]\times\Omega$,

$$Ext\, u(t,x) = \hat{E}_t(\hat{E}_x u)(t,x) = (\hat{E}_x u)(t,x) = u(t,x), \tag{A.33}$$

so (i) holds. Moreover,

$$\|Ext\, u\|_{H_t^{\alpha_2}(\mathbb{R};H_x^{\alpha_1}(\mathbb{R}^d))} \le C_T \|\hat{E}_x u\|_{H_t^{\alpha_2}(0,T;H_x^{\alpha_1}(\mathbb{R}^d))} \le C_T C_\Omega \|u\|_{H_t^{\alpha_2}(0,T;H_x^{\alpha_1}(\Omega))}, \tag{A.34}$$

which yields (ii) with $C(\Omega,T) := C_T \cdot C_\Omega$. ■

A.3.3 Inner Product in $H_t^1 H_x^1$

The Hilbert space structure of $H_t^1 H_x^1 := H^1(0,T;H^1(\Omega))$ plays a crucial role in our analysis. We use the norm

$$\|u\|_{H_t^1 H_x^1} := \left(\int_0^T \|u(\cdot,t)\|_{H^1(\Omega)}^2 + \left\|\frac{\partial u}{\partial t}(\cdot,t)\right\|_{H^1(\Omega)}^2 dt\right)^{1/2}. \tag{A.35}$$

The associated inner product is

$$\langle u,v\rangle_{H_t^1 H_x^1} := \int_0^T \int_\Omega uv + \nabla u\cdot\nabla v + \partial_t u\, \partial_t v + \nabla(\partial_t u)\cdot\nabla(\partial_t v)\, dx\, dt. \tag{A.36}$$

We also define the $L^2(\Omega_T)$ inner product as

$$\langle u,v\rangle_{L^2(\Omega_T)} := \int_0^T \int_\Omega uv\, dx\, dt, \tag{A.37}$$

so that the $H_t^1 H_x^1$ inner product can be rewritten as

$$\langle u,v\rangle_{H_t^1 H_x^1} = \langle u,v\rangle_{L^2(\Omega_T)} + \langle \nabla u, \nabla v\rangle_{L^2(\Omega_T)} + \langle \partial_t u, \partial_t v\rangle_{L^2(\Omega_T)} + \langle \nabla(\partial_t u), \nabla(\partial_t v)\rangle_{L^2(\Omega_T)}. \tag{A.38}$$

B Background for Linear Parabolic PDEs

In this section, we provide overview of linear parabolic PDEs theory.

B.1 Dirichlet Boundary Problem

Given $f \in L^2(\Omega_T)$ and $g \in L^2(\Omega)$, the Dirichlet boundary value problem for linear parabolic PDEs is defined as

$$\begin{cases} u_t + Lu = f, & in\ \Omega\times(0,T], \\ u = 0, & on\ \partial\Omega\times[0,T], \\ u = g, & on\ \Omega\times\{t=0\}, \end{cases} \tag{B.1}$$

where L is an elliptic operator

$$L := -\sum_{i,j=1}^n a^{ij}(x,t)u_{x_i x_j} + \sum_{i=1}^n b^i(x,t)u_{x_i} + c(x,t)u,\ \ a^{ij}, b^i, c \in L^\infty(\Omega_T). \tag{B.2}$$

Furthermore, the stronger *uniform ellipticity* condition requires that there exists some constant $\theta > 0$ such that

$$\sum_{i,j=1}^n a^{ij}(x,t)\xi_i\xi_j \geqslant \theta|\xi|^2 \text{ for all } (x,t)\in\Omega_T, \xi\in\mathbb{R}^d. \tag{B.3}$$

Because solutions to (B.1) may lack of classical smoothness, particularly when the forcing term f or initial condition g have singularities, it is natural to study solutions in the weak sense.

B.2 WEAK SOLUTION

We say u is a weak solution to (B.1) if

$$(i) \quad u \in L^2(0,T;H_0^1(\Omega)), \text{ with } u' \in L^2(0,T;H^{-1}(\Omega)), \tag{B.4}$$

$$(ii) \quad (u',v) + B[u,v;t] = (f,v), \text{ for all test function } v \in H_0^1(\Omega) \text{ and } a.e \text{ time } 0 \leqslant t \leqslant T, \tag{B.5}$$

$$(iii) \quad u(0) = g, \tag{B.6}$$

where $B[\cdot,\cdot]$ is the time-dependent bilinear form

$$B[u,v;t] := \int_\Omega \sum_{i,j=1}^n a^{ij}(\cdot,t)u_{x_i}u_{x_j} + \sum_{i=1}^n b^i(\cdot,t)u_{x_i}v + c(\cdot,t)uvdx \tag{B.7}$$

for any $u, v \in H_0^1(\Omega)$ and a.e $0 \leqslant t \leqslant T$.

Next we present a classical result on the regularity of solutions to linear parabolic PDEs. This result highlights the smoothing property of parabolic equations: the solution u gains one additional regularity in space compared to the initial data.

B.3 REGULARITY THEORY

Theorem B.1 (Higher Regularity Theory for Linear Parabolic PDEs)**.** . *Assume*

$$g \in H^{2m+1}(\Omega), \frac{d^p}{dt^p}f \in L^2(0,T;H^{2m-2p}(\Omega)) \quad (p = 0,...,m). \tag{B.8}$$

Suppose also that the following p-th order compatibility conditions hold:

$$\begin{cases} g_0 := g \in H_0^1(\Omega), \ g_1 := f(0,x) - Lg_0 \in H_0^1(\Omega), \\ \\ ,...., \ g_m := \dfrac{d^{p-1}}{dt^{p-1}}f(0,x) - Lg_{p-1} \in H_0^1(\Omega). \end{cases} \tag{B.9}$$

Then

$$\frac{d^p u}{dt^p} \in L^2(0,T;H^{2m+2-2p}) \text{ holds for solution } u \text{ to PDEs } (B.1). \tag{B.10}$$

Proof of Theorem B.1. The proof is standard and follows from Theorem 6 in Chapter 7.1 of (Evans, 2010). ■

Theorem B.2 (Infinite differentiability)**.** . *Assume $g \in C^\infty(\bar{\Omega}), f \in C^\infty(\bar{\Omega}_T)$ and compatibility condition $(B.9)$ hold for $m = 0,1,2,3,....$ Then the parabolic initial/boundary-value problem $(B.1)$ has a unique solution*

$$u \in C^\infty(\bar{\Omega}_T). \tag{B.11}$$

Proof of Theorem B.2. See Theorem 7 in Chapter 7 of (Evans, 2010). ■

We conclude this section by an important characterization of solution to linear parabolic PDEs, which is essential to construction of our lower bound and space-time trade-off in approximation error analysis.

B.4 DUHAMEL PRINCIPLE

Theorem B.3 (Duhamel's Principle for General Linear Parabolic PDE)**.** *Let $\Omega \subseteq \mathbb{R}^n$ be a domain with boundary $\partial\Omega$, and let L is an elliptic operator defined in (B.2)*

$$\begin{cases} u_t(x,t) - Lu(x,t) = f(x,t), & (x,t) \in \Omega \times (0,T], \\ u(x,t) = 0, & (x,t) \in \partial\Omega \times [0,T], \\ u(x,0) = 0, & x \in \Omega. \end{cases} \tag{B.12}$$

Then the solution to the inhomogeneous problem is given by the Duhamel formula

$$u(x,t) = \int_0^t (P^s f)(x,\, s)\, \mathrm{d}s = \int_0^t v(x,\, s)\, \mathrm{d}s. \tag{B.13}$$

where P^s is the solution operator of the homogeneous problem with $s > 0$

$$\begin{cases} v_t(x,t) - Lv(x,t) = 0, & (x,t) \in \Omega \times (s,T], \\ v(x,t) = 0, & (x,t) \in \partial\Omega \times [s,T], \\ v(x,s) = f(x,s), & x \in \Omega. \end{cases} \tag{B.14}$$

Proof of Theorem B.3. See (Umarov, 2012, Section 2.4) and (John, 1982, Chapter 5, Section 1). ■

An immediate corollary followed by the Duhamel principle is the solution to non-zero initial value problem with Dirichlet boundary. Duhamel Principle tells us that the construction of solution u to linear parabolic equation in our setting is enough to look at the homogeneous solution w generated by initial data $g(x)$ when $t = 0$ and v generated by forcing function $f(x,t)$ when $t > 0$.

Corollary B.4. *Assume the conditions in Theorem B.3 holds. Then the solution u to the PDE problem (B.1) is given by*

$$u(x,t) = g(x) + \int_0^t w(x,\, s)\, \mathrm{d}s \tag{B.15}$$

where $w(x,t)$ is the solution to homogeneous initial value problem with Dirichlet boundary condition

$$\begin{cases} w_t(x,t) - Lw(x,t) = 0, & (x,t) \in \Omega \times (0,T], \\ w(x,t) = 0, & (x,t) \in \partial\Omega \times [0,T], \\ w(x,0) = g(x). \end{cases} \tag{B.16}$$

Corollary B.5 (Scaling Invariant Property)**.** *Let $L = \triangle$ and $\alpha > 0$. u satisfies the scaling invariant property such that $u(\frac{x}{\alpha}, \frac{t}{\alpha^2})$ is a solution to the PDE problem (B.1) if and only if $u(x,t)$ is a solution to the PDE problem (B.1).*

Proof of Corollary B.5. This proof follows directly from the fact that $w(\frac{x}{\alpha}, \frac{t}{\alpha^2})$ and $v(\frac{x}{\alpha}, \frac{t}{\alpha^2})$ solves $u_t - \triangle u = 0$ if and only if $w(x,t)$ and $v(x,t)$ solves $u_t - \triangle u = 0$. Therefore, it is immediately followed by direct computation that $u(x,t)$ also enjoys this scaling invariance property. ■

C DERIVATION FOR OUR LOSS FUNCTION

Our loss function is motivated by solving linear parabolic PDEs using its corresponding convex optimization formulation. Let $f \in \mathcal{F}$, $g \in \mathcal{G}$ be right-hand side and boundary value of the PDE and define elliptic operator [see Chapter 6 in Evans (2010)]

$$L := -\sum_{i,j=1}^{n} a^{ij}(x,t) u_{x_i x_j} + \sum_{i=1}^{n} b^i(x,t) u_{x_i} + c(x,t)u. \tag{C.1}$$

Furthermore, there exists some constant $\theta > 0$ such that

$$\sum_{i,j=1}^{n} a^{ij}(x,t)\xi_i\xi_j \geqslant \theta|\xi|^2 \text{ for all } (x,t) \in \Omega_T, \xi \in \mathbb{R}^d. \tag{C.2}$$

Solving the boundary valued problem of linear parabolic PDEs is defined as finding $u^* \in \mathcal{U}$ satisfying the following constraint

$$\begin{cases} u_t^* + Lu^* = f & in\ \Omega \times (0,T], \\ u^*(x,t) = 0 & on\ \partial\Omega \times (0,T], \\ u^*(x,0) = g & on\ \Omega. \end{cases} \tag{C.3}$$

Note $\mathcal{F}$ and $\mathcal{G}$ and $\mathcal{U}$ are three distinctive functional spaces where differentiability and integrability could be totally different. Its relaxation is given as the equation below, and we leave the detailed discussion to Section C.1 and C.2

$$-\epsilon u_{tt}^* + u_t^* + Lu^* = f \quad in\ \Omega \times (0,T]. \tag{C.4}$$

The solution to the relaxed problem enjoys the following uniform in time approximation property. (Mielke & Stefanelli, 2011, Theorem 1) (Lions & Magenes, 1972, Section 7, Theorem 7.2)

$$\text{As } \epsilon_n \to 0,\ u_{\epsilon_n}^* \to u^*\ in\ L^2(\Omega)\ , \forall\, t > 0. \tag{C.5}$$

This approximation theory allows us to solve linear parabolic PDEs in its relaxed version, and surprisingly the variational structure underlying the relaxed problem can be formulated as convex optimization problem for some strongly convex energy functional $I_\epsilon^{\text{parabolic}}(u)$ in some funtional space $\mathcal{U}_{parabolic}$

$$\min_{u \in \mathcal{U}_{parabolic}} I_\epsilon^{\text{parabolic}}(u). \tag{C.6}$$

However, $I_\epsilon^{parabolic}(u)$, directly minimizing $I_\epsilon^{parabolic}$ is not applicable for serious drawbacks. First, $I_\epsilon^{parabolic}(u)$, though strongly convex, is degenerate when ϵ approaches to 0, leading to vanishing lower curvature bound in derived stability estimate. Therefore, we introduce additional data-fitting function $D(u, u^*)$ over Sobolev space to improve lower curvature bound in derived stability estimate.

C.1 Euler-Lagrange equation for the weighted energy functional (2.1)

In this section, we derive the Euler–Lagrange equations associated with the weighted energy functional $I_\epsilon(u)$. As shown in Theorem C.1, minimizing $I_\epsilon(u)$ is equivalent to solving initial value problem of a regularized heat equation with zero Dirichlet condition in equation . The well-posedness ensures that the minimizer of $I_\epsilon(u)$ is unique. Building upon this, we also show that the minimizer of the loss functional $L_{\epsilon,\beta}(u)$ is the solution to the (exact) heat equation (1.1), as shown in Theorem 2.1.

We incorporate regularity theory from linear parabolic PDEs in accordance with hyperbolic PDEs. Assume that forcing function $f \in \mathcal{F}$ and initial data $u_0^* \in H^{2s+1}(\Omega) \cap H_0^1(\Omega)$ satisfy the compatibility condition in (B.9), we consider the minimizer u living in the search space $\mathcal{U}$

$$\mathcal{U} := \left\{ u(x,t) \in H_t^p(0,T; H_x^{2s+2-2p}(\Omega)) \cap L_t^2(0,T; H_0^1(\Omega)) \;\middle|\; u(x,0) = u_0^*(x) \right\}. \tag{C.7}$$

Moreover, we denote $u^* \in \mathcal{U}$ as ground truth solution that solves initial value problem for heat equation with zero Dirichlet boundary in (1.1). Our search functional space $\mathcal{U}$, forcing function f, and initial data u_0^* satisfy functional setting where Euler-Lagrange equation has unique minimizer. In literature, it is often required though with some variation that $f \in L^2(0,T; L^2(\Omega))$ and $u_0^* \in H_0^1(\Omega)$ and $u \in L^2(0,T; H_0^1(\Omega))$. See Mielke & Stefanelli (2011); Mielke & Ortiz (2008); Stefanelli (2025).

C.2 Euler-Lagrange Equation for the weighted energy functional $I_\epsilon(u)$

The following Theorem C.1 gives the first-order optimality condition for the weighted energy functional $I_\epsilon(u)$ in (2.1).

Theorem C.1 (Minimizer of the WEF $I_\epsilon(u)$)**.** *There exists a unique minimizer $u \in \mathcal{U}$ of $I_\epsilon(u)$ such that u is a solution to linear hyperbolic PDEs*

$$\begin{cases} -\epsilon u_{tt} + u_t - \triangle u = f, \quad (x,t) \in \Omega \times (0,T], \\ u(x,t) = 0, \quad (x,t) \in \partial\Omega \times [0,T], \\ u(x,0) = u_0^*(x), \quad x \in \Omega. \end{cases} \tag{C.8}$$

Remark C.2. In (Lions & Magenes, 1972, Section 7.2), equation (C.8) is called elliptic regularization of heat equation. It is shown there as well that regularized solution u satisfies classical energy estimate for linear parabolic PDEs theory so that the regularized solution u_ϵ is unique. The general theory of solution to the above equations is discussed in (Temam, 1997, Chapter 2).

Proof of Theorem C.1. Recall the definition of the weighted energy functional (WEF) :

$$I_\epsilon(u) = \int_0^T e^{-\frac{t}{\epsilon}} \int_\Omega \left[\frac{1}{2}|u_t|^2 + \frac{1}{2\epsilon}|\nabla u|^2 - \frac{1}{\epsilon} f u\right] dxdt. \tag{C.9}$$

Note that the Euler-Lagrangian equation depends on the definition of test function. Here we choose test function $v \in C_c^\infty([0,T]\times\Omega)$. For any test function $v \in C_c^\infty([0,T]\times\Omega)$, we know $v(\cdot,0) = v(\cdot,T) = 0$. Then,

$$I_\epsilon(u+sv) = \int_0^T e^{-\frac{t}{\epsilon}} \int_\Omega \frac{1}{2}\big|u_t + sv_t\big|^2 + \frac{1}{2\epsilon}\big|\nabla u + s\nabla v\big|^2 - \frac{1}{\epsilon} f\big(u+sv\big)\, dxdt. \tag{C.10}$$

We apply Fubini theorem and integration by parts with respect to time t:

$$\frac{d}{ds} I_\epsilon(u+sv)\Big|_{s=0} = \int_0^T e^{-\frac{t}{\epsilon}} \int_\Omega u_t v_t + uv - \frac{1}{\epsilon}\triangle uv - \frac{1}{\epsilon} fv\, dxdt \tag{C.11}$$

$$= \int_0^T e^{-\frac{t}{\epsilon}} \int_\Omega \frac{1}{\epsilon} u_t v - \frac{1}{\epsilon}\triangle uv - \frac{1}{\epsilon} fv - vu_{tt} dxdt \tag{C.12}$$

$$- e^{-\frac{T}{\epsilon}} \int_\Omega f(x,T)v(x,T)dx + u_t(x,0)v(x,0)dx. \tag{C.13}$$

Making appropriate substitutions and matching terms from the above, we obtain

$$\frac{d}{ds} I_\epsilon(u+sv)\Big|_{s=0} = e^{-\frac{T}{\epsilon}} \int_\Omega \big(-\triangle u(x,T) + u_t(x,T) - f(x,T)\big)v(x,T)dx \tag{C.14}$$

$$+ \int_0^T \int_\Omega e^{-\frac{t}{\epsilon}}\big(-u_{tt} + \frac{1}{\epsilon}u_t - \frac{1}{\epsilon}\triangle u - \frac{1}{\epsilon} f\big)v dxdt \tag{C.15}$$

$$+ \int_\Omega -u_t(x,0)v(x,0)dx \tag{C.16}$$

$$= \int_0^T \int_\Omega e^{-\frac{t}{\epsilon}}\big(-u_{tt} + \frac{1}{\epsilon}u_t - \frac{1}{\epsilon}\triangle u - \frac{1}{\epsilon} f\big)v dxdt. \tag{C.17}$$

We set $\frac{d}{ds} I_\epsilon(u+sv) = 0$ and obtain the corresponding Euler-Lagrangian equations for the functional weighted energy, and the boundary condition is satisfied because $u \in L_t^2 H_{0,x}^1$.

$$\begin{cases} -\epsilon u_{tt} + u_t - \triangle u = f, \quad (x,t) \in \Omega\times(0,T], \\ u(x,t) = 0, \quad (x,t) \in \partial\Omega \times [0,T], \\ u(x,0) = u_0^*(x), \quad x \in \Omega. \end{cases} \tag{C.18}$$

■

C.3 Minimizer of the Total Loss Function $L_{\epsilon,\beta}(u)$

Next, we characterize the optimality condition for the total loss function $L_{\epsilon,\beta}(u)$ in (2.4). Recall that Theorem 2.1 states any minimizer $u \in \mathcal{U}$ the total loss function $L_{\epsilon,\beta}(u)$ is the solution to heat equation in (1.1). Below we provide the proof.

Proof of Theorem 2.1. Recall that

$$L_{\epsilon,\beta}(u) := ||u-u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} + \beta\left[I_\epsilon(u) + \int_0^T \int_\Omega e^{-\frac{t}{\epsilon}} u_{tt}^* u dxdt - \int_\Omega e^{-\frac{T}{\epsilon}} u_t^*(x,T)u(x,T)dx\right]. \tag{C.19}$$

According to the inner product in $H_t^1(0,T;H_x^1(\Omega))$, we readily compute first variation of loss function $L_{\epsilon,\beta}(u)$. For any test function $v \in C_c^\infty(\Omega \times [0,T])$,

$$\frac{d}{ds}L_{\epsilon,\beta}(u+sv)\Big|_{s=0} = 2\langle u-u^*, v\rangle_{H_t^1 H_x^1} + \beta\int_0^T\int_\Omega e^{-\frac{t}{\epsilon}}\big(-u_{tt}+\frac{1}{\epsilon}u_t-\frac{1}{\epsilon}\triangle u-\frac{1}{\epsilon}f\big)v dxdt \tag{C.20}$$

$$+\beta\int_0^T\int_\Omega e^{-\frac{t}{\epsilon}}u_{tt}^* v dxdt - \beta\int_\Omega u_t(x,0)v(x,0)dx - \beta\int_\Omega e^{-\frac{T}{\epsilon}}u_t^*(x,T)u(x,T)dx \tag{C.21}$$

$$= 2\langle u-u^*, v\rangle_{H_t^1 H_x^1} + \beta\int_0^T\int_\Omega e^{-\frac{t}{\epsilon}}\big(u_{tt}^*-u_{tt}+\frac{1}{\epsilon}u_t-\frac{1}{\epsilon}\triangle u-\frac{1}{\epsilon}f\big)v dxdt \tag{C.22}$$

$$= 2\int_0^T\int_\Omega (u-u^*)v dxdt + 2\int_0^T\int_\Omega(\nabla u-\nabla u^*)\nabla v dxdt \tag{C.23}$$

$$+2\int_0^T\int_\Omega (u_t-u_t^*)v_t dxdt + 2\int_0^T\int_\Omega(\nabla u_t-\nabla u_t^*)\nabla v_t dxdt \tag{C.24}$$

$$+\beta\int_0^T\int_\Omega e^{-\frac{t}{\epsilon}}\big(u_{tt}^*-u_{tt}+\frac{1}{\epsilon}u_t-\frac{1}{\epsilon}\triangle u-\frac{1}{\epsilon}f\big)v dxdt \tag{C.25}$$

$$= \int_0^T\int_\Omega\Bigg[2\bigg(u-u^*-\triangle u+\triangle u^*-u_{tt}+u_{tt}^*+\triangle u_{tt}-\triangle u_{tt}^*\bigg) \tag{C.26}$$

$$+\frac{\beta}{\epsilon}e^{-\frac{t}{\epsilon}}\bigg(\epsilon(u_{tt}^*-u_{tt})+u_t-\triangle u-f\bigg)\Bigg]v dxdt. \tag{C.27}$$

Second equality follows by $v(x,0)=0$ and $v(x,T)=0$; third equality follows by inner product in $H_t^1 H_x^1$;last equality follows by integration by parts and vanishing boundary of v. Setting $\frac{d}{ds}L_{\epsilon,\beta}(u+sv)=0$, we obtain

$$2\bigg[u-u^*-\triangle u+\triangle u^*-u_{tt}+u_{tt}^*+\triangle u_{tt}-\triangle u_{tt}^*\bigg]+\frac{\beta}{\epsilon}e^{-\frac{t}{\epsilon}}\bigg(\epsilon(u_{tt}^*-u_{tt})+u_t-\triangle u-f\bigg)=0. \tag{C.28}$$

Let $w := u-u^*$, we have (x,t)-a.s in Ω_T that

$$2\bigg[w-\triangle w-w_{tt}+\triangle w_{tt}\bigg]+\frac{\beta}{\epsilon}e^{-\frac{t}{\epsilon}}(-\epsilon w_{tt}+w_t-\triangle w+u_t^*-\triangle u^*-f)=0. \tag{C.29}$$

Since u^* is true solution to heat equation, $u_t^*-\triangle u^*-f=0$, and

$$\tfrac{\beta}{\epsilon}e^{-\frac{t}{\epsilon}}(\epsilon w_{tt}-w_t+\triangle w) = 2\bigg[w-\triangle w-w_{tt}+\triangle u_{tt}\bigg]\text{, }(x,t)\text{-}a.s\text{ in }\Omega_T. \tag{C.30}$$

We denote inner product in $L_2(\Omega_T)$

$$\langle u,v\rangle_{L^2(\Omega_T)} := \int_0^T\int_\Omega uv\,dxdt. \tag{C.31}$$

Taking inner product with respect to w in $L_2(\Omega_T)$ on both sides and applying integration by parts, we have

$$\langle\tfrac{\beta}{\epsilon}e^{-\frac{t}{\epsilon}}(\epsilon w_{tt}-w_t+\triangle w),w\rangle_{L_2(\Omega_T)} = 2\langle w-\triangle w-w_{tt}+\triangle u_{tt},w\rangle_{L^2(\Omega_T)}. \tag{C.32}$$

Making some rearrangement, we obtain

$$0 \geqslant \Bigg[(2+\beta e^{-\frac{T}{\epsilon}})\langle w_t,w_t\rangle_{L_2(\Omega_T)}+(2+\tfrac{\beta}{\epsilon}e^{-\frac{T}{\epsilon}})\langle\nabla w,\nabla w\rangle_{L_2(\Omega_T)} \tag{C.33}$$

$$+2\langle w,w\rangle_{L_2(\Omega_T)}+2\langle\nabla w_t,\nabla w_t\rangle_{L_2(\Omega_T)}\Bigg]. \tag{C.34}$$

Since β, $\langle w_t, w_t\rangle_{L_2(\Omega_T)}$, $\langle \nabla w, \nabla w\rangle_{L_2(\Omega_T)}$, $\langle w, w\rangle_{L_2(\Omega_T)}$, and $\langle \nabla w_t, \nabla w_t\rangle_{L_2(\Omega_T)}$ are nonnegative, we have

$$w_t = 0, \quad \nabla w = 0, \quad \nabla w_t = 0, \quad w = 0, \quad (x,t)\, a.s \text{ in } \Omega_T. \tag{C.35}$$

Since $w = u^* - u$, we have

$$u^* = u \ , (x,t)\text{-}a.s \text{ in } \Omega_T. \tag{C.36}$$

Therefore, the minimizer to the total loss function $L_{\epsilon,\beta}(u)$ is the solution to heat equation. ■

C.4 Empirical Loss Function

Our estimator $\hat{u}_\theta$ is defined as $\hat{u}_\theta := \underset{u\in\mathcal{U}_J}{\arg\min}\ \hat{L}_{\epsilon,\beta}(u)$, where $\hat{L}_{\epsilon,\beta}(u)$ is empirical counterpart of $L_{\epsilon,\beta}(u)$. In particular, $\hat{L}_{\epsilon,\beta}(u)$ is defined as

$$\hat{L}_{\epsilon,\beta}(u) := \frac{1}{nm}\hat{D}(u,u^*) + \lambda\Big(\hat{I}_\epsilon(u) + \hat{M}_\epsilon(u,u^*)\Big) \tag{C.37}$$

where

$$\hat{D}(u,u^*) := |\Omega_T| \sum_{i=1}^{n} \Bigg(|u(X_{im},t_m) - u^*(X_{im},t_m)|^2 + |\partial_t u(X_{im},t_m) - \partial_t u^*(X_{im},t_m)|^2 \tag{C.38}$$

$$+ |\nabla u(X_{im},t_m) - \nabla u^*(X_{im},t_m)|^2 + |\nabla\partial_t u(X_{im},t_m) - \nabla\partial_t u^*(X_{im},t_m)|^2 \Bigg), \tag{C.39}$$

$$\hat{I}_\epsilon(u) := |\Omega_T| \sum_{j=1}^{m} e^{-\frac{t_j}{\epsilon}} \sum_{i=1}^{n} \left(\frac{1}{2}|\partial_t u|^2(x_{ij},t_j) + \frac{1}{2\epsilon}|\nabla u(x_{ij},t_j)|^2 - \frac{1}{\epsilon} f(x_{ij},t_j)u(x_{ij},t_j) \right) \tag{C.40}$$

$$+ m|\Omega| e^{-\frac{t_m}{\epsilon}} \sum_{i=1}^{n} \left(\frac{1}{2}|\nabla u(x_{im},t_m)|^2 - f(x_{im},t_m)u(x_{im},t_m) \right), \tag{C.41}$$

$$\hat{M}_\epsilon(u^*) := |\Omega_T| \sum_{j=1}^{m} e^{-\frac{t_j}{\epsilon}} \sum_{i=1}^{n} \partial_{tt} u^*(X_{ij},t_j) u(X_{ij},t_j) - m|\Omega| e^{-\frac{t_m}{\epsilon}} \sum_{i=1}^{n} u(x_{im},t_m) u_t^*(x_{im},T_m). \tag{C.42}$$

We are given with the training data on space-time point clouds on which ground truth solution u^* and forcing function f are defined , namely $S := \{(X_{ij}, t_j, u^*_{ij}, \nabla u^*_{ij}, \partial_t u^*_{ij}, \nabla\partial_t u^*_{ij}, f_{ij})\}_{i=1,j=1}^{n,m}$, where derivative labels $\nabla u^*_{ij}, \partial_t u^*_{ij}, \nabla\partial_t u^*_{ij}$ are obtained through numerical computation for $H_t^1 H_x^1$ evaluation norm.

D Truncated Local Basis in Space and Time

Among various representations for parametrizing $\hat{u}_\theta$, we choose $\hat{u}_\theta := \hat{u}_J$, the projection over a (J,J)-truncated space $\mathcal{U}_J$ spanned by spatio-temporal local basis due to its approximation power in functional space $\mathcal{U}$ and more importantly its analytical simplicity. One might use other estimators, such as neural network RHKS, as long as the estimator's analytical approximation results are well-established.

$$\mathcal{U}_J := \left\{ \hat{u}_J(x,t) \,\middle|\, \hat{u}_J(x,t) = \sum_{|\xi|_\infty \leqslant J, \tau \leqslant J} a_{\xi,\tau}\phi_{\xi,\tau}(x,t), \hat{u}_J \in \mathcal{U} \right\}. \tag{D.1}$$

The essential idea of elliptic regularization is to view time as an additional dimension of space with extra penalty scheduler term $e^{-\frac{t}{\epsilon}}$. In this viewpoint, we can write local basis as direct product of two local basises $\phi_\xi(x)$ and $\phi_\tau(t)$ in space and time respectively

Our proof needs $\phi_\tau(t)$ and $\phi_\xi(x)$ to satisfy the following propertieis:

$$(1) \quad \phi_\xi(x) \text{ and } \phi_\tau(t) \text{ are orthonormal basis in } L^2(\mathbb{R}^d) \text{ and } L^2(\mathbb{R}) \text{ respectively,} \tag{D.2}$$

$$(2) \quad \phi_\xi(x) \text{ and } \phi_\tau(t) \text{ have compact support in } \mathbb{R}^d \text{ and in } \mathbb{R} \text{ respectively .} \tag{D.3}$$

There are only few family of local bases satisfying (1)-(2) while demonstrating nice approximation property in Sobolev space. Local basis representation like Legendre polynomial and Fourier representation will fail locally compact support property. For this representation, we construct wavelet bases using Dauchecies wavelets to establish approximation in Besov spaces whose special case is Sobolev space, following the definitions given in Triebel (2006). In this paper, the reader can assume our local bases are wavelets, and the omission of thorough construction of such bases, however, will not jeopardize our argument as long as (1) and (2) are applied. For example, we can construct wavelet bases using Dauchecies wavelets to establish approximation in Besov spaces whose special case is Sobolev space, following the definitions given in Triebel (2006)[Section 3], and we refer the reader to this reference for further details on wavelets. Denote by $\psi_{\mathfrak{m}} \in C^r(\mathbb{R})$ and $\psi_{\mathfrak{F}} \in C^r(\mathbb{R})$ a wavelet on compact support and scaling function, respectively. This implies that

$$\psi_k^j = \begin{cases} \psi_{\mathfrak{F}}(x-k), & j=0,\, k \in \mathbb{Z}, \\ 2^{(j-1)/2}\psi_{\mathfrak{m}}(2^{j-1}x-k), & j \in \mathbb{N},\, k \in \mathbb{Z}. \end{cases} \tag{D.4}$$

is an orthonormal basis of $L^2(\mathbb{R})$.

To obtain a basis of $L^2(\mathbb{R}^d)$, for $j \in \mathbb{N}$, set

$$G^j = \{\mathfrak{F}, \mathfrak{m}\}^d \setminus \{(\mathfrak{F}, \ldots, \mathfrak{F})\}, \quad G^0 = \{(\mathfrak{F}, \ldots, \mathfrak{F})\}, \tag{D.5}$$

and for $g \in G^j \cup G^0$,

$$\Psi_k^g(x) = \prod_{i=1}^d \psi_{g_i}(x_i - k_i), \quad k \in \mathbb{Z}^d. \tag{D.6}$$

This gives the orthonormal basis

$$\Psi_k^{j,g} = \begin{cases} \Psi_k^g(x), & j=0,\, g \in G^0,\, k \in \mathbb{Z}^d, \\ 2^{(j-1)d/2}\Psi_k^g(2^{j-1}x), & j \in \mathbb{N},\, g \in G^j,\, k \in \mathbb{Z}^d. \end{cases} \tag{D.7}$$

Wavelet coefficients, defined as the expansion coefficients with respect to the above basis for $L^2(\mathbb{R}^d)$ functions, can be used to characterize the *Besov spaces*, whose special cases are Sobolev spaces.

Let $1 \le p, q \le \infty$, $\alpha \ge 0$, and let the regularity of the above wavelets satisfy

$$r > \alpha \vee \left(\frac{2d}{p} + \frac{d}{2} - s\right). \tag{D.8}$$

If $f(x)$ has wavelet representation, we have

$$f = \sum_{j=0}^{\infty} \sum_{g \in G^j} \sum_{k \in \mathbb{Z}^d} \gamma_k^{j,g} \Psi_k^{j,g}, \tag{B.1}$$

we define the Sobolev norm as special case of Besov norm $H^s(\Omega)$ by

$$\|f\|_{H^\alpha(\mathbb{R}^d)} = \|f\|_{B^\alpha_{2,2}} := \|\gamma\|_{B^\alpha_{2,2}} := \left[\sum_{j=0}^{\infty} 2^{2j\alpha} \sum_{g \in G^j} \sum_{k \in \mathbb{Z}^d} |\gamma_k^{j,g}|^2\right]^{1/2}. \tag{D.9}$$

In particular, by the orthonormality of the wavelets $\{\Psi_k^{j,g}\}_{j,g,k}$,

$$\|f\|_{B^0_{2,2}} = \|\gamma\|_{\ell^2} = \|f\|_{L^2(\mathbb{R}^d)}. \tag{D.10}$$

With this construction, we can replace $\phi_\tau(x)$ with wavelet bases $\Psi_k^{j,g}$, so ξ is indexed by both wavelet level and group G. When we treat $\xi = 2^j$ and $|a_\xi|^2 = \sum_{g \in G^j} \sum_{k \in \mathbb{Z}^d} |\gamma_k^{j,g}|^2$, one can show the

above Sobolev norm in terms of wavelet expansion is as same as standard Sobolev norm and moreover we can achieve the same estimate

$$\|f\|^2_{H^\alpha(\mathbb{R}^d)} := \sum_{j=0}^{\infty} 2^{2j\alpha} \sum_{g\in G^j} \sum_{k\in\mathbb{Z}^d} |\gamma_k^{j,g}|^2 = \sum_{|\xi|=2^j, j\geqslant 0} |\xi|^{2\alpha} a_\xi^2 \leqslant \sum_{\xi\geqslant 0} (1+|\xi|^\alpha)^2 a_\xi^2. \tag{D.11}$$

Following time-dependent definition of Sobolev space (see Chapter 5.10 of Evans (2010)), one can proceed to apply similar construction for basis $\phi_\tau(x)$ on time dimension. Since we treat time as an additional dimension of space, we obtain

$$||f||^2_{H^{\alpha'}(\mathbb{R};H^\alpha(\mathbb{R}^d))} \lesssim \sum_{\xi=0}^{\infty} (1+|\xi|^\alpha)^2 (1+\tau^{\alpha'})^2 a^2_{\xi,\tau}, \tag{D.12}$$

where $a_{\xi,\tau} = a_\xi a_\tau$.

E Proof of Main Results

In this Appendix, we provide the proofs of our main theoretical results in the paper.

E.1 Stability of the total loss function $L_{\epsilon,\beta}(u)$ in (2.4)

Proof of Proposition 2.2. Recall

$$L_{\epsilon,\beta}(u) := ||u-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} + \beta\bigg(I_\epsilon(u) + \int_0^T\int_\Omega e^{-\frac{t}{\epsilon}} u^*_{tt} u dxdt - \int_\Omega e^{-\frac{T}{\epsilon}} u^*_t(x,T)u(x,T)dxdt\bigg). \tag{E.1}$$

We can write the excess risk as following

$$L_{\epsilon,\beta}(u) - L_{\epsilon,\beta}(u^*) = ||u-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} + \beta\bigg(I_\epsilon(u) - I_\epsilon(u^*) + \int_0^T\int_\Omega e^{-\frac{t}{\epsilon}} u^*_{tt}(u^*-u)dxdt \tag{E.2}$$

$$+ \int_\Omega e^{-\frac{T}{\epsilon}} u^*_t(x,T)(u^*(x,T)-u(x,T))dxdt\bigg). \tag{E.3}$$

Step 1: Compute $I_\epsilon(u) - I_\epsilon(u^*) + \int_0^T\int_\Omega e^{-\frac{t}{\epsilon}} u^*_{tt}(u^*-u)dxdt$.

Recall $f(x,t) = u^*_t(x,t) - \triangle u^*(x,t)$, $(x,t)\in\Omega_T$. Apply integration by parts, we get

$$I_\epsilon(u) - I_\epsilon(u^*) = \int_0^T e^{-\frac{t}{\epsilon}} \int_\Omega \frac{1}{2}|u_t|^2 + \frac{1}{2\epsilon}|\nabla u|^2 - \frac{1}{\epsilon} f u dxdt \tag{E.4}$$

$$- \int_0^T e^{-\frac{t}{\epsilon}} \int_\Omega \frac{1}{2}|u^*_t|^2 + \frac{1}{2\epsilon}|\nabla u^*|^2 - \frac{1}{\epsilon} f u^* dxdt \tag{E.5}$$

$$= \int_0^T e^{-\frac{t}{\epsilon}} \int_\Omega \frac{1}{2}|u_t|^2 + \frac{1}{2\epsilon}|\nabla u|^2 - \frac{1}{\epsilon}(u^*_t - \triangle u^*) u dxdt \tag{E.6}$$

$$+ \int_0^T e^{-\frac{t}{\epsilon}} \int_\Omega -\frac{1}{2}|u^*_t|^2 - \frac{1}{2\epsilon}|\nabla u^*|^2 + \frac{1}{\epsilon}(u^*_t - \triangle u^*) u^* dxdt \tag{E.7}$$

$$= \int_0^T e^{-\frac{t}{\epsilon}} \int_\Omega \frac{1}{2}|u_t|^2 + \frac{1}{2\epsilon}|\nabla u|^2 - \frac{1}{\epsilon} u^*_t u - \frac{1}{\epsilon}\nabla u^* \nabla u dxdt \tag{E.8}$$

$$+ \int_0^T e^{-\frac{t}{\epsilon}} \int_\Omega -\frac{1}{2}|u^*_t|^2 - \frac{1}{2\epsilon}|\nabla u^*|^2 + \frac{1}{\epsilon} u^*_t u^* + \frac{1}{\epsilon}|\nabla u^*|^2 dxdt. \tag{E.9}$$

By a second integration by parts to $u^*_t u$ and $|u^*_t|^2$, we get

$$I_\epsilon(u) - I_\epsilon(u^*) = \int_0^T\int_\Omega e^{-\frac{t}{\epsilon}} \bigg(\frac{1}{2}|u_t|^2 - \frac{1}{2}|u^*_t|^2 + u^*_{tt}(u^*-u) + |u^*_t|^2 - u^*_t u_t\bigg) dxdt \tag{E.10}$$

$$+ \int_\Omega e^{-\frac{t}{\epsilon}} u^*_t \bigg(u-u^*\bigg) dx\bigg|_{t=0}^T + \int_0^T\int_\Omega \frac{1}{2\epsilon} e^{-\frac{t}{\epsilon}} |\nabla u - \nabla u^*|^2 dx. \tag{E.11}$$

By definition of $\mathcal{U}$, $u(x,0) = u^*(x,0)$, it follows that

$$I_\epsilon(u) - I_\epsilon(u^*) = \int_0^T \int_\Omega e^{-\frac{t}{\epsilon}} u^*_{tt}(u^* - u)dxdt + \int_\Omega e^{-\frac{T}{\epsilon}} u^*_t(x,T)\big(u(x,T) - u^*(x,T)\big)dx \tag{E.12}$$

$$+ \int_0^T e^{-\frac{t}{\epsilon}} \int_\Omega \frac{1}{2\epsilon}|\nabla u - \nabla u^*|^2 + \frac{1}{2}|u_t - u^*_t|^2 dxdt. \tag{E.13}$$

Then, we obtain

$$I_\epsilon(u) - I_\epsilon(u^*) + \int_0^T \int_\Omega e^{-\frac{t}{\epsilon}} u^*_{tt}(u - u^*)dxdt + \int_\Omega e^{-\frac{T}{\epsilon}} u^*_t(x,T)\big(u^*(x,T) - u(x,T)\big)dx \tag{E.14}$$

$$= \int_0^T \int_\Omega \frac{1}{2} e^{-\frac{t}{\epsilon}}|u_t - u^*_t|^2 dxdt + \int_0^T \int_\Omega \frac{1}{2\epsilon} e^{-\frac{t}{\epsilon}}|\nabla u - \nabla u^*|^2 dx. \tag{E.15}$$

Step 2: Bound $I_\epsilon(u) - I_\epsilon(u^*) + \int_0^T \int_\Omega e^{-\frac{t}{\epsilon}} u^*_{tt}(u^* - u)dxdt + \int_\Omega e^{-\frac{T}{\epsilon}} u^*_t(x,T)\big(u^*(x,T) - u(x,T)\big)dx$ from up and below.

To bound first integral in (E.15), we may assume without loss of generality that $u(x,t) - u^*(x,t)$ is centered in the following sense

$$\int_0^T e^{-\frac{t}{\epsilon}}\Big(u(x,t) - u^*(x,t)\Big)dt = 0. \tag{E.16}$$

According to weighted Poincaré inequality in Lemma F.9 with respect to the time variable, there exist constant $C(d,\epsilon,\Omega) > 0$ such that

$$\frac{1}{C^2(d,\epsilon,\Omega)} \int_0^T \int_\Omega e^{-\frac{t}{\epsilon}}|u - u^*|^2 dxdt \leqslant \int_0^T \int_\Omega e^{-\frac{t}{\epsilon}}|u_t - u^*_t|^2 dxdt. \tag{E.17}$$

Then it follows that

$$\int_0^T \int_\Omega \frac{1}{2} e^{-\frac{t}{\epsilon}}|u_t - u^*_t|^2 dxdt = \int_0^T \int_\Omega \frac{1}{4} e^{-\frac{t}{\epsilon}}|u_t - u^*_t|^2 dxdt + \int_0^T \int_\Omega \frac{1}{4} e^{-\frac{t}{\epsilon}}|u_t - u^*_t|^2 dxdt \tag{E.18}$$

$$\geqslant \frac{1}{C^2(d,\epsilon,\Omega)} \int_0^T \int_\Omega e^{-\frac{t}{\epsilon}}|u - u^*|^2 dxdt + \int_0^T \int_\Omega \frac{1}{4} e^{-\frac{t}{\epsilon}}|u_t - u^*_t|^2 dxdt \tag{E.19}$$

$$= \min\Big\{\frac{e^{-\frac{T}{\epsilon}}}{C^2(d,\epsilon,\Omega)}, \frac{e^{-\frac{T}{\epsilon}}}{4}\Big\} \int_0^T \int_\Omega \Big(|u - u^*|^2 + |u_t - u^*_t|^2\Big)dxdt \tag{E.20}$$

$$= \min\Big\{\frac{e^{-\frac{T}{\epsilon}}}{C^2(d,\epsilon,\Omega)}, \frac{e^{-\frac{T}{\epsilon}}}{4}\Big\}||u - u^*||^2_{H^1_t(0,T;L^2_x(\Omega))}. \tag{E.21}$$

We obtain

$$\int_0^T \int_\Omega \frac{1}{2} e^{-\frac{t}{\epsilon}}|u_t - u^*_t|^2 dxdt \leqslant \int_0^T \int_\Omega \frac{1}{2} e^{-\frac{t}{\epsilon}}\Big(|u_t - u^*_t|^2 + |u - u^*|^2\Big)dxdt \leqslant \frac{1}{2}||u - u^*||^2_{H^1_t(0,T;L^2_x(\Omega))}. \tag{E.22}$$

To bound the second integral in (E.15) , we again assume without loss of generality that $u(x,t) - u^*(x,t)$ is centered in the following sense

$$\int_\Omega u(x,t) - u^*(x,t)dx = 0. \tag{E.23}$$

Let's call $P_\epsilon(u) := \int_0^T \int_\Omega \frac{1}{2\epsilon} e^{-\frac{t}{\epsilon}}\Big(|\nabla u - \nabla u^*|^2 + |u - u^*|^2\Big)dxdt$.

According to standard Poincare inequality to space derivative, we obtain

$$\int_0^T \int_\Omega \frac{1}{2\epsilon} e^{-\frac{t}{\epsilon}} |\nabla u - \nabla u^*|^2 dx dt \leqslant P_\epsilon(u) \leqslant \frac{1}{2\epsilon} ||u - u^*||^2_{L^2_t(0,T;H^1_x(\Omega))}, \tag{E.24}$$

and

$$\int_0^T \int_\Omega \frac{1}{2\epsilon} e^{-\frac{t}{\epsilon}} |\nabla u - \nabla u^*|^2 dx dt \geqslant P_\epsilon(u) \geqslant \frac{e^{-\frac{T}{\epsilon}}}{4\epsilon} ||u - u^*||^2_{L^2_t(0,T;H^1_x(\Omega))}. \tag{E.25}$$

We obtain

$$\frac{e^{-\frac{T}{\epsilon}}}{4\epsilon} ||u - u^*||^2_{L^2_t(0,T;H^1_x(\Omega))} \leqslant \int_0^T \int_\Omega \frac{1}{2\epsilon} e^{-\frac{t}{\epsilon}} |\nabla u - \nabla u^*|^2 dx dt \leqslant \frac{1}{2\epsilon} ||u - u^*||^2_{L^2_t(0,T;H^1_x(\Omega))}. \tag{E.26}$$

Step 3: Combine (E.22) and (E.26) together.

Let's denote

$$E_{\epsilon,\beta}(u) := \beta(I_\epsilon(u) - I_\epsilon(u^*)) + \beta \int_0^T \int_\Omega e^{-\frac{t}{\epsilon}} u^*_{tt}(u^* - u) dx dt \tag{E.27}$$

$$+ \int_\Omega e^{-\frac{T}{\epsilon}} u^*_t(x,T)\big(u^*(x,T) - u(x,T)\big) dx. \tag{E.28}$$

Then, we have

$$E_{\epsilon,\beta}(u) \geqslant \frac{\beta e^{-\frac{T}{\epsilon}}}{4\epsilon} ||u - u^*||^2_{L^2_t(0,T;H^1_x(\Omega))} + \min\Big\{ \frac{\beta e^{-\frac{\epsilon}{T}}}{C^2(d,\epsilon,\Omega)}, \frac{\beta e^{-\frac{\epsilon}{T}}}{4} \Big\} ||u - u^*||^2_{H^1_t(0,T;L^2_x(\Omega))}, \tag{E.29}$$

$$E_{\epsilon,\beta}(u) \leqslant \frac{\beta}{2} ||u - u^*||^2_{H^1_t(0,T;L^2_x(\Omega))} + \frac{\beta}{2\epsilon} ||u - u^*||^2_{L^2_t(0,T;H^1_x(\Omega))}. \tag{E.30}$$

Step 4: Embedding into $H^1_t([0,T]; H^1_x(\Omega))$.

By the definition of time-dependent Sobolev space (see [Evans, 2010, Chapter 5]) and standard H^1 embedding into L^2, we can absorb lower and upper bound into $||u - u^*||^2_{H^1_t([0,T];H^1_x(\Omega))}$. It is straightforward to establish upper bound

$$||u - u^*||^2_{H^1_t([0,T];H^1_x(\Omega))} + E_{\epsilon,\beta}(u) \leqslant \Big(1 + \frac{\beta}{2} + \frac{\beta}{2\epsilon}\Big) ||u - u^*||^2_{H^1_t([0,T];H^1_x(\Omega))}. \tag{E.31}$$

To obtain lower bound, we have

$$||u - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} + E_{\epsilon,\beta}(u) \geqslant \frac{1}{3} ||u - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} + \frac{1}{3} ||u - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \tag{E.32}$$

$$+ \frac{1}{3} ||u - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} + \frac{\beta e^{-\frac{T}{\epsilon}}}{4\epsilon} ||u - u^*||^2_{L^2_t(0,T;H^1_x(\Omega))} \tag{E.33}$$

$$+ \min\Big\{ \frac{\beta e^{-\frac{\epsilon}{T}}}{C^2(d,\epsilon,\Omega)}, \frac{\beta e^{-\frac{\epsilon}{T}}}{4} \Big\} ||u - u^*||^2_{H^1_t(0,T;L^2_x(\Omega))} \tag{E.34}$$

$$\geqslant \frac{1}{3} ||u - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} + \frac{1}{3} ||\nabla u - \nabla u^*||^2_{H^1_t(0,T;L^2_x(\Omega))} \tag{E.35}$$

$$+ \min\Big\{ \frac{\beta e^{-\frac{\epsilon}{T}}}{C^2(d,\epsilon,\Omega)}, \frac{\beta e^{-\frac{\epsilon}{T}}}{4} \Big\} ||u - u^*||^2_{H^1_t(0,T;L^2_x(\Omega))} \tag{E.36}$$

$$+ \frac{1}{3} ||u_t - u^*_t||^2_{L^2_t(0,T;H^1_x(\Omega))} + \frac{\beta e^{-\frac{T}{\epsilon}}}{4\epsilon} ||u - u^*||^2_{L^2_t(0,T;H^1_x(\Omega))} \tag{E.37}$$

$$\geqslant (\frac{1}{3} + \min(\frac{1}{3}, \frac{\beta e^{-\frac{T}{\epsilon}}}{C^2(d,\epsilon,\Omega)}, \frac{\beta e^{-\frac{T}{\epsilon}}}{4})) ||u - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))}. \tag{E.38}$$

Finally, we obtain

$$(\frac{1}{3}+\min(\frac{1}{3},\frac{\beta e^{-\frac{T}{\epsilon}}}{C^2(d,\epsilon,\Omega)},\frac{\beta e^{-\frac{T}{\epsilon}}}{4}))||u-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))}\leqslant L_{\epsilon,\beta}(u)-L_{\epsilon,\beta}(u^*) \tag{E.39}$$

and

$$L_{\epsilon,\beta}(u)-L_{\epsilon,\beta}(u^*)\leqslant\Big(1+\frac{\beta}{2}+\frac{\beta}{2\epsilon}\Big)||u-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))}. \tag{E.40}$$

■

E.2 Upper bound on estimation error

Proving the upper bound on the estimation error is quite involved. So we outline the plan the proof and divide it into several steps.

To begin this section, First, we need to prove high probability bound on excess risk in Proposition E.1 and incorporate this bound using localization technique introduced in van de Geer (1987; 2002) along with strong convexity of our loss function. Last step is to plug in approximation error (E.158) to establish classical bias-variance trade-off.

Proposition E.1 (High Probability Bound on Excess Risk)**.** *Under Assumptions A1 - A4 in Appendix A.2 and define $\mathcal{U}_J$ as in (3.1). For any localization parameter $\rho>0$, define*

$$\mathcal{U}_J(\rho):=\Big\{u\in\mathcal{U}_J\;\Big|\;S_{*,\epsilon}(u)\leqslant\rho\Big\}. \tag{E.41}$$

Then there exists constant $C_1(d,\Omega,T,\epsilon,\beta)$ such that with probability $1-e^{-s}$,

$$\sup_{u_J\in\mathcal{U}_J(\rho)}|\hat{S}_{*,\epsilon}(u_J)-S_{*,\epsilon}(u_J)|\lesssim\psi(\rho,\tau,J)+\sqrt{\frac{2\rho B(\epsilon,\beta,\Omega,T,d)s}{nm}}+\tau_1(\rho,J), \tag{E.42}$$

where

$$\psi(\rho,\tau,J):=\frac{1}{\sqrt{nm}}J^{\frac{d+2}{2}}\sqrt{\rho B(\epsilon,\beta,\Omega,T,d)\log(1+\sqrt{\frac{C_2\rho}{B(\epsilon,\beta,\Omega,T,d)}})}+C_1(d,\Omega,T,\epsilon,\beta)\sqrt{\frac{\rho}{nm}} \tag{E.43}$$

and

$$\tau_1(\rho,J):=\begin{cases}\frac{J^{d+1}s}{nm}+\frac{RC(\Omega,T)}{nm}J^{2(d+2)}\log(1+\sqrt{\frac{C_2\rho}{B(\epsilon,\beta,\Omega,T,d)}}), & q\leqslant\frac{d}{2},\\ \frac{Rs}{nm}+\frac{RC(\Omega,T)}{nm}J^{d+2}\log(1+\sqrt{\frac{C_2\rho}{B(\epsilon,\beta,\Omega,T,d)}}), & q>\frac{d}{2}.\end{cases} \tag{E.44}$$

$B(\epsilon,\beta,\Omega,T,d):=(\frac{1}{3}+\frac{\beta}{\beta_0})^{-1}$, $\beta_0:=\frac{1}{C^2(d,\epsilon,\Omega)e^{T/\epsilon}}$ *and $C(d,\epsilon,\Omega)$ is a Cheeger constant.*

Proof of Proposition E.1. For simplicity of notation, we denote the excess risks as

$$S_*(u):=L_{\epsilon,\beta}(u)-L_{\epsilon,\beta}(u^*),\;\hat{S}_*(u):=\hat{L}_{\epsilon,\beta}(u)-\hat{L}_{\epsilon,\beta}(u^*) \tag{E.45}$$

and the space of localized functional class $\mathcal{U}_J$ as

$$\mathcal{U}_J(\rho):=\Big\{u_J\in\mathcal{U}_J\;\Big|\;S_{*,\epsilon}(u_J)\leqslant\rho\Big\}. \tag{E.46}$$

Our goal is to control excess risk $L_{\epsilon,\beta}(\hat{u}_J)-L_{\epsilon,\beta}(u^*)$ for a parametrized estimator $\hat{u}_J\in\mathcal{U}_J$, where $\mathcal{U}_J$ is a class of functions that have Fourier series representations with truncation of the level J. First, we decompose the excess risk $L_{\epsilon,\beta}(\hat{u}_J)-L_{\epsilon,\beta}(u^*)$ into telescoping sums, and we view the telescoping sums as a localized version of the empirical process and control empirical process for $u_J\in\mathcal{U}_J(\rho)$.

To start, we center all functions appropriately and consider the extension of a restriction of the functions u_J and u^* to compact space Ω over finite time horizon $[0, T]$, which allows us to use Fourier expansions for harmonic analysis. Note that u_J and u^* are defined in all of space-time $\mathbb{R}^d \times \mathbb{R}$, but we do not need to control their norms over the entire space-time domain. The existence of such extension is mainly attributed to extension Theorem A.2, which induces space-time expansion Lemma A.3.

With notation of linear extension operator ext defined in Theorem A.3, we use the following notation

$$ext\, u_J = u_J,\ ext\, u^* = u^*. \tag{E.47}$$

For any $u_J \in \mathcal{U}_J(p)$, by optimality of $\hat{u}_J$ we have

$$L_{\epsilon,\beta}(\hat{u}_J) - L_{\epsilon,\beta}(u_J) \tag{E.48}$$

$$= L_{\epsilon,\beta}(\hat{u}_J) - L_{\epsilon,\beta}(u^*) + \hat{L}_{\epsilon,\beta}(u^*) - \hat{L}_{\epsilon,\beta}(\hat{u}_J) + \hat{L}_{\epsilon,\beta}(\hat{u}_J) - \hat{L}_{\epsilon,\beta}(u^*) - L_{\epsilon,\beta}(u_J) + L_{\epsilon,\beta}(u^*) \tag{E.49}$$

$$= L_{\epsilon,\beta}(\hat{u}_J) - L_{\epsilon,\beta}(u^*) - \hat{L}_{\epsilon,\beta}(\hat{u}_J) - \hat{L}_{\epsilon,\beta}(u^*) + \hat{L}_{\epsilon,\beta}(\hat{u}_J) - \hat{L}_{\epsilon,\beta}(u^*) - L_{\epsilon,\beta}(u_J) - L_{\epsilon,\beta}(u^*) \tag{E.50}$$

$$\leqslant \left(S_*(\hat{u}_J) - \hat{S}_*(\hat{u}_J) \right) + \left(\hat{S}_*(u_J) - S_*(u_J) \right). \tag{E.51}$$

From the above inequality, it is enough to bound $S_*(\hat{u}_J) - \hat{S}_*(u_J)$ to control $L_{\epsilon,\beta}(\hat{u}_J) - L_{\epsilon,\beta}(u_J)$. To proceed, we consider bounding a localized version of the above empirical process

$$\sup_{u_J \in \mathcal{U}_J(\rho)} S_*(u_J) - \hat{S}_*(u_J). \tag{E.52}$$

We want to first control the expectation of the supremum. The expectation control has two components: the first component contains the bounding tail error arising from the truncation of the Fourier series at the truncation of level J and the second part contains statistical error from truncated Fourier series itself. Once we obtain upper bound for the expectation, we are able to produce a high-probability bound via Talagrand's concentration inequality, cf. Theorem F.5.

For convenience, let's write $h = ext\,(u_J - u^*)$ and its local basis Representation as

$$h = \sum_{0 \leqslant \tau, |\xi|_\infty \leqslant \infty} a_{\tau,\xi} \phi_{\tau,|\xi|_\infty}(x,t) \tag{E.53}$$

$$= \sum_{0 \leqslant \tau \leqslant J^2, 0 \leqslant |\xi|_\infty \leqslant J} a_{\tau,\xi} \phi_{\tau,|\xi|_\infty}(x,t) + \sum_{\tau > J^2, |\xi|_\infty > J} a_{\tau,\xi} \phi_{\tau,|\xi|_\infty}(x,t) \tag{E.54}$$

$$= P_{J,J^2} h + P_{>J,>J^2} h. \tag{E.55}$$

Plugging local basis representation, we obtain

$$\mathbb{E}\left[\sup_{u_J \in \mathcal{U}_J(\rho)} |S_{*,\epsilon}(u_J) - \hat{S}_{*.\epsilon}(u_J)| \right] \tag{E.56}$$

$$= \mathbb{E}\left[\sup_{u_J \in \mathcal{U}_J(\rho)} |L_{\epsilon,\beta}(u_J) - \hat{L}_{\epsilon,\beta}(u_J) + \hat{L}_{\epsilon,\beta}(u^*) - L_{\epsilon,\beta}(u^*)| \right] \tag{E.57}$$

$$= \mathbb{E}\left[\sup_{u_J \in \mathcal{U}_J(\rho)} |(L_{\epsilon,\beta} - \hat{L}_{\epsilon,\beta})(u_J - u^*)| \right] \tag{E.58}$$

$$\leqslant \underbrace{\mathbb{E}\left[\sup_{u_J \in \mathcal{U}_J(\rho)} |(L_{\epsilon,\beta} - \hat{L}_{\epsilon,\beta}) P_{J,J^2} h| \right]}_{I_1} + \underbrace{\mathbb{E}\left[\sup_{u_J \in \mathcal{U}_J(\rho)} |(L_{\epsilon,\beta} - \hat{L}_{\epsilon,\beta}) P_{>J,>J^2} h| \right]}_{I_2}. \tag{E.59}$$

Here we want to apply Symmetrization technique to I_1 and I_2. Since the sampling scheme (1.2) and (1.3) treats time dimension t as another additional spatial dimension, we flatten $d+1$-d spatial-temporal training data matrix $\{(x_{ij}, t_j)\}_{i=1,j=1}^{n,m}$ to 1-d data vector $\{z_i\}_{i=1}^{nm}$ such that

$$z_1 = (x_{11}, t_1), ..., z_m = (x_{m1}, t_1), z_{m+1} = (x_{12}, t_2), ..., z_{2m} = (x_{m2}, t_2), ..., z_{mn} = (x_{mn}, t_n).$$

In this way, standard symmetrization argument with nm copies of independent Rademacher random variables σ_i is applied

$$I_1 \leqslant \underbrace{\frac{1}{nm}\mathbb{E}\Big[\sup_{u_J\in\mathcal{U}_J(\rho)}\sum_{i=1}^{nm}\sigma_i P_{J,J^2}h(z_i)\Big]}_{T_1},\quad I_2 \leqslant \underbrace{\frac{1}{nm}\mathbb{E}\Big[\sup_{u_J\in\mathcal{U}_J(\rho)}\sum_{i=1}^{nm}\sigma_i P_{>J,>J^2}h(z_i)\Big]}_{T_2}. \tag{E.60}$$

The rest of the subsection is devoted to first bound T_2 and then bound T_1.

Control of T_2:

$$T_2 := \frac{1}{nm}\mathbb{E}\Big[\sup_{u_J\in\mathcal{U}_J(\rho)}\sum_{i=1}^{nm}\sigma_i P_{>J,>J^2}h(z_i)\Big] = \frac{1}{nm}\mathbb{E}\Big[\sup_{u_J\in\mathcal{U}_J(\rho)}\sum_{i=1}^{nm}\sigma_i \sum_{\tau>J^2,|\xi|_\infty>J} a_{\tau,\xi}\phi_{\tau,\xi}(x,t)\Big]. \tag{E.61}$$

By Cauchy-Schwarz inequality, we obtain

$$\sum_{i=1}^{nm}\sigma_i \sum_{\tau>J^2,|\xi|_\infty>J} a_{\tau,\xi}\phi_{\tau,\xi}(x,t) \leqslant \underbrace{\Big(\sum_{i=1}^{nm}\sigma_i^2\Big)^{\frac{1}{2}}}_{T_{2,1}} \underbrace{\Big(\sum_{i=1}^{nm}\big(\sum_{\tau>J^2,|\xi|>J} a_{\tau,\xi}\phi_{\tau,\xi}(x,t)\big)^2\Big)^{\frac{1}{2}}}_{T_{2,2}} \tag{E.62}$$

$$T_{2,2} = \Big(\sum_{\tau>J^2,|\xi|_\infty>J} a_{\tau,\xi}^2\phi_{\tau,\xi}^2(z_i) + \sum_{1\leqslant j\neq k\leqslant nm} a_{\tau,\xi}^2\phi_{\tau,\xi}(z_j)\phi_{\tau,\xi}(z_k)\Big)^{\frac{1}{2}}. \tag{E.63}$$

By Jensen inequality, we can pass expectation over σ inside $T_{2,1}$ and expectation over z inside $T_{2,2}$. Note σ_i and σ_j are independent variables with $\mathbb{E}\sigma_i = \mathbb{E}\sigma_j = 0$, and $||\phi_{\tau,\xi}||_{L_2}^2 = 1$.

$$\mathbb{E}_{\sigma,z}\Big[\ T_{2,2}\ \Big] \leqslant \Big(\sum_{\tau>J^2,|\xi|_\infty>J} a_{\tau,\xi}^2||\phi_{\tau,\xi}||_{L_2}^2 + \sum_{\tau,\tau'>J^2}\sum_{|\xi|_\infty,|\xi'|_\infty>J} a_{\tau,\xi}a_{\tau',\xi'}||\phi_{\tau,\xi}\phi_{\tau',\xi}(z_k)||_{L_1}\Big)^{\frac{1}{2}} \tag{E.64}$$

$$= \Big(\sum_{\tau>J^2,|\xi|_\infty>J} a_{\tau,\xi}^2\Big)^{\frac{1}{2}} \quad - \text{ Orthogonality of } \phi_{\tau,\xi}(z) \tag{E.65}$$

$$\leqslant ||P_{>J,>J^2}h||_{H_t^1(\mathbb{R};H_x^1(\mathbb{R}^d))} \tag{E.66}$$

$$\leqslant C(\Omega,T)||P_{>J,>J^2}h||_{H_t^1(0,T;H_x^1(\Omega))} \quad - \text{ space-time extension Lemma A.3} \tag{E.67}$$

$$\lesssim RJ^{-(2p+q-3)}\sqrt{\rho}. \quad - \text{ stability estimate Proposition 2.2} \tag{E.68}$$

Combining $T_{2,1}$ and $T_{2,2}$ all together, we obtain

$$T_2 = T_{2,1} * \mathbb{E}_{\sigma,z}[T_{2,2}] \lesssim J^{-(2p+q-3)}\sqrt{\frac{\rho}{nm}}. \tag{E.69}$$

Control of T_1:
In this section, we stick with $h := ext\,(u_J - u^*)$. We use a chaining bound for metric entropy to bound T_1 using $L_t^2(\mathbb{R};L_x^2(\mathbb{R}^d))$ bracketing number, and note $L_t^2(\mathbb{R};L_x^2(\mathbb{R}^d))$ is equivalent to $L_{t,x}^2(\mathbb{R}\times\mathbb{R}^d)$. In particular, we construct a bracketing number for J finitely many Fourier Coefficients $a_{\xi,\tau}$ in L^2 balls.

For $h_1, h_2 \in \mathcal{U}_J(\rho)$,

$$\big|\big|\, P_{J,J^2}h_1 - P_{J,J^2}h_2 \,\big|\big|^2_{L^2_t(0,T;L^2_x(\Omega))} = \big|\big| \sum_{1\leqslant|\xi|_\infty\leqslant J, 1\leqslant\tau\leqslant J^2} a^{(1)}_{\xi,\tau}\phi_{\xi,\tau}(x,t) - a^{(2)}_{\xi,\tau}\phi_{\xi,\tau}(x,t) \,\big|\big|^2_{L^2_t(0,T;L^2_x(\Omega))} \tag{E.70}$$

$$\lesssim \big|\big| \sum_{1\leqslant|\xi|_\infty\leqslant J, 1\leqslant\tau\leqslant J^2} (\, a^{(1)}_{\xi,\tau} - a^{(2)}_{\xi,\tau}\,)\phi_{\xi,\tau}(x,t) \,\big|\big|^2_{L^2_t(0,T;L^2_x(\Omega))} \tag{E.71}$$

$$\leqslant \sum_{1\leqslant|\xi|_\infty\leqslant J, 1\leqslant\tau\leqslant J^2} \big|\, a^{(1)}_{\xi,\tau} - a^{(2)}_{\xi,\tau}\,\big|^2. \tag{E.72}$$

Hence we obtain

$$\big|\big|\, P_Jh_1 - P_Jh_2 \,\big|\big|^2_{L^2_t(0,T;L^2_x(\Omega))} \lesssim \sum_{1\leqslant|\xi|_\infty\leqslant J, 1\leqslant\tau\leqslant J^2} |a^{(1)}_{\xi,\tau} - a^{(2)}_{\xi,\tau}|^2. \tag{E.73}$$

The inequality above suggests a control of finitely many spatial-temporal cofficients $a_{\xi,\tau}$ in L^2 ball of coefficient space is sufficient to construct an ϵ-net for $h \in \mathcal{U}_J(\rho)$.

We want to construct L^2 packing number for the space of finitely many Fourier coefficients $a_{\xi,\tau}$. Using standard L^2-packing for L^2 balls argument in Lemma F.6 and note the dimension of coefficients are at most J^{d+2}, we have

$$\log N(\mathcal{U}_J(\rho), ||\cdot||_{L^2_t(0,T;L^2_x(\Omega))}, \epsilon) \lesssim \log(\frac{3}{\delta})^{J^{d+2}} = J^{d+2}\log(\frac{1}{\epsilon}). \tag{E.74}$$

With the above packing number at hand, we're ready to apply Bernstein Chaining in Theorem F.4. Before applying Bernstein Chaining, we nee to check finite second moment condition and finite L_∞ norm condition.

From the lower bound of stability estimates in Proposition 2.2 and Lemma F.10, we can verify second moment condition for $P_Jh \in \mathcal{U}_J(\rho)$

$$||P_Jh||_{L^2_t(\mathbb{R};L^2_x(\mathbb{R}^d))} \leqslant ||h||_{L^2_t(\mathbb{R};L^2_x(\mathbb{R}^d))} \lesssim C(\Omega,T)\sqrt{\rho B(\epsilon,\beta,\Omega,T,d)}, \tag{E.75}$$

where $B(\epsilon,\beta,\Omega,T,d) := (1+\frac{\beta}{\beta_0})^{-1}$, $\beta_0 = e^{T/\epsilon}C^2(d,\epsilon,\Omega)$, and $C(d,\epsilon,\Omega)$ is the Cheeger constant here.

To verify L_∞ condition, we use Lemma F.10 to obtain

$$||P_Jh||_\infty \lesssim \begin{cases} R, & q > \frac{d}{2}, \\ J^{d+2}, & q \leqslant \frac{d}{2}. \end{cases} \tag{E.76}$$

Next we will incorporate bracketing numbers into the Dudley integral theorem (cf. Theorem F.3) to get

$$D_{[]} := D_{[]}\Big(\sqrt{\rho B(\epsilon,\beta,\Omega,T,d)},\ \mathcal{U}_J(\rho),\ ||\cdot||_{L^2_{t,x}(\mathcal{P})}\Big) \tag{E.77}$$

$$= \int_0^{\sqrt{\rho B(\epsilon,\beta,\Omega,T,d)}} \sqrt{logN_{[]}\Big(\epsilon, \mathcal{U}_J(\rho), ||\cdot||_{L^2_{t,x}(\mathcal{P})}\Big)}\, d\epsilon \tag{E.78}$$

$$\lesssim \int_0^{\sqrt{\rho B(\epsilon,\beta,\Omega,T,d)}} J^{\frac{d+2}{2}} \sqrt{\log(1+\frac{1}{\epsilon})}\, d\epsilon \tag{E.79}$$

and

$$D_{[]} \lesssim J^{(d+2)/2}\sqrt{\rho B(\epsilon,\beta,\Omega,T,d)\log(1+\sqrt{\frac{C_2\beta_0}{\rho\beta}})}, \tag{E.80}$$

where C_2 is some appropriately chosen constant.

Plugging second moment condition (E.75) and L^∞ condition (E.76) into Theorem F.4, we obtain

$$T_1 :\lesssim \frac{1}{\sqrt{nm}} D_{[]} \left(1 + \frac{D_{[]}}{\sqrt{nm}||P_J h||^2_{L^2_t([0,T];L^2_x(\Omega))}} ||P_J h||_\infty \right). \tag{E.81}$$

We obtain

$$T_1 \lesssim \frac{1}{\sqrt{nm}} J^{\frac{d+2}{2}} \sqrt{\rho B(\epsilon, \beta, \Omega, T, d) \log(1 + \sqrt{\frac{C_2 \rho}{B(\epsilon, \beta, \Omega, T, d)}})} + \tau(\rho, J), \tag{E.82}$$

where

$$\tau(\rho, J) \lesssim \begin{cases} \frac{R}{nm} J^{2(d+2)} \log(1 + \sqrt{\frac{C_2 \rho}{B(\epsilon, \beta, \Omega, T, d)}}), & q \leqslant \frac{d}{2}, \\ \frac{R}{nm} J^{d+2} \log(1 + \sqrt{\frac{C_2 \rho}{B(\epsilon, \beta, \Omega, T, d)}}), & q > \frac{d}{2}. \end{cases} \tag{E.83}$$

Final Statistical Estimation Error Bound:
From control of T_1 (E.82) and control of T_2 (E.69), we obtain

$$T_1 \lesssim \frac{1}{\sqrt{nm}} J^{\frac{d+2}{2}} \sqrt{\rho B(\epsilon, \beta, \Omega, T, d) \log(1 + \sqrt{\frac{C_2 \rho}{B(\epsilon, \beta, \Omega, T, d)}})} + \tau(\rho, J), \tag{E.84}$$

$$T_2 \lesssim C(d, \Omega, T, \epsilon, \beta) \sqrt{\frac{\rho}{nm}}, \tag{E.85}$$

$$T_1 + T_2 = \psi(\rho, \tau, J) + \tau(\rho, J). \tag{E.86}$$

Combining the bound on T_1 and T_2 all together yields

$$\mathbb{E}\left[\sup_{u_J \in \mathcal{U}, S_{*,\epsilon}(u_J) \leqslant \rho} |S_{*,\epsilon}(u_J) - \hat{S}_{*.\epsilon}(u_J)| \right] \lesssim \psi(\rho, \tau, J) + \tau(\rho, J), \tag{E.87}$$

where

$$\psi(\rho, \tau, J) := \frac{1}{\sqrt{nm}} J^{\frac{d+2}{2}} \sqrt{\rho B(\epsilon, \beta, \Omega, T, d) \log(1 + \sqrt{\frac{C_2 \rho}{B(\epsilon, \beta, \Omega, T, d)}})} + C(d, \Omega, T, \epsilon, \beta) \sqrt{\frac{\rho}{nm}}. \tag{E.88}$$

Finally, we apply Talagrand's concentration inequality Theorem F.5 to bound $I_\epsilon(h) - \hat{I}_\epsilon(h)$, which upper bounds excess risk. In order to apply the concentration inequality, we again need to verify the same second moment condition (E.75) and L_∞ condition (E.76), which were verified earlier:

$$||h||_{L^2_{t,x}(\mathbb{R} \times \mathbb{R}^d)} \lesssim \sqrt{\rho B(\epsilon, \beta, \Omega, T, d)}, \ ||P_J h||_\infty \lesssim \begin{cases} R, & q > \frac{d}{2}, \\ J^{d+2}, & q \leqslant \frac{d}{2}. \end{cases} \tag{E.89}$$

With probability $1 - e^{-s}$,

$$\sup_{u_J \in \mathcal{U}_{\mathcal{J}}(\rho)} |\hat{S}_{*,\epsilon}(u_J) - S_{*,\epsilon}(u_J)| \lesssim \psi(\rho, \tau, J) + \sqrt{\rho B(\epsilon, \beta, \Omega, T, d) \cdot \frac{2s}{nm}} + \tau_1(\rho, J), \tag{E.90}$$

where

$$\tau_1(\rho, J) := \begin{cases} \frac{J^{d+1} s}{nm} + \frac{R}{nm} J^{2(d+2)} \log(1 + \sqrt{\frac{C_2 \rho}{B(\epsilon, \beta, \Omega, T, d)}}), & q \leqslant \frac{d}{2}, \\ \frac{Rs}{nm} + \frac{R}{nm} J^{d+2} \log(1 + \sqrt{\frac{C_2 \rho}{B(\epsilon, \beta, \Omega, T, d)}}), & q > \frac{d}{2}. \end{cases} \tag{E.91}$$

■

Remark E.2. we obtain L_∞ bound on h would be through the fact that L^∞ norm agrees with C^0 norm on compact set Ω_T. By embedding theorem F.8, we can use H_x^q norm to bound C_x^0 and therefore L_x^∞ when spatial regularity $q > \frac{d}{2}$, while temporal domain lies in 1d and therefore it's naturally embedded into L_t^∞. However, this only works when spatial regularity q is large enough against dimension d.

Next we use localization technique to prove the following proposition.

Proposition E.3 (Localization)**.** *Under Assumptions A1- A4 in Appendix A.2 hold and define $\mathcal{U}_J$ as in (3.1). Then, there exist constant $C_2 := C_2(\Omega, T)$ such that estimator $\hat{u}_J$ defined in (2.7) satisfies with probability at least $1 - e^{-s}$,*

If $q < \frac{d}{2}$,

$$||\hat{u}_J - u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} \lesssim \left[1 + (1 + \frac{\beta}{2\epsilon})B(\epsilon\beta, \Omega, T, d)\right] ||u_J - u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} \tag{E.92}$$

$$+ (1 + \frac{\beta}{2\epsilon})B^2(\epsilon, \beta, \Omega, T, d)\frac{J^{2(d+2)}}{nm}\log(1 + C_2 nm) \tag{E.93}$$

$$+ \frac{J^{2(d+2)}s + 1}{B(\epsilon, \beta, \Omega, T, d)nm}. \tag{E.94}$$

Moreover,

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H_t^1([0,T];H_x^1(\Omega))} \lesssim \left[1 + (1 + \frac{\beta}{2\epsilon})B(\epsilon, \beta, \Omega, T, d)\right] ||u_J - u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} \tag{E.95}$$

$$+ (1 + \frac{\beta}{2\epsilon})B^2(\epsilon, \beta, \Omega, T, d)\frac{J^{2(d+2)}}{nm}\log(1 + C_2 nm) \tag{E.96}$$

$$+ \frac{1}{B(\epsilon, \beta, \Omega, T, d)nm}. \tag{E.97}$$

On the other hand when $q \geqslant \frac{d}{2}$, with the probability $1 - e^{-s}$,

$$||\hat{u}_J - u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} \lesssim \left[1 + (1 + \frac{\beta}{2\epsilon})B(\epsilon, \beta, \Omega, T, d)\right] ||u_J - u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} \tag{E.98}$$

$$+ B(\epsilon, \beta, \Omega, T, d)(1 + \frac{\beta}{2\epsilon})\frac{J^{d+2}}{nm}\log(1 + C_2 nm) + \frac{1+s}{nm}. \tag{E.99}$$

Moreover,

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} \lesssim_R \left[1 + (1 + \frac{\beta}{2\epsilon})B(\epsilon, \beta, \Omega, T, d)\right] ||u_J - u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} \tag{E.100}$$

$$+ B(\epsilon, \beta, \Omega, T, d)(1 + \frac{\beta}{2\epsilon})\frac{J^{d+2}}{nm}\log(1 + C_2 nm) + \frac{1}{nm} \tag{E.101}$$

Proof of Proposition E.3. Let $\hat{u}_J := \underset{u \in \mathcal{U}_J(\rho)}{\operatorname{argmin}} \hat{L}_{\epsilon,\beta}(u)$ and $u_J := \underset{u \in \mathcal{U}_J(\rho)}{\operatorname{argmin}} L_{\epsilon,\beta}(u)$.

Fix $r > 0$ to be determined later, we set

$$\hat{u}_s := s\hat{u}_J + (1 - s)u_J, \;\; s = \frac{\sqrt{r}}{\sqrt{r} + ||\hat{u}_J - u_J||_{H_t^1(0,T;H_x^1(\Omega))}}. \tag{E.102}$$

Since $s \in [0, 1]$ and $\mathcal{U}_J$ is convex, we have $\hat{u}_s \in \mathcal{U}_J$ and $\hat{u}_s$ is localized in the following sense

$$||\hat{u}_s - u_J||_{H_t^1(0,T;H_x^1(\Omega))} \leqslant \frac{\sqrt{r}||\hat{u}_s - u_J||_{H_t^1(0,T;H_x^1(\Omega))}}{\sqrt{r} + ||\hat{u}_s - u_J||_{H_t^1(0,T;H_x^1(\Omega))}} \leqslant \sqrt{r}. \tag{E.103}$$

Using stability estimate in Proposition 2.2 and the triangle inequality, we therefore have $\hat{u}_s \in \mathcal{U}_J(\rho)$.

$$S_{*,\epsilon}(\hat{u}_s) \leqslant \left(1 + \frac{\beta}{2\epsilon}\right)|| \hat{u}_s - u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} \leqslant \left(1 + \frac{\beta}{2\epsilon}\right)\left(r + || u_J - u^*||^2_{H_t^1(0,T;H_x^1(\Omega))}\right) := \rho. \tag{E.104}$$

By Convexity of $\hat{S}_{*,\epsilon}$, we have

$$\hat{S}_{*,\epsilon}(\hat{u}_s) \leqslant \hat{S}_{*,\epsilon}(u_J) + s\bigg(\hat{S}_{*,\epsilon}(\hat{u}_J) - \hat{S}_{*,\epsilon}(u_J)\bigg). \tag{E.105}$$

By the definition of empirical minimizer $\hat{u}$, we have

$$\hat{L}_\epsilon(\hat{u}_J) \leqslant \hat{L}(u_J). \tag{E.106}$$

It follows

$$\hat{S}_{*,\epsilon}(\hat{u}_J) - \hat{S}_{*,\epsilon}(u_J) = \hat{L}_\epsilon(\hat{u}_J) - \hat{L}(u_J) \leqslant 0. \tag{E.107}$$

Therefore, the empirical excess risk of $\hat{u}_s$ is smaller than u_J

$$\hat{S}_{*,\epsilon}(\hat{u}_s) \leqslant \hat{S}_{*,\epsilon}(u_J). \tag{E.108}$$

Next we want to bound excess risk $S_{*,\epsilon}$ using generaliztion error

$$\hat{S}_{*,\epsilon}(\hat{u}_s) - S_{*,\epsilon}(\hat{u}_s) + S_{*,\epsilon}(\hat{u}_s) \leqslant \hat{S}_{*,\epsilon}(u_J) + S_{*,\epsilon}(u_J) - S_{*,\epsilon}(u_J), \tag{E.109}$$

$$S_{*,\epsilon}(\hat{u}_s) \leqslant S_{*,\epsilon}(u_J) + \hat{S}_{*,\epsilon}(u_J) - S_{*,\epsilon}(u_J) + S_{*,\epsilon}(\hat{u}_s) - \hat{S}_{*,\epsilon}(\hat{u}_s), \tag{E.110}$$

$$S_{*,\epsilon}(\hat{u}_s) \leqslant S_{*,\epsilon}(u_J) + 2 \sup_{u\in\mathcal{U}_J} \big|S_{*,\epsilon}(u) - \hat{S}_{*,\epsilon}(u)\big|. \tag{E.111}$$

Using lower bound of stability estimate in Proposition 2.2 and Proposition E.1, we obtain

$$||u_J - \hat{u}_s||^2_{H^1_t(0,T;H^1_x(\Omega))} \leqslant 2||u_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} + 2||\hat{u}_s - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \tag{E.112}$$

$$\leqslant B(\epsilon,\beta,\Omega,T,d)\bigg[2S_{*,\epsilon}(u_J) + 2S_{*,\epsilon}(\hat{u}_s) - e^{-\frac{T}{\epsilon}}||\nabla u(\cdot,T) - \nabla u^*(\cdot,T)||^2_{L^2(\Omega)}\bigg] \tag{E.113}$$

$$\leqslant 4B(\epsilon,\beta,\Omega,T,d)\bigg[S_{*,\epsilon}(u_J) + \sup_{u\in\mathcal{U}_J}(S_{*,\epsilon}(u) - \hat{S}_{*,\epsilon}(u)) - E_T(\epsilon)\bigg], \tag{E.114}$$

where we denote $E_T(\epsilon) := \frac{e^{-\frac{T}{\epsilon}}}{2}||\nabla u(\cdot,T) - \nabla u^*(\cdot,T)||^2_{L^2(\Omega)}$.

With $s = \frac{1}{2}$, we have $||\hat{u}_s - u_J||_{H^1_t([0,T];H^1_x(\Omega))} \geqslant \frac{\sqrt{r}}{2}$ iff $||\hat{u}_J - u_J||_{H^1_t([0,T];H^1_x(\Omega))} \geqslant \sqrt{r}$. Using upper and lower bound of stability estimate in Proposition 2.2 and the relation (E.104) between ρ and r leads to the following

$$\mathbb{P}\Big(||\hat{u}_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \geqslant r + ||u_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))}\Big) \tag{E.115}$$

$$\leqslant \mathbb{P}\Big(||\hat{u}_s - u_J||^2_{H^1_t(0,T;H^1_x(\Omega))} \geqslant \frac{r}{4}\Big) \tag{E.116}$$

$$\leqslant \mathbb{P}\bigg(\sup_{u\in\mathcal{U}_J(\rho)}(S_{*,\epsilon}(u) - \hat{S}_{*,\epsilon}(u)) \gtrsim \frac{r}{16B(\epsilon,\beta,\Omega,T,d)} - S_{*,\epsilon}(u_J) + E_T(\epsilon)\bigg) \tag{E.117}$$

$$\leqslant \mathbb{P}\bigg(\sup_{u\in\mathcal{U}_J(\rho)}(S_{*,\epsilon}(u) - \hat{S}_{*,\epsilon}(u)) \gtrsim \frac{\rho - (1+\frac{\beta}{2\epsilon})||u_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))}}{16(1+\frac{\beta}{2\epsilon})B(\epsilon,\beta,\Omega,T,d)} - S_{*,\epsilon}(u_J) + E_T(\epsilon)\bigg)\,. \tag{E.118}$$

In observation of the high probability bound of the excess risk, we can use the previous result on probabilistic bound of $S_{*,\epsilon}(u) - \hat{S}_{*,\epsilon}(u)$. In order to apply this previous result, we want the following to hold:

$$\frac{\rho}{16(1+\frac{\beta}{2\epsilon})B(\epsilon,\beta,\Omega,T,d)} \geqslant S_{*,\epsilon}(u_J) + \sqrt{\rho B(\epsilon,\beta,\Omega,T,d)\frac{2s}{nm}} + \frac{||u_J - u^*||^2_{H^1_t([0,T];H^1_x(\Omega))}}{16B(\epsilon,\beta,\Omega,T,d)} \tag{E.119}$$

$$- E_T(\epsilon) + 2\psi(\rho,\tau,J) + \tau_1(\rho,J). \tag{E.120}$$

In order to make the inequality hold, we substitute the upper bound of stability estimate in Proposition 2.2. The terminal term $E_T(\epsilon)$ will vanish due to the cancellation by the same term appearing in the upper bound. So it follows

$$\frac{\rho}{16(1+\frac{\beta}{2\epsilon})B(\epsilon,\beta,\Omega,T,d)} \geqslant (1+\frac{\beta}{2\epsilon})||u_J-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \tag{E.121}$$

$$+2\psi(\rho,\tau,J)+\sqrt{\rho B(\epsilon,\beta,\Omega,T,d)\frac{2s}{nm}}+\tau_1(\rho,J). \tag{E.122}$$

The balance between r and J and β is different for different regularity regime: $q \leqslant \frac{d}{2}$ and $q > \frac{d}{2}$. We first determine r when $q < \frac{d}{2}$. According to the relation (E.104) between r and ρ, we can choose r to be

$$r \gtrsim (1+\frac{\beta}{2\epsilon})B^2(\epsilon,\beta,\Omega,T,d)J^{2(d+2)}\frac{\log(1+C_2\sqrt{nm})}{nm} \tag{E.123}$$

$$+(1+\frac{\beta}{2\epsilon})B(\epsilon,\beta,\Omega,T,d)||u_J-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))}+\frac{J^{2(d+2)}s+1}{B(\epsilon,\beta,\Omega,T,d)nm}. \tag{E.124}$$

Applying probabilistic bound for excess risk $S_{*,\epsilon}(u)-\hat{S}_{*,\epsilon}(u)$ with r in (E.123) and (E.124), we obtain with probability $1-e^{-s}$,

$$||\hat{u}_J-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim ||u_J-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))}+r \tag{E.125}$$

$$\lesssim \left[1+(1+\frac{\beta}{2\epsilon})B(\epsilon,\beta,\Omega,T,d)\right]||u_J-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \tag{E.126}$$

$$+(1+\frac{\beta}{2\epsilon})B^2(\epsilon,\beta,\Omega,T,d)\frac{J^{2(d+2)}}{nm}\log(1+C_2nm) \tag{E.127}$$

$$+\frac{J^{2(d+2)}s}{B(\epsilon,\beta,\Omega,T,d)nm}+\frac{1}{B(\epsilon,\beta,\Omega,T,d)nm}. \tag{E.128}$$

Taking expectation over s, we obtain

$$\mathbb{E}||\hat{u}_J-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim (1+\frac{\beta}{2\epsilon})B^2(\epsilon,\beta,\Omega,T,d)\frac{J^{2(d+2)}}{nm}\log(1+C_2nm) \tag{E.129}$$

$$+\left[1+(1+\frac{\beta}{2\epsilon})B(\epsilon,\beta,\Omega,T,d)\right]||u_J-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \tag{E.130}$$

$$+\frac{1}{B(\epsilon,\beta,\Omega,T,d)nm}. \tag{E.131}$$

When $q \geqslant \frac{d}{2}$, we mimic previous step to choose different r as

$$r \gtrsim B(\epsilon,\beta,\Omega,T,d)(1+\frac{\beta}{2\epsilon})J^{d+2}\frac{\log(1+C_2\sqrt{nm})}{nm} \tag{E.132}$$

$$+(1+\frac{\beta}{2\epsilon})B(\epsilon,\beta,\Omega,T,d)||u_J-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))}+\frac{Rs+1}{B(\epsilon,\beta,\Omega,T,d)nm}. \tag{E.133}$$

Similarly, with r in (E.132) and (E.133), we obtain with probability $1-e^{-s}$,

$$||\hat{u}_J-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim_R \left[1+(1+\frac{\beta}{2\epsilon})B(\epsilon,\beta,\Omega,T,d)\right]||u_J-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \tag{E.134}$$

$$+B(\epsilon,\beta,\Omega,T,d)(1+\frac{\beta}{2\epsilon})\frac{J^{d+2}}{nm}\log(1+C_2nm)+\frac{1+s}{nm}. \tag{E.135}$$

and we obtain in expectation

$$\mathbb{E}||\hat{u}_J-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim \left[1+(1+\frac{\beta}{2\epsilon})B(\epsilon,\beta,\Omega,T,d)\right]||u_J-u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \tag{E.136}$$

$$+B(\epsilon,\beta,\Omega,T,d)(1+\frac{\beta}{2\epsilon})\frac{J^{d+2}}{nm}\log(1+C_2nm)+\frac{1}{nm}. \tag{E.137}$$

Next, we want to control the approximation error $\inf_{u_J \in \mathcal{U}_J} ||u_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))}$ that appeared in Proposition E.3. Since our solution u^* is defined in compact domain space-time domain $\Omega \times [0,T]$, it is sufficient to control $\bar{u}_J := \prod_J ext\, u^*$, which is a projection of u^* over J truncated local basis extended to entire space-time domain $\mathbb{R}^d$ and $\mathbb{R}$. In light of extension Lemma A.2, ext u^* preserves Sobolev norm up to some constant depending on Ω and $[0,T]$, and together with assumption A3 in Appendix A.2, we can show under Sobolev norm $||\cdot||_{H^1_t(0,T;H^1_x(\Omega))}$, approximation errors are controllable.

Lemma E.4. *Given $p, q \in \mathbb{Z}^+$ and $J, J' \in \mathbb{Z}^+$. For any function $f(x,t) \in H^p_t(\mathbb{R}; H^q_x(\mathbb{R}^d))$, there exists Spatial-Temporal approximation $f_{J,J'}(x,t)$ in the space $\mathcal{F}_{J,J'} \cap H^p_t(\mathbb{R}; H^q_x(\mathbb{R}^d))$ such that the following approximation inequalities hold for $0 \leqslant p_1 \leqslant p, 0 \leqslant q_1 \leqslant q$,*

$$||f - f_{J,J'}||^2_{H^{p_1}_t(\mathbb{R};H^{q_1}_x(\mathbb{R}^d))} \lesssim J^{-2(q-q_1)} J'^{-2(p-p_1)} ||f||^2_{H^p_t(\mathbb{R};H^q_x(\mathbb{R}^d))}. \tag{E.138}$$

In particular, if $J' = J^2$,

$$||f - f_{J,J'}||^2_{H^{p_1}_t(\mathbb{R};H^{q_1}_x(\mathbb{R}^d))} \lesssim J^{-2(q+2p-q_1-2p_1)} ||f||^2_{H^p_t(\mathbb{R};H^q_x(\mathbb{R}^d))}, \tag{E.139}$$

where

$$\mathcal{F}_{J,J'} := \left\{ f \;\middle|\; f \in V_{J,J'} \cap H^p_t(\mathbb{R}; H^q_x(\mathbb{R}^d)) \right\}. \tag{E.140}$$

Proof of Lemma E.4. For any $f \in H^p_t(\mathbb{R}; H^q_x(\mathbb{R}^d))$, there exists Spatial-temporal representation in terms of L_2 basis in space-time domain Ω_T,

$$f(x,t) = f_{J,J'}(x,t) + f_{J,>J'}(x,t) + f_{>J,J'}(x,t) + f_{>J,>J'}(x,t). \tag{E.141}$$

Where $f_{J,J'}(x,t)$ is the truncated term and the rest are tail terms and $\phi_{\xi,\tau}(x,t)$ is basis of L^2 of product space of space and time

$$f_{J,J'}(x,t) := \sum_{|\xi|_\infty \leqslant J} \sum_{|\tau| \leqslant J'} a_{\xi,\tau} \phi^\xi_\tau(t) \phi_\xi(x), \quad f_{J,>J'}(x,t) := \sum_{|\xi|_\infty \leqslant J} \sum_{|\tau| > J'} a_{\xi,\tau} \phi^\xi_\tau(t) \phi_\xi(x), \tag{E.142}$$

$$f_{>J,J'}(x,t) := \sum_{|\xi|_\infty > J} \sum_{|\tau| \leqslant J'} a_{\xi,\tau} \phi^\xi_\tau(t) \phi_\xi(x), \quad f_{>J,>J'}(x,t) := \sum_{|\xi|_\infty > J} \sum_{|\tau| > J'} a_{\xi,\tau} \phi^\xi_\tau(t) \phi_\xi(x). \tag{E.143}$$

According to the definition of Sobolev norm (A.21) and definition of time-dependent Sobolev space in Appendix A.3.1, we can rewrite the tailed error of Fourier approximation of $f(x,t)$ as the following

$$||f(\cdot,t) - f_{J,J'}(\cdot,t)||^2_{H^{q_1}_x(\mathbb{R}^d)} = ||f_{>J,J'}(\cdot,t) + f_{J,>J'}(\cdot,t) + f_{>J,>J'}(\cdot,t)||^2_{H^{q_1}_x(\mathbb{R}^d)} \tag{E.144}$$

$$\lesssim \Big(\sum_{|\xi|_\infty > J} \sum_{|\tau| > J'} + \sum_{|\xi|_\infty > J} \sum_{|\tau| \leqslant J'} + \sum_{|\xi|_\infty \leqslant J} \sum_{|\tau| > J'} \Big) \Big(1 + |\xi|^{q_1}\Big)^2 a^2_{\xi,\tau} \phi^2_\tau(t) \Big). \tag{E.145}$$

We then calculate $H^{p_1}_t(\mathbb{R}; H^{q_1}_x(\mathbb{R}^d))$ norm and obtain

$$||f - f_{J,J'}||^2_{H^{p_1}_t(\mathbb{R};H^{q_1}_x(\mathbb{R}^d))} \lesssim \Big(\sum_{|\xi|_\infty > J} \sum_{|\tau| > J'} + \sum_{|\xi|_\infty > J} \sum_{|\tau| \leqslant J'} + \sum_{|\xi|_\infty \leqslant J} \sum_{|\tau| > J'} \Big) \Big(1 + \tau^{p_1}\Big)^2 \Big(1 + \xi^{q_1}\Big)^2 a^2_{\xi,\tau} \Big). \tag{E.146}$$

Using the highest order term τ^{2p_1} and ξ^{2q_1} to bound above equality, we obtain

$$||f - f_{J,J'}||^2_{H_t^{p_1}(\mathbb{R};H_x^{q_1}(\mathbb{R}^d))} \lesssim \Bigg(\sum_{|\xi|_\infty > J} \sum_{|\tau| > J'} + \sum_{|\xi|_\infty > J} \sum_{|\tau| \leqslant J'} + \sum_{|\xi|_\infty \leqslant J} \sum_{|\tau| > J'} \tau^{2p_1} \xi^{2q_1} a^2_{\xi,\tau} \Bigg) \tag{E.147}$$

$$= \Bigg(\sum_{|\xi|_\infty > J} \sum_{|\tau| > J'} + \sum_{|\xi|_\infty > J} \sum_{|\tau| \leqslant J'} + \sum_{|\xi|_\infty \leqslant J} \sum_{|\tau| > J'} \tau^{-2(p-p_1)} \xi^{-2(q-q_1)} \tau^{2p} \xi^{2q} a^2_{\xi,\tau} \Bigg) \tag{E.148}$$

$$\lesssim J^{-2(p-p_1)} J'^{-2(q-q_1)} ||f||^2_{H_t^{p_1}(\mathbb{R};H_x^{q_1}(\mathbb{R}^d))}. \tag{E.149}$$

Finally, we obtain the desired result

$$||f - f_{J,J'}||^2_{H_t^{p_1}(\mathbb{R};H_x^{q_1}(\mathbb{R}^d))} \lesssim J^{-2(p-p_1)} J'^{-2(q-q_1)} ||f||^2_{H_t^{p_1}(\mathbb{R};H_x^{q_1}(\mathbb{R}^d))}. \tag{E.150}$$

■

If we can show $\bar{u}_J \in \mathcal{U}_J$, then it immediately follows from Lemma E.4 and space-time extension Lemma A.3

$$\inf_{u_J \in \mathcal{U}_J} ||u_J - u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} \leqslant ||\bar{u}_J - ext\, u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} \tag{E.151}$$

$$\leqslant ||\bar{u}_J - ext\, u^*||^2_{H_t^1(\mathbb{R};H_x^1(\mathbb{R}^d))} \lesssim R^2 J^{-2(q+2p-3)}. \tag{E.152}$$

It remains to show $\bar{u}_J \in \mathcal{U}_J$. We note extension operator can preserve time-dependent Sobolev norm from Theorem A.3 up to some constant $C(\Omega, T)$. Therefore, $\bar{u}_J$ satisfies boundedness assumption of time-dependent Sobolev norm in assumption A3 in Appendix A.2, and $\bar{u}_J \in \mathcal{U}_J$.

Finally, we incorporate approximation errors into bound of excessive risk in expectation, leading to bias-variance trade-off.

On one hand, when $q \geqslant \frac{d}{2}$,

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} \lesssim_R \left[1 + (1 + \frac{\beta}{2\epsilon}) B(\epsilon, \beta, \Omega, T, d)\right] J^{-2(2p+q-3)} \tag{E.153}$$

$$+ B(\epsilon, \beta, \Omega, T, d)(1 + \frac{\beta}{2\epsilon}) \frac{J^{d+2}}{nm} \log(1 + C_2 nm) + \frac{1}{nm}. \tag{E.154}$$

On other hand, when $q < \frac{d}{2}$,

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} \lesssim_R \left[1 + (1 + \frac{\beta}{2\epsilon}) B(\epsilon, \beta, \Omega, T, d)\right] J^{-2(2p+q-3)} \tag{E.155}$$

$$+ (1 + \frac{\beta}{2\epsilon}) B^2(\epsilon, \beta, \Omega, T, d) \frac{J^{2(d+2)}}{nm} \log(1 + C_2 nm) \tag{E.156}$$

$$+ \frac{1}{B(\epsilon, \beta, \Omega, T, d) nm}. \tag{E.157}$$

Bias decreases as approximation accuracy controlled by J increases, while variance becomes larger when introducing high-frequency statistics information. Therefore, we need to find trade-off between bias and variance to balance variance and bias together with regularization parameter.

Similarly, we also need to balance regularization $B(\epsilon, \beta, \Omega, T)$ and $1 + \frac{\beta}{\epsilon}$ on either variance or bias, and we derive regularization threshold $\beta^* \approx \log(1 + C_2 N)^{\frac{b}{a+2b}} N^{\frac{a}{a+2b}} \beta_0$ in (E.165) of Appendix E.3. We notice β^* is a small constant and therefore regularization falls into the regime of bias regularization for the most of times. If $0 < \beta < \beta^*$, β is in the regime of regularizing variance; if $\beta > \beta^*$, β is in the regime of regularizing bias.

Now, we're ready to prove upper bound on estimation error.

PROOF OF THEOREM 3.1

In this section, we discuss the bias-variance trade-off for two different regularity condition: $q \leqslant \frac{d}{2}$ and $q > \frac{d}{2}$. First, we focus on the regularization regime when $q \leqslant \frac{d}{2}$. According to Lemma E.4 with $J' = J^2$ and assumption A1 of Appendix A.2, we have

$$\inf_{u_J} ||u_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim J^{-2(2p+q-3)} R^2. \tag{E.158}$$

Substituting approximation error (E.158) into Proposition E.1, we obtain the following

When $q < \frac{d}{2}$,

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim_R \underbrace{(1+\frac{\beta}{2\epsilon})B^2(\epsilon,\beta,\Omega,T,d)\frac{J^{2(d+2)}}{nm}\log(1+C_2nm)}_{\text{Variance}} \tag{E.159}$$

$$+ \underbrace{\left[1+(1+\frac{\beta}{2\epsilon})B(\epsilon,\beta,\Omega,T,d)\right]J^{-2(2p+q-3)}}_{\text{Bias}} + \frac{1}{B(\epsilon,\beta,\Omega,T,d)nm}. \tag{E.160}$$

When $q \geqslant \frac{d}{2}$,

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim \underbrace{B(\epsilon,\beta,\Omega,T,d)(1+\frac{\beta}{2\epsilon})\frac{J^{d+2}}{nm}\log(1+C_2nm)}_{\text{bias}} \tag{E.161}$$

$$+ \underbrace{\left[1+(1+\frac{\beta}{2\epsilon})B(\epsilon,\beta,\Omega,T,d)\right]J^{-2(2p+q-3)}R^2}_{\text{variance}} + \frac{1}{nm}. \tag{E.162}$$

E.3 REGIME OF REGULARIZATION ($q < \frac{d}{2}$)

Here we want to find the regimes for the regularization parameter β with respect to variance scaling $J^{2(d+2)}$. Let $N := nm$ denote the total sample size and the excess risk upper bound by . Consider the above excess-risk bound (E.160), we set $a := 2(d+2)$ and $b := 2(2p+q-3)$,

$$\mathcal{E}(J,\beta) \lesssim_R (1+\frac{\beta}{2\epsilon})B^2(\epsilon,\beta,\Omega,T,d)\,\frac{J^a}{N}\log(1+C_2N) \tag{E.163}$$

$$+ \left[1+(1+\frac{\beta}{2\epsilon})B(\epsilon,\beta,\Omega,T,d)\right]J^{-b} + \frac{1}{B(\epsilon,\beta,\Omega,T,d)nm}, \tag{E.164}$$

where the β-dependent variance term is

$$B(\epsilon,\beta,\Omega,T,d) = \Big(\frac{1}{3} + \frac{\beta e^{-T/\varepsilon}}{C^2(d,\varepsilon,\Omega)}\Big)^{-1} = \Big(\frac{1}{3}+\frac{\beta}{\beta_0}\Big)^{-1}, \qquad \beta_0 := C^2(d,\varepsilon,\Omega)\,e^{T/\varepsilon}. \tag{E.165}$$

Thus β decreases the variance term $B(\beta)$, but increases the bias term $\big(1+\beta/(2\varepsilon)\big)$.

E.3.1 OPTIMAL TRUNCATION LEVEL FOR A GIVEN β

Balancing variance and bias in (E.163) yields

$$J_{1,*}(\beta) \asymp \left[\frac{N}{\log(1+C_2N)}\cdot\frac{1}{(1+\frac{\beta}{2\epsilon})B^2(\epsilon,\beta,\Omega,T,d)}\cdot\left(1+(1+\frac{\beta}{2\epsilon})B(\epsilon,\beta,\Omega,T,d)\right)\right]^{\frac{1}{a+b}} \tag{E.166}$$

and the corresponding optimized bound becomes

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim_R \Big(1+\frac{\beta_0}{2\epsilon}\Big)\cdot\Big(\frac{\log(1+C_2N)}{N}\Big)^{\frac{b}{a+b}}\cdot(\frac{\beta_0}{\beta_0+\beta})^{\frac{b}{a+b}} + (1+\frac{\beta}{\beta_0})\frac{1}{N}. \tag{E.167}$$

E.3.2 CASE I (FROM UNREGULARIZED TO MILDLY REGULARIZED REGIME): $0 \leqslant \beta \leqslant \beta_0$

In this regime where energy regularization can only improve the rate up to some constant, we obtain sub-optimal rate

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim_R \left(1 + \frac{\beta_0}{2\epsilon}\right) \cdot \left(\frac{\log(1 + C_2 N)}{N}\right)^{\frac{b}{a+b}}. \tag{E.168}$$

E.3.3 CASE II (OPTIMAL RATE WITH FINE-TUNED β): $\beta = \beta_*$

when we balance β, we obtain optimal rate

$$\beta^* \approx \log(1 + C_2 N)^{\frac{b}{a+2b}} N^{\frac{a}{a+2b}} \beta_0 \tag{E.169}$$

and plugging β^* back, we obtain the following rate

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim_R \left(\frac{\sqrt{\log(1 + C_2 N)}}{N}\right)^{\frac{2b}{a+2b}}. \tag{E.170}$$

In this regime, regularization reduces the variance with mildly inflating the bias term.

E.3.4 CASE III (VARIANCE-IMPROVING REGIME): $\beta_0 < \beta < \beta^*$

In this scenario, variance is reduced to improve the rate

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim_R \left(1 + \frac{\beta_0}{2\epsilon}\right) \cdot \left(\frac{\log(1 + C_2 N)}{N}\right)^{\frac{b}{a+b}} \cdot (\frac{\beta_0}{\beta_0 + \beta})^{\frac{b}{a+b}}. \tag{E.171}$$

E.3.5 CASE V (BIAS-INFLATING REGIME): $\beta^* < \beta < \infty$

In this scenario, variance is inflated to make the rate worse

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim_R (1 + \frac{\beta}{\beta_0})\frac{1}{N}. \tag{E.172}$$

Next we consider the regularization regime where $q > \frac{d}{2}$. We take similar step to set variance equal to bias

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim \left[1 + (1 + \frac{\beta}{2\epsilon})B(\epsilon, \beta, \Omega, T, d)\right] J^{-2(2p+q-3)} \tag{E.173}$$

$$+ B(\epsilon, \beta, \Omega, T, d)(1 + \frac{\beta}{2\epsilon})\frac{J^{d+2}}{nm} \log(1 + C_2 nm) + \frac{1}{nm}. \tag{E.174}$$

E.4 REGIME OF REGULARIZATION ($q \geqslant \frac{d}{2}$)

when $q > \frac{d}{2}$, variance scaling becomes $J^{2(d+2)}$. With the same notation as before, we set $N := nm$ the total sample size, $\frac{a}{2} := (d + 2)$, and $b := 2(2p + q - 3)$,

$$\mathcal{E}(J, \beta) \lesssim B(\epsilon, \beta, \Omega, T, d)(1 + \frac{\beta}{2\epsilon})\frac{J^{\frac{a}{2}}}{N} \log(1 + C_2 N) + \left[1 + (1 + \frac{\beta}{2\epsilon})B(\epsilon, \beta, \Omega, T, d)\right] J^{-b} + \frac{1}{nm}. \tag{E.175}$$

E.4.1 OPTIMAL TRUNCATION LEVEL FOR A GIVEN β

Balancing variance and bias in (E.175) yields

$$J_{2,*}(\beta) \asymp \left[\frac{N}{\log(1 + C_2 N)} \cdot \frac{1 + \frac{\beta}{\beta_0}}{1 + \frac{\beta}{2\epsilon}} \cdot \left(1 + \frac{1 + \frac{\beta}{2\epsilon}}{1 + \frac{\beta}{\beta_0}}\right)\right]^{\frac{2}{a+2b}} \tag{E.176}$$

and the corresponding optimized bound becomes

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim_R \left(\frac{\log(1 + C_2 N)}{N}\right)^{\frac{2b}{a+2b}} \cdot (1 + \frac{1 + \frac{\beta}{2\epsilon}}{1 + \frac{\beta}{\beta_0}})^{\frac{2b}{a+2b}} \cdot \left(\frac{1 + \frac{\beta}{2\epsilon}}{1 + \frac{\beta}{\beta_0}}\right)^{\frac{a}{a+2b}} + \frac{1}{N}. \tag{E.177}$$

For any $\beta > 0$, the rate remains optimal up to some constant

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H^1_t(0,T;H^1_x(\Omega))} \lesssim_R \left(\frac{\log(1+C_2N)}{N}\right)^{\frac{2b}{a+2b}} \cdot (1+\frac{\beta_0}{2\epsilon})^{\frac{a}{a+2b}} \cdot \left(\frac{\beta_0}{2\epsilon}\right)^{\frac{2b}{a+2b}}. \tag{E.178}$$

Finally, we wrap up our proof of convergence rates in (E.3) and (E.4) by using scaling relation $\beta = \lambda \cdot nm$ between scaled regularization parameter β and unscaled regularization parameter λ in Section 3 to change β back to λ. ■

E.5 Sampling Balance

Next, we will show it is necessary to balance space-time sample complexity using parabolic anisotropic balancing as reflected in multi-level truncation J for space and J^2 for time.

Proposition E.5 (Sampling Balance)**.** *Given functional space V_{J,J^2} defined in A.4, the evaluation map of our empirical loss $\hat{L}_{\epsilon,\beta}$ is injective only if*

$$n \geqslant J^2. \tag{E.179}$$

Furthermore, it is necessary to require

$$n \gtrsim m^{2/d} \tag{E.180}$$

up to logarithmic factor for theorem 3.1 hold.

Proof. Lower bound on the temporal sample size. We first show lower bound on temporal sample complexity n. Given temporal samples $t_1, \ldots, t_n$, it is enough to look at the evaluation map $ev: u \in V_{J,J^2} \longmapsto \left(\partial_t^k D_x^\alpha u(X_{ij}, t_j)\right)_{i\leqslant m,\ j\leqslant n}$. It is clear the evaluation mapping ev is linear on the finite-dimensional space V_{J,J^2}, so it fails to be injective if its kernel is nontrivial.We will show the contrapositive is true.

We choose $\psi = \sum_{\tau=1}^K c_\tau \phi_\tau \in \Psi_K$ with its coefficient vector $c \in \mathbb{R}^K$, with $\psi^{(k)}(t_j) = 0$ and $\psi(0) = 0$, for $1 \leqslant j \leqslant n, 0 \leqslant k \leqslant r$. It forms a homogeneous linear system $Ac = 0$ with evaluation matrix

$$A = \begin{pmatrix} \phi_1(t_1) & \phi_2(t_1) & \cdots & \phi_K(t_1) \\ \phi_1'(t_1) & \phi_2'(t_1) & \cdots & \phi_K'(t_1) \\ \vdots & \vdots & & \vdots \\ \phi_1(t_n) & \phi_2(t_n) & \cdots & \phi_K(t_n) \\ \phi_1'(t_n) & \phi_2'(t_n) & \cdots & \phi_K'(t_n) \\ \phi_1(0) & \phi_2(0) & \cdots & \phi_K(0) \end{pmatrix} \in \mathbb{R}^{\left(2n+1\right)\times K}$$

. The matrix has $2n+1$ rows, so rank$A \leqslant 2n+1 < K$, and the rank–nullity theorem gives

$$\dim \ker A \ = \ K - \text{rank}A \ \geqslant \ K - 2n - 1 \ \geqslant \ 1$$

The corresponding $\psi \in \Psi_K \setminus \{0\}$ vanishes, together with its first r derivatives, at every temporal position t_j. Pick any spatial basis element ϕ_ξ with $|\xi|_\infty \leqslant J$ and set $v := \psi\, \phi_\xi \in V_{J,J^2} \setminus \{0\}$; by the tensor structure, for all $0 \leqslant k \leqslant r$ and $\alpha \in \mathbb{N}^d$,

$$\partial_t^k D_x^\alpha v\,(X_{ij}, t_j) = \psi^{(k)}(t_j)\, D_x^\alpha \phi_\xi(X_{ij}) = 0,$$

so $v \in \ker \text{ev}$ and ev is not injective. Since $K \asymp J^2$, (E.179) follows.

Next, we show upper bound on the spatial sample size m. Using $n \geqslant J^2$, we fix ϕ_τ at the temporal level $2^\ell \asymp J^2$ and ϕ_ξ with $|\xi|_\infty \leqslant J$, $supp\, \phi_\xi \subset \Omega^\circ$, and consider the squared basis direction $h_J := a_J^2 \phi_\xi^2 \phi_\tau^2$ in the loss class, with admissible amplitude a_J. We deonote P_X, P_T for the uniform laws on Ω and $[0,T]$, and $\widehat{P}$, P as the empirical and population means on Ω_T, decompose

$$(\widehat{P} - P)h_J = W_J + B_J, \tag{E.181}$$

$$W_J := \frac{1}{n}\sum_{j=1}^n \Big[\frac{1}{m}\sum_{i=1}^m h_J(X_{ij}, t_j) - P_X h_J(\cdot, t_j)\Big], \tag{E.182}$$

$$B_J := \frac{1}{n}\sum_{j=1}^n P_X h_J(\cdot, t_j) - P h_J. \tag{E.183}$$

Define the events

$$E_1 := \Big\{|B_J| \geqslant c_1 a_J^2 \sqrt{J^2/n}\Big\}, \qquad E_2 := \Big\{|W_J| < \tfrac{c_1}{2}\, a_J^2 \sqrt{J^2/n}\Big\}.$$

By lemma F.11,

$$\mathbb{E}W_J^2 \lesssim\ a_J^4\, \frac{J^{d+2}}{nm}, \qquad \mathbb{P}(E_1) \geqslant c_2$$

using $\int \phi_\xi^4 \asymp J^d$ and $\int \phi_\tau^4 \asymp J^2$ and Chebyshev's inequality,

$$\mathbb{P}(E_2^c)\ \leqslant\ \frac{4\,\mathbb{E}W_J^2}{c_1^2\, a_J^4\, J^2/n}\ \lesssim\ \frac{J^d}{m}\ \leqslant\ \frac{c_2}{2} \qquad \text{once } m \geqslant C_0 J^d$$

for a sufficiently large constant $C_0(c_1, c_2)$. Hence

$$\mathbb{P}(E_1 \cap E_2)\ \geqslant\ \mathbb{P}(E_1) - \mathbb{P}(E_2^c)\ \geqslant\ c_2 - \frac{c_2}{2}\ =\ \frac{c_2}{2},$$

and on $E_1 \cap E_2$ the reverse triangle inequality gives

$$\big|(\widehat{P} - P)h_J\big|\ =\ |B_J + W_J|\ \geqslant\ |B_J| - |W_J|\ >\ c_1 a_J^2 \sqrt{J^2/n} - \tfrac{c_1}{2}\, a_J^2 \sqrt{J^2/n}\ =\ \tfrac{c_1}{2}\, a_J^2 \sqrt{J^2/n}\,.$$

The direction h_J lies in the localized fixed-J class over which our inequality bounds the empirical process by $C\, a_J^2 \sqrt{J^{d+2} \log(1 + C_2 N)/N}$ if

$$\sqrt{\frac{J^2}{n}}\ \lesssim\ \sqrt{\frac{J^{d+2} \log(1 + C_2 nm)}{nm}},$$

This implies $m\ \lesssim\ J^d \log(1 + C_2 nm)$, and furthermore $n\ \gtrsim\ J^2\ \gtrsim\ \Big(\frac{m}{\log(1+C_2 nm)}\Big)^{2/d}$, ■

E.6 Lower bound on estimation error

Proof of Theorem 3.3. The main idea of our proof directly refers to standard construction tools from Tsybakov (2009) with additional tool from PDE theory Theorem B.3. By definition of $\mathcal{U}$ and u^* in (A.4), for any ground truth solution $u^* \in \mathcal{U}$, we can always find data $f = Lu^* \in \mathcal{F}$ with which we can construct estimator for $\hat{u}$. The construction is following: We follow our convention to treat time domain as another dimension so we write $d+1$ hypercube as $\mathcal{D}_T = [0,1]^d \times [0,1]$. According to regularity theory (Theorem B.1), given forcing function $f \in \mathcal{F}$ and initial condition $u_0^* \in H^{2s+1}([0,1]^d)$, there exists solution $h \in \mathcal{U}$ defined in (A.4) corresponding to this f. Our h here is not necessarily a smooth function in the regular sense, but it is smooth in the Sobolev sense and its smoothness is determined by the functional space of forcing $\mathcal{F}$ and the initial data u_0^*. In particular, h has the following smoothness property

$$\frac{d^p h}{dt^p} \in L^2(0,1; H^{2s+2-2p}([0,1]^d)), 0 \leqslant p \leqslant s,\ s \geqslant 1, s, q \in \mathbb{Z}. \tag{E.184}$$

This implies $\frac{dh}{dt}$ and $\triangle h$ exists in $L^2_{t,x}([0,1]^d \times [0,1])$ sense and $h \in H^1(0,T; H^1([0,1]^d)) := H_t^1 H_x^1$. Moreover, since f and is non-zero, then h can't be identitically zero for any $(x,t) \in [0,1]^d \times [0,1]$. For f_0, there exists unique solution $u_{f_0} := h$.We also note the difference between f_0 and other f_k is smooth perturbation, because its convolution with heat kernel, by Duhamel's Principle B.3, is smooth. Therefore, we then pick some function $g(x,t)$ for this perturbation such that $\frac{\partial g}{\partial t}$ and ∇g are not identically zero.

Let r be the number of grid size on hypercube $[0,1]^d$ to be determined later. According to scaling invariance property Corollary B.5 of h, $g(x,t)$ is a solution to heat equation if only if $g(\frac{x}{r}, \frac{t}{r^2})$ is a solution to heat equation. Therefore, we should have r^2 number of grid on unit time horizon $[0,1]$ while r number of grid on spacial hypercube $[0,1]^d$ so that temporal scaling by $\frac{1}{r^2}$ can match with spacial scaling $\frac{1}{r}$ in upcoming construction of hypothesis test.

We define our hypothesis as the following

$$u_{f_k}(x,t) = h(x,t) + \sum_{j\in[r^2]}\sum_{i\in[r]^d} c\,\tau_{i,j}^{(k)}\, w^{2p+q} g(\frac{x-x^i}{w}, \frac{t-t^j}{w^2}),\ w=\frac{1}{2r},\ k=1,...,2^{\frac{r^{d+2}}{8}}, \tag{E.185}$$

$$u_{f_0}(x,t) = h(x,t), \tag{E.186}$$

where c, λ are some constants.

It is readily to check the collection of functions $\{g(\frac{x-x^i}{w}, \frac{t-t^j}{w^2})\}_{i,j}$ has disjoint compact support for pair (i,j) and $h(\frac{x-x^i}{w}, \frac{t-t^j}{w^2})$ satisfies heat equation given initial condition u_0^* and f. Therefore, u_k satisfies the heat equation as well.

We discretize the space $[0,1]^d$ by defining a collection of vectors $\{\, x^i : x^i = (x_1^i, ..., x_d^i), i \in [r]^d \,\} \subset [0,1]^d$ to have coordinates $x_l^i := \frac{i_l - 1}{r}$, $i_l \in [r]$, $l = 1, ..., d$. Again, we similarly define another collection of numbers $\{t^j : t^j = \frac{j-1}{r^2},\ j \in [r^2]\} \subset [0,1]$. According to the Varshamov-Gilbert lemma, there exists $\frac{r^{d+2}}{8}$ many 0-1 sequences $\tau^{(1)}, ..., \tau^{(2^{r^{d+2}/8})} \in \{0,1\}^{r^{d+2}}$ such that $|| \tau^k - \tau^{k'} ||^2 \geqslant \frac{r^{d+2}}{8}, k \neq k'$. Then we cast solving parabolic PDEs Problem into high-dimensional multiple hypothesis testing on hypercube $[0,1]^d \times [0,1]$ which considers all mappings Φ from nm sampled data on domain Ω_T to the hypothesis

$$\Phi : (x,t) \in (\mathbb{R})^{d+1^{\otimes nm}} \times \mathbb{R}^{\otimes nm} \to U := \big\{ u_k(x,t) \mid k = 1, 2, ..., 2^{\frac{r^{d+2}}{8}} \big\}. \tag{E.187}$$

We recall Fano's inequality

$$P(\hat{u} \neq u) \geqslant 1 - \frac{I(U;X) + \log 2}{\log(|U|)}, \qquad \text{for any estimator } \hat{u}. \tag{E.188}$$

We obtain a constant lower bound of $P(u_k' \neq u_k)$ by bounding KL-divergence from above

$$I(U;X) \leqslant \frac{1}{|U|} \sum_{(x,t)} \sum_{u_k \neq u_0} D_{KL}(\mathbb{P}_{u_k} || \mathbb{P}_{u_0}). \tag{E.189}$$

where P_{u_k} is a joint distribution of the sampled data $((x,t), f_k(x,t))$ for hypothesis u_k corresponding to noisy observation of force function f_k, and $(x,t) \sim P_0$, P_0 is a product of Lebesgue measure on $[0,1]^d$ and Lebesgue measure on $[0,1]$; $y_k = f_k(x,t) + \epsilon$ where ϵ is an i.i.d Gaussian noise for each location point (x,t), and so $y_k \sim Q_k$, Q_k is probability measure of standard normal random variable with mean f_k. Therefore, we can write $W_{f_k} := P_0 \otimes Q_k$, where P_0 is independent from Q_k across tensor product

Note $\triangle u_k - \frac{\partial u}{\partial t} = f_k$ and $\triangle u_0 - \frac{\partial u_0}{\partial t} = \triangle h - \frac{\partial h}{\partial t} = f_0$.

$$D_{KL}(W_{f_k} || W_{f_0}) = D_{KL}(P_0 \otimes Q_k || P_0 \otimes Q_0) \tag{E.190}$$

$$= \mathbb{E} \log(\frac{dQ_k}{dQ_0}) \tag{E.191}$$

$$= ||f_k - f_0||^2_{L^2_{t,x}([0,1]\times[0,1]^d)} \tag{E.192}$$

$$= ||(\triangle u_{f_k} - \frac{\partial u_{f_k}}{\partial t}) - (\triangle u_{f_0} - \frac{\partial u_{f_0}}{\partial t})||^2_{L^2_{t,x}([0,1]\times[0,1]^d)} \tag{E.193}$$

$$= \Big|\Big| \sum_{i\in[r^{d+1}]} c\,\tau_i^{(k)}\, w^{2p+q-1} \Big(g_t(\frac{x-x^i}{w}, \frac{t-t^j}{w^2}) - \triangle g(\frac{x-x^i}{w}, \frac{t-t^j}{w^2}) \Big) \Big|\Big|^2_{L^2_{t,x}([0,1]\times[0,1]^d)} \tag{E.194}$$

$$\leqslant c^2 w^{2(2p+q-3)} \big|\big| g_t - \triangle g \big|\big|^2_{L^2_{t,x}([0,1]\times[0,1]^d)} \tag{E.195}$$

$$\lesssim \frac{c^2}{r^{2(2p+q-3)}}. \tag{E.196}$$

Next, we want to use energy estimate from Theorem 5 in chapter 7.1 of Evans (2010), which upper bounds u_k using f_k. Note $u(0,x) = 0$ in our setting, so we have

$$||u_k - u_0||^2_{L^2([0,1];L^2([0,1]^d))} \leqslant C||f_k - f_0||^2_{L^2([0,1];L^2([0,1]^d))} \lesssim \frac{c^2}{r^{2(2p+q-3)}}. \tag{E.197}$$

With $r = [(nm)^{\frac{1}{2(2p+q-3)+d+2}}]$,

$$\mathbb{P}(u_k \neq u_k') \geqslant 1 - \frac{nmD_{KL}(\mathbb{P}_u||\mathbb{P}_{u'})}{\log(|U|)} \geqslant 1 - \frac{\frac{8c^2}{r^{2(2p+q-3)}}}{r^{d+2}\log 2} = 1 - \frac{8c^2}{r^{2(2p+q-3)+d+2}\log 2}. \quad \text{(E.198)}$$

We then choose large c and λ such that $\mathbb{P}(u_k \neq u_k') \geqslant \frac{1}{2}$.Finally, we check separation condition under the evaluation metric $H_t^1 H_x^1$ for hypothesis u_k and u_k'

$$||u_k - u_k'||^2_{H_t^1 H_x^1} = H(\tau^k, \tau^{k'})c^2 w^{2(2p+q-3)} || \, g(\frac{x-x^i}{w}, \frac{t-t^j}{w^2})) \, ||^2_{L^2_{t,x}([0,1]^d \times [0,1])} w^{d+2} \gtrsim r^{-2(2p+q-3)}. \quad \text{(E.199)}$$

Plugging in $r = [(nm)^{\frac{1}{2(2p+q-3)+d+2}}]$ and $2p+q = 2s+2$, we obtain

$$\inf_{\hat{u}} \sup_{u^* \in H^s(\Omega)} \big|\big| \, \hat{u} - u^* \, \big|\big|^2_{H_t^1(0,1;H_x^1([0,1]^d)} \gtrsim (nm)^{-\frac{2(2p+q-3)}{2(2p+q-3)+d+2}} = (nm)^{-\frac{4s-2}{4s+d}}. \quad \text{(E.200)}$$

■

E.7 Excess Risk Bound

Proof of Theorem 3.4. We fix $\epsilon > 0$, $s \geqslant 1$ and $2p+q = 2s+2$ for $0 \leqslant p \leqslant s$. According to Theorem 3.1, regardless of the relation between p and d, as long as we choose

$$\lambda^* := \log(1 + C_2 nm)^{\frac{4s-2}{4s+d}} (nm)^{-\frac{2+d}{4s+d}} e^{T/\epsilon}. \quad \text{(E.201)}$$

Our estimation rate becomes optimal as follows

$$\mathbb{E}||\hat{u}_J - u^*||^2_{H_t^1(0,T;H_x^1(\Omega))} \lesssim \Big(\frac{\log(1+C_2 nm)}{nm}\Big)^{\frac{4s-2}{4s+d}}. \quad \text{(E.202)}$$

Thanks to the stability estimate in Proposition 2.2, excess risk is upper bounded by estimation error

$$\mathbb{E} \; L_{\epsilon,\beta}(\hat{u}_J) - L_{\epsilon,\beta}(u^*) \lesssim \frac{e^{T/\epsilon}}{\epsilon}\Big(\frac{\log(1+C_2 nm)}{nm}\Big)^{\frac{4s-2}{4s+d}}. \quad \text{(E.203)}$$

■

E.8 Generalization Error Bound

To prove generalization error bound, We define functional classes that are closely associated with our loss function $L_{\epsilon,\beta}(u)$

$$\begin{aligned}
\mathcal{G}_{\text{data}} &:= \Big\{ g \,\Big|\, g = u - u^* + \nabla(u-u^*) + \partial_t(u-u^*) + \nabla\partial_t(u-u_t^*) \Big\}, \\
\mathcal{G}_{\text{energy}} &:= \Big\{ g \,\Big|\, g = \frac{e^{-t/\epsilon}}{2}|u_t|^2 + \frac{1}{2\epsilon}|\nabla u|^2 - \frac{e^{-t/\epsilon}}{\epsilon} f u \Big\}, \\
\mathcal{G}_{\text{correction}} &:= \Big\{ g \,\Big|\, g = e^{-T/\epsilon} u_t^*(x,T) u(x,T) + e^{-t/\epsilon} u_{tt}^* u \Big\}.
\end{aligned} \quad \text{(E.204)}$$

where $u^*, u \in \mathcal{U}$ and $u_t^*(x,T)$ and $u(x,T)$ are evaluated at $t = T$

We relate the above functional classes $\mathcal{G}_{data}$, $\mathcal{G}_{energy}$ and $\mathcal{G}_{correction}$ to their Rademacher Complexities $\mathcal{R}$

Lemma E.6. *Given the functional spaces defined above, the generalization error bound is upper bound by their Rademacher Complexities as follows*

$$\mathbb{E}\Delta\mathcal{E}_{S,gen} \leqslant 2[\,\mathcal{R}_N(\mathcal{G}_{data}) + \mathcal{R}_N(\mathcal{G}_{energy}) + \mathcal{R}_N(\mathcal{G}_{correction})\,]. \quad \text{(E.205)}$$

Proof of Lemma E.6. Fix $\epsilon > 0$ and $\beta = \beta_{reg}$ and denote $N = nm$.

$$\mathbb{E}\Delta\mathcal{E}_{S,gen} \leqslant \mathbb{E}\sup_{u\in\mathcal{U}}|L_{\epsilon,\beta}(\hat{u}_S) - \hat{L}_{\epsilon,\beta}(\hat{u}_S)| \tag{E.206}$$

$$\leqslant \mathbb{E}\sup_{u\in\mathcal{U}}|\mathcal{E}_{data} - \mathcal{E}_{data,n}| + \mathbb{E}\sup_{u\in\mathcal{U}}|\mathcal{E}_{energy} - \mathcal{E}_{energy,n}| + \mathbb{E}\sup_{u\in\mathcal{U}}|\mathcal{E}_{data} - \mathcal{E}_{data,n}| \tag{E.207}$$

$$\leqslant \mathbb{E}\left|\frac{1}{nm}\sum_{i=1,j=1}^{nm} g_{data}(x_{ij},t_j) - \mathbb{E}[g_{data}]\right| \tag{E.208}$$

$$+ \mathbb{E}\left|\frac{1}{nm}\sum_{i=1,j=1}^{nm} g_{correction}(x_{ij},t_j) - \mathbb{E}[g_{correction}]\right| \tag{E.209}$$

$$+ \mathbb{E}\left|\frac{1}{nm}\sum_{i=1,j=1}^{nm} g_{energy}(x_{ij},t_j) - \mathbb{E}[g_{energy}]\right|. \tag{E.210}$$

Thanks to Lemma $F.2$, we have

$$\mathbb{E}\Delta\mathcal{E}_{S,gen} \leqslant 2\mathcal{R}_N(\mathcal{G}_{data}) + 2\mathcal{R}_N(\mathcal{G}_{energy}) + 2\mathcal{R}_N(\mathcal{G}_{correction}). \tag{E.211}$$

Next we apply Dudley's theorem (cf. Theorem F.3) to bound the ϵ_1-covering number of the $\mathcal{R}_N(\mathcal{G}_{data})$, $\mathcal{R}_N(\mathcal{G}_{energy})$, and $\mathcal{R}_N(\mathcal{G}_{correction})$. ■

According to the Lemma E.6, it is essential to bound Rademacher complexities to control the generalization error bound $\mathcal{E}_{S,gen}$. The idea is to bound the Rademacher complexities in terms of the metric entropy by using the well-known Dudley theorem Dudley (1967).

Lemma E.7. *Let $\mathcal{N}(\mathcal{F},\rho,\epsilon_1)$ be ϵ_1-covering number for functional space $\mathcal{F}$ under evaluation metric ρ. When $\rho := ||\cdot||_{L^2_t(0,T;L^2_x(\Omega))}$, then*

$$\log N(\mathcal{G}_{data},\rho,\epsilon_1) \lesssim J^{d+2}\left[\log\frac{1}{\epsilon_1} + 2d\log J\right], \tag{E.212}$$

$$\log N(\mathcal{G}_{energy},\rho,\epsilon_1) \lesssim_{R,F} J^{d+2}\left[\log\frac{\beta}{\epsilon\cdot\epsilon_1} + 4dq\log J\right], \tag{E.213}$$

$$\log N(\mathcal{G}_{correction},\rho,\epsilon_1) \lesssim_{R,T} J^{d+2}\left[\log\frac{\beta}{\epsilon\cdot\epsilon_1} + 4dq\log J\right]. \tag{E.214}$$

Next we use covering numbers in Lemma E.7 along with Dudley theorem to prove the following theorem.

Proof of Lemma E.7. Our estimator is defined as $\hat{u} := \hat{u}_\theta = \hat{u}_J$, which is discussed in (2.7) and (3.1)). Before we dive into our proof, we want to define the following relevant functional spaces:

$$\mathcal{A} := \{\, a \mid a := a_{\xi,\tau}, ||a_{\xi,\tau}||_2 \leqslant R\,\}, \tag{E.215}$$

$$\mathcal{G}_{data,1} := \{\, g \mid g := \nabla u_t - \nabla u_t^*, u^*, u \in \mathcal{U}\,\}, \tag{E.216}$$

$$\mathcal{G}_{energy,1} := \{\, g \mid g := \frac{\beta\cdot e^{-\frac{t}{\epsilon}}}{2}|u_t|^2, u \in \mathcal{U}_J\,\}, \tag{E.217}$$

$$\mathcal{G}_{energy,2} := \{\, g \mid g := \frac{\beta\cdot e^{-\frac{t}{\epsilon}}}{2\epsilon}|\nabla u|^2, u \in \mathcal{U}_J\,\}, \tag{E.218}$$

$$\mathcal{G}_{energy,2} := \{\, g \mid g := -\frac{\beta\cdot e^{-\frac{t}{\epsilon}}}{\epsilon} f u, u \in \mathcal{U}_J, \sup_{t,x}|f| \leq F\,\}, \tag{E.219}$$

$$\mathcal{G}_{correction,1} := \{\, g \mid g := \beta e^{-\frac{T}{\epsilon}} u_t^*(x,T)u(x,T),\ u^*, u \in \mathcal{U}_J\,\}, \tag{E.220}$$

$$\mathcal{G}_{correction,2} := \{\, g \mid g := \beta e^{-\frac{t}{\epsilon}} u_{tt}^* u,\ u^*, u \in \mathcal{U}_J\,\}. \tag{E.221}$$

We first obtain covering number for $\mathcal{G}_{data}$. It is sufficient to focus on $\nabla\hat{u}_t - \nabla u_t^*$ in data fitting term because it contains largest metric entropy compared to other terms due to loss of space-time

derivatives.

$$\left|(\nabla\hat{u}_t^1-\nabla u_t^*)-(\nabla\hat{u}_t^2-\nabla u_t^*)\right|=\left|\nabla\hat{u}_t^1-\nabla\hat{u}_t^2\right|\leqslant\sum_{1\leqslant|\xi|,\tau\leqslant J}|a_{\xi,\tau}^1-a_{\xi,\tau}^2|\cdot|\xi|\cdot|\tau|\cdot\phi_{\xi,\tau}(x,t). \tag{E.222}$$

According to orthogonality property of local basis $\phi_{\xi,\tau}(x,t)$, we have

$$||\nabla\hat{u}_t^1-\nabla\hat{u}_t^2||^2_{L_t^2(\mathbb{R};L_x^2(\mathbb{R}^d))}\leqslant\sum_{1\leqslant|\xi|_\infty\leqslant J,1\leqslant\tau\leqslant J^2}|a_{\xi,\tau}^1-a_{\xi,\tau}^2|^2\cdot|\xi|^2\cdot|\tau|^2 \tag{E.223}$$

$$\leqslant J^{2(d+2)}\sum_{1\leqslant|\xi|_\infty\leqslant J,1\leqslant\tau\leqslant J^2}|a_{\xi,\tau}^1-a_{\xi,\tau}^2|^2. \tag{E.224}$$

The inequality above suggests a control of finitely many spatial-temporal cofficients $a_{\xi,\tau}\in\mathcal{A}$ in L^2 ball of coefficient space is sufficient to construct an ϵ_1-net for $g\in\mathcal{G}_{data,1}$.Let $\delta:=\frac{\epsilon_1}{J^{d+1}}$, we want to construct L^2 packing number for the space of finitely many Fourier coefficients $a_{\xi,\tau}$. Using standard L^2-packing for L^2 balls argument in Lemma F.6 and note the dimension of coefficients of local basis representation are at most J^{d+1}, we have

$$\log N(\mathcal{A},||\cdot||_2,\delta)\lesssim\log(\frac{J^{d+2}}{\epsilon_1})^{J^{d+2}}=J^{d+2}\left[\log\frac{1}{\epsilon_1}+(d+2)\log J\right]. \tag{E.225}$$

In this way, we obtain δ-covering number for the functional space $\mathcal{G}_{data,1}$

$$\log N(\mathcal{G}_{data,1},||\cdot||_{L_t^2(0,T;L_x^2(\Omega))},\epsilon_1)\lesssim\log N(\mathcal{A},||\cdot||_2,\delta)\lesssim J^{d+2}\left[\log\frac{1}{\epsilon_1}+(d+2)\log J\right]. \tag{E.226}$$

Next we omit the same procedure applied to $\mathcal{G}_{energy}$ and $\mathcal{G}_{correction}$. In similar fashion, we have the following elementary estimates

$$|g_{energy,1}|=\left|\frac{\beta\cdot e^{-\frac{t}{\epsilon}}}{2}|\hat{u}_t^1|^2-\frac{\beta\cdot e^{-\frac{t}{\epsilon}}}{2}|\hat{u}_t^2|^2\right|^2\leqslant\frac{\beta^2}{4}\left|\hat{u}_t^1-\hat{u}_t^2\right|^2\left(\frac{3}{2}|\hat{u}_t^1|^2+\frac{3}{2}|\hat{u}_t^2|^2\right), \tag{E.227}$$

$$|g_{energy,2}|=\left|\frac{\beta\cdot e^{-\frac{t}{\epsilon}}}{2\epsilon}|\nabla\hat{u}^1|^2-\frac{\beta\cdot e^{-\frac{t}{\epsilon}}}{2\epsilon}|\nabla\hat{u}_t^2|^2\right|^2\leqslant\frac{\beta^2}{4\epsilon^2}\left|\nabla\hat{u}^1-\nabla\hat{u}^2\right|^2\left(\frac{3}{2}|\nabla\hat{u}^1|^2+\frac{3}{2}|\nabla\hat{u}^2|^2\right), \tag{E.228}$$

$$|g_{energy,3}|=\left|\frac{\beta\cdot e^{-\frac{t}{\epsilon}}}{\epsilon}fu^1-\frac{\beta\cdot e^{-\frac{t}{\epsilon}}}{\epsilon}fu^2\right|^2\leqslant\frac{\beta^2}{\epsilon^2}\cdot f\cdot\left|\hat{u}^1-\hat{u}^2\right|^2. \tag{E.229}$$

Thanks to Sobolev Multiplication theorem Theorem F.7, when $q\geqslant\frac{d}{2}+1$, we can obtain for any t,

$$||g_{energy,1}(\cdot,t)||_{L_x^2(\mathbb{R}^d)}\leqslant\frac{3\beta}{4}||\hat{u}_t^1(\cdot,t)-\hat{u}_t^2(\cdot,t)||_{H_x^q(\mathbb{R}^d)}(||\hat{u}_t^1(\cdot,t)||_{H_x^q(\mathbb{R}^d)}+||\hat{u}_t^2(\cdot,t)||_{H_x^q(\mathbb{R}^d)}), \tag{E.230}$$

$$||g_{energy,2}(\cdot,t)||_{L_x^2(\mathbb{R}^d)}\leqslant\frac{3\beta}{4\epsilon}||\hat{u}^1(\cdot,t)-\hat{u}^2(\cdot,t)||_{H_x^q(\mathbb{R}^d)}(||\hat{u}^1(\cdot,t)||_{H_x^q(\mathbb{R}^d)}+||\hat{u}^2(\cdot,t)||_{H_x^q(\mathbb{R}^d)}). \tag{E.231}$$

Applying uniform control of Sobolev norm in time t and orthogonality property of local basis $\phi_{\xi,\tau}(x,t)$, we have

$$||g_{energy,1}||^2_{L_t^2(\mathbb{R};L_x^2(\mathbb{R}^d))}\leqslant\frac{9\cdot R^2\cdot\beta^2}{16}||\hat{u}^1-\hat{u}^2||^2_{H_t^1(\mathbb{R};H_x^q(\mathbb{R}^d))} \tag{E.232}$$

$$\leqslant\frac{9\cdot R^2\cdot\beta^2}{16}J^{2+4dq}\sum_{1\leqslant|\xi|_\infty,\tau\leqslant J}|a_{\xi,\tau}^1-a_{\xi,\tau}^2|^2, \tag{E.233}$$

$$||g_{energy,2}||^2_{L^2_t(\mathbb{R};L^2_x(\mathbb{R}^d))} \leqslant \frac{9\cdot R^2\cdot\beta^2}{16\epsilon^2}||\hat{u}^1-\hat{u}^2||^2_{L^2_t(\mathbb{R};H^q_x(\mathbb{R}^d))} \tag{E.234}$$

$$\leqslant \frac{9\cdot R^2\cdot\beta^2}{16\epsilon^2}J^{4dq}\sum_{1\leqslant|\xi|_\infty,\tau\leqslant J}|a^1_{\xi,\tau}-a^2_{\xi,\tau}|^2 \tag{E.235}$$

$$, \tag{E.236}$$

$$||g_{energy,3}||^2_{L^2_t(\mathbb{R};L^2_x(\mathbb{R}^d))} \leqslant \frac{\beta^2}{\epsilon^2}\cdot F^2||\hat{u}^1-\hat{u}^2||^2_{L^2_t(\mathbb{R};L^2_x(\mathbb{R}^d))} \tag{E.237}$$

$$\leqslant \frac{\beta^2}{\epsilon^2}\cdot F^2J^{2d+4}\sum_{1\leqslant|\xi|_\infty,\tau\leqslant J}|a^1_{\xi,\tau}-a^2_{\xi,\tau}|^2. \tag{E.238}$$

It follows as before

$$\log N(\mathcal{G}_{energy,1},||\cdot||_{L^2_t(0,T;L^2_x(\Omega))},\epsilon_1)\lesssim J^{d+2}\bigg[\log\frac{3R\cdot\beta}{4\epsilon_1}+(2+dq)\log J\bigg], \tag{E.239}$$

$$\log N(\mathcal{G}_{energy,2},||\cdot||_{L^2_t(0,T;L^2_x(\Omega))},\epsilon_1)\lesssim J^{d+2}\bigg[\log\frac{3R\cdot\beta}{4\epsilon\cdot\epsilon_1}+dq\log J\bigg], \tag{E.240}$$

$$\log N(\mathcal{G}_{energy,3},||\cdot||_{L^2_t(0,T;L^2_x(\Omega))},\epsilon_1)\lesssim J^{d+1}\bigg[\log\frac{F\cdot\beta}{\epsilon\cdot\epsilon_1}+(d+2)\log J\bigg]. \tag{E.241}$$

we then bound $\mathcal{G}_{correction}$

$$|g_{correction,1}|=\bigg|e^{-\frac{T}{\epsilon}}u^*_t(x,T)\hat{u}^1(x,T)-e^{-\frac{T}{\epsilon}}u^*_t(x,T)\hat{u}^2(x,T)\bigg|^2 \tag{E.242}$$

$$\leqslant\bigg|\hat{u}^1_t(x,T)-\hat{u}^2_t(x,T)\bigg|^2\cdot\bigg|u^*_t(x,T)\bigg|^2, \tag{E.243}$$

$$|g_{correction,2}|=\bigg|e^{-\frac{t}{\epsilon}}u^*_{tt}\hat{u}^1-e^{-\frac{t}{\epsilon}}u^*_{tt}\hat{u}^2\bigg|^2\leqslant e^{-\frac{2t}{\epsilon}}\bigg|\hat{u}^1-\hat{u}^2\bigg|^2\cdot\bigg|u^*_{tt}\bigg|^2. \tag{E.244}$$

Again, we apply the Sobolev multiplication theorem (cf. Theorem F.7) and use upper bound on Sobolev norm uniformly in time t to get

$$||\,g_{correction,1}\,||_{L^2_t(0,T;L^2_x(\mathbb{R}^d))}\leqslant\beta\cdot T||\,\hat{u}^1_t(\cdot,T)-\hat{u}^2_t(\cdot,T)\,||_{H^q_x(\mathbb{R}^d)}||\,u^*_t(\cdot,T)\,||_{H^q_x(\mathbb{R}^d)} \tag{E.245}$$

$$\leqslant\beta\cdot R\cdot T||\,\hat{u}^1_t(\cdot,T)-\hat{u}^2_t(\cdot,T)\,||_{H^q_x(\mathbb{R}^d)}, \tag{E.246}$$

$$||g_{correction,2}||_{L^2_t(0,T;L^2_x(\mathbb{R}^d)}\leqslant\beta\cdot||\hat{u}^1-\hat{u}^2||_{L^2_t(0,T;H^q_x(\mathbb{R}^d))}\cdot||u^*_{tt}||_{L^2_t(0,T;H^q_x(\mathbb{R}^d))} \tag{E.247}$$

$$\leqslant R||\hat{u}^1-\hat{u}^2||_{L^2_t(0,T;H^q_x(\mathbb{R}^d))}. \tag{E.248}$$

Here we use property D.3 of our local basis $\phi_\tau(t)$ to get $\sup_t\phi_\tau\leqslant C(T)$, where $C(T)$ is some constant depending on T

$$||g_{correction,1}||^2_{L^2_t(0,T;L^2_x)}\leqslant R^2T^2C(T)^2\beta^2J^{2+2dq}\sum_{1\leqslant|\xi|_\infty,\tau\leqslant J}|a^1_{\xi,\tau}-a^2_{\xi,\tau}|^2, \tag{E.249}$$

$$||g_{correction,2}||^2_{L^2_t(0,T;L^2_x)}\leqslant R^2\beta^2J^{2dq}\sum_{1\leqslant|\xi|_\infty,\tau\leqslant J}|a^1_{\xi,\tau}-a^2_{\xi,\tau}|^2. \tag{E.250}$$

It follows

$$\log N(\mathcal{G}_{correction,1},||\cdot||_{L^2_t(0,T;L^2_x(\Omega))},\epsilon_1)\lesssim J^{d+2}\bigg[\log\frac{R\cdot T\cdot C(T)\cdot\beta}{4\epsilon\cdot\epsilon_1}+(2+dq)\log J\bigg], \tag{E.251}$$

$$\log N(\mathcal{G}_{correction,2},||\cdot||_{L^2_t(0,T;L^2_x(\Omega))},\epsilon_1)\lesssim J^{d+2}\bigg[\log\frac{R\cdot\beta}{4\epsilon_1}+dq\log J\bigg]. \tag{E.252}$$

Finally, we obtain the following covering numbers for $\mathcal{G}_{data}$, $\mathcal{G}_{energy}$, and $\mathcal{G}_{correction}$ respectively

$$\log N(\mathcal{G}_{data}, ||\cdot||_{L_t^2(0,T;L_x^2(\Omega))}, \epsilon_1) \lesssim J^{d+2}\left[\log\frac{1}{\epsilon_1} + (1+d)\log J\right], \tag{E.253}$$

$$\log N(\mathcal{G}_{energy}, ||\cdot||_{L_t^2(0,T;L_x^2(\Omega))}, \epsilon_1) \lesssim C_1(F)J^{d+2}\left[\log\frac{\beta}{\epsilon_1} + (1+dq)\log J\right], \tag{E.254}$$

$$\log N(\mathcal{G}_{correction}, ||\cdot||_{L_t^2(0,T;L_x^2(\Omega))}, \epsilon_1) \lesssim C_2(R,T)J^{d+2}\left[\log\frac{\beta}{\epsilon\cdot\epsilon_1} + (1+dq)\log J\right], \tag{E.255}$$

where C_1 is some constant depending on F and R; C_2 is some constant depending on R and T. ■

Finally, we present our proof to Theorem 3.5.

Proof of Theorem 3.5. First, we define population risk and its empirical parts:

$$\mathcal{E}_{data} := \mathbb{E}\Big[u(x,t) - u^*(x,t) + \nabla(u(x,t) - u^*(x,t)) \tag{E.256}$$

$$+ (u_t(x,t) - u_t^*(x,t)) + \nabla u_t(x,t) - \nabla u_t^*(x,t))\Big], \tag{E.257}$$

$$\mathcal{E}_{data,n} := \frac{1}{nm}\sum_{i=1,j=1}^{n,m}\Big[u(x_{ij},t_j) - u^*(x_{ij},t_j) + \nabla(u(x_{ij},t_j) - u^*(x_{ij},t_j)) \tag{E.258}$$

$$+ (u_t(x_{ij},t_{ij}) - u_t^*(x_{ij},t_j)) + \nabla(u_t(x_{ij},t_j) - u_t^*(x,t))\Big], \tag{E.259}$$

$$\mathcal{E}_{energy} := \mathbb{E}\left[\frac{\beta\cdot e^{-\frac{t}{\epsilon}}}{2}\left(|u_t|^2(x,t) + \frac{1}{2\epsilon}|\nabla u|^2(x,t) - f(x,t)u(x,t)\right)\right], \tag{E.260}$$

$$\mathcal{E}_{energy,n} := \frac{\beta}{nm}\sum_{i=1,j=1}^{n,m}\cdot\left[\frac{e^{-\frac{t}{\epsilon}}}{2}\left(|u_t|^2(x_{ij},t_j) + \frac{1}{2\epsilon}|\nabla u|^2(x_{ij},t_j) - f(x_{ij},t_j)u(x_{ij},t_j)\right)\right], \tag{E.261}$$

$$\mathcal{E}_{correction} := \beta\cdot\mathbb{E}\left[e^{-\frac{T}{\epsilon}}u_t^*(x,t_m)u(x,t_m) + e^{-\frac{t}{\epsilon}}u_{tt}^*(x,t)u(x,t)\right], \tag{E.262}$$

$$\mathcal{E}_{correction,n} := \frac{\beta}{nm}\sum_{i=1,j=1}^{n,m}\left[e^{-\frac{T}{\epsilon}}u_t^*(x_{ij},t_m)u(x_{ij},t_m) + e^{-\frac{t}{\epsilon}}u_{tt}^*(x_{ij},t_j)u(x_{ij},t_j)\right]. \tag{E.263}$$

It is essential to prove Lemma E.6 and Lemma E.7 before proving Theorem 3.5.

With L^2-covering numbers at our hand, we can directly apply Dudley theorem to bound Rademacher complexities. Before applying Lemma F.10, we need to find $\sup_{f\in\mathcal{F}}\mathbb{E}(f^2)$. For $\mathcal{G}_{data}$, we have

$$\mathbb{E}[g_{data,1}^2] \lesssim \rho. \tag{E.264}$$

For $\mathcal{G}_{energy}$, since $||f||_{L^\infty} \leqslant F$ and when $q \geqslant \frac{d}{2}+1$, we can apply embedding Theorem F.10 to spatial variable and obtain $H^q(\mathbb{R}^d) \subset W^{1,4}(\mathbb{R}^d)$. Furthermore, when $p \geqslant 2$, we have $u_t(t,\cdot) \in L^4([0,T])$ by standard Sobolev embedding theorem as well, and it follows that

$$\mathbb{E}[g_{energy}^2] \lesssim \int_0^T\int_{\mathbb{R}^d}(1+\frac{e^{\frac{-2t}{\epsilon}}}{4})|u_t^4| + (1+\frac{1}{4\epsilon^2})|\nabla u|^4 + (1+\frac{e^{\frac{-2t}{\epsilon}}}{\epsilon^2})f^4 + (1+\frac{e^{\frac{-2t}{\epsilon}}}{\epsilon^2})u^4 dxdt \tag{E.265}$$

$$\lesssim \frac{5}{4}||u_t||_{L_t^4(0,T;L_x^4(\mathbb{R}^d))}^4 + (1+\frac{5}{4\epsilon^2})||u||_{L_t^4(0,T;W_x^{1,4}(\mathbb{R}^d))}^4 + (1+\frac{1}{\epsilon^2})F^4 \tag{E.266}$$

$$\lesssim (1+\frac{5}{4\epsilon^2})(R^4+F^4). \tag{E.267}$$

For $\mathcal{G}^2_{correction}$, it follows that

$$\mathbb{E}[\mathcal{G}^2_{correction}] \leqslant \int_0^T \int_\Omega \frac{3}{4}|u_{tt}|^4 + \frac{3}{2}|u|^4 dxdt + T\int_\Omega u_t^4(x,T)dx \tag{E.268}$$

$$\lesssim \frac{3}{4}||u_{tt}||^4_{L^4_t(0,T;L^4_x(\mathbb{R}^d))} + \frac{3}{2}||u||^4_{L^4_t(0,T;L^4_x(\mathbb{R}^d))} + T||u_t(\cdot,T)||^4_{L^4_x(\mathbb{R}^d)} \tag{E.269}$$

$$\lesssim \frac{3}{2}R^4 + T\max_{t\in[0,T]}||u_t(\cdot,t)||^4_{L^4_x(\mathbb{R}^d)} = \max\{\frac{3}{2},T\}(R^4 + M^4(T)). \tag{E.270}$$

Finally, we can calculate with $C_1(R,F) := (1+\frac{5}{4\epsilon^2})(R^4+F^4)$ and $C_2(R,F) = (1+\frac{5}{4\epsilon^2})(R^4+F^4)$,

$$\mathcal{R}_n(\mathcal{G}_{data}) \leq \inf_{0\leq\delta}\left(4\delta + \frac{12}{\sqrt{nm}}\int_\delta^{R^2}\sqrt{J^{d+1}\left[\log\frac{1}{\epsilon_1} + (2+d)\log J\right]}d\varepsilon_1\right), \tag{E.271}$$

$$\mathcal{R}_n(\mathcal{G}_{energy}) \leq \inf_{0\leq\delta}\left(4\delta + \frac{12}{\sqrt{nm}}\int_\delta^{C_1(R,F)}\sqrt{J^{d+1}\left[\log\frac{\beta}{\epsilon\cdot\epsilon_1} + (2+dq)\log J\right]}d\varepsilon_1\right), \tag{E.272}$$

$$\mathcal{R}_n(\mathcal{G}_{correction}) \leq \inf_{0\leq\delta}\left(4\delta + \frac{12}{\sqrt{nm}}\int_\delta^{C_2(R,F)}\sqrt{J^{d+1}\left[\log\frac{\beta}{\epsilon\cdot\epsilon_1} + (2+dq)\log J\right]}d\varepsilon_1\right). \tag{E.273}$$

It follows that

$$\mathcal{R}_n(\mathcal{G}_{data}) \lesssim \sqrt{\frac{\rho}{nm}}J^{\frac{d+1}{2}}\left[\sqrt{\log(\frac{1}{\rho}+1)} + \sqrt{(2+d)\log J}\right] \tag{E.274}$$

$$\lesssim C_3(R)\sqrt{\frac{\rho}{nm}}J^{\frac{d+1}{2}}\sqrt{\log J\log(1+\frac{1}{\rho})}, \tag{E.275}$$

$$\mathcal{R}_n(\mathcal{G}_{data}) \lesssim \sqrt{\frac{\beta}{nm}}J^{\frac{d+2}{2}}\sqrt{(1+\frac{5}{4})(R^4+F^4)}\Bigg[\sqrt{\log(\frac{1}{(1+\frac{5}{4})(R^4+F^4)\epsilon}+1)} \tag{E.276}$$

$$+\sqrt{(2+dq)\log J}\Bigg] \tag{E.277}$$

$$\lesssim C_4(\epsilon,d,q,R,F)\sqrt{\frac{\beta}{nm}}J^{\frac{d+2}{2}}\sqrt{\log J}, \tag{E.278}$$

$$\mathcal{R}_n(\mathcal{G}_{data}) \lesssim \sqrt{\frac{\beta}{nm}}J^{\frac{d+1}{2}}\sqrt{\max\{\frac{3}{2},T\}(R^4+M^4(T))}\Bigg[\sqrt{\log(\frac{1}{\max\{\frac{3}{2},T\}(R^4+M^4(T))\epsilon}+1)} \tag{E.279}$$

$$+\sqrt{(2+dq)\log J}\Bigg] \tag{E.280}$$

$$\lesssim C_5(\epsilon,d,q,R,T)\sqrt{\frac{\beta}{nm}}J^{\frac{d+2}{2}}\sqrt{\log J}. \tag{E.281}$$

We need to appropriately balance the radius of localization ball ρ and regularizing parameter β. We pick $\rho \approx \frac{1}{nm}$ and $\beta \approx \frac{\log nm}{nm}\log J$, and then

$$\mathcal{R}_n(\mathcal{G}) \lesssim C(\epsilon,d,q,R,T,F)J^{\frac{d+2}{2}}\log J\frac{\log nm}{nm}. \tag{E.282}$$

■

F USEFUL TECHNICAL TOOLS

Theorem F.1 (Bernstein Inequality)**.** *Let X_i,...,X_N be i.i.d random variables. Suppose that $\mathbb{E}X_i = 0$, $|X_i| \leqslant M$ almost surely, and $\frac{1}{N}\sum_{i=1}^{n} Var(X_i) \leqslant \sigma^2$. For any $t \geqslant 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geqslant t\right) \leqslant \exp\left(-\frac{nt^2}{2\sigma^2 + \frac{2}{3}Mt}\right). \tag{F.1}$$

Proof of Theorem F.1. van der Vaart (1996, Lemma 2.2.9). ■

Lemma F.2 (Maximal inequality of empirical process)**.** *Let Ω be the domain and $\mathcal{F}$ be a set of functions f and $\mathcal{P}_\Omega$ is probability measure from which X is sampled. Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i)\right| \leqslant 2\mathcal{R}_{n,\mathcal{F}}. \tag{F.2}$$

Proof of Lemma F.2. See Proposition 4.11 in (Wainwright, 2019). ■

Theorem F.3 (Dudley's Theorem)**.** *Let $\mathcal{F}$ be a function class such that $\sup_{f\in\mathcal{F}} \|f\|_{L_2(P_n)} \leq M$. Then the Rademacher complexity $\mathcal{R}_n(\mathcal{F})$ satisfies*

$$\mathcal{R}_n(\mathcal{F}) \leq \inf_{0\leq\delta}\left\{4\delta + \frac{12}{\sqrt{n}}\int_{\delta}^{\sup_{f\in\mathcal{F}}\mathbb{E}(f^2)} \sqrt{\log N\big(\varepsilon,\, \mathcal{F},\, \|\cdot\|_{L^2(P_n)}\big)}\, d\varepsilon\right\}, \tag{F.3}$$

where $N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)$ denotes the ε-covering number of $\mathcal{F}$ with respect to the L^∞-norm.

Proof of Theorem F.3. See Theorem 4.28 in (Ma, 2022). ■

Theorem F.4 (Bernstein Chaining)**.** *Let $\mathcal{F}$ be a class of measurable functions such that $\mathbb{E}[f^2] < \sigma^2$ and $||f||_\infty < L$ for all $f \in \mathcal{F}$. Then,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\sqrt{n}\big(\hat{P} - P\big)f|\right] \lesssim \mathcal{D}_{[\,]}(\sigma, \mathcal{F}, L_2(P))\left(\mathcal{D}_{[\,]}(\sigma, \mathcal{F}, L_2(P)) + \frac{\mathcal{D}_{[\,]}(\sigma, \mathcal{F}, L_2(P))}{\sigma^2\sqrt{n}}L\right). \tag{F.4}$$

Proof. Apply van der Vaart (1996, Lemma 3.4.2) to $\mathcal{F}/L$ and rescale. ■

Proof of Theorem F.4. See Lemma 4.3.2 in (van der Vaart & Wellner, 2007). ■

Theorem F.5 (Talagrand's Inequality)**.** *Let $\mathcal{F}$ be a class of measurable functions such that $\mathbb{E}[f^2] < \sigma^2$ and $||f||_\infty < L$ for all $f \in \mathcal{F}$. Then*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\sqrt{n}|(\hat{P} - P)f| \geqslant 2\mathbb{E}[\sup_{f\in\mathcal{F}}|(\hat{P} - P)f| + \sigma\sqrt{2s} + \frac{L}{\sqrt{n}}s\right) \leqslant e^{-s}. \tag{F.5}$$

Proof of Theorem F.5. See Bousquet (2002, theorem 2.3). ■

Lemma F.6. *(Covering numbers for L_2 balls) Let $p \in \mathbb{N}$ and $B_2(r)$ be the L_2 balls of $\mathbb{R}^d$ with radius r such that $\epsilon \leqslant r$. Then*

$$N(B_2(r), ||\cdot||_2, \epsilon) \leqslant \left(\frac{3r}{\epsilon}\right)^d. \tag{F.6}$$

Proof of Lemma F.6. This lemma follows directly from Corollary 4.2.13 in (Vershynin, 2018) by rescaling. ■

Theorem F.7 (Sobolev Multiplication Theorem)**.** *Let $\Omega \subseteq \mathbb{R}^d$ be a Lipschitz domain, and let $s > \frac{d}{p}$ with $1 \leqslant p \leqslant \infty$. For any $u, v \in W^{s,p}(\Omega)$, there exists a constant $C = C(n, s, p, \Omega)$ such that:*

$$\|uv\|_{W^{s,p}(\Omega)} \leqslant C\|u\|_{W^{s,p}(\Omega)}\|v\|_{W^{s,p}(\Omega)}. \tag{F.7}$$

Proof of Theorem F.7. This theorem follows from Theorem 2.8.3 in (Triebel, 2006). ■

Lemma F.8 (Morrey's embedding)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with C^1 boundary. Let $1 \leq p < \infty$ and $k, m \in \mathbb{N}_0$ satisfy $m > d/p$. Then the embedding*

$$W^{k+m,p}(\Omega) \hookrightarrow C^k(\overline{\Omega})$$

is continuous and compact. Then

$$W^{k+m,p} \hookrightarrow C^k(\bar{\Omega}). \tag{F.8}$$

In particular, the embedding is compact if $m - \frac{d}{p} > k$.

Proof of Lemma F.8. This lemma is a consequence of Section 5.6.2 in (Evans, 2010) and Theorem 4.12 in (Adams, 1975). ■

Lemma F.9 (Weighted Poincaré Inequality)**.** *Let Ω be bounded Lipschitz domain and $w(x)$ be non-negative weight on Ω. There exist constant $C := C(d, \Omega, w) > 0$ such that for any $f \in H^1(\Omega)$, we have*

$$\|(f - f_{w,\Omega})w(\cdot)\|_{L^2(\Omega)} \leqslant C\|w(\cdot)\nabla f\|_{L^2(\Omega)}, \tag{F.9}$$

where $f_{w,\Omega} := \frac{\int_\Omega w(x)f(x)dx}{\int_\Omega w(x)dx}$.

Proof of Lemma F.9. The lemma can be found in (Deny & Lions, 1954) and Section 3 of (Turesson, 2000). This variant of Poincare Inequality is also called the Deny-Lions Lemma. Detailed proof for the case $p = q = 2$ can be found in (Rathmair, 2019). ■

Lemma F.10. *Given $p, q \geqslant 0$. Assume $f \in H_t^p(\mathbb{R}; H_x^q(\mathbb{R}^d)) \cap \mathcal{F}_{J,J}$ and f admits spatial-temporal Fourier series with smooth basis dense in L^2 on compact space-time domain $\Omega_T = \Omega \times [0, T]$, then the following estimates hold*

$$||P_J f||^2_{L_t^2(\mathbb{R};L_x^2(\mathbb{R}^d))} \leqslant C(\Omega, T)||f||^2_{H_t^1([0,T];H_x^1(\Omega))}, \tag{F.10}$$

$$||P_J f||_\infty \leqslant C(\Omega)R, \qquad \text{if } q > \frac{d}{2}, \tag{F.11}$$

$$||P_J f||_\infty \leqslant C(\Omega, T)J^{d+2}, \qquad \text{if } q \leqslant \frac{d}{2}, \tag{F.12}$$

where $C_1(\Omega, T)$ is a constant depending on Ω_T .

Proof of Lemma F.10. To prove (F.1), we note

$$||P_J f||^2_{L_t^2(\mathbb{R};L_x^2(\mathbb{R}^d))} \leqslant ||f||^2_{H_t^1(\mathbb{R};H_x^1(\mathbb{R}^d))} \tag{F.13}$$

$$\leqslant C(\Omega, T)||f||^2_{H_t^1([0,T];H_x^1(\Omega))} \quad - \text{ Lemma A.3 .} \tag{F.14}$$

To prove (70), we only need to rewrite

$$P_J f = \sum_{1\leqslant\tau,|\xi|\leqslant J} |\xi|^p\tau^q \frac{1}{|\xi|^p\tau^q} a_{\xi,\tau}\phi_{\xi,\tau}(x, t). \tag{F.15}$$

From Morrey's embedding in case of $q > \frac{d}{2}$ and the Sobolev norm of f is bounded by R and the extension Theorem A.2,

$$||P_J f(\cdot, t)||_\infty \leqslant C(\Omega)||P_J f(\cdot, t)||_{H^q(\Omega)} \leqslant C(\Omega)R, \quad \forall t \in [0, T]. \tag{F.16}$$

Otherwise, if $q \leqslant \frac{d}{2}$, we note there are finitely many coefficients $a_{\xi,\tau}$ and each $\phi_{\xi,\tau}(x, t)$ has local compact support

$$P_J f = \sum_{1\leqslant\tau\leqslant J^2,1\leqslant|\xi|\leqslant J} a_{\xi,\tau}\phi_{\xi,\tau}(x, t) \leqslant ||a_{\xi,\tau}||_\infty \sum_{1\leqslant\tau\leqslant J^2,1\leqslant|\xi|\leqslant J} \phi_{\xi,\tau}(x, t) \leqslant C(\Omega, T)J^{d+2}. \tag{F.17}$$

■

Lemma F.11. *Let $\phi_\tau(t) = 2^{\ell/2}\varphi(2^\ell t - k_0)$ be an L^2-normalized compactly supported wavelet at level $2^\ell \asymp J^2$, let $T_1, \dots, T_n$ be i.i.d. uniform on $[0, T]$, and set $Y_j := \phi_\tau(T_j)^2$, $B := \frac{1}{n}\sum_{j=1}^n \big(Y_j - \mathbb{E}Y_1\big)$. Then*

$$\mathit{Var}(Y_1) \asymp J^2, \qquad \mathbb{E}\big|Y_1 - \mathbb{E}Y_1\big|^4 \lesssim J^6, \tag{F.18}$$

and there exist constants $c_1, c_2 > 0$ such that, whenever $n \geqslant J^2$,

$$\mathbb{P}\Big(|B| \geqslant c_1\sqrt{J^2/n}\Big) \;\geqslant\; c_2. \tag{F.19}$$

Proof. By the change of variables $u = 2^\ell t - k_0$, $\int \phi_\tau^{2q} = 2^{\ell(q-1)}\|\varphi\|_{L^{2q}}^{2q}$ for $q \in \{1, 2, 4\}$, whence $\mathbb{E}Y_1 = T^{-1}$, $\mathbb{E}Y_1^2 \asymp 2^\ell \asymp J^2$, and $\mathbb{E}Y_1^4 \asymp 2^{3\ell} \asymp J^6$; (F.18) follows. For i.i.d. centered variables,

$$\mathbb{E}B^2 = \frac{\mathrm{Var}(Y_1)}{n}, \qquad \mathbb{E}B^4 = \frac{\mathbb{E}(Y_1 - \mathbb{E}Y_1)^4}{n^3} + \frac{3(n-1)}{n^3}\mathrm{Var}(Y_1)^2 \lesssim \frac{J^6}{n^3} + \frac{J^4}{n^2}.$$

If $n \geqslant J^2$, then $J^6/n^3 \leqslant J^4/n^2$, so $\mathbb{E}B^4 \leqslant \kappa\,(\mathbb{E}B^2)^2$ for an absolute constant κ. The Paley–Zygmund inequality applied to B^2 gives, for $\theta \in (0, 1)$,

$$\mathbb{P}\big(B^2 \geqslant \theta\,\mathbb{E}B^2\big) \geqslant (1-\theta)^2\frac{(\mathbb{E}B^2)^2}{\mathbb{E}B^4} \geqslant \frac{(1-\theta)^2}{\kappa};$$

taking $\theta = \frac{1}{2}$ and using $\mathbb{E}B^2 \asymp J^2/n$ yields (F.19). ∎

G More Experiments

G.0.1 Implementation Details

The spatial domain Ω is set as $[-1, 1]^d$ and the finite time horizon is $[0, 1]$. All simulations are implemented in Python 3.11, Pytorch 2.8.0, Numpy 2.3.0, and Scipy 1.16.2. Numerical derivatives are computed using findiff library 0.12.1 with an accuracy of order 6. In low dimensions, numerical integration is performed using Scipy's integration tools in combination with vectorized Einstein summation through NumPy's einsum function. In higher dimensions, where traditional Numerical integration becomes computationally expensive, so we employ a Monte Carlo method by randomly sampling 4096 points to approximate the integral.

G.0.2 Data Preparation

Our estimator defined in (2.7) assumes the access to training data on space-time cloud points $S := \{(X_{ij}, t_j, u^*_{ij}, \nabla u^*_{ij}, \partial_t u^*_{ij}, \nabla\partial_t u^*_{ij}, f_{ij})\}_{i=1,j=1}^{n,m}$. Following the sampling procedure described in (1.2), we uniformly sample N_t time locations from interval $[0, 1]$ and at each time location t_j, we then uniformly sample N_x space location x_{ij} from domain Ω. To compute the data fidelity term under the $H_t^1 H_x^1$ norm, we require not only ground truth solution u^* as our primary label but also derivative labels u^*_t, u^*_x and ∇u^*_t through numerical derivative computation. Training models with additional derivative information is commonly refered to Sobolev Training (Czarnecki et al., 2017; Lu et al., 2022; Mroueh et al., 2019). These derivative labels are generated using numerical solver of PDEs and the findiff library. In our experiments, accurate computation of derivative labels is critical. We notice imprecise derivatives can significantly degrade performance and even destabilize the learning process. To obtain high fidelity derivative datas, we propose two common sources to obtain derivatives: (1) numerically solved PDEs solutions (2) known analytical solutions. We test data from both sources. In Section 4, our training data is obtained from numerically solved PDEs solutions, with derivatives calculated by findiff, and our examples all have known analytical solutions, with derivatives calculated by autograd. For each random seed, we sample 50 time coordinates uniformly from $[0, 1]$. At every sampled time, we independently draw 80 spatial locations uniformly from $[-1, 1]$, giving 4,000 co-located training points.

G.0.3 Training

Our loss function is the empirical counterpart of the population loss function defined in (2.4). In practice, we observe model correction term has negligible impact on training outcomes. To

accommodate for solving different PDEs, one might need to modify energy regularization term accordingly. The model architecture is based on a standard multilayer perceptron (MLP), following design choices commonly adopted in recent state-of-the-art deep learning approaches for physics learning tasks. We believe such architecture is not optimal architecture suitable for our framework, but we stick with MLP for fair experiments. In this work, we adopt MLP structure with fully connected layers and Tanh activations. Optimization is carried out using the Adam optimizer Kingma & Ba (2015) as implemented in torch.optim with a fixed learning rate $1e^{-3}$. We optimize the empirical loss using Adam with a fixed learning rate of $1e^{-3}$, which we empirically observe it achieves unstable convergence with fixed learning rate Figure 29. Adaptive or smaller learning rate will make convergence stable Figure 31. Understanding the convergence behavior and robustness of different optimizers in solving PDEs remains an important question, which leave as a direction for future investigation.

G.0.4 Model Validation

Our dataset consists of $N_t \times N_x$ spatio-temporal samples, each associated with corresponding ground-truth labels. For evaluation purposes, we use mainly relative L_2 norm for evaluation metric. The accuracy metric is the relative space-time L^2 error

$$\mathrm{RelErr}_{L^2}(\widehat{u}, u^*) = \frac{\|\widehat{u} - u^*\|_{L^2(\Omega\times[0,T])}}{\|u^*\|_{L^2(\Omega\times[0,T])}}.$$

For these benchmarks, all reported scalar values are medians over 10 independent random seeds.However, to assess the effectiveness of our proposed method, we compare its performance against a vanilla Physics-Informed Neural Network (PINN) baseline. Both models are evaluated on the test set using the mean squared error (MSE) between predicted and ground-truth solution values as the primary performance metric. It has been reported PINNs usually have low MSE error but high L_2 relative error, which is also verified in table 4.

G.1 Parabolic PDEs

Table 1: Overview of the additional PDEs experiments .

Benchmark	PDE type	Dimension
Heat equation	Linear parabolic	$d = 1, 3$
Sine–Gordon equation	Nonlinear hyperbolic	$d = 1, 3$
LP1–LP3	Variable-coefficient parabolic	$d = 3$
Viscous HJ equation	Nonlinear parabolic	$d = 1, 3, 10, 20, 50$

We are motivated by a toy example solving the Cauchy problem for the 1D heat equation. The equation has a uniquely determined solution $u(x,t) = e^{-t}\sin(x)$ for $u_t - \Delta u = 0$ with $u_0 = \sin(x)$. We compare several methods against the vanilla PINN baseline before and after introducing new labeled data.

In standard PINNs Karniadakis et al. (2021), we choose weight $w = 1.0$ for the residuals of the initial conditions $\mathcal{R}_{\text{ic}}$. Based on the standard PINN, introducing a data-fitting term alongside $\mathcal{R}_{\text{pde}}$ and $\mathcal{R}_{\text{ic}}$ leads to significant improvement. Meanwhile, Sobolev training Czarnecki et al. (2017) uses derivative labels to provide extra information. While Son et al. (2021) proposes incorporating higher derivatives in the PDE loss and derivative labels in boundary/initial conditions, their experiments primarily show improvement via supervised components. We investigate whether Sobolev training is applicable to the *unsupervised* component of PINNs. Our results suggest it is not directly applicable here, as forcing higher-order derivatives in $\mathcal{R}_{\text{pde}}$ creates a solution search class that is too large due to the infinite number of solutions satisfying higher-order constraints.

Since our approach aims to provide data-driven alternative to solve for non-smooth solutions, we also compare our method against other variants of PINNs that improve accuracy to learn weak solutions for PDEs. We pick **W-PINNs**De Ryck et al. (2024), **hp-vPINNs** Kharazmi et al. (2021), **SA-PINNs** McClenny & Braga-Neto (2023), and **adversarial training** for PINNs Wang et al. (2022a) as these methods stand out among many invariants of PINNs that address discontinuity of shocks waves and

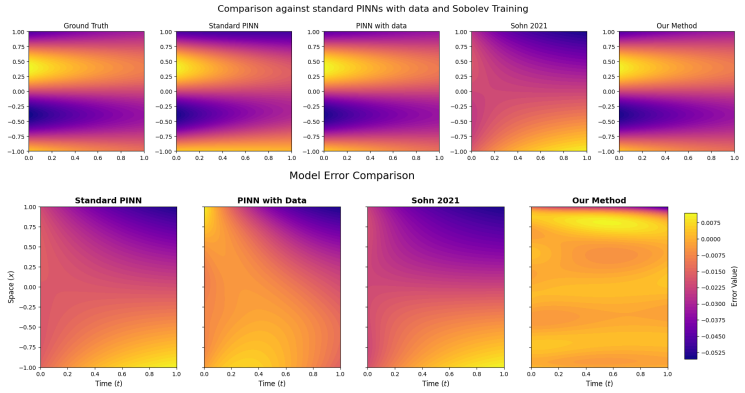

Figure 5: Error plot.

variational forms of PDEs to solve PDEs. We briefly walk through the key-features of these methods: **hp-vPINNs** utilize domain decomposition to physically isolate sharp gradients and smooth out using test functions, preventing error propagation across the solution domain. Alternatively, **W-PINNs** and **SA-PINNs** fall under the umbrella of adversarial frameworks to force network attention onto high-error regions dynamically by continuously optimizing trainable residual weights for weak-form residuals on collocation points. Adversarial training for PINNs Wang et al. (2022a) extend this adversarial philosophy by abandoning fixed points entirely, utilizing gradient ascent to physically shift collocation coordinates into worst-case regions to evaluate loss in L^∞ norm.

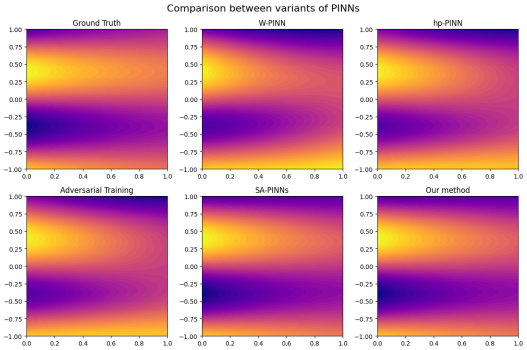

Figure 6: Learned solutions for PINN variants for 1d heat equation.

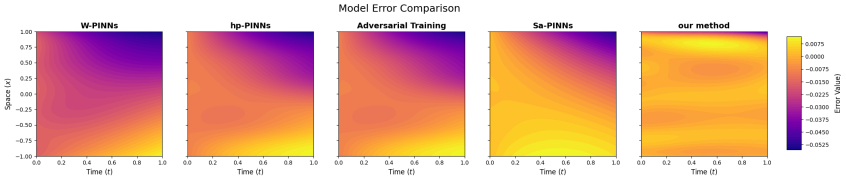

Figure 7: Error plot for PINN variants.

Variable-coefficient and nonlinear parabolic equations. We next consider three linear variable-coefficient parabolic equations and one nonlinear viscous Hamilton–Jacobi equation in 3d dimensional setting. These experiments examine whether the implementation can accommodate diffusion, reaction, and drift coefficients that are more general than those of the heat equation. The four equations are defined as follows.

LP1: Variable diffusion–reaction equation.

$$u_t - \partial_x(a_1(x)u_x) + c_1(x)u = f,$$

where

$$a_1(x) = 1 + \frac{1}{2}\sin(\pi x), \qquad c_1(x) = \frac{1}{2} + x^2.$$

LP2: Space–time-dependent diffusion–reaction equation.

$$u_t - \partial_x(a_2(x,t)u_x) + c_2(x,t)u = f,$$

where

$$a_2(x,t) = 1 + \frac{1}{4}\sin(\pi x)\cos(2\pi t),$$

and

$$c_2(x,t) = \frac{1}{2} + \frac{1}{4}(1+t)\cos^2(\pi x).$$

LP3: Reversible variable-drift equation.

$$u_t - \mu(x)^{-1}\partial_x(\mu(x)a_3(x)u_x) + c_3(x)u = f,$$

where

$$a_3(x) = 1 + \frac{x^2}{4}, \qquad \mu(x) = e^{-x^2/4}, \qquad c_3(x) = \frac{1}{4}(1+x^2).$$

HJ: Viscous quadratic Hamilton–Jacobi equation.

$$u_t - 0.1u_{xx} + \frac{1}{2}u_x^2 = f.$$

For LP1–LP3, we use the common manufactured solution

$$u^*(x,t) = G(t)P(x),$$

where

$$G(t) = e^{-t}\left[1 + \frac{1}{4}\sin(2\pi t)\right], \qquad P(x) = (1-x^2)\left[1 + \frac{1}{5}\sin(\pi x)\right].$$

The required derivatives are

$$G'(t) = e^{-t}\left[\frac{\pi}{2}\cos(2\pi t) - 1 - \frac{1}{4}\sin(2\pi t)\right],$$

$$P'(x) = -2x\left[1 + \frac{1}{5}\sin(\pi x)\right] + \frac{\pi}{5}(1-x^2)\cos(\pi x),$$

and

$$P''(x) = -2\left[1 + \frac{1}{5}\sin(\pi x)\right] - \frac{4\pi x}{5}\cos(\pi x) - \frac{\pi^2}{5}(1-x^2)\sin(\pi x).$$

For the one-dimensional Hamilton–Jacobi equation, we use

$$u^*_{\mathrm{HJ}}(x,t) = \frac{1}{2}(1-e^{-t})(1-x^2).$$

The initial and boundary conditions are prescribed consistently with the corresponding solutions.

Substituting the manufactured solutions into the equations gives

$$f_{\mathrm{LP1}}(x,t) = G'(t)P(x) - G(t)\left[\left(1+\frac{1}{2}\sin(\pi x)\right)P''(x) + \frac{\pi}{2}\cos(\pi x)P'(x)\right] + \left(\frac{1}{2}+x^2\right)G(t)P(x),$$

$$f_{\mathrm{LP2}}(x,t) = G'(t)P(x) - G(t)\left[\left(1+\frac{1}{4}\sin(\pi x)\cos(2\pi t)\right)P''(x) + \frac{\pi}{4}\cos(\pi x)\cos(2\pi t)P'(x)\right] + \left[\frac{1}{2}+\frac{1}{4}(1+t)\cos^2(\pi x)\right]G(t)P(x),$$

$$f_{\mathrm{LP3}}(x,t) = G'(t)P(x) - G(t)\left[\left(1+\frac{x^2}{4}\right)P''(x) - \frac{x^3}{8}P'(x)\right] + \frac{1}{4}(1+x^2)G(t)P(x),$$

$$f_{\mathrm{HJ}}(x,t) = \frac{1}{2}e^{-t}(1-x^2) + 0.1(1-e^{-t}) + \frac{1}{2}(1-e^{-t})^2x^2.$$

High-dimensional viscous Hamilton–Jacobi equations. To study **dimensional scalability** on a controlled nonlinear family, we consider

$$u_t - 0.1\Delta u + \frac{1}{2}\|\nabla u\|_2^2 = f$$

on

$$\Omega_d \times [0,1] = [-1,1]^d \times [0,1], \qquad d \in \{3,10,20,50\}.$$

We use the manufactured solution

$$u_d^*(x,t) = \frac{1}{2}(1-e^{-t})\left(d - \|x\|_2^2\right),$$

which gives

$$f_d(x,t) = \frac{1}{2}e^{-t}\left(d-\|x\|_2^2\right) + 0.1d(1-e^{-t}) + \frac{1}{2}(1-e^{-t})^2\|x\|_2^2.$$

The initial condition is zero, and the boundary condition is prescribed by the trace of u_d^* on $\partial\Omega_d$.

G.2 Comparison with Other Methods on LP1–LP3 and HJB

For comparison baselines, we consider **W-PINNs** (De Ryck et al., 2024), **hp-vPINNs** (Kharazmi et al., 2021), **SA-PINNs** (McClenny & Braga-Neto, 2023), and **adversarial training** for PINNs (Wang et al., 2022a) as these methods stand out among many invariants of PINNs that address discontinuity of shocks waves and variational forms to solve PDEs. Figures 8–10 compare our method with five PINN-based baselines on 3d LP1–LP3 benchmarks. Our method achieves the lowest mean error in both metrics across parabolic equations. Notably, on LP3, it reduces test MSE and relative L^2 error by 27.2% and 16.5%, respectively, compared with the strongest baseline, Adversarial PINN.

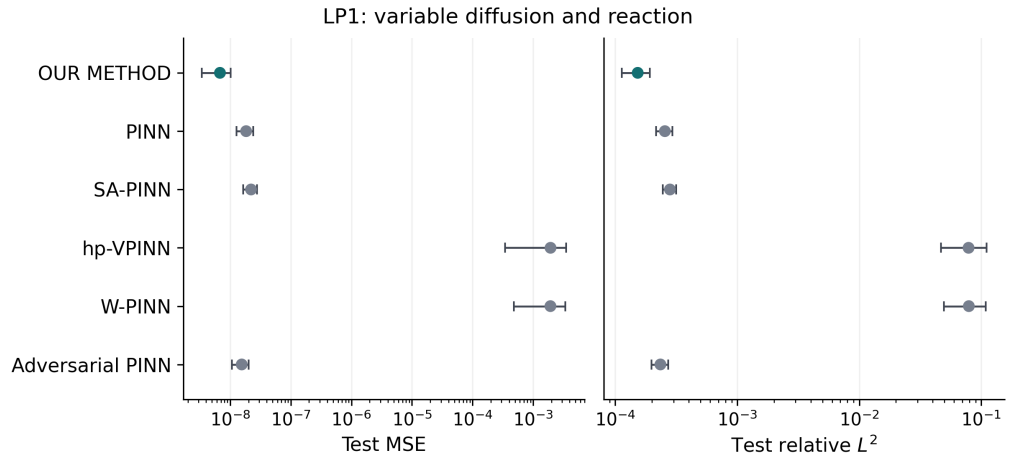

Figure 8: Comparison with other methods on LP1. Points show mean test errors and bars indicate one sample standard deviation over five seeds.

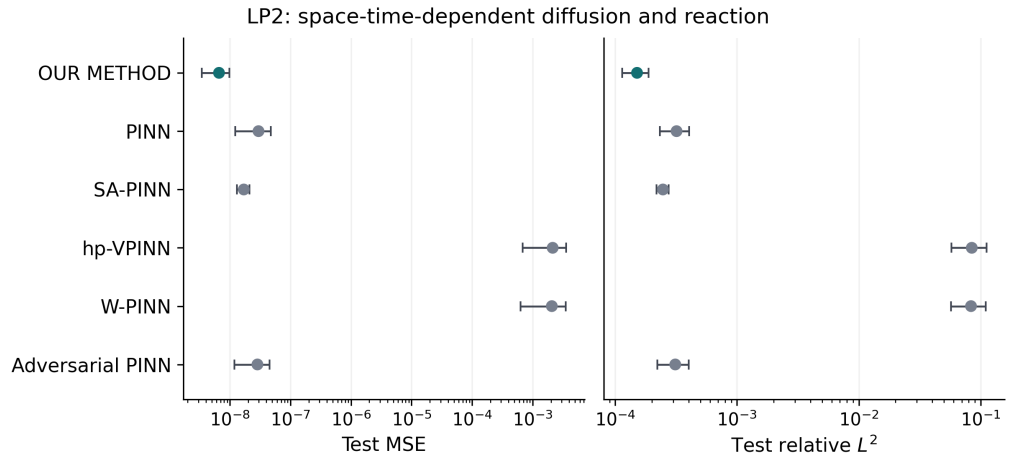

Figure 9: Comparison with other methods on LP2. Points show mean test errors and bars indicate one sample standard deviation over five seeds.

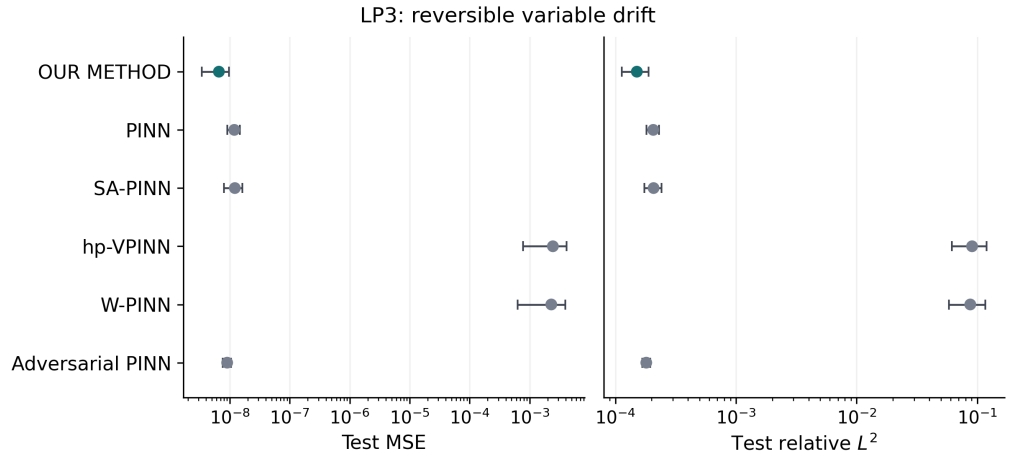

Figure 10: Comparison with other methods on LP3. Points show mean test errors and bars indicate one sample standard deviation over five seeds.

Viscous Hamilton Jacobi Equation in various dimensions. Figures 11–15 show the same six methods for the viscous Hamilton–Jacobi family (HJB, denoted HJ above) in $d \in \{1, 3, 10, 20, 50\}$.

Figures 10–15 compare our method with five PINN-based baselines using test MSE and relative L^2 error. Results are averaged over ten random seeds, with error bars indicating one sample standard deviation. Our method achieves the lowest mean error in all five HJB. Across $d \in \{1, 3, 10, 20, 50\}$, our method reduces test MSE by 79.5%–89.4% and relative L^2 error by 55.3%–67.4% compared with the best competing method in each configuration. At $d = 50$, it achieves a relative L^2 error of 0.964%, compared with 2.959% for SA-PINN, while reducing test MSE by 89.4%. Although its errors increase from $d = 3$ to $d = 50$, our method maintains an accuracy advantage throughout the dimensions. Besides, our results show all PINNs baselines and our method suffer from the curse of dimensionality.

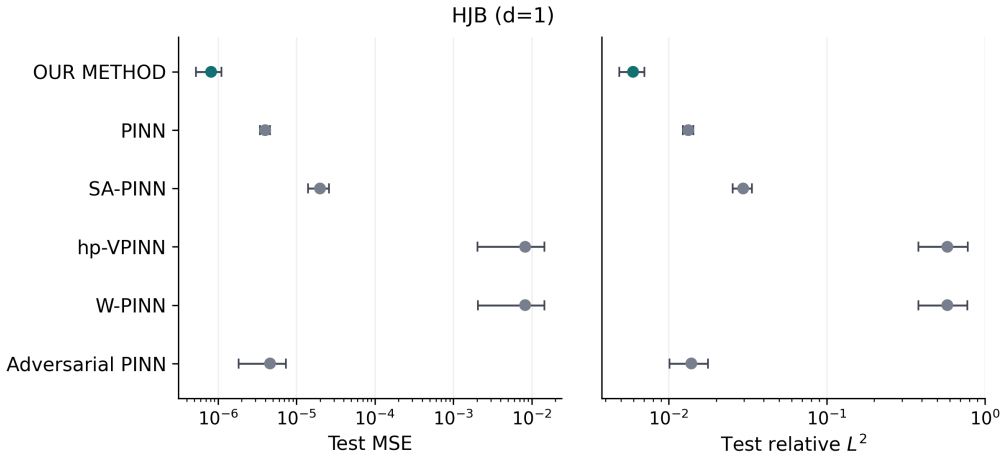

Figure 11: Comparison with other methods on HJB in $d = 1$. Points show mean test errors and bars indicate one sample standard deviation over ten seeds.

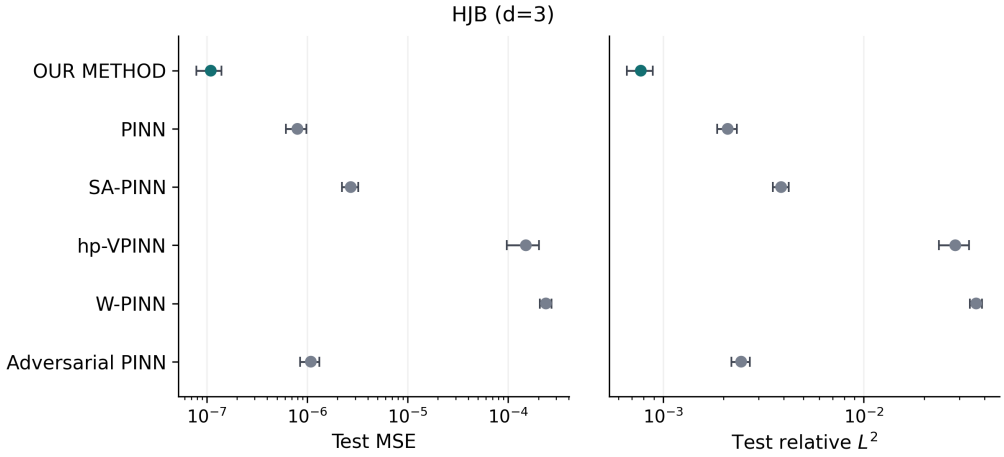

Figure 12: Comparison with other methods on HJB in $d = 3$. Points show mean test errors and bars indicate one sample standard deviation over ten seeds.

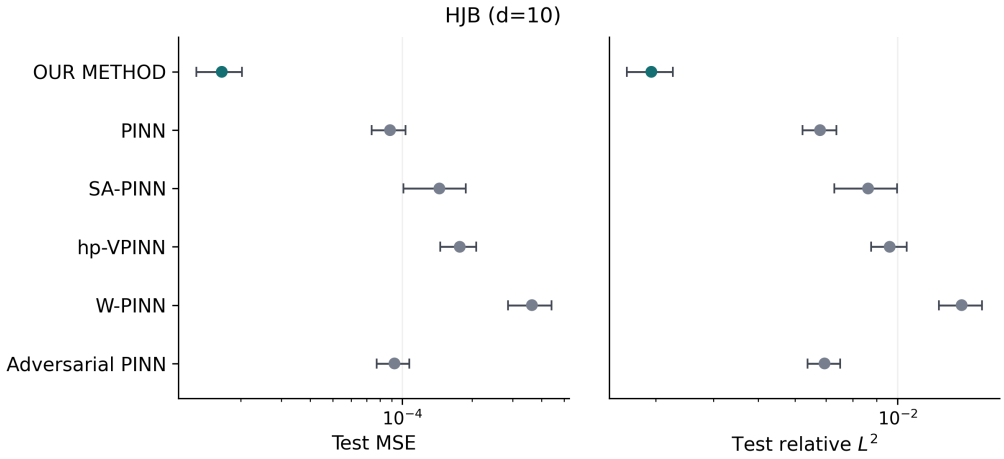

Figure 13: Comparison with other methods on HJB in $d = 10$. Points show mean test errors and bars indicate one sample standard deviation over ten seeds.

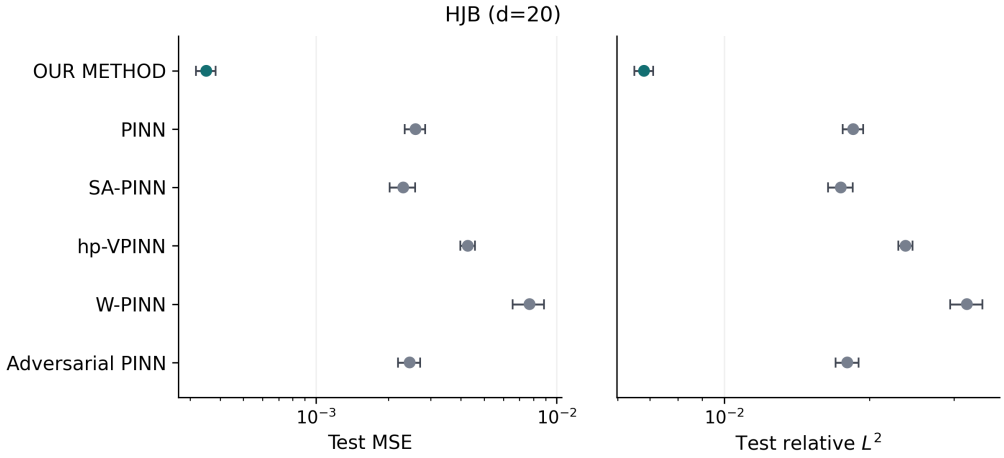

Figure 14: Comparison with other methods on HJB in $d = 20$. Points show mean test errors and bars indicate one sample standard deviation over ten seeds.

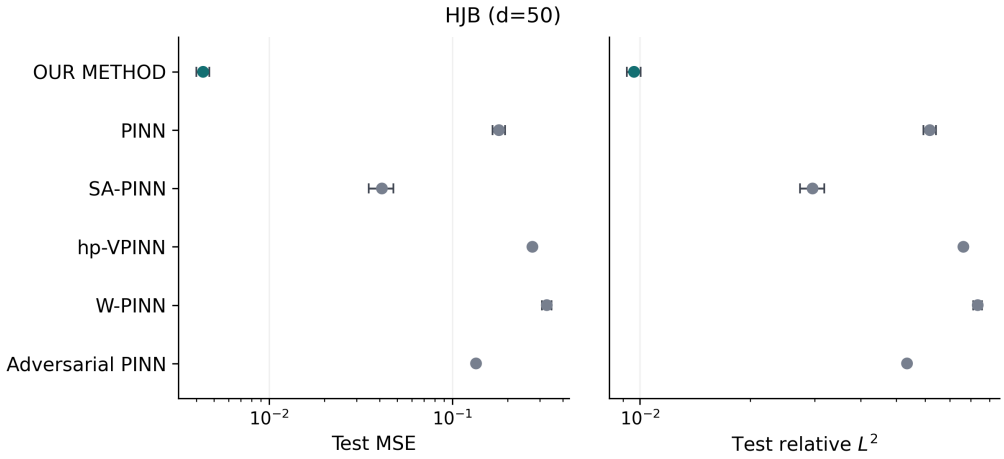

Figure 15: Comparison with other methods on HJB in $d = 50$. Points show mean test errors and bars indicate one sample standard deviation over ten seeds.

G.3 Regularization Improvement for Some Parabolic PDEs

We examine regularization effects on heat equations and other parabolic equations LP1-LP3. We compare the purely Sobolev fitting term $\widehat{D}_{H_t^1 H_x^1}$ with the regularized term $\widehat{D}_{H_t^1 H_x^1} + \beta(I_\epsilon + M_\epsilon)$, with the optimal selected epsilon-beta pair. Our first example includes fixed-size comparison using fixed regularization β, and then we will exam how unregularized estimator versus regularized estimator using fixed regularization and adaptive regularization across various sample size.

Fixed-size paired comparison. We compare the two regularized and unregularized methods using $n = 128$ and $m = 128$ space time sampling locations and forcing function with 5% noise level.

Example	Sobolev H^1 MSE	Regularized Sobolev H^1 MSE	MSE gain (%)
Heat	$(4.44 \pm 3.65) \times 10^{-6}$	$(9.45 \pm 7.28) \times 10^{-7}$	78.7
LP1	$(1.89 \pm 2.09) \times 10^{-5}$	$(1.79 \pm 1.22) \times 10^{-6}$	90.5
LP2	$(3.21 \pm 2.97) \times 10^{-7}$	$(3.32 \pm 1.93) \times 10^{-8}$	89.7
LP3	$(1.40 \pm 1.13) \times 10^{-7}$	$(3.21 \pm 1.72) \times 10^{-9}$	97.7

Fixed-size comparison. At a fixed sample size of $n = 128$ and $m = 128$ and a noise level of 5%, we compare the mean MSE over 100 runs. As shown in Figure 16, regularization reduces the mean MSE by 78.7%, 90.5%, 89.7%, and 97.7% for Heat, LP1, LP2, and LP3, respectively, relative to the unregularized Sobolev baseline.

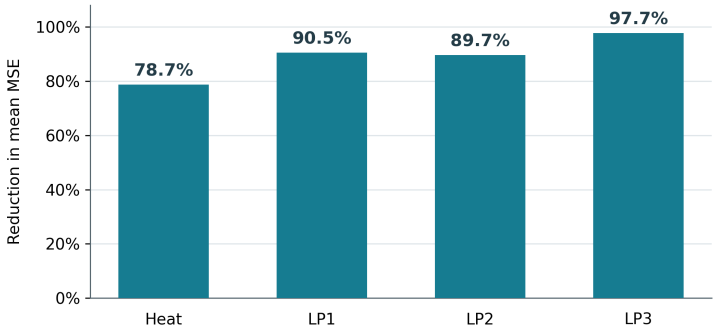

Figure 16: MSE reduction from regularization for parabolic equations at $n, m = 128$ and 5% noise.

We next evaluate generalization gap over various space-temporal temporal size $N = 64, 128, 256, 512, 1024, 2048, 4096, 16{,}384, 65{,}536, 262{,}144, 10^6$.

Generalization Error We define E_{test} be testing error and E_{train} training error . We report absolute generalization gap as

$$G_{generalization} = |E_{\text{test}} - E_{\text{train}}|$$

Figures 17 and 18 compare the test MSE and the mean absolute clean-reference generalization gap, respectively, as the sample size increases. In particular, we implement unregularized baseline, regularization using fixed β and adaptive β proposed in our framework. For heat and LP1–LP3, regularization consistently reduces both test mse and generalization gap relative to the unregularized Sobolev baseline across various validation sample sizes. What we found is that constant β only improves the rate by only constant, while adaptive regularization as proposed in our scheme results in change of rate order.

Relation to the theoretical parameter choice. Under the manuscript's rescaling $\lambda = \beta/N$, our choice for $\beta := (1/N)^{\frac{1}{2}}$ differs only a bit from the sample-dependent optimal choice in the supercritical regime of Theorem 3.1, which gives

$$\beta^*(N) = N\lambda^*(N) = e^{T/\epsilon} N^{\frac{d+2}{4s+d}} \left[\log(1 + C_2 N)\right]^{\frac{2s-1}{4s+d}}.$$

We admit that our experiments do not exactly verify theoretically optimal regularization, and we notice sample-size dependent tuning regularization parameter in general improve the rate of convergence.

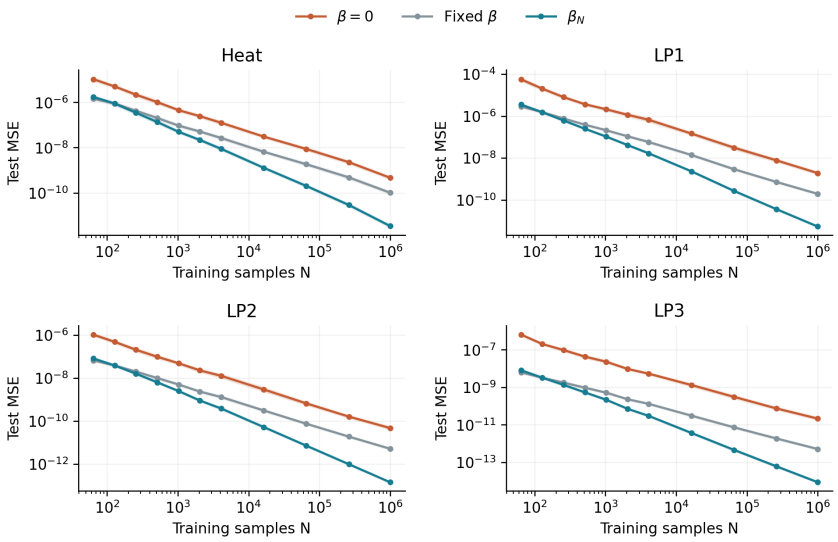

Figure 17: Test MSE versus sample size for parabolic equations with and without regularization.

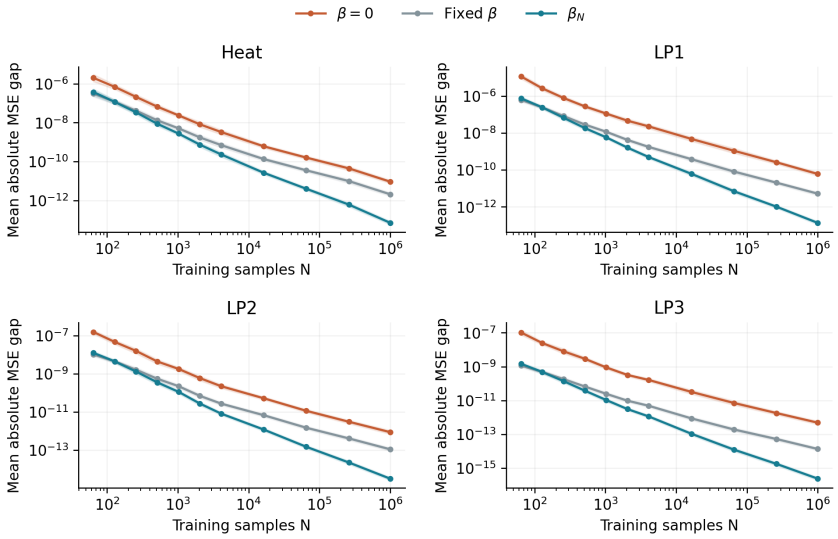

Figure 18: Mean absolute generalization gap versus sample size for parabolic equations with and without regularization.

G.3.1 Noise ablation studies

Noise in the solution and forcing labels. We now keep the parabolic equations in previous experiments, network architecture, sampling procedure, and optimization settings fixed, and vary only the observation noise. We distinguish two observation models:

1. additive noise in the solution and forcing labels, while keeping the derivative labels exact; and
2. additive perturbations applied directly to the derivative labels, while keeping the solution labels exact.

The second experiment is not equivalent to differentiating noisy solution observations. This distinction is discussed explicitly below.

Table 2: Median test relative L^2 error over ten paired seeds for the heatmap-selected mixed-H^1 estimator under additive solution and forcing noise. Derivative labels and validation/test values are exact.

Equation	ρ_u	$\rho_f = 0$	$\rho_f = 0.05$	$\rho_f = 0.10$
LP1	0	1.616×10^{-4}	1.609×10^{-4}	1.608×10^{-4}
LP1	0.05	1.644×10^{-4}	1.641×10^{-4}	1.636×10^{-4}
LP1	0.1	2.070×10^{-4}	2.073×10^{-4}	2.072×10^{-4}
LP2	0	1.604×10^{-4}	1.593×10^{-4}	1.597×10^{-4}
LP2	0.05	1.659×10^{-4}	1.657×10^{-4}	1.654×10^{-4}
LP2	0.1	2.091×10^{-4}	2.074×10^{-4}	2.067×10^{-4}
LP3	0	1.593×10^{-4}	1.591×10^{-4}	1.595×10^{-4}
LP3	0.05	1.660×10^{-4}	1.666×10^{-4}	1.663×10^{-4}
LP3	0.1	2.097×10^{-4}	2.094×10^{-4}	2.090×10^{-4}
HJ	0	5.569×10^{-3}	6.127×10^{-3}	5.839×10^{-3}
HJ	0.05	6.434×10^{-3}	6.075×10^{-3}	5.797×10^{-3}
HJ	0.1	8.666×10^{-3}	1.015×10^{-2}	9.574×10^{-3}

Table 3 extracts the diagonal entries of Table 2. For median error E_{ρ_u,ρ_f}, the change relative to the clean result is $\Delta_\rho = 100(E_{\rho,\rho}/E_{0,0} - 1)\%$.

Table 3: Fresh reruns of the clean and jointly noisy conditions with ten paired seeds. These values are the corresponding diagonal entries of Table 2; percentage changes use unrounded medians.

Equation	$(0,0)$	$(0.05,0.05)$	$(0.10,0.10)$	$\Delta_{0.05}$	$\Delta_{0.10}$
LP1	1.616×10^{-4}	1.641×10^{-4}	2.072×10^{-4}	$+1.50\%$	$+28.17\%$
LP2	1.604×10^{-4}	1.657×10^{-4}	2.067×10^{-4}	$+3.28\%$	$+28.85\%$
LP3	1.593×10^{-4}	1.666×10^{-4}	2.090×10^{-4}	$+4.59\%$	$+31.22\%$
HJ	5.569×10^{-3}	6.075×10^{-3}	9.574×10^{-3}	$+9.09\%$	$+71.90\%$

Across all nine noise settings, the median relative error for LP1–LP3 ranges from 0.0159% to 0.0210%. Across all nine noise settings, the median relative error for HJ ranges from 0.5569% to 1.0154%. The largest signed relative change along the jointly noisy diagonal is $+71.90\%$ for HJ at $\rho_u = \rho_f = 0.10$. The absolute errors and relative changes should be considered together, because a small clean error can produce a large percentage increase.

H PARAMETER SENSITIVITY ANALYSIS

In this section, we present careful analysis on the relation between regularization parameter β and PDE regularization parameter ϵ and total sample size N to examine how N, ϵ, and β jointly affect the benefit of energy regularization over mixed-H^1 data fitting term. Figures 19, 21, 23, and 25 show relative L^2 error and test MSE; the companion figures report the signed generalization gap and the test-MSE ratio to the fitting-only baseline. The latter ratio makes improvements over $\beta = 0$ directly visible.

Interaction between ϵ and β. In the second heatmaps of Figures 22 and 24, the rows with very small β remain close to a ratio of one, while intermediate weights produce localized improvements. For example, in the upper noise row at $N = 100$, $\epsilon = 0.5$, and $\beta = 10^{-2}$, the LP2 and LP3 ratios are approximately 0.63 and 0.64, corresponding to test-MSE reductions of 37.3% and 36.2%. The corresponding relative-L^2 panels in Figures 21 and 23 also improve, by 17.1% and 16.8%. At the same (N, β), the LP2 ratio is approximately one when $\epsilon = 0.01$, whereas the $\beta = 1$ rows show substantial error increases throughout both LP ratio heatmaps. Thus, the benefit depends on the combination of ϵ and β; increasing the energy weight alone does not ensure improvement. This matches with our upper bound where β depends on both $epsilon$ and N.

Dependence on sample size and noise location. Reading the $\beta = 0$ rows from left to right in Figures 19, 21, 23, and 25 shows substantially smaller baseline errors at the largest sample sizes. The orange outlines at $N = 10^6$ in Figures 22 and 24 select $\beta = 0$ for every displayed ϵ and both noise settings, indicating additional improvement for these LP examples at large N. For HJB-3D, the upper noise row of the ratio heatmap in Figure 26 retains a ratio of about 0.92 at $N = 10^6$, $\epsilon = 0.5$, and $\beta = 0.1$. The two noise rows also differ: at $N = 100$ with the same $(\epsilon, \beta) = (0.5, 0.1)$, the HJB ratio is 0.87 for noisy f but 1.98 for noisy u. The corresponding accuracy panels in Figure 25 also shows this reversal relation. These observations support adapting the parameters to the equation, sample size, and noise location.

Prediction accuracy and generalization gap. The first heatmap in Figure 20 shows why the gap must be read together with the MSE ratio. In its upper noise row at $N = 100$ and $\epsilon = 0.2$, increasing β from 0 to 0.1 changes the mean signed gap from approximately 1.8×10^{-8} to -9.4×10^{-7}, while the companion ratio rises to about 24.2. A more negative gap therefore need not indicate better prediction or a smaller train–test discrepancy. In contrast, the upper noise row of Figure 26 at $N = 100$, $\epsilon = 0.5$, and $\beta = 0.1$ shows both a lower test-MSE ratio and a positive mean gap closer to zero than the baseline.

Consistency with the theoretical discussion. The improved regions in Figures 22 and 24 provide empirical support for finite-sample improvement from fine-tuned weighted energy guidance. Their limited gains at the largest N are qualitatively compatible with the subcritical discussion following Theorem 3.1, where energy guidance need not improve the leading-order rate for sufficiently smooth targets. The HJB observations in Figure 26 illustrate the empirical extension beyond the heat-equation setting of that theorem. This agreement remains qualitative: the heatmaps measure value-based errors and a clean value-MSE gap for neural approximations, whereas the theory concerns its specified estimator and the full regularized loss.

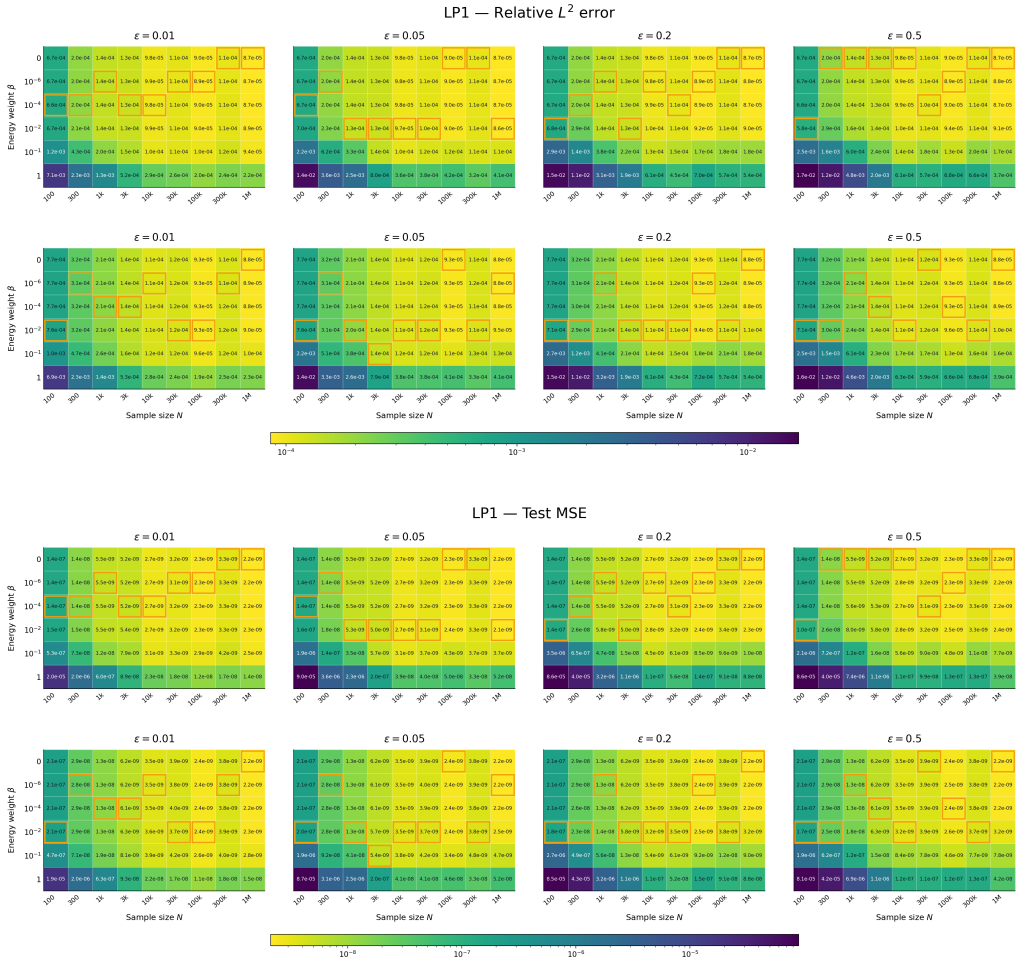

Figure 19: LP1: relative L^2 error (first heatmap) and test MSE (second heatmap). Within each heatmap, the upper row has noisy f and clean u; the lower row has clean f and noisy u. Orange outlines mark the validation-selected β for each fixed (ϵ, N).

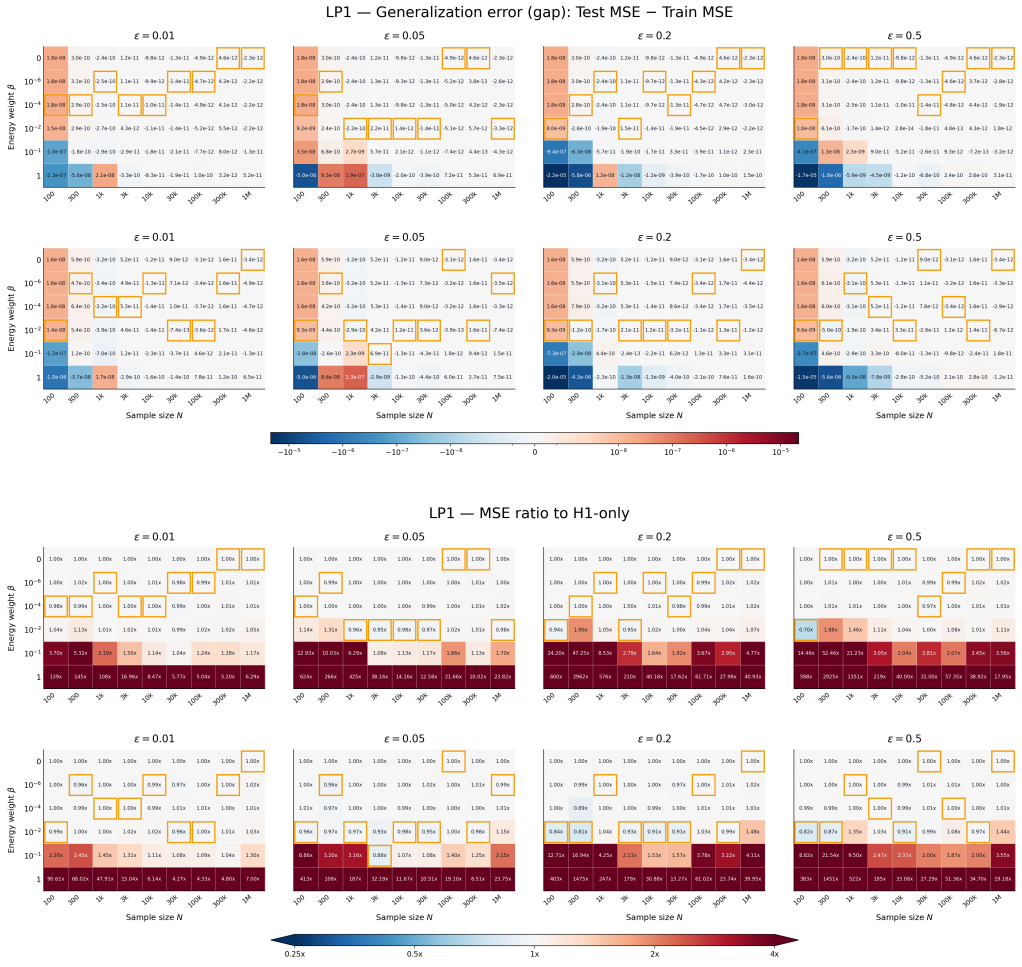

Figure 20: LP1: signed value-MSE generalization gap, $G = E_{\text{test}} - E_{\text{train}}$ (first heatmap), and test-MSE ratio to the paired mixed-H^1 fitting baseline (second heatmap). Gap cells are means of signed gaps across seeds; ratios compare mean test MSEs. Ratios below one indicate improvement. Noise rows and orange outlines follow Figure 19.

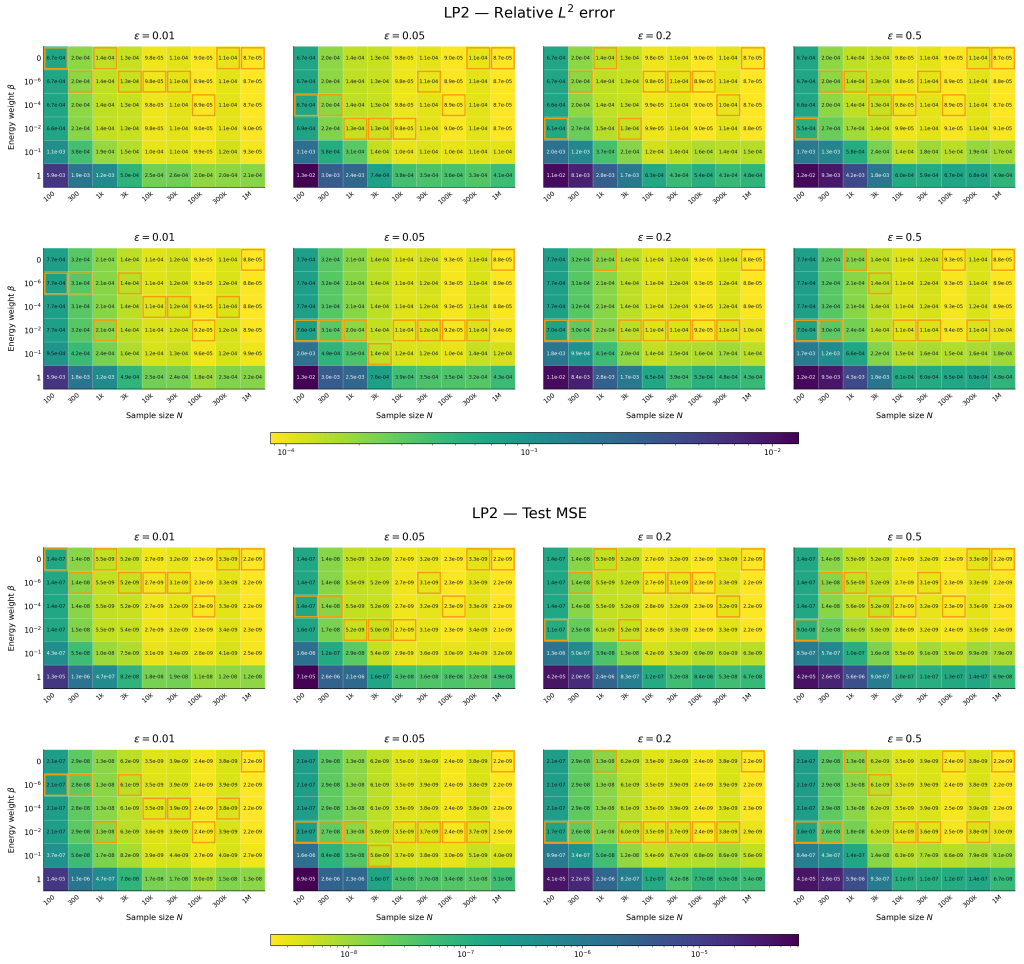

Figure 21: LP2: relative L^2 error (first heatmap) and test MSE (second heatmap). Within each heatmap, the upper row has noisy f and clean u; the lower row has clean f and noisy u. Orange outlines mark the validation-selected β for each fixed (ϵ, N).

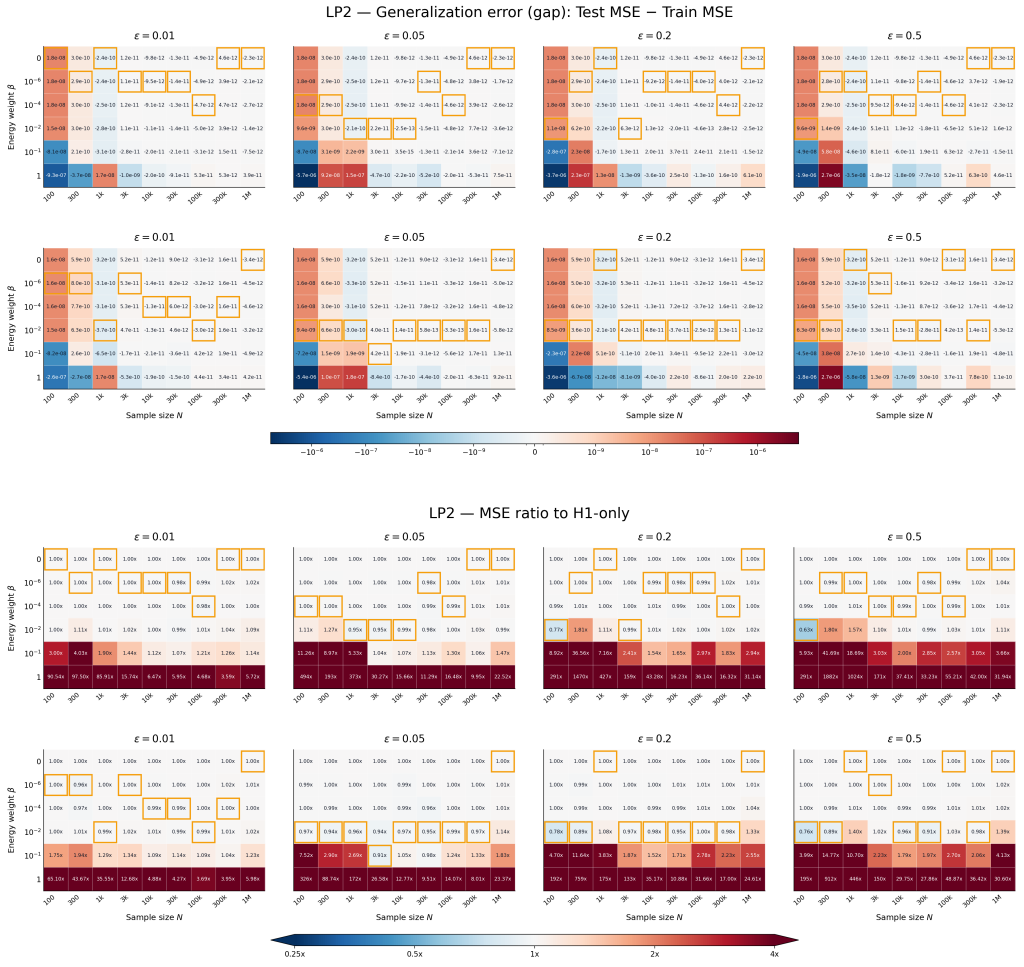

Figure 22: LP2: signed value-MSE generalization gap, $G = R_{\text{test}} - R_{\text{train}}$ (first heatmap), and test-MSE ratio to the paired mixed-H^1 fitting baseline (second heatmap). Gap cells are means of signed gaps across seeds; ratios compare mean test MSEs. Ratios below one indicate improvement. Noise rows and orange outlines follow Figure 21.

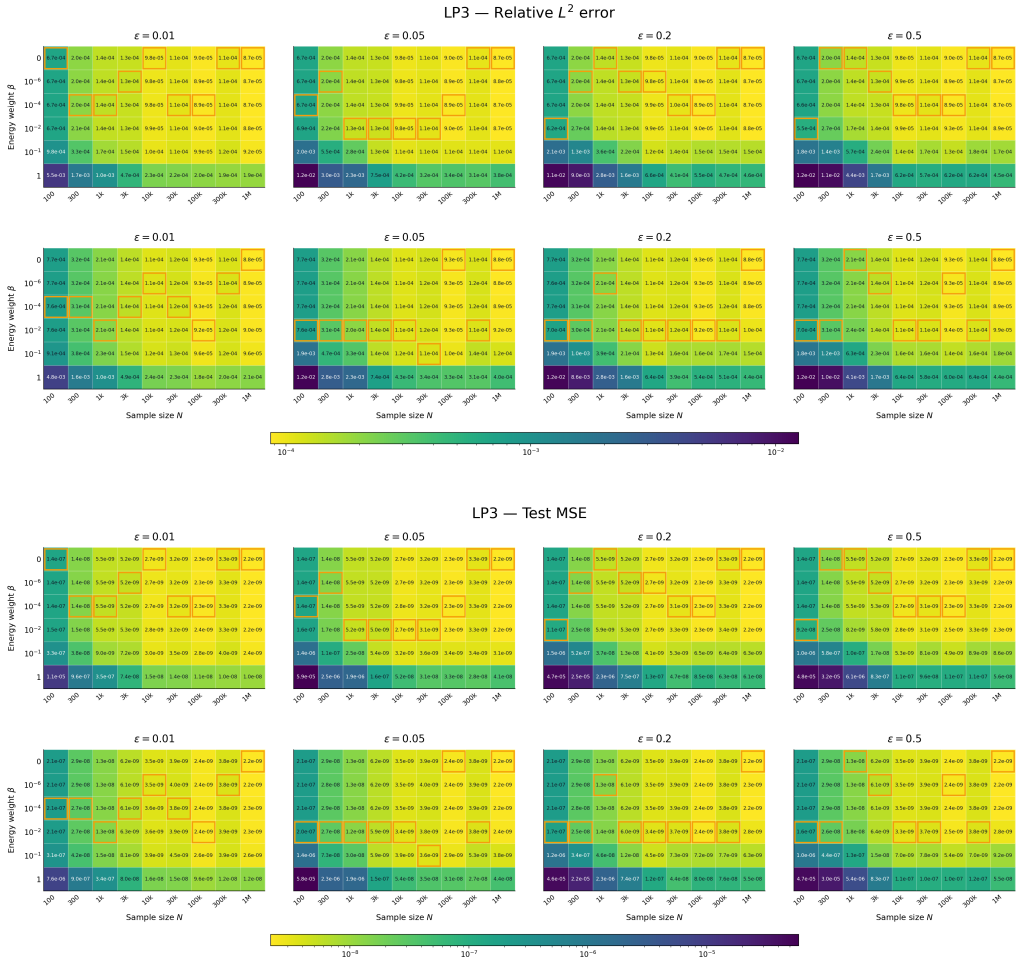

Figure 23: LP3: relative L^2 error (first heatmap) and test MSE (second heatmap). Within each heatmap, the upper row has noisy f and clean u; the lower row has clean f and noisy u. Orange outlines mark the validation-selected β for each fixed (ϵ, N).

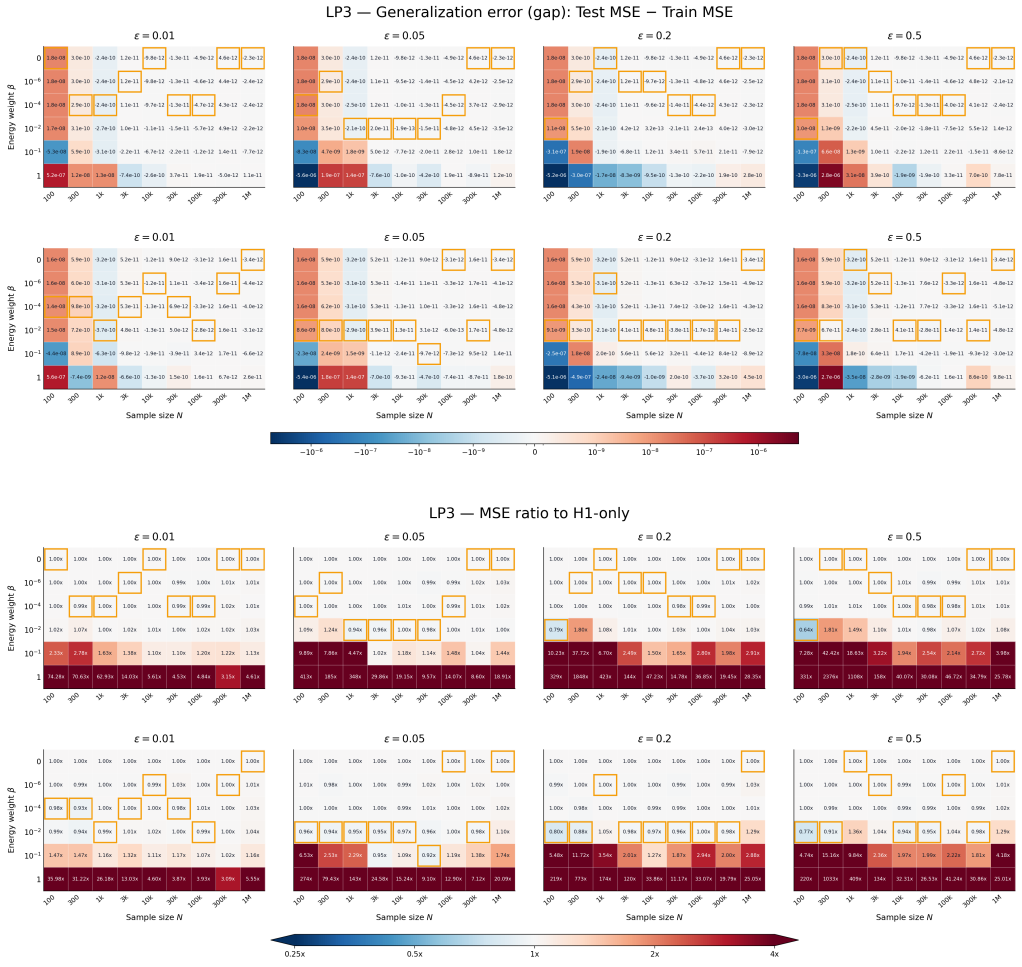

Figure 24: LP3: signed value-MSE generalization gap, $G = R_{\text{test}} - R_{\text{train}}$ (first heatmap), and test-MSE ratio to the paired mixed-H^1 fitting baseline (second heatmap). Gap cells are means of signed gaps across seeds; ratios compare mean test MSEs. Ratios below one indicate improvement. Noise rows and orange outlines follow Figure 23.

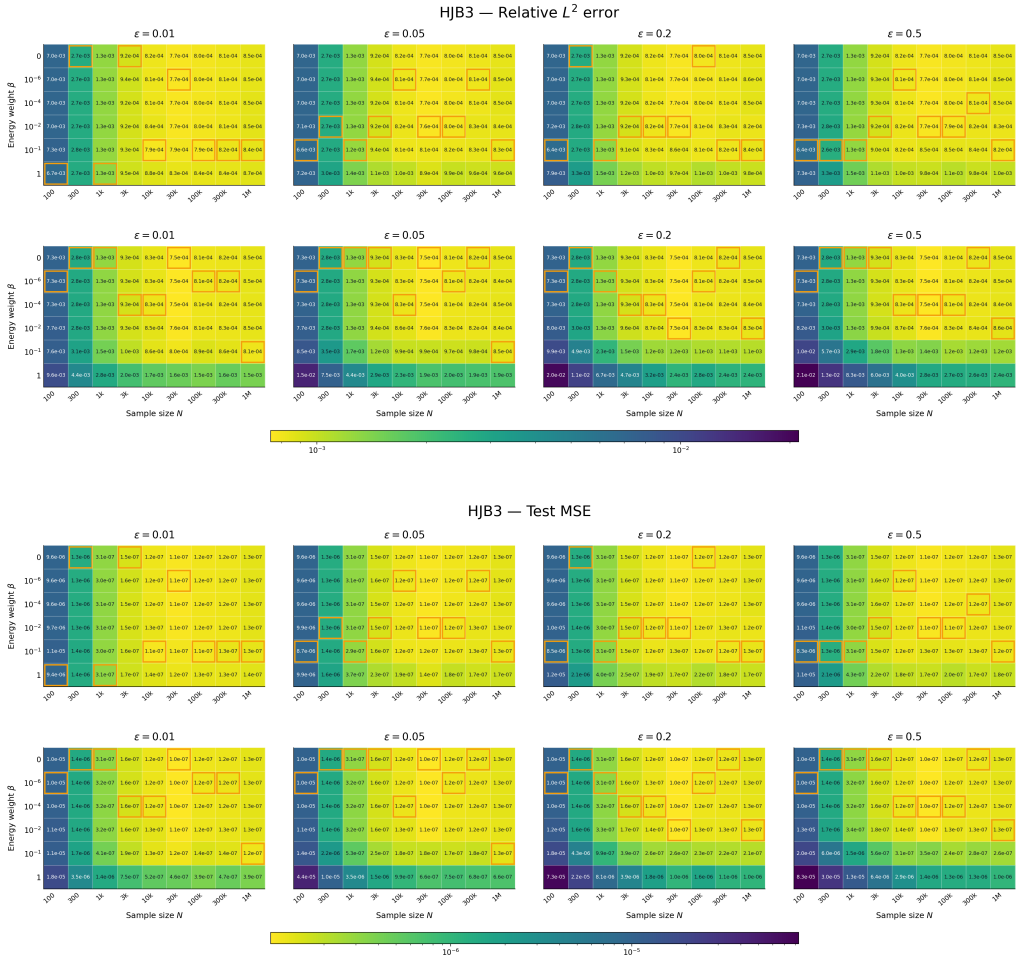

Figure 25: HJB-3D: relative L^2 error (first heatmap) and test MSE (second heatmap). Within each heatmap, the upper row has noisy f and clean u; the lower row has clean f and noisy u. Orange outlines mark the validation-selected β for each fixed (ϵ, N).

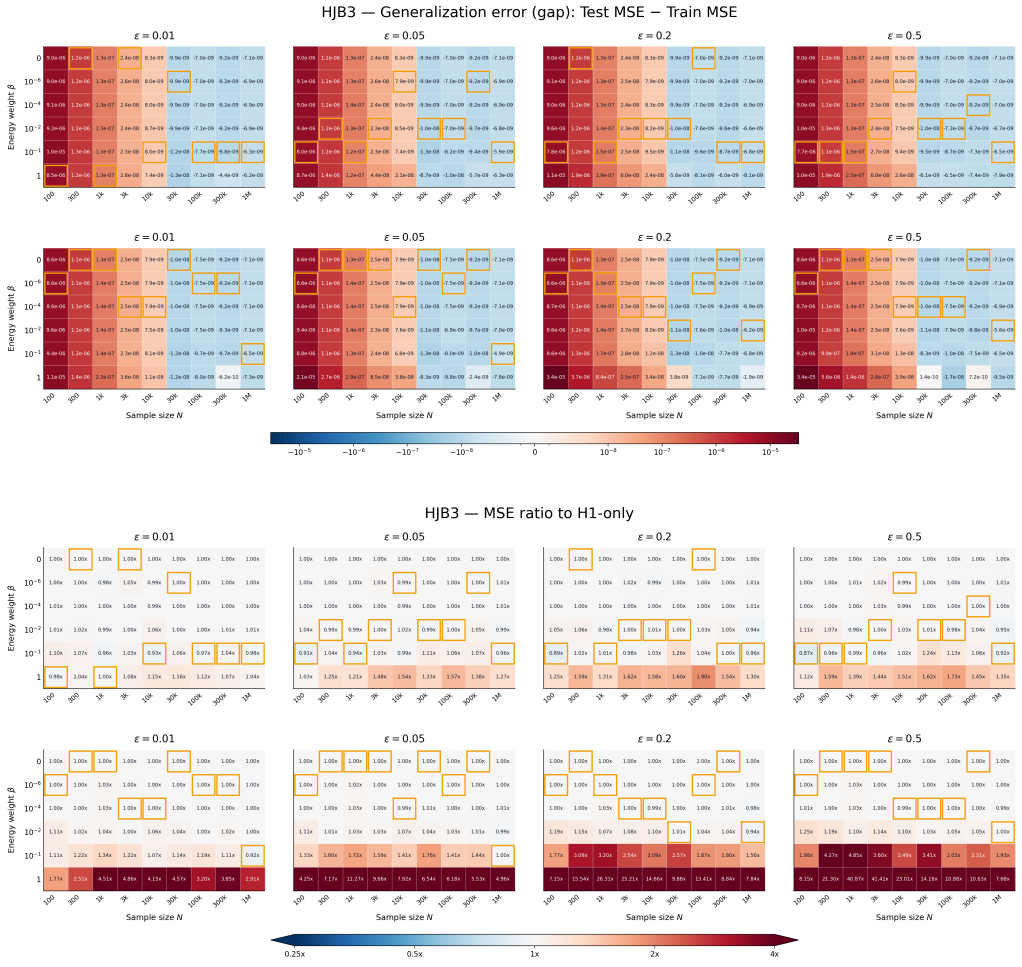

Figure 26: HJB-3D: signed value-MSE generalization gap, $G = R_{\text{test}} - R_{\text{train}}$ (first heatmap), and test-MSE ratio to the paired mixed-H^1 fitting baseline (second heatmap). Gap cells are means of signed gaps across seeds; ratios compare mean test MSEs. Ratios below one indicate improvement. Noise rows and orange outlines follow Figure 25.

I Extension to Hyperbolic Equations

We next use one family of hyperbolic equations to examine the extendability of our framework to see whether the weighted-energy construction can be adapted computationally to a different class of evolutionary PDEs. Further investigation of our method's extendability is worth to investigate in the future.

Nonlinear hyperbolic equations: Sine–Gordon. As a nonlinear hyperbolic example, we consider the Sine–Gordon equation. Variational approaches based on weighted inertia–energy functionals have been developed for nonlinear wave equations; following this principle Akagi et al. (2025a), we use

$$I_\epsilon^{\mathrm{hyp}}(u) := \int_0^T \int_\Omega e^{-t/\epsilon} \left(\frac{\epsilon^2 \rho}{2} |u_{tt}|^2 + \frac{1}{2} |\nabla u|^2 + F(u) \right) dx\, dt. \tag{I.1}$$

The corresponding limiting wave equation has the form

$$\begin{cases} \rho u_{tt} - \Delta u + F'(u) = 0, & (x,t) \in \Omega \times (0,T), \\ u(x,0) = u_0(x), & x \in \Omega, \\ u_t(x,0) = u_1(x), & x \in \Omega. \end{cases} \tag{I.2}$$

For the Sine–Gordon equation, we set

$$\rho = 1, \qquad F(u) = 1 - \cos(u), \qquad F'(u) = \sin(u),$$

so that

$$u_{tt} - \Delta u + \sin(u) = 0.$$

Let

$$s_d(x,t) = \frac{d^{-1/2} \sum_{j=1}^d x_j - vt}{\sqrt{1-v^2}}, \qquad v = 0.3.$$

A planar traveling-wave solution is

$$u_d^*(x,t) = 4 \arctan(\exp(s_d(x,t))). \tag{I.3}$$

The initial displacement and velocity are prescribed from this exact solution:

$$u_0(x) = 4 \arctan(\exp(s_d(x,0))),$$

and

$$u_1(x) = -\frac{2v}{\sqrt{1-v^2}} \operatorname{sech}(s_d(x,0)).$$

We use this experiment only as a computational extension of the energy-guided construction. The statistical theory in the main text remains specific to the parabolic setting.

One-dimensional Sine–Gordon equation. Figure 27 compares the learned solution produced by our method with the manufactured traveling-wave solution.

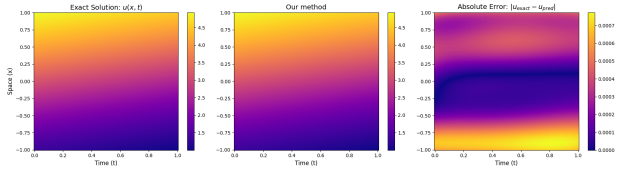

Figure 27: Learned solution from our method and the exact traveling-wave solution for the one-dimensional Sine–Gordon equation.

We use $\epsilon = 10^{-5}$ in (I.1). We compare three basic objectives: the vanilla PINN loss, the weighted-energy functional without labeled solution data, and our complete objective combining data fitting with energy guidance. Figure **??** shows that the weighted-energy functional alone is not sufficient to

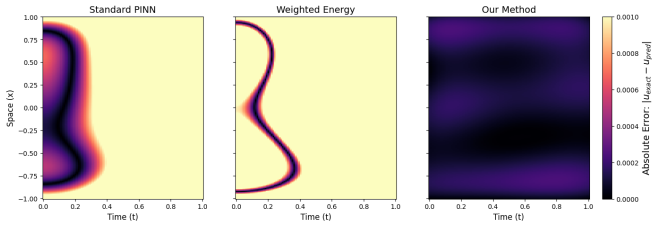

Figure 28: Error comparison for the one-dimensional Sine–Gordon equation: vanilla PINN, the weighted-energy objective without solution labels, and our complete energy-guided estimator.

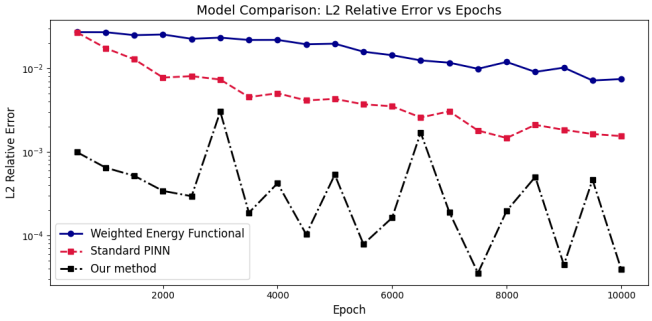

Figure 29: Relative L^2 error over training for the three objectives on the one-dimensional Sine–Gordon equation with 5,000 training samples.

identify an accurate approximation under this training configuration. Its combination with labeled data produces a substantially more accurate solution.

The same conclusion is visible in the relative L^2 error curves in Figure 29. With 5,000 training samples, combining the data-fitting and weighted-energy terms gives the smallest final relative error among the three objectives.

The training curves occasionally contain sharp error spikes. These spikes are observed in both the one- and three-dimensional experiments and are primarily associated with optimization instability in the multi-term objective. In particular, the data-fitting, derivative-label, and energy-guidance terms can generate gradients with substantially different scales. Consequently, a parameter update that decreases the training objective can temporarily increase the test error.

Figure 30 examines the effects of batch size and learning rate. Smaller batches can initially reduce the test error more rapidly than full-batch training, but they also introduce larger stochastic fluctuations. Reducing the learning rate from 10^{-3} to 10^{-4} substantially stabilizes the training trajectory. Appropriate weighting of the derivative and energy terms provides another possible stabilization mechanism, although it introduces the loss-balancing problem commonly encountered in PINN training Wang et al. (2021).

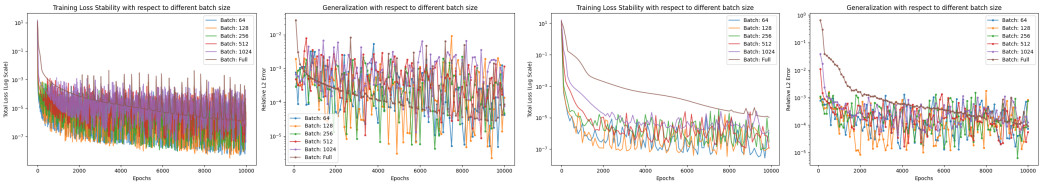

Figure 30: Batch-size and learning-rate ablations for the one-dimensional Sine–Gordon equation. Left: batch-size comparison with learning rate 10^{-3}. Right: training with the reduced learning rate 10^{-4}.

After reducing the learning rate, the comparison in Figure 31 shows more stable convergence of the proposed estimator relative to the baseline methods.

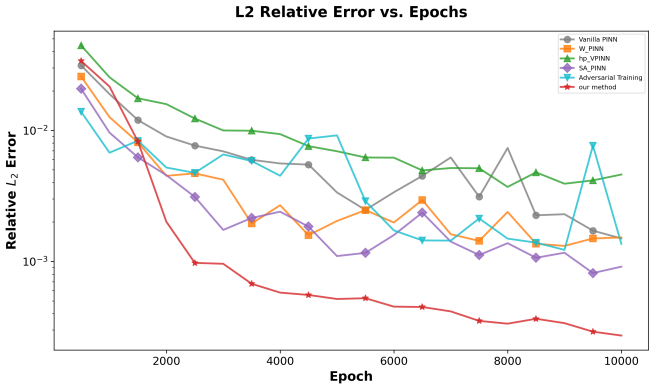

Figure 31: Method comparison for the one-dimensional Sine–Gordon equation with learning rate 10^{-4}.

Three-dimensional Sine–Gordon equation. We also evaluate the planar traveling-wave solution (I.3) in three spatial dimensions. Figure 32 presents representative temporal and spatial slices of the learned solution and its error relative to the manufactured solution.

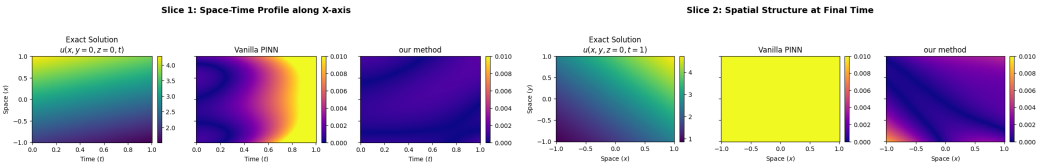

Figure 32: Representative slices for the three-dimensional Sine–Gordon equation. The ground-truth slices are shown together with the corresponding errors of the learned solutions.

The one- and three-dimensional experiments indicate that the energy-guided objective can be adapted computationally to this nonlinear hyperbolic example. They also show that the physical energy term alone is not sufficient under the present data and optimization configuration; the improvement is obtained by combining the energy structure with labeled solution information. It is clear from Table 4 that our method achieves the lowest mean squared error (MSE) and L2 relative error against various PINNs methods that struggles to handle singularity. Moreover, comparisons between our learned solution and ground-truth for temporal evolution snapshots when $z = 0$ are shown in Figure 33.

Method	**Rel. L^2 error**	**MSE**
Vanilla PINNs Karniadakis et al. (2021)	4.01×10^{-3}	1.51×10^{-4}
W-PINNs Wang et al. (2021)	5.32×10^{-3}	2.65×10^{-4}
SA-PINNs McClenny & Braga-Neto (2023)	3.85×10^{-3}	1.39×10^{-4}
Adversarial Training Wang et al. (2022a)	8.25×10^{-2}	6.37×10^{-2}
Our method	$\mathbf{8.89 \times 10^{-4}}$	$\mathbf{7.39 \times 10^{-6}}$

Table 4: Test error comparison for the 3D Sine–Gordon

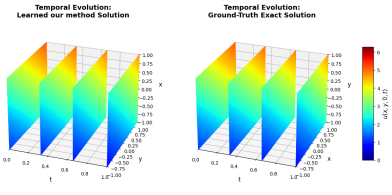

Figure 33: Solution snapshots for the 3D Sine–Gordon equation at $z = 0$.