Neural Stein Critics With Staged $L^2$-Regularization

Matthew Repasky, Xiuyuan Cheng, and Yao Xie, Member, IEEE

Abstract—Learning to differentiate model distributions from observed data is a fundamental problem in statistics and machine learning, and high-dimensional data remains a challenging setting for such problems. Metrics that quantify the disparity in probability distributions, such as the Stein discrepancy, play an important role in high-dimensional statistical testing. In this paper, we investigate the role of $L^2$ regularization in training a neural network Stein critic so as to distinguish between data sampled from an unknown probability distribution and a nominal model distribution. Making a connection to the Neural Tangent Kernel (NTK) theory, we develop a novel staging procedure for the weight of regularization over training time, which leverages the advantages of highly-regularized training at early times. Theoretically, we prove the approximation of the training dynamics by the kernel optimization, namely the “lazy training”, when the $L^2$ regularization weight is large, and training on $n$ samples converge at a rate of $O(n^{-1/2})$ up to a log factor. The result guarantees learning the optimal critic assuming sufficient alignment with the leading eigen-modes of the zero-time NTK. The benefit of the staged $L^2$ regularization is demonstrated on simulated high dimensional data and an application to evaluating generative models of image data.

Index Terms—Stein discrepancy, goodness-of-fit test, neural tangent kernel, lazy training, generative models.

I. INTRODUCTION

UNDERSTANDING the discrepancy between probability distributions is a central problem in machine learning and statistics. In training generative models, learning to minimize such a discrepancy can be used to construct a probability density model given observed data, such as in the case of generative adversarial networks (GANs) trained to minimize $f$-divergences [1], Wasserstein GANs [2], and score matching techniques [3], [4]. Generally, GANs and other generative models require discriminative critics to distinguish between data and a distribution [5]; such critics have the ability to tell the location where the model and data distributions differ as well as the magnitude of departure. Recent developments in the training of generative models have facilitated advancements in out-of-distribution detection [6], in which such models learn to predict the higher likelihood for in-distribution samples. There exists a wide array of integral probability metrics that quantify distances on probability distributions [7], including the Stein discrepancy [8]. In particular, the computation of the Stein discrepancy only requires knowledge of the score function (the gradient of the log density) of the model distribution. Thus, it avoids the need to integrate the normalization constant in high dimensions. This makes the Stein discrepancy useful for evaluating some deep models like energy-based models.

Implicit in the minimization of the discrepancy between a model distribution and observed data is the concept of goodness-of-fit (GoF). In the GoF problem and the closely related two-sample test problem, the goal of the analysis is to approximate and estimate the discrepancy between two probability distributions. Integral probability metrics are widely used for such problems. For example, Reproducing Kernel Hilbert space (RKHS) Maximum Mean Discrepancy (MMD) [9], a kernel-based approach, is used for two-sample testing among other testing tasks. Kernels parameterized by deep neural networks have been adopted recently in [10] to improve the testing power. For GoF tests, where the task is to detect the departure of an unknown distribution of observed data from a model distribution, methods using the Stein discrepancy metrics have been developed. The Stein discrepancy is also calculated using kernel methods [11], [12], [13] and more recently using deep neural network-aided techniques [14]. We provide more background information related to the Stein discrepancy and its role in GoF testing in Section II.

In machine learning, a wide array of modern generative model architectures exist. Energy-based models (EBMs) are a particularly useful subset of generative models. Such models can be described by an energy function which describes a probability density up to a normalizing constant [15]. While such models provide flexibility in representing a probability density, the normalizing constant (which requires an integration over the energy function to compute) is required to compute the likelihood of data given the model. The Stein discrepancy provides a metric for evaluating EBMs without knowledge of this normalization constant [14]. Another popular class of generative models is flow-based models [16], [17]. Flow-based modeling approaches, such as RealNVP and Glow [18], provide reversible and efficient transformations representing complex distributions, yielding simple
log-likelihood computation. In Section V-B, we outline our approach for evaluating generative EBM models using neural Stein critics.

In this paper, we introduce a method for learning the Stein discrepancy via a novel staged regularization strategy when training neural network Stein critics. We consider the \( L^2 \)-regularization of the neural Stein critic, which has been adopted in past studies of neural network Stein discrepancy [14], [19]. Another motivation to use \( L^2 \)-regularization is due to the fact that the (population) objective of the \( L^2 \)-regularized Stein discrepancy is equivalent to the mean-squared error between the trained network critic and the optimal one (up to an additive constant); see (11). It has also been shown previously that the Stein discrepancy evaluated at the optimal critic under \( L^2 \) regularization reveals the Fisher divergence [19]. We analyze the role of the regularization strength parameter in \( L^2 \) neural Stein methods, emphasizing its impact on neural network optimization, which was overlooked in previous studies. On the practical side, our study shows the benefit of softening the impact of this regularization over the course of training, yielding critics which fit more quickly at early times, followed by stable convergence with weaker regularization. An example is shown in Figure 1: (A)-(C) illustrate that the target critic changes in magnitude throughout training as the weight of regularization is decreased, and (D)-(F) shows the rough approximation to the optimum at early times followed by more nuanced changes in the later stages of training.

The proposed staging of \( L^2 \)-regularization is closely connected to the so-called “lazy training” phenomenon of neural networks [20], which approximately solves a kernel regression problem corresponding to the Neural Tangent Kernel (NTK) [21] at the early training times. Theoretically, we prove the kernel learning dynamic of \( L^2 \) neural Stein methods when the regularization weight is large - the main observation is that the role played by the regularization weight parameter is equivalent to a scaling of the neural network function, which itself leads to kernel learning in the case of strong regularization. This theoretical result motivates the usage of a large penalty weight in the beginning phase of training before decreasing it to a smaller value later.

For GoF problems, the trained neural Stein critic provides model comparison capabilities that assess the accuracy of a model’s approximation of the true distribution, allowing for identifying the locations of distribution departure in observed data. This naturally leads to applications for GoF testing and evaluation of generative models. In summary, the contributions of the current work are as follows:

1) We introduce a new method for training neural Stein critics, which incorporates a staging of the weight of the \( L^2 \) regularization over the process of mini-batch training.

2) We prove the NTK kernel learning (lazy-training dynamic) of neural Stein critic training with large \( L^2 \) regularization weight, providing a theoretical justification of the benefit of using strong \( L^2 \) regularization in the early training phase. The analysis reveals a convergence at the rate of \( O(n^{-1/2}) \) up to a log factor, \( n \) being the sample size, when training with finite-sample empirical loss.

3) The advantage of the proposed neural Stein method is demonstrated in experiments, exhibiting improvements over fixed-regularization neural Stein critics and the kernelized Stein discrepancy on simulated data. The neural Stein critic is applied to evaluating generative models of image data.

A. Related Works

The Stein discrepancy has been widely used in various problems in machine learning, including measuring the quality of sampling algorithms [8], evaluating generative models by diffusion kernel Stein discrepancy which unifies score matching with minimum Stein discrepancy estimators [22], GoF testing, among others. For GoF testing, a kernel Stein discrepancy (KSD) approach has allowed for closed-form computation of the discrepancy metric [11], [13]. Similar metrics have been used in the GoF setting, such as the finite set Stein discrepancy (FSSD), which behaves as the KSD but can be computed in linear time [23]. Our work leverages neural networks, which potentially have large expressiveness, to parameterize the critic function space and studies the influence of \( L^2 \) regularization from a training dynamic point of view.

Recent studies have developed alternatives to kernelized approaches to computing the Stein discrepancy. Using neural networks to learn Stein critic functions, [19] applied neural Stein in the training of high-quality samplers from unnormalized densities. The neural network Stein critic has also been applied to the GoF hypothesis test settings to evaluate EBMs [14]. These methods impose a boundedness constraint on the \( L^2 \) norm of the functions represented by the neural network using a regularization term added to the training objective. The optimal critic associated with this method yields a Stein discrepancy equivalent to the Fisher divergence and provides an additional benefit in that the critic can be used as a diagnostic tool to identify regions of poor fit [19]. The staging scheme we introduce in this work is motivated by the observed connection between large \( L^2 \) penalty and lazy training. In practice, it yields an improvement upon past techniques used to learn the \( L^2 \)-penalized critic.

Many methods have been developed to train and evaluate generative models without knowledge of the likelihood of a model. Early works used the method of Score Matching, which minimizes the difference in score function between the data and model distributions using a proxy objective [3]. Methods building on this approach are known as score-based methods. Score-based methods include approaches that can estimate the normalizing constant for computation of likelihoods, as in the case of Noise-Contrastive Estimation [24], and can conduct score matching using deep networks with robust samplers, as in the case of Noise Conditional Score Networks with Langevin sampling [4]. Our approach potentially provides a more efficient training scheme to obtain discriminative Stein discrepancy critics, leading to a metric representing the discrepancy between distributions. We also experimentally observe that the trained Stein critic function indicates the differential regions between distributions of high dimensional data.
The Stein divergence can be interpreted as a divergence to measure the discrepancy between two distributions. The estimation error analysis of neural network function class-based divergence measure has been studied in [25], [26], and [27], where the global optimizer within the neural network function class on the empirical loss is assumed without addressing the optimization guarantee. To incorporate the analysis of neural network training dynamics, the current paper utilizes theoretical understandings developed by the NTK theory [21], [28], [29]. Particularly, we utilize ideas surrounding the “lazy training” phenomenon of neural networks, suggesting the loss through training decays rapidly with a relatively small change to the parameters of the network model, which results in a kernel regression optimization dynamic [20]. For neural network hypothesis testing, the NTK learning was used in computing neural network MMD for two-sample testing [30].

As for when neural network training falls under the lazy training regime, [20] showed that it depends on a choice of scaling of the network mapping. In this paper, we adopt a staging of the weight of $L^2$ regularization, which can utilize the NTK kernel learning regime at early periods of training, and also, in practice, go beyond kernel learning at later phases.

B. Notation

Notations in this paper are mostly standard, with a few clarifications as follows: We use $\partial_x$ with subscript $x$ to denote partial derivative, e.g., $\partial_x f(x, \theta)$ means $\frac{\partial}{\partial \theta} f(x, \theta)$. We use $\mathbb{E}_{x \sim p}$ to denote integral over the measure $p(x)dx$, that is, $\mathbb{E}_{x \sim p}(f) = \int_X f(x)p(x)dx$. We may omit the variable in a short-hand notation of integrals, that is, we write $\int_X f$ for $\int_X f(x)dx$. The symbol $\cdot$ stands for vector-vector inner-product (also used in the divergence operator $\nabla \cdot$), and $\circ$ stands for matrix-vector multiplication.

II. BACKGROUND

We begin by providing necessary preliminaries of the Stein Discrepancy, $L^2$ Stein critics, and GoF testing.

A. Stein Discrepancy

The recent works on Stein discrepancy in machine learning are grounded in the theory of Stein’s operator, the study of which dates back to statistical literature of the 1970s [31]. Let $\mathcal{X}$ be a domain in $\mathbb{R}^d$, and we consider probability densities on $\mathcal{X}$. Given a density $q$ on $\mathcal{X}$, for a sufficiently regular vector field $f : \mathcal{X} \rightarrow \mathbb{R}^d$, the Stein operator $T_q$ [31], [32] applied to $f$ is defined as

$$T_q f(x) := s_q(x) \cdot f(x) + \nabla \cdot f(x),$$

(1)

where $s_q$ is the score function of $q$, defined as

$$s_q := \nabla \log q.$$ 

(2)

Note that the divergence of $f$, denoted by $\nabla \cdot f(x)$, is the same as the trace of the Jacobian of $f$ evaluated at $x$, and throughout this paper $\cdot$ stands for vector inner-product.

Given another probability density $p$ on $\mathcal{X}$ and a bounded function class $\mathcal{F}$ of sufficient regularity, the Stein discrepancy [8] which measures the difference between $p$ and $q$ is defined as

$$SD_{\mathcal{F}}(p, q) := \sup_{f \in \mathcal{F}} SD[f], \quad SD[f] := \mathbb{E}_{x \sim p} T_q f(x).$$

(3)

In this paper, we call $f$ the “critic”, and we call $SD[f]$ the Stein discrepancy evaluated at the critic $f$. We further consider the class of $f$ that satisfy a mild boundary condition, namely $\mathcal{F}_0(p) := \{ f, p f \partial X = 0 \}$. When $X$ is unbounded, $\partial X$ is understood as the infinity boundary condition. When $X$ is bounded, sufficient regularity of $\partial X$ is assumed so that the divergence theorem can apply. When $p = q$, one can verify that for any $f \in \mathcal{F}_0(p)$ and sufficiently regular, $SD[f] = 0$, which is known as the Stein’s identity. The other direction, namely for some $\mathcal{F} \subset \mathcal{F}_0(p)$, $SD_{\mathcal{F}}(p, q) = 0 \Rightarrow p = q$, is also established under certain conditions [33]. This explains the name “critic” because the $f$ which makes $SD[f]$ significantly large can be viewed as a test function (vector field) that indicates the difference between $p$ and $q$.

To proceed, we introduce the following assumption on the densities $p$ and $q$:

Assumption 1: The two densities $p$ and $q$ are supported and non-vanishing on $\mathcal{X}$ (vanishing at the boundary of $\mathcal{X}$), differentiable on $\mathcal{X}$, and the score functions $s_p$ and $s_q$ are in $\mathcal{F}_0(p)$.

When the function class $\mathcal{F}$ is set to be the unit ball in an RKHS, the definition (3) yields the kernelized Stein discrepancy (KSD) [11], [13]. A possible limitation of the KSD approach is the sampling complexity and computational scalablility in high dimension. In this work, we consider critics of regularized $L^2$ norm to be parametrized by neural networks as previously studied in [19] and [14], which have potential advantages in model expressiveness and computation.

B. $L^2$ Stein Critic

This paper focuses on when the critic $f$ in Stein discrepancy (3) is at least squared integrable on $(\mathcal{X}, p(x)dx)$. Consider the critic function $f : \mathcal{X} \rightarrow \mathbb{R}^d$, and the $d$ coordinates of $f$ are denoted as $f(x) = (f_1(x), \ldots, f_d(x))^T$. We first introduce notations of the $L^2$ space of vector fields. Define the inner-product and $L^2$-norm of vector fields on $(\mathcal{X}, p(x)dx)$ as the following: for $v, w : \mathcal{X} \rightarrow \mathbb{R}^d$, let

$$\langle v, w \rangle_p := \int_X v(x) \cdot w(x)p(x)dx,$$

(4)

and then the $L^2$ norm is defined as

$$\|v\|^2_p := \langle v, v \rangle_p.$$

(5)

We stress that, throughout this paper, subscript $p$ in the norm means 2-norm weighted by the measure $p(x)dx$. We denote all critics $f : \mathcal{X} \rightarrow \mathbb{R}^d$ such that $\|f\|^2_p < \infty$ the space of $L^2(p) := L^2(\mathcal{X}, p(x)dx)$.

The Stein discrepancy over the class of $L^2$ critics with bounded $L^2$ norm is defined, for some $r > 0$, as

$$SD_r(p, q) = \sup_{f \in \mathcal{F}_0(p), \|f\|^2_p \leq r} SD[f].$$

(6)
where SD[f] is defined as in (3). Define

$$f^* := s_q - s_p,$$  \hspace{1cm} (7)

the following lemma characterizes the solution of (6), and the proof (left to Appendix A) also derives a useful equality when f and $f^*$ are in $L^2(p) \cap \mathcal{F}_0(p)$, see also e.g. [11, Lemma 2.3]:

$$\text{SD}[f] = E_{x \sim p} T_q f(x) = (f^*, f)_p.$$  \hspace{1cm} (8)

Lemma 2.1: For any $r > 0$, suppose $s_q$ and $s_p$ are in $L^2(p)$ and $\mathcal{F}_0(p)$, then $SD_r(p, q) = r \|f\|_p < \infty$ and, if $\|f\|_p > 0$, the supremum of (6) is achieved at $f = (r/\|f\|_p) f^*$. Lemma 2.1 implies that if $\|f^*\|_p = 0$, then $SD_r(p, q) = 0$. Thus, we take the following assumption, which restricts the case where the $L^2$ Stein critic can achieve a positive discrepancy when $q \neq p$.

Assumption 2: The score functions $s_p$ and $s_q$ are in $L^2(p)$, and when $q \neq p$, $\|s_q - s_p\|_p > 0$.

C. Goodness-of-Fit Tests

In GoF testing, we are presented with samples $X = \{x_i\}_i$ drawn from an unknown distribution $p$, and we wish to assess whether this sample is likely to have come from the model distribution $q$. That is, we may define the null and alternative hypotheses as follows:

$$H_0 : p = q, \hspace{1cm} H_1 : p \neq q.$$  \hspace{1cm} (9)

A GoF test is conducted using a test statistic $\hat{T} = \hat{T}(X)$ computed using observed samples in $X$. After specifying a number $t_{\text{thresh}}$, which is called the “test threshold”, the null hypothesis is rejected if $\hat{T} > t_{\text{thresh}}$. There are different approaches to specifying a test threshold, for example, if prior knowledge of the distribution of $\hat{T}$ under $H_0$ is available, it can be used to choose $t_{\text{thresh}}$. In the experiments of this work, we compute $t_{\text{thresh}}$ using a bootstrap strategy from samples from the model distribution $q$.

The selection of $t_{\text{thresh}}$ is to control the Type-I error, defined as $Pr[\hat{T} > t_{\text{thresh}}] \mid H_0$. The randomness of Pr is with respect to data $X$. The goal is to guarantee that $Pr[\hat{T} > t_{\text{thresh}}] \leq \alpha$, which is called the “significance level” (typically, $\alpha = 0.05$). The Type-II error measures the probability that the null hypothesis is improperly accepted as true, that is, $Pr[\hat{T} \leq t_{\text{thresh}}] \mid H_1$. Finally, the “test power” is defined as one minus the Type-II error. For the application to GoF testing, the current work develops a test statistic computed using a Stein critic parameterized and learned by a neural network, computed from a training-testing split of the dataset. More details will be introduced in Section V-A.

III. METHOD

A. Neural $L^2$ Stein Critic

Replacing the $L^2$-norm constraint in (6) to be a regularization term leads to the following minimization over a certain class of $f$ (inside $L^2(p) \cap \mathcal{F}_0(p)$) as

$$\mathcal{L}_\lambda[f] := -\text{SD}[f] + \frac{\lambda}{2} \|f\|_p^2 = E_{x \sim p} \left(-T_q f(x) + \frac{\lambda}{2} \|f(x)\|_p^2\right),$$  \hspace{1cm} (10)

where $\lambda > 0$ is the penalty weight of the $L^2$ regularization. Under Assumption 1-2, the equality (8) gives that

$$\mathcal{L}_\lambda[f] = -(f^*, f)_p + \frac{\lambda}{2} (f, f)_p = \frac{1}{2\lambda} (\|f - f^*\|_p^2 - \|f^*\|_p^2).$$  \hspace{1cm} (11)

Since $\|f^*\|_p^2$ is a constant independent of $f$, (11) immediately gives that $\mathcal{L}_\lambda[f]$ is minimized at

$$f^\lambda := f^* = \frac{1}{\lambda} (s_q - s_p),$$  \hspace{1cm} (12)

see also Theorem 4.6 of [19]. The expression (12) reveals an apparent issue: if one only considers global functional minimization, then the choice of $\lambda$ plays no role but contributes a scalar normalization of the optimal critic, and consequently does not affect the learned Stein critic in practice, e.g., in computing test statistics. A central observation of the current paper is that the choice of $\lambda$ plays a role in optimization, specifically, in training neural network parameterized Stein critics.

Following [14] and [19], we parameterize the critic by a neural network mapping $f(x, \theta)$ parameterized by $\theta$, and we assume that $f(\cdot, \theta) \in L^2(p)$ for any $\theta$ being considered. We denote this $f(\cdot, \theta)$ the “neural Stein critic”. The population loss of $\theta$ follows from (10) as

$$L_\lambda(\theta) = \mathcal{L}_\lambda[f(\cdot, \theta)].$$  \hspace{1cm} (13)

The neural Stein critic is trained by minimizing the empirical version of $L_\lambda(\theta)$ computed from finite i.i.d. samples $x_i \sim p$, i.e., given $n_{\text{tr}}$ training samples, the empirical loss is defined as

$$\hat{L}_\lambda(\theta) := \frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} (-T_q f(x_i, \theta) + \frac{\lambda}{2} \|f(x_i, \theta)\|_p^2).$$

Remark 3.1 ($\lambda$ as Scaling Parameter): We note a “$\lambda$-scaling” view of the $L^2$-regularized objective $L_\lambda(\theta)$, namely, it is equivalent to a scaling of the parameterized function $f(x, \theta)$ by $\lambda$. Specifically, consider $L_1[f]$ as the “scale-free” ($\lambda$-agnostic) functional minimizing objective, then by definition, we have that

$$\mathcal{L}_\lambda[f] = \frac{1}{\lambda} \mathcal{L}_1[\lambda f].$$  \hspace{1cm} (14)

This $\lambda$-scaling view of the $L^2$ Stein discrepancy allows us to connect to the lazy training of neural networks [20], which indicates that, with large values of $\lambda$, the early stage of the training of the network approximates the optimization of a kernel solution. We detail the theoretical proof of such an approximation in Section IV. This view motivates our proposed usage of large $\lambda$ in early training, which will be detailed in Section III-B.

We close the current subsection by introducing a few notations to clarify the $\lambda$-scaling indicated by (12). Suppose the neural Stein critic $f(\cdot, \theta)$ trained by minimizing $L_\lambda(\theta)$ approximates the minimizer of $L_\lambda$, namely the “optimal critic” $f^\lambda$, then both $f(\cdot, \theta)$ and the value of $\text{SD}[f(\cdot, \theta)]$ would scale like $1/\lambda$. We call $f^\lambda$ defined in (7) the “scaleless optimal critic function”. The expression (12) also suggests that, if the neural
Stein critic successfully approximates the optimal, we would expect
\[ \lambda f(\cdot, \theta) \approx f^*, \]
which will be confirmed by the theory in Section IV. We thus call \( \lambda f(\cdot, \theta) \) the “scaleless neural Stein critic”.

B. Staged-\( \lambda \) Regularization in Training

In practice, the neural network is optimized via the \( \lambda \)-dependent loss \( L_\lambda \). We propose a staging of the regularization weight \( \lambda \) that begins with a large value followed by a gradual decrease until some terminal value. Specifically, we adopt a log-linear staged regularization scheme by which we decrease \( \lambda \) via the application of a multiplicative factor \( \beta < 1 \) after a certain interval of training time measured by the number of batches \( B_w \). In addition to \( B_w \), this staging scheme uses three parameters: the initial weight \( \lambda_{\text{init}} \), the decay factor \( \beta \in (0, 1) \), and the terminal weight \( \lambda_{\text{term}} \) (at which point the decay ceases). We denote the discrete-time regularization staging with these parameters as \( \Lambda(B_i; \lambda_{\text{init}}, \lambda_{\text{term}}, \beta) \), where \( B_i = i \cdot B_w \) stands for the \( i \)-th interval of batches, \( i \in \mathbb{N} \), and \( \lambda \) will be set to the value on the \( i \)-th interval. \( i = 0 \) refers to the period before training has begun such that \( \Lambda(B_0) = \lambda_{\text{init}} \) and, for \( i > 0 \),
\[ \Lambda(B_i; \lambda_{\text{init}}, \lambda_{\text{term}}, \beta) = \max \{ \lambda_{\text{init}} \cdot \beta^i, \lambda_{\text{term}} \}. \] (15)
That is, when \( i \) increments of \( B_w \) number of batches have occurred, \( \lambda_{\text{init}} \) is annealed by a factor of \( \beta \), \( \beta \) can be any positive number in \( (0, 1) \) and we typically choose \( \beta \) to be about 0.7 to 0.9.

The beginning of the training with a large \( \lambda \) is motivated by the analysis in Section IV, which suggests that with large \( \lambda \) at early times of training, the training of the neural Stein critic can be approximately understood from the perspective of kernel regression optimization, rapidly reaching its best approximation in \( \sim 1/\lambda \) time (up to a log factor), see Theorem 4.6 (and Theorem 4.10 for the analysis with finite-sample loss). In many cases, the kernel learning solution may not be sufficient to approximate the optimal critic \( f^* \), and this calls for going beyond the NTK kernel learning in training neural networks, which remains a theoretically challenging problem. On the other hand, the proper regularization strength \( \lambda \), depending on the problem (the distributions \( p \) and \( q \) and the sample size), may be much smaller than the initial large \( \lambda \), which we set to be \( \lambda_{\text{term}} \). We empirically observed that gradually reducing the \( \lambda \) in later training phases will allow the neural Stein critic to progressively fit to \( f^* \), as illustrated in Figure 1. Meanwhile, we also observed that using large \( \lambda \) at the beginning phase of training and then gradually tuning down to small \( \lambda \) performs better than using small and fixed \( \lambda \) throughout, see Section VI. These empirical results suggest the benefit of fully exploiting the kernel learning (by using large \( \lambda \)) in the early training phase and annealing to small \( \lambda \) in the later phase of training.

To further study other possible staging methods, we investigated an adaptive scheme of annealing \( \lambda \) over training time by monitoring the validation error, and the details are provided in Appendix B-E. On simulated high dimensional Gaussian mixture data, the adaptive staging gives performance similar to the scheme (15), and the resulting trajectory of \( \lambda \) also resembles the exponential decay as in (15), see Figure 12. Thus we think the scheme (15) can be a typical heuristic choice of the staging, and other choices of annealing \( \lambda \) are also possible.

C. Evaluation and Validation Metrics

We introduce the mean-squared error metric to quantitatively evaluate how the learned neural Stein critic \( f(x, \theta) \) approximates the optimal critic. In our experiments on synthetic data, we can use the knowledge of \( f^* \) (which requires knowledge of \( s_p \)) and compute an estimator of \( \text{MSE}_q := \mathbb{E}_{x \sim q} \| \lambda f(x, \theta) - f^*(x) \|^2 \) by the sample average on a set of \( n_{\text{te}} \) test samples \( x_i \sim q \), that is,
\[ \hat{\text{MSE}}_q[f(\cdot, \theta)] = \frac{1}{n_{\text{te}}} \sum_{i=1}^{n_{\text{te}}} \| \lambda f(x_i, \theta) - f^*(x_i) \|^2. \] (16)
Similarly, if additional testing or validation samples from data distribution \( p \) are available, we can compute an estimator of \( \text{MSE}_p := \mathbb{E}_{x \sim p} \| \lambda f(x, \theta) - f^*(x) \|^2 \) by the sample average denoted as \( \text{MSE}_p \).

We note that the relationship between \( \text{MSE}_p \) and \( L_\lambda[f(\cdot, \theta)] \) as in (11) leads to an estimator of \( \text{MSE}_p \) (up to a constant) that can be computed without the knowledge of \( f^* \) as follows. Define \( \text{MSE}^{(m)}_{p}[f(\cdot, \theta)] := 2\lambda L_\lambda[f(\cdot, \theta)] = 2\lambda \mathbb{E}_{x \sim p} \left(-T_p f(x, \theta) + \frac{1}{2} \| f(x, \theta) \|^2 \right) \), and (11) gives that \( 2\lambda L_\lambda[f(\cdot, \theta)] = \text{MSE}_p[f(\cdot, \theta)] - \| f^* \|^2_p \). Thus, we can use the sample-average estimator of \( \text{MSE}^{(m)}_{p} \) to estimate \( \text{MSE}_p \) up to the unknown constant \( \| f^* \|^2_p \). That is, given \( n_{\text{val}} \) validation samples \( x_i \sim p \), the estimator can be computed as
\[ \hat{\text{MSE}}^{(m)}_{p}[f(\cdot, \theta)] := \frac{2\lambda}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} \left(-T_p f(x_i, \theta) + \frac{\lambda}{2} \| f(x_i, \theta) \|^2 \right). \] (17)
The superscript \( ^{(m)} \) stands for “monitor”, and in experiments we will use \( \text{MSE}^{(m)}_{p} \) to monitor the training progress of neural Stein critic by a stand-alone validation set.

IV. THEORY OF LAZY TRAINING

In this section, we show the theoretical connection between choosing large \( L^2 \)-regularization weight \( \lambda \) at the beginning of training and “lazy learning” - referring to when the training dynamic resembles a kernel learning determined by the model initialization [20]. The main result, Theorem 4.6 (for population loss training) and Theorem 4.10 (for empirical loss training), prove the kernel learning of neural Stein critic at large \( \lambda \) and theoretically support the benefit of using large \( \lambda \) in the beginning phase of training.

To recall the setup, we assume the score function of model distribution \( s_q \) is accessible, and the score of data distribution \( s_p \) is unknown. We aim to train a neural network critic to infer the unknown scaleless optimal critic \( f^* \) from data. All the proofs are in Section VIII and Appendix A.
A. Evolution of Network Critic Under Gradient Descent

Throughout this section, we derive the continuous-time optimization dynamics of gradient descent (GD). The continuous-time GD dynamic reveals the small learning rate limit of the discrete-time GD, and the analysis may be extended to minibatch-based Stochastic Gradient Descent (SGD) method [34] used in practice. We start from training with the population loss, and the finite-sample analysis will be given in Section IV-D.

Consider a neural Stein critic \( f(\cdot, \theta) \) parameterized by \( \theta \) which maps from \( \mathcal{X} \subset \mathbb{R}^d \) to \( \mathbb{R}^d \). We assume \( \theta \in \Theta \), which is some bounded set in \( \mathbb{R}^{M_\Theta} \), where \( M_\Theta \) is the total number of trainable parameters. The notation \( \langle \cdot, \cdot \rangle_\Theta \) stands for the inner-product in \( \mathbb{R}^{M_\Theta} \), and \( \| \cdot \|_\Theta \) the Euclidean norm. The subscript \( \Theta \) may be omitted when there is no confusion.

**Assumption 3:** The network function \( f(x, \theta) \) is differentiable on \( \mathcal{X} \times \Theta \), and for any \( \theta \in \Theta \), \( f(\cdot, \theta) \in L^2(p) \cap \mathcal{F}_0(p) \).

The boundedness of \( \Theta \) is valid in most applications, and our theory will restrict \( \theta \) inside a Euclidean ball in \( \mathbb{R}^{M_\Theta} \), see more in Assumption 4. Recall that for regularization weight \( \lambda > 0 \), the training objective \( L_\lambda(\theta) \) is defined as in (13). Suppose the neural network parameter \( \theta \) evolving over training time is denoted as \( \theta(t) \) for \( t > 0 \). The GD dynamic is defined by the ordinary differential equation

\[
\dot{\theta}(t) = -\partial_\theta L_\lambda(\theta(t)),
\]

starting from some initial value of \( \theta(0) \). The following lemma gives the expression of (18).

**Lemma 4.1:** For \( \lambda > 0 \), the GD dynamic of \( \theta(t) \) of minimizing \( L_\lambda(\theta) \) can be written as

\[
\dot{\theta}(t) = -\mathbb{E}_{x \sim p} \partial_\theta f(x, \theta(t)) \cdot (\lambda f(x, \theta(t)) - f^*(x)).
\]

Next, we derive the evolution of the network critic over time. We start by defining

\[
u(x, t) := f(x, \theta(t)),
\]

and by chain rule, we have that

\[
\partial_\theta \nu(x, t) = \langle \partial_\theta f(x, \theta(t)), \dot{\theta}(t) \rangle_\Theta.
\]

Combining Lemma 4.1 and (21) leads to the evolution equation of \( \nu(x, t) \) to be derived in the next lemma. To proceed, we introduce the definition of the finite-time (matrix) neural tangent kernel (NTK) \( \mathbf{K}_s(x, x') \) as

\[
[\mathbf{K}_s(x, x')]_{ij} = \langle \partial_\theta f_i(x, \theta(t)), \partial_\theta f_j(x', \theta(t)) \rangle_\Theta,
\]

where \( f_i \) denotes the \( i \)-th coordinate of \( f \). With the notation of \( \mathbf{K}_s(x, x') \) we have the following lemma.
Lemma 4.2: The dynamic of \( u(x,t) \) follows that
\[
\partial_t u(x,t) = -E_{x' \sim p} K_t(x,x') \circ (\lambda u(x',t) - f^\star(x')).
\]
(23)

The evolution equation (23) is exact for \( u(x,t) \), but the kernel \( K_t(x,x') \) changes with \( t \) as the training progresses. Next in Section IV-B, we will replace \( K_t(x,x') \) with the zero-time kernel throughout in the evolution equation, resulting in the more analyzable kernel learning solution.

B. Kernel Learning With Zero-Time NTK
Theoretical studies of NTK to analyze neural network learning [21], [28], [29] typically use two approximations: (i) zero-time approximation, namely \( K_t(x,x') \approx K_0(x,x') \) where both \( K_t \) and \( K_0 \) are finite-width NTK, and (ii) infinite-width approximation, namely at initialization, \( K_0(x,x') \approx K^{(\infty)}_0(x,x') \) as the widths of hidden layers increase, where \( K^{(\infty)}_0 \) is the limiting kernel at infinite width. As has been pointed out in [20], the reduction to kernel learning in “lazy training” does not necessarily require model over-parameterization - corresponding to large width (number of neurons) in a neural network - but can be a consequence of a scaling of the network function. Here we show the same phenomenon where the scaling factor is the \( L^2 \) regularization parameter \( \lambda \), see (14). That is, we show the approximation (i) only and prove lazy training for finite-width neural networks.

We derive the property of the zero-time NTK kernel learning in this subsection, and prove the validity of the approximation (i) in the next subsection. To proceed, consider the kernel \( K_i(x,x') \) defined in (22) at time zero, which can be written as
\[
[K_0(x,x')]_{ij} = \langle \partial_x f_i(x,\theta(0)), \partial_x f_j(x',\theta(0)) \rangle_{\Theta},
\]
(24)

for \( i, j = 1, \cdots, d \). The kernel \( K_0(x,x') \) only depends on the initial network weights \( \theta(0) \), which is usually random and independent from the data samples. We call \( K_0(x,x') \) the zero-time finite-width (matrix) NTK. Following the NTK analysis of neural network training, we will show in Section IV-C that the evolution dynamic of the network function \( u(x,t) \) can be approximated by that of a kernel regression optimization - the so-called “lazy-training” dynamic - which is expressed by replacing the finite-time NTK with the zero-time NTK. For the dynamic in (23), the lazy-training dynamic counterpart is defined by the evolution of another solution \( \bar{u}(x,t) \) by replacing the kernel \( K_t \) with \( K_0 \) in (23) starting from the same initial value, namely, \( \bar{u}(x,0) = u(x,0) \) and \( \partial_t \bar{u}(x,t) = -E_{x' \sim p} K_0(x,x') \circ (\lambda \bar{u}(x',t) - f^\star(x')). \)

(25)

For simplicity, one assumes that at initialization, the network function is zero mapping, that is, both \( u(x,0) \) and \( \bar{u}(x,0) \) are zero. The argument generalizes to when the initial network function is small in magnitude, cf. Remark 4.1, which can be obtained, e.g., by initializing neural network weights with small values.

To analyze the dynamic of (25), we introduce the eigen-decomposition of the kernel in the next lemma.

Lemma 4.3: Suppose \[ \| \partial_x f_i(x,\theta(0)) \|_{\Theta} \] for \( i = 1, \cdots, d \) are squared integrable on \( \mathcal{X} \). The kernel \( K_0(x,x') \) on \( \mathcal{X} \times \mathcal{X} \) has a finite collection of \( M \) eigen-functions \( v_k : \mathcal{X} \to \mathbb{R}^d, k = 1, 2, \cdots, \), associated with positive eigenvalues, in the sense that
\[
\int_{\mathcal{X}} K_0(x,x') \circ v_k(x') p(x') dx' = \mu_k v_k(x),
\]
(26)

where \( \mu_1 \geq \cdots \geq \mu_M > 0 \). The eigen-functions are orthonormal in \( L^2(p) \), namely, \( \langle v_k, v_l \rangle_p = \delta_{kl} \), and for any \( v \) orthogonal to \( \text{span}\{v_1, \cdots, v_M\} \), \( \int_{\mathcal{X}} K_0(x,x') \circ v(x') p(x') dx' = 0 \).

The square integrability of \( \partial_t f_i(x,\theta(0)) \) can be guaranteed by a boundedness condition on \( \partial_{\theta f} \), see more in Assumption 4 below. The finite rank of the kernel, as shown in the proof, is due to the fact that we use a neural network of finite width. Below, we will assume that the optimal critic \( f^\star \) can be efficiently expressed by the span of finitely many leading eigen-functions \( v_k \). To show that the kernel spectrum is expressible enough to approximate an \( f^\star \), one may combine our analysis with NTK approximation (ii): the expressiveness of the limiting kernel \( K^{(\infty)}_0 \) at infinite width can be theoretically characterized in certain settings, meaning that its eigen-functions \( v_k^{(\infty)} \) collectively can span a rich functional space. By the approximation \( K_0 \approx K^{(\infty)}_0 \) in spectrum, the expressiveness of the span of \( v_k \) can also be shown. Such an extension is postponed here.

By Lemma 4.3, the eigen-functions \( \{v_1, \cdots, v_M\} \) form an orthonormal set with respect to the inner-product \( \langle \cdot, \cdot \rangle_p \). Thus for any integer \( m \leq M \), the scaleless optimal critic \( f^\star \in L^2(p) \) (by Assumption 2) can be orthogonally decomposed into two parts \( f^\star_1 \) and \( f^\star_2 \) such that
\[
f^\star = f^\star_1 + f^\star_2,
\]
where \( f^\star_1 \in \text{span}\{v_1, \cdots, v_m\} \) and \( f^\star_2 \in \text{span}\{v_{m+1}, \cdots, v_M\} \). Making use of the orthogonal decomposition and by assuming that \( f^\star \) has a significant projection on the eigen-space of the kernel \( K_0(x,x') \) associated with large eigenvalues, the following proposition derives the optimization guarantee of \( u(x,t) \).

Proposition 4.4: Under the condition of Lemma 4.3 and notations as therein, suppose for \( \delta > 0 \) and some integer \( m \leq M \), \( \mu_1 \geq \cdots \geq \mu_m > \delta > 0 \). Let \( f^\star_2 = f^\star_1 + f^\star_2 \) be the orthogonal decomposition as above. Then, for \( \lambda > 0 \), starting from \( \bar{u}(x,0) = 0 \), for all \( t > 0 \),
\[
||\lambda \bar{u}(\cdot,t) - f^\star||_p \leq e^{-\lambda \delta} ||f^\star_1||_p + ||f^\star_2||_p.
\]
(27)

In particular, if for some \( 0 < \epsilon < 1 \), \( ||f^\star_1||_p \leq \epsilon ||f^\star||_p \), then we have
\[
||\lambda \bar{u}(\cdot,t) - f^\star||_p \leq 2\epsilon ||f^\star||_p, \quad \text{when } t \geq \frac{1}{\lambda} \frac{\log(1/\epsilon)}{\delta}.
\]
(28)

C. Approximation by Lazy-Training Dynamic
The network function \( f(x,\theta) \) maps from \( \mathcal{X} \times \Theta \) to \( \mathbb{R}^d \), and \( \partial_{\theta f}(x,\theta) \) is a \( d \times M_\Theta \) matrix. We denote by \( || \cdot || \) the vector 2-norm and the matrix operator norm. We denote by \( B_r \) the open Euclidean ball of radius \( r \) (centered at the origin) in \( \mathbb{R}^{M_\Theta} \).

Assumption 4: Under Assumption 3, there are positive constants \( r \) and \( L_1, L_2 \) such that \( B_r \subset \Theta \) and
(C1) For any \( \theta \in B_r \), \( \sup_{x \in \mathcal{X}} \| \partial_\theta f(x, \theta) \| \leq L_1 \).
(C2) For any \( \theta_1, \theta_2 \in B_r \), \( \sup_{x \in \mathcal{X}} \| \partial_\theta f(x, \theta_1) - \partial_\theta f(x, \theta_2) \| \leq L_2 \| \theta_1 - \theta_2 \| .

Proposition 4.5: Under Assumptions 1-4, suppose \( \theta(0) \in B_{r/2} \), and \( u(x, 0) = u(0) = 0 \), then for any \( \lambda > 0 \),

(i) For any \( t \geq 0 \) such that \( B_{r/2 + \sqrt{t/(2\lambda)}} \| f^* \|_p \subseteq \Theta \),

\[
\| \theta(t) - \theta(0) \| \leq \sqrt{\frac{t}{2\lambda}} \| f^* \|_p .
\] (29)

(ii) For any \( t \leq \frac{1}{2} \left( \frac{r}{\| f^* \|_p} \right)^2 \lambda \) (recall that \( \| f^* \|_p > 0 \) by Assumption 2),

\[
\| \lambda u(\cdot, t) - \lambda u(\cdot, t) \|_p \leq \frac{4\sqrt{2}}{3} L_1 L_2 \sqrt{\lambda} t^{3/2} \| f^* \|_p^2 .
\] (30)

While the idea of proving Proposition 4.5 largely follows that of Theorem 2.2 in [20], we adopt a slightly improved analysis. Specifically, when we choose \( t \sim 1/\lambda \), both bounds in (29) and (30) reduce to \( O(1/\lambda) \), which echoes Theorem 2.2 in [20]. (Note that our normalization of objective multiplies another factor of \( \lambda \), and thus our \( t \) corresponds to \( T = t \lambda \) for time \( T \) in [20].) Here we would like to derive the approximation up to time \( t \sim \log(1/\epsilon)/\lambda \), corresponding to time \( T \sim \log(1/\epsilon) \) in [20] instead of \( 1 \) time. Technically, our analysis also improves the bounding constant in Theorem 2.2 of [20] by removing a factor of \( (e^{CT} - 1) \), which will become a factor of \( e^{-C} \) when \( T \sim \log(1/\epsilon) \). Thus our improvement is important to apply to the case when \( \epsilon \) is small.

The improvement is by the special property of mean-squared loss, which is equivalent to the neural Stein minimizing loss \( L_\lambda \) up to a constant, cf. (11). We include a proof in Section VIII for completeness.

We are ready to derive the main theorem of this section by combining Propositions 4.4 and 4.5.

Theorem 4.6: Under Assumptions 1-4, let the decomposition of \( f^* \) into \( f_1^* \) and \( u_2^* \) be as in Proposition 4.4 and satisfy the conditions therein and for some \( 0 < \epsilon < 1 \), \( \| f_2^* \|_p \leq \epsilon \| f^* \|_p \). Suppose \( \theta(0) \in B_{r/2} \) and \( u(x, 0) = 0 \), then when \( \lambda > \frac{2 \log(1/\epsilon)}{\delta} 1/2 \| f^* \|_p \), for

\[
t = \frac{t_0 \log(1/\epsilon)}{\lambda \delta} , \quad 1 \leq t_0 \leq \frac{r}{\| f^* \|_p} \frac{\delta}{2 \log(1/\epsilon)} \lambda^2 ,
\] (31)

we have

\[
\| \lambda u(\cdot, t) - f^* \|_p \leq \left( 2\epsilon + \frac{C_1}{\lambda} \left( \frac{t_0 \log(1/\epsilon)}{\delta} \right)^{3/2} \right) \| f^* \|_p .
\] (32)

where \( C_1 = \frac{4\sqrt{2}}{3} L_1 L_2 \| f^* \|_p \) is a constant determined by \( f^* \) and Assumption 4(C1)(C2).

The needed upper bound of \( t \), which \( \sim \lambda \) is technical (to ensure that \( \theta(t) \) stays inside \( B_r \)), the theorem suggests that when the rescalable optimal critic \( f^* \) can be represented using the leading eigen-modes of the zero-time NTK up to an \( O(\epsilon) \) residual, training the neural Stein critic for \( \sim \log(1/\epsilon)/\lambda \) time achieves an approximation of \( f^* \) with a relative error of \( O(\epsilon, 1/\lambda) \) up to a factor involving \( \log(1/\epsilon) \). The theoretical bound in Theorem 4.6 does not depend on data dimension \( d \) or the domain \( \mathcal{X} \) explicitly, however, such dependence are indirect through the constants \( L_1 \) and \( L_2 \) in (C1)(C2). The same applies to the finite-sample analysis in Theorem 4.10 with respect to the constants in Assumptions 4 and 5.

Remark 4.1: (Small Initialization) The result extends to when the initial network function \( u(x, 0) \) is zero non-zero by satisfying \( \| \lambda u(x, 0) \|_p \leq \epsilon \| f^* \|_p \). By considering the evolution of \( u(x, t) \) starting from \( u(x, 0) = u(0) = 0 \), one can extend Proposition 4.5 where the bounds in (29) and (30) are multiplied by \( O(1) \) constant factors (due to that \( \| \lambda u(x, 0) - f^* \|_p \leq (1 + \epsilon) \| f^* \|_p \) when using the argument in (52)). Proposition 4.4 also extends by considering the evolution equation (51) from \( w(x, 0) = \lambda u(x, t) - f^* (x) \), and then \( \lambda u(x, t) \) contributes to another \( \epsilon \| f^* \|_p \) in the bound (28).

D. Training With Finite Samples

In this subsection, we extend the analysis to training using empirical training loss \( L_\lambda(\theta) \) with \( n \) training samples, which can be written as

\[
\hat{L}_\lambda(\theta) = \mathbb{E}_{x \sim p} \left( -T_q f(x, \theta) + \frac{\lambda}{2} \| f(x, \theta) \|^2 \right) ,
\] (33)

where \( \mathbb{E}_{x \sim p} \) denotes the sample average over i.i.d. samples \( x_i \sim p \), i.e., \( \mathbb{E}_{x \sim p} \theta(x) = \frac{1}{n} \sum_{i=1}^n \theta(x_i) \). Again, using continuous-time evolution, the GD dynamic of \( \theta(t) \) is defined by

\[
\dot{\theta}(t) = -\partial_\theta \hat{L}_\lambda(\theta(t)), \quad \theta(0) = \theta(0) ,
\] (34)

where \( \theta(0) \) is some random initialization of the parameters. We define

\[
\hat{u}(x, t) := f(x, \hat{\theta}(t)) ,
\]

assume zero-initialization \( \hat{u}(x, 0) = 0 \), and similarly as in (21) we have

\[
\partial_\theta \hat{u}(x, t) = \langle \partial_\theta f(x, \hat{\theta}(t)), \partial_\theta \hat{u}(x, \hat{\theta}(t)) \rangle \|_{\Theta} .
\] (35)

As the counterpart of (22), we introduce the finite-time empirical NTK as

\[
[\hat{K}_i(x, x')]_{ij} := \langle \partial_\theta f_i(x, \hat{\theta}(t)), \partial_\theta f_j(x', \hat{\theta}(t)) \rangle \|_{\Theta} ,
\]

\( i, j = 1, \cdots, d \), which we also denote as

\[
\hat{K}_i(x, x') = \langle \partial_\theta f_i(x, \hat{\theta}(t)), \partial_\theta f(x', \hat{\theta}(t)) \rangle ,
\] (36)

\( \partial_\theta f(x, \theta) \in \mathbb{R}^{d \times M_n} \). Since \( \theta(0) = \theta(0) \), we have that

\[
\hat{K}_0(x, x') = K_0(x, x') = \partial_\theta f(x, \theta(0)) \|_{\Theta} ,
\] (37)

which is a kernel matrix independent from training data, and this fact is important for our analysis. Using the definition of \( K_i(x, x') \), the dynamic (35) has the following equivalent form.

Lemma 4.7: The dynamic of \( \hat{u}(x, t) \) follows that

\[
\partial_t \hat{u}(x, t) = -\mathbb{E}_{x' \sim p} \left( \hat{K}_i(x, x') \circ (\lambda \hat{u}(x', t) - s_q(x')) - \nabla_{x'} \hat{K}_i(x, x') \right) .
\] (38)

Our analysis will compare the kernel \( \hat{K}_i(x, x') \) to \( K_0(x, x') \), which allows to compare \( \hat{u}(x, t) \) to \( u(x, t) \) where we will also need to control the error by replacing \( \mathbb{E}_{x' \sim p} \).
with $\mathbb{E}_{x \sim p}$. The kernel comparison relies on showing that $\|\tilde{\theta}(t) - \theta(0)\|$ is small, which we derive in the next lemma after introducing additional technical assumptions on the score functions $s_p, s_q$, the function $f(x, \theta)$ and its derivatives. For a set $B \subset \mathbb{R}^d$, we denote by $\bar{B}$ the closure of the set.

**Assumption 5:** Suppose $f(x, \theta)$ is $C^2$ on $\mathcal{X} \times \Theta$. For the $B_r$ as in Assumption 4.

(C3) There is $L_3 > 0$ such that, for any $\theta \in B_r$, $\sup_{x \in X} \|\nabla_x \cdot \partial_x f(x, \theta)\| \leq L_3$.

(C4) There is $L_4 > 0$ such that, for any $\theta_1, \theta_2 \in B_r$, $\sup_{x \in X} \|\nabla_x \cdot \partial_x f(x, \theta_1) - \nabla_x \cdot \partial_x f(x, \theta_2)\| \leq L_4 \|\theta_1 - \theta_2\|$.

(C5) There are positive constants $b(0), b(1)$ and $b_p$ such that, for any $\theta \in B_r$, $\sup_{x \in X} \|\nabla x \cdot \partial_x f(x, \theta)\| \leq b(0)$, and for any $\theta \in B_r$, $\sup_{x \in X} \|\nabla x \cdot \partial_x f(x, \theta)\| \leq b_p$. There are positive constants $C_x$ and $b_\mathcal{X}$ and a positive $\gamma \leq 1/2$, such that, w.p. $\geq 1 - 2n^{-10}$,

$$\max\{ \sup_{\theta \in \bar{B}_r} \|\nabla_x \cdot \partial_x f(x, \theta)\|, \sup_{\theta \in \bar{B}_r} \|\nabla_x \cdot \partial_x f(x, \theta) \cdot s_p(x)\| \} \leq C_x \gamma^2 + b_\mathcal{X} \sqrt{\log n} \sqrt{n} \cdot (39)$$

(C6) There is a constant $b_q > 0$ such that $\|s_q\|_p \leq b_q$, and for any $\theta \in B_r$, $\sup_{x \in X} \|f(x, \theta) \cdot s_q(x)\| \leq b_q$, and

$$\sup_{x \in X} \|\partial_x f(x, \theta)^T s_q(x)\| \leq L_1 b_q.$$ The random variables $\|s_q(x)\|^2$ and $\|f^*(x)\|^2$ with $x \sim p$ are sub-exponential. Specifically, the constant $b_q$ and another constant $b_q > 0$ satisfy that, when $n$ is large (s.t. $\sqrt{\log n} \sim n$ is less than a constant possibly depending on $b_q$ and $b_q$), w.p. $\geq 1 - n^{-10}$,

$$\|\nabla x \cdot \partial_x f(x, \theta)\| \leq \sqrt{20b_q^2 \log \log n}, \quad \sup_{x \in X} \|\nabla x \cdot \partial_x f(x, \theta)\|^2 \leq \sqrt{20b_q^2 \log \log n}.$$

In the below, we adopt big-O notation to facilitate exposition and $\bar{O}$ stands for the involvement of a log factor. The constant dependence can be tracked in the proof. We derive a non-asymptotic result which holds at a sufficiently large finite sample size $n$.

**Remark 4.2 (Uniform-Law and Rates):** The condition (C5) gives the standard uniform-law bounds which can be derived by Rademacher complexity of the relevant function classes over $\theta \in \bar{B}_r$, see e.g. [35], where we treat $s_p$ as a fixed bounded function on $\mathcal{X}$. The $C_x \gamma^2$ term corresponds to the Rademacher complexity which is bounded by the covering complexity of the function class, and the exponent $\gamma$ is determined by the covering number bound that usually involves the dimensionality of the domain and the regularity of the function class. The constant $b_\mathcal{X}$ corresponds to the boundedness of the functions, namely $b(1)$ and $b(0)b_p$. We note that while (C5) gives an overall $\bar{O}(\gamma n^{-1})$ bound, it will only be used in the middle-step analysis (Lemma 4.8) and our final finite-sample bound in Theorem 4.10 achieves the parametric rate of $\bar{O}(n^{-1/2})$.

**Remark 4.3 (Boundedness Condition):** In (C5)-(C6), the uniform boundedness of $f(x, \theta) : s_p(x), f(x, \theta) : s_q(x)$ and $\|\partial_x f(x, \theta)^T s_p(x)\|$ on $\mathcal{X}$ can be fulfilled if $f(x, \theta)$ and $\partial_x f(x, \theta)$ vanish sufficiently fast when approaching $\partial \mathcal{X}$, and thus allowing the score functions $s_p$ and $s_q$ to be potentially unbounded in $L^\infty(\mathcal{X})$. The score functions still need to be in $L^2(p)$ and the tails cannot be too heavy to guarantee that $\|s_q(x)\|^2$ and $\|f^*(x)\|^2$ are sub-exponential. The Bernstein-type concentration bound in (C6) follows from the standard property of sub-exponential random variable, see e.g. [35, Ch. 2].

**Lemma 4.8:** Under Assumptions 1-5, suppose $\theta(0) \in B_{r/2}$, and $\tilde{u}(x, 0) = 0$. Then for any $\lambda > 0$, there is positive integer $n_\lambda$ s.t. when $n > n_\lambda$, under a good event which happens w.p. $\geq 1 - 3n^{-10}$, for any $t \geq 0$ s.t. $\|f^\ast\| \sqrt{\lambda}/\sqrt{r} \leq \lambda$, $\|\tilde{\theta}(t) - \theta(0)\| \leq \sqrt{\lambda}/r|x|$, and $\tilde{\theta}(t) \in B_r$.

Based on the lemma, we derive the comparison of $\tilde{u}(x, t)$ with $\bar{u}(x, t)$ in the following proposition.

**Proposition 4.9:** Under the same assumption as in Lemma 4.8, $\bar{u}(x, 0) = \tilde{u}(x, 0) = 0$. For $\lambda > 0$ and any $t \leq \left( \frac{\sqrt{\lambda}}{r} \right)^2 \lambda$, when $n$ is sufficiently large (depending on $\lambda, t, \|f^\ast\| \sqrt{\log M_{\theta}}$), under the intersection of the good event in Lemma 4.8 and another good event which happens w.p. $\geq 1 - 4n^{-10}$,

$$\|\lambda \bar{u}(t, t) - \lambda \bar{u}(t, t)\| \leq C_2 (1 + \lambda t) \left( \lambda^1/2 \mu^3/2 \|f^\ast\| \right) \left( 1 + \lambda t \right) \lambda \sqrt{\log \log n + \log M_{\theta}} \right) \frac{n}{\lambda} \right) \frac{n}{\lambda}$$

where $C_2$ is an $O(1)$ constant bounded by a multiple of $(1 + b_q + \gamma)$ and the constant factor depends on $L_1, L_2, L_3, L_4$.

**Remark 4.4 (Largeness of $n$ and the relation to $\|f^\ast\|\|p\|$):** The largeness requirement of $n$ depends on $\lambda, t, \|f^\ast\| \sqrt{\log M_{\theta}}$, and specifically, $n > \max\{n_1, n_2, n_3, n_4\}$ which are defined in (69)-(73) respectively: $n_1$ depends on $\lambda$, $\|f^\ast\| \sqrt{\log n}$ and constants $c_2, b_2, c_2, b_2$, and when $\|f^\ast\|$ is small it calls for larger $n$, specifically $\|f^\ast\|^2 > \gamma n^{-\gamma}$ as the leading term, up to constant; $n_3$ depends on $\log M_{\theta}$; $n_4$ depends on $\lambda$ and $\log M_{\theta}$. The construction of the constant $C_2$ can be found in the proof. The first term in the r.h.s. of (40) is the analog of (30) in Proposition 4.5(ii). When $\|f^\ast\| \sqrt{\log M_{\theta}}$ is small, $C_2$ stays bounded (as long as $b_q$ is bounded) and the first term in (40) is proportional to $\|f^\ast\|$. The second term of the order $\bar{O}(n^{-1/2})$ is due to the finite-sample training, and when $\|f^\ast\|$ is small, it suggests that larger $n$ is needed so as to make the second term balance with the first term.

The main theorem for finite-sample lazy training follows by combining Propositions 4.4 and 4.9.

**Theorem 4.10:** Under Assumptions 1-5, suppose the decomposition of $f^\ast$ satisfies the same condition as in Theorem 4.6 with respect to $0 < \epsilon < 1$, $\theta(0) \in B_{r/2}$, and $\bar{u}(x, 0) = 0$. For $\lambda > 2(\log(1/\epsilon))/2 \|f^\ast\| \sqrt{\log M_{\theta}}$, let

$$t = \frac{t_0 \log(1/\epsilon)}{\lambda \delta}, \quad 1 \leq t_0 \leq \left( \frac{r}{\|f^\ast\|^2} \right)^{\delta} \frac{\delta}{4 \log(1/\epsilon)} \lambda X^2.$$ (41)

Then, when $n$ is sufficiently large (depending on $\lambda, \|f^\ast\| \sqrt{\log M_{\theta}}$ and $t_0 \log(1/\epsilon)$), w.p. $\geq 1 - 7n^{-10}$,

$$\|\lambda \bar{u}(t, t) - f^\ast\| \leq \left( 2\epsilon + \frac{C_2}{\lambda} \kappa_1 \right) \|f^\ast\| + \epsilon \sqrt{\log \log n + \log M_{\theta}}.$$ (42)
where $C_2$ is as in Proposition 4.9 and $\kappa_1$, $\kappa_2$ are constant factors involving powers of $t_0 \log(1/\epsilon)$.

The result in Theorems 4.6 and 4.10 suggests that by using a larger $\lambda$ at the beginning of the training process, the training can achieve the NTK kernel learning solution more rapidly. In practice, staying with large $\lambda$ too long would lead to the worsening of the model, and we propose the gradual annealing scheme of $\lambda$ as described in Section III-B so as to combine the benefits of both large $\lambda$ in the beginning and small $\lambda$ in the later phases of training. The theoretical benefit of using large $\lambda$ at the beginning phase of training is supported by experiments in later phases of training. The theoretical benefit of using large $\lambda$ at the beginning and small $\lambda$ in the later phases of training is supported by experiments in Section VI.

V. APPLICATIONS TO TESTING AND MODEL EVALUATION

A. Goodness-of-Fit (GoF) Testing

In a GoF test, we are given $n_{\text{sample}}$ data samples $x_i \sim p$ and a model distribution $q$, and assume we can sample from $q$ as well as access its score function $s_q$. To apply the neural Stein test, we conduct a training-test split of the samples, where the two splits have $n_{\text{tr}}$ and $n_{\text{GoF}}$ samples respectively, and $n_{\text{sample}} = n_{\text{tr}} + n_{\text{GoF}}$. We first train a neural Stein critic $f(x, \theta)$ from the training split $\{x_i\}_{i=1}^{n_{\text{tr}}}$, and then we compute the following test statistic on the test split $\{x_i\}_{i=1}^{n_{\text{GoF}}}$

$$
\hat{T} = \frac{1}{n_{\text{GoF}}} \sum_{i=1}^{n_{\text{GoF}}} T_q f(x_i, \theta),
$$

which can be viewed as a sample-average estimator of $\text{SD}[f(\cdot, \theta)]$ as defined in (3).

To assess the null hypothesis as in (9), we will adopt a bootstrap strategy to compute the test threshold $t_{\text{thresh}}$ by drawing samples from $q$, to be detailed in Section V-A.1. We also derive GoF test consistency analysis in Section V-A.2.

1) Bootstrap Strategy to Compute Test Threshold: The bootstrap strategy draws independent samples $y_i \sim q$ to simulate the distribution of the test statistic under $H_0$. We denote the test statistic as $\hat{T}_{\text{null}}$, which can be computed from a set of $n_{\text{boot}}$ replicas. Three sets of 10,000 test statistics are computed this way the “efficient null statistics”. We observed that setting $r_{\text{pool}} = 50$ is usually sufficient to render a null statistic distribution that resembles that of the fresh statistics. With $n_{\text{boot}} = 500$, this yields about ten times speedup in the computation of the bootstrap.

To illustrate the validity of the bootstrap strategy, we apply the method to a critic trained on 1D Gaussian mixture data as described in Appendix B-B. In this case, a partially trained critic, displayed in Figure 2(A), is used to compute the test statistics. We set $n_{\text{GoF}} = 100$, and to better illustrate the empirical distribution of the test statistics we use $n_{\text{boot}} = 10,000$ replicas. Three sets of 10,000 test statistics are computed: 1) using $n_{\text{GoF}}$ fresh samples from $p$ to compute each $\hat{T}$, 2) using $n_{\text{GoF}}$ fresh samples from $q$ to compute $\hat{T}_{\text{null}}$, and 3) efficient bootstrap null statistics computed from a pre-generated pool with $r_{\text{pool}} = 50$. The empirical distributions are visualized by histograms in Figure 2(B). The plot shows a clear disparity between the distribution of test statistics under $H_1$ and the two distributions of test statistics under $H_0$, and the validity of the efficient bootstrap procedure.
is demonstrated by the similarity of the histograms of the fresh and efficient null statistics.

2) Theoretical Guarantee of Test Power: Let the trained neural Stein critic be \( \hat{f}(x) := f(x, \theta(t)) \). We first observe the asymptotic normality of the test statistics \( T \) and \( T_{null} \): by definition, they are sample averages over i.i.d. test samples \( x_i \sim p(y_i \sim q) \) independent from the training split. Thus, conditioning on training from \( n_t \) samples, the test statistic is an independent sum averaged over i.i.d. random variables

\[
\xi_i := T_q \hat{f}(x_i) = s_q \cdot \hat{f}(x_i) + \nabla \cdot \hat{f}(x_i), \quad i = 1, \ldots, n_{\text{GoF}}.
\]

By Assumption 5(C5)(C6) (and that \( \hat{\theta}(t) \in B_{\rho} \) by Lemma 4.8), \( \xi_i \) are uniformly bounded (either evaluated on \( x_i \) or \( y_i \) in \( \lambda \)). Thus, by Central Limit Theorem, as \( n_{\text{GoF}} \to \infty \),

- \( \sqrt{n_{\text{GoF}}} \bar{T} \) converges in distribution to \( N(\mu_1, \sigma_1^2) \) where \( \mu_1 = \mathbb{E}_{x \sim p} T_q \hat{f}(x) \), and \( \sigma_1^2 = \text{Var}_{x \sim p}(T_q \hat{f}(x)) \),

- \( \sqrt{n_{\text{GoF}}} T_{null} \) converges in distribution to \( N(\mu_0, \sigma_0^2) \) where \( \mu_0 = \mathbb{E}_{y \sim q} T_q \hat{f}(y) = 0 \), and \( \sigma_0^2 = \text{Var}_{y \sim q}(T_q \hat{f}(y)) \).

The closeness to normal density is typically observed when \( n_{\text{GoF}} \) is a few hundred, see Figure 2(B) where \( n_{\text{GoF}} = 100 \). The uniform boundedness of \( \xi_i \) implies that \( \sigma_1^2 \) and \( \sigma_0^2 \) are at most \( O(1) \) constants. This indicates that as long as the training obtains a critic \( \hat{f}(x) \) that makes \( \mu_1 = \mathbb{E}_{x \sim p} T_q \hat{f}(x) > 0 \), then the GoF test can successfully reject the null when \( n_{\text{GoF}} \) is large enough.

We also derive the finite-sample test power guarantee of the neural Stein GoF test in the following corollary by incorporating the learning guarantee in Section IV, and the proof is left to Section VIII.

**Corollary 5.1:** Consider a significance level \( \alpha \) and target Type-II error \( \beta \), \( 0 < \alpha, \beta < 1 \). Suppose the conditions in Theorem 4.10 are satisfied with \( \epsilon < 1/2 \), \( \lambda > 0 \) as required and \( 2\epsilon + C_2 \rho_1 \lambda^{-1} < 0.8 \), and \( n_{tr} \) is large enough for the good event (call it \( E_{tr} \)) to hold, \( C_2 \rho_2 \sqrt{\log n_{tr} + \log M_0} / n_{tr} < 0.1 \|f^*\|_p \) , and \( 7n_{tr}^{10} < 0.1 \min\{\alpha, \beta\} \). Then the GoF test using \( \hat{f}(x) = f(x, \theta(t)) \) achieves a significance level \( \alpha \) and a test power at least \( 1 - \beta \), if

\[
\sqrt{2(b(0)b_q + b(1))} \left( 1 + \sqrt{\log \frac{1}{\alpha}} + \sqrt{\log \frac{1}{\beta}} \right) \lambda n_{\text{GoF}}^{-1/2} < 0.1 \|f^*\|_p^2.
\]

In particular, the test power \( \to 1 \) as \( n_{tr}, n_{\text{GoF}} \to \infty \).

**B. Evaluation of EBM Generative Models**

We consider the application of the Stein discrepancy in evaluating generative models. The model evaluation problem is to detect how the model density \( q \) (by a given generative model) is different from the unknown data density \( p \). While this can be formulated as a GoF testing, in machine learning applications it is of interest where and how much the two densities differ rather than rejecting or accepting the null hypothesis. To this end, we note that the trained Stein critic can be used as an indicator to reveal where \( p \) and \( q \) locally differ.

In the case of EBMs, the model probability density takes the form as \( q(x) = \exp(-E_\phi(x))/Z \), where \( Z \) is the normalizing constant and the real-valued energy function \( E_\phi(x) \) is parameterized by \( \phi \), which is another model such as a neural network (so the set of parameters \( \phi \) differs from the neural Stein critic parametrization \( \theta \)). The score function of \( q \), therefore, equals the gradient of the energy function, i.e.,

\[
s_q = -\nabla E_\phi(x), \tag{44}
\]

which can be computed from the parameterized form of \( E_\phi(x) \). In the case that the energy function is represented by a deep generative neural network, the gradient (44) can be computed by back-propagation, which is compatible with the auto-differentiation implementation of widely-used deep network platforms. In practice, the evaluation metric of a trained EBM can be computed on a holdout validation dataset.

Below we give the expression of the score function for Gaussian-Bernoulli Restricted Boltzmann Machines (RBMs) [14], which is a specific type of EBM. The energy of an RBM with latent Bernoulli variable \( h \) is defined as \( E(x, h|B, b, c) = -\frac{1}{2} x^T B h - B^T x - c^T h + \frac{1}{2} \|x\|^2 \). Therefore, the score function has the expression \( s_q(x) = b - x + B \cdot \tanh(B^T x + c) \). In Section VI-C, we evaluate Gaussian-Bernoulli RBMs using a Stein discrepancy test computed via neural Stein critic functions, and we also compare neural Stein critics trained using different regularization strategies.

**VI. EXPERIMENT**

In this section, we present numerical experiments applying the proposed neural Stein method on differentiating a data distribution \( p \) (from which we have access to a set of data samples) and a model distribution \( q \) (of which the score function is assumed to be known). We compare the proposed neural Stein method with staged regularization to that with fixed \( L^2 \) regularization, as well as to a kernel Stein method previously developed in the literature.

In Sections VI-A and VI-B, we consider a set of Gaussian mixture models for both \( p \) and \( q \). In Section VI-C, the data are sampled from the MNIST handwritten digits dataset [36], and the model distribution is a Gaussian-Bernoulli RBM neural network model. Codes to reproduce the results in this section can be found at the following repository: [https://github.com/mrepasky3/Staged_L2_Neural_Stein_Critics](https://github.com/mrepasky3/Staged_L2_Neural_Stein_Critics).

**A. Gaussian Mixture Data**

1) Simulated Datasets: We compare the performance of a variety of neural Stein critics in the scenario in which the distributions are bimodal, \( d \)-dimensional Gaussian mixtures, following an example studied in [10]. The model \( q \) has equally-weighted components with means \( \mu_1 = 0_d \) and \( \mu_2 = 0.5 \times I_d \), both having identity covariance. The data distribution \( p \) has the following form:

\[
p = \frac{1}{2} \mathcal{N}\left( \mu_1, \begin{bmatrix} 1 & \rho_1 & 0_{d \times 2}^T \\ \rho_1 & 1 & 0_{d \times 2}^T \\ 0_{d \times 2} & 0_{d \times 2} & I_d \\ \end{bmatrix} \right) + \frac{1}{2} \mathcal{N}\left( \mu_2, \begin{bmatrix} \omega^2 & \omega \rho_2 & \omega \rho_2 \\ \omega \rho_2 & 1 & 0_{d \times 2} \\ \omega \rho_2 & 0_{d \times 2} & I_d \\ \end{bmatrix} \right), \tag{45}
\]
where \( \rho_1 = -\rho_2 \) represents the covariance shift with respect to the model distribution. We also introduce a parameter \( \omega \), which scales the covariance matrix of the second component of the mixture. We examine this scenario in three settings of increasingly high dimensions. In each setting of this section, we fix the parameters \( \rho_1 = 0.5 \) and \( \omega = 0.8 \).

2) Experimental Setup:

a) Neural network training: All neural Stein critics trained are two-hidden-layer MLPs with Swish activation [37], where each hidden layer includes 512 hidden units. The weights of the linear layers of the models are initialized using the standard PyTorch weight initialization, and the biases of the linear layers are initialized as zero. The Adam optimizer (using the default momentum parameters by PyTorch) was used for network optimization. The critics are trained using \( \nabla \cdot \) which scales the covariance matrix of the second component \( \rho \) where \( \lambda \) refers to the significance level. We use \( 500 \) samples from the data distribution \( p \); the learning rate is fixed at \( 10^{-3} \), and the batch size is 200 samples. All models are trained for 60 epochs. To compute the divergence \( \nabla \cdot f(x, \theta) \) in dimension greater than two, we use Hutchinson’s unbiased estimator of the trace Jacobian [38] following [14]. In these experiments, the staging of \( \lambda \) according to the staged regularization strategies occurs every \( B_w = 10 \) batches, which is equivalent to the end of every epoch. For each choice of regularization strategy, we train 10 neural Stein critic network replicas. We compare the neural Stein critics trained using fixed-\( \lambda \) regularization vs. staged regularization schemes.

b) Computation of MSE: Having knowledge of the score functions of both \( p \) and \( q \), we are able to compute the \( \text{MSE}_q \) (16) using \( n_{te} = 20,000 \) samples from the model \( q \). Over the course of training, we also computed \( \text{MSE}_p \) (17) using \( n_{val} = 1,000 \) samples from the data distribution \( p \). This monitor is used to select the “best” model over the course of training, where the model with the lowest \( \text{MSE}_p \) value is selected.

c) Computation of test power: After a trained neural critic is obtained, we perform the GoF hypothesis testing as described in Section V-A, including computing the test threshold \( t_{\text{thresh}} \) by the efficient bootstrap strategy. We set the significance level \( \alpha = 0.05 \), the number of bootstrap \( n_{\text{boot}} = 500 \), and the efficient bootstrap ratio number \( r_{\text{pool}} = 50 \). We use \( n_{\text{GoF}} \) specific for each data distribution example, see below.

By definition, the test power is the probability at which the null hypothesis \( H_0 : p = q \) is correctly rejected. For each trained neural critic from a given training split of \( p \) samples, we estimate the test power of this critic empirically by conducting \( n_{\text{run}} \) times of the GoF tests (including independent realizations of the test split of \( p \) samples and computing the \( t_{\text{thresh}} \) by bootstrap) and counting the frequency when the null hypothesis is rejected. Because the training also contributes to the randomness of the quality of the GoF test, we repeat the procedure for \( n_{\text{replica}} \) replicas (including training the model on an independent realization of the training split in each replica), and report the mean and standard deviation of the estimated test power from the replicas. We use \( n_{\text{run}} = 500 \) and \( n_{\text{replica}} = 10 \) in our experiments.

d) Setup per example: We introduce the specifics of the experimental setup for Gaussian mixtures in 2D, 10D, and 25D below.

(a) 2-dimensional mixture: The selected fixed-\( \lambda \) regularization strategies in 2 dimensions are \( \lambda \in \{1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}, 1 \times 10^{0}\} \). We compare these fixed schemes to a few staged regularization strategies. Using the notation of Equation (15), these are the \( \Lambda(1 \times 10^{0}, 5 \times 10^{-2}, 0.95) \) and \( \Lambda(1 \times 10^{0}, 5 \times 10^{-2}, 0.90) \) staged regularization schemes. In 2D, the number of test samples for the GoF testing is \( n_{\text{GoF}} = 75 \).

(b) 10-dimensional mixture. For this setting, we examine fixed \( \lambda \in \{2.5 \times 10^{-4}, 1 \times 10^{-3}, 1 \times 10^{-2}, 6.4 \times 10^{-2}, 2.56 \times 10^{-1}, 1.024 \times 10^{0}\} \). We analyze the staging schemes \( \Lambda(5 \times 10^{-1}, 1 \times 10^{-3}, 0.80) \) and \( \Lambda(5 \times 10^{-1}, 1 \times 10^{-3}, 0.85) \). For the GoF power analysis, we use \( n_{\text{GoF}} = 200 \) test samples from the data distribution \( p \).

(c) 25-dimensional mixture: The fixed regularization weights are selected as \( \lambda \in \{2.5 \times 10^{-4}, 1 \times 10^{-3}, 4 \times 10^{-3}, 1 \times 10^{-2}, 6.4 \times 10^{-2}\} \). The staging schemes analyzed in this case are \( \Lambda(4 \times 10^{-1}, 5 \times 10^{-2}, 0.80) \), \( \Lambda(4 \times 10^{-1}, 5 \times 10^{-2}, 0.85) \), and \( \Lambda(4 \times 10^{-1}, 5 \times 10^{-2}, 0.90) \). We use \( n_{\text{GoF}} = 500 \) test samples from the data distribution \( p \) for the GoF tests in 25D.

3) Results:

a) 2-dimensional mixture: The value of \( \overline{\text{MSE}}_q \) computed via Equation (16) throughout training for all 2D critics can be seen in Figure 3(A), where the mean and standard deviation are plotted across the 10 networks for each staging strategy. We begin by examining the training behavior for fixed regularization weight \( \lambda \). Rapid descent in \( \text{MSE}_q \) is observed for large \( \lambda \) early in training while small \( \lambda \) make slower and steadier progress. The staged regularization strategies exploit this rapid descent at early times, followed by steady late-stage training. However, note that the advantage of staging in 2 dimensions only yields a marginal benefit in MSE. In Figure 4, we visualize the critic vector field plotted through training for the \( \Lambda(1 \times 10^{0}, 5 \times 10^{-2}, 0.90) \) staging scheme. Rapid fitting of some regions can be seen at the beginning of training in Figure 4(A), followed by more nuanced adjustments in Figures 4(B) and 4(C), resulting in a good fit to the theoretically optimal critic function in Figure 4(D). Furthermore, consider the test power results displayed in Figure 5(A). While the staged regularization cases have among the lowest \( \text{MSE}_q \) values, similar testing power is achieved by the fixed-\( \lambda \) settings. For more detail regarding the 2-dimensional experiment, see Table I in Appendix B-A, which shows the average GoF hypothesis testing power at the approximate “best” training epoch as chosen by finding the average lowest monitor \( \overline{\text{MSE}}_p \) calculated using Equation (17). The table also displays the average \( \overline{\text{MSE}}_q \) as calculated by Equation (16) at the “best” epoch for each regularization scheme.

b) 10-dimensional mixture: As in 2D, the \( \overline{\text{MSE}}_q \) is plotted for critics over the course of training in Figure 3(B). The curves corresponding to some fixed \( \lambda \) strategies, such as \( \lambda = 6.4 \times 10^{-2}, 2.56 \times 10^{-1}, \) and \( 1.024 \times 10^{0} \) are omitted since they quickly diverge in \( \overline{\text{MSE}}_q \). In higher dimensions, the
difference in using a wider range of regularization weights $\lambda$ is more pronounced. Larger values result in networks that rapidly fit a (relatively) poor representation of the optimal critic $f^*$, followed by dramatic overfitting. Smaller choices of fixed $\lambda$ delay both phenomena, fitting to a critic which has a lower value of $\text{MSE}_q$ at best. The staged regularization strategies exploit both types of training dynamic, descending in $\text{MSE}_q$ more rapidly than most fixed-$\lambda$ strategies while achieving lower value (and hence better fit) than any fixed strategy, on average. Examining the results in Figure 5(C) further exemplifies the increase to 25 dimensions. The advantage of staged regularization becomes more pronounced with increasing dimension.

B. Comparison to Kernel Stein Discrepancy

Using the simulated Gaussian mixture data as in Section VI-A, we compare the testing power of the neural Stein critic GoF hypothesis test to that of KSD. To do so, we perform the GoF hypothesis test by computing the KSD test statistic outlined in [11] and [13], whereby the Stein discrepancy is computed using a critic restricted to an RKHS. Following common practice in the literature, we construct this RKHS to be defined by a radial basis function (RBF) with a bandwidth equal to the median of the data Euclidean distances in a given GoF test. We also compare our method to a KSD test with RBF bandwidth which is selected to maximize the power. To compute the KSD test, we adopt the implementation of [13]. Further details of the KSD GoF test, including selecting the most effective bandwidth, are given in Appendix B-C, with GoF testing power for a range of bandwidths displayed in Figure 9.

It would be natural to ask how the neural network test compares with other traditional parametric and non-parametric tests. There are cases where the generalized likelihood ratio (GLR) test can be computed from parametric models and would be the optimal test. When a parametric model is not available, traditional non-parametric tests like Kolmogorov-Smirnov may be restricted to low-dimensional data. We thus focus on comparison with the KSD test which is a kernel-based non-parametric test generally applicable to high-dimensional data.

1) Experimental Setup: Using the model distribution $q$ and data distribution $p$ as defined in Section VI-A.1, we construct a two-component Gaussian mixture in 50D to compare the power and computation time of the neural Stein discrepancy GoF hypothesis test (Section V-A) to the KSD test (Appendix B-C). The model distribution $q$ remains as the isotropic, two-component Gaussian mixture, while the data distribution $p$ has the form of (45) with covariance shift $\rho_1 = 0.5$ and covariance scaling $\omega = 0.5$.

a) Computation of test power: For all neural Stein critic GoF tests, a total of $n_{tr} + n_{GoF} = n_{sample}$
samples from \( p \) are used, where \( n_{\text{sample}} \) takes on values in \{100, 200, 300, 500, 1000, 1500, 2500\}. In each case, \( n_{\text{str}} \) and \( n_{\text{GoF}} \) are equal to \( n_{\text{sample}}/2 \). The 50%/50% training/testing split was chosen by comparing the testing power for various training/testing splits. Our findings indicate a range of training/testing splits exists for which the testing power performance is comparable. While the training duration contributes most significantly to the computation time, the test achieves high power once the training split reaches 50% of the sample size. That is, the 50%/50% train/test split is a generic split choice in this range of high power splits, and therefore we use this split in the results outlined in Section VI-B.2 and in Figure 6. The details of this finding can be found in Appendix B-D, Figure 10, and an application to a simpler setting is displayed in Figure 11. Furthermore, the training data are partitioned into an 80%/20% train/validation split. The testing procedure and test power computation are as described in Section VI-A.2. Specifically, the significance level \( \alpha = 0.05 \) in all tests, the number of bootstraps \( n_{\text{boot}} = 500 \), and the efficient bootstrap ratio is \( r_{\text{pool}} = 50 \).

The KSD tests [13] are computed over the same range of samples size. The KSD test is conducted with an RBF kernel using the median data distance heuristic for bandwidth, and we also examine the selected bandwidth chosen to maximize the test power. To compute \( t_{\text{thresh}} \), the KSD test uses a “wild bootstrap” procedure [13]. We provide details of bandwidth selection and wild bootstrap in Appendix B-C. The number of wild bootstrap samples used for KSD is equal to the number of bootstrap samples used for the neural Stein test, namely \( n_{\text{boot}} = 500 \).

The test power is computed using \( n_{\text{run}} = 400 \) and \( n_{\text{replica}} = 5 \) for each KSD and neural Stein GoF test.

b) Comparison of test time: We compare the time taken to perform one KSD GoF test vs. training a neural Stein critic and performing a neural GoF test. All reported times are elapsed time on a laptop with an Intel Core i7-1165G7 processor with 16 GB of RAM. We record and report the computation times of the tests over a range of sample sizes.

For the neural test, we measure the duration time of training followed by the computation time of a single test statistic and the computation time of the bootstrap, averaged over \( n_{\text{replica}} = 5 \) replicas with \( n_{\text{run}} = 10 \) and \( n_{\text{boot}} = 500 \).

To determine the computation time of the KSD test, we measure the duration of 10 tests using the best-selected bandwidth, which can be broken down into the time taken to compute an individual test statistic and the computation time of the wild bootstrap, averaged over \( n_{\text{replica}} = 5 \), \( n_{\text{run}} = 10 \), and \( n_{\text{boot}} = 500 \).

c) Neural network training: The neural Stein critics are learned using two-hidden-layer (512 nodes) MLPs with Swish activation, initialized using PyTorch standard initialization with biases set to 0. We use the Adam optimizer with a learning rate set to \( 5 \times 10^{-3} \), minibatch size equal to 40, and the critics are trained for 60 epochs. The \( \Lambda(4 \times 10^{-1}, 5 \times 10^{-4}, 0.85) \) staging is used in the training of the neural Stein critics, where the frequency of \( \lambda \) staging, \( B_w \) batches, is again chosen to be equivalent to the number of training batches per
epoch. The networks are trained using \( n_{\text{tr}} \) number of training samples from the data distribution \( p \), which are split into an 80%/20% train/validation split for model selection. As in Section VI-A.2, the networks are selected to minimize the monitor \( \text{MSE}_p \) of Equation (17).

2) Results: The result of the testing power comparison on the 50D data can be found in Figure 6. First, examining the comparison in power between the methods in Figure 6(A), the neural Stein critic GoF test (solid black line) remains more powerful than the KSD GoF tests for all sample sizes. The KSD GoF test using the median data distances heuristic bandwidth (solid red line) does not achieve comparable power to the staged-regularization neural Stein critic approach, even for as many as 2500 samples. While the KSD test using the best-selected bandwidth (dashed red line) achieves much higher power than the heuristic approach, the power of this method still falls below that of the neural Stein GoF test. Next, Figure 6(B) highlights the quadratic time complexity of the KSD GoF test, resulting in a dramatic increase in computation time that surpasses the overall time to train a neural Stein critic and compute a test for samples sizes larger than 1,000. Furthermore, the breakdown into test statistic computation time and bootstrap computation time in Figure 6(C) reveals that the time to compute an individual test statistic for the neural Stein approach becomes more time-efficient than KSD when \( n_{\text{sample}} \) is just larger than 1500. While the test statistic computation and wild bootstrap of KSD contribute substantially to the overall computation time of KSD, the neural Stein GoF test computation time is dominated by the training period.

Our findings indicate that the neural Stein critic clearly outperforms KSD in this setting for a larger sample size, with higher power and lower computation time. This result highlights the deeper expressivity of the \( L^2 \) function space compared to kernel methods in addition to the benefit of the linear time complexity of the neural Stein critic GoF test, as opposed to the quadratic complexity of KSD.

C. MNIST Handwritten Digits Data

The results of Section VI-A indicate that the staging of regularization when training neural Stein critics yields greater benefit in higher dimensions. Therefore, we extend to a real-data example in an even higher dimension: the MNIST handwritten digits dataset [36]. We compare the fixed and staged regularization strategies to train critics that discriminate between an RBM and a mixture model of MNIST digits.

1) Authentic and Synthetic MNIST Data: To construct the model distribution \( q \) for the MNIST setting, we follow the approach of [14]. We use a Gaussian-Bernoulli RBM that models the MNIST data distribution, trained using a learned neural Stein critic to minimize the Stein discrepancy between the true MNIST density and the RBM. We declare the model distribution \( q \) to be this 728-dimensional RBM. The data distribution \( p \) is a mixture model composed of 97% the RBM and 3% true digits “1” from the MNIST dataset. Therefore, any disparity in the distributions of \( p \) and \( q \) are caused by this infusion of digits from MNIST into \( p \).

2) Learned Neural Stein Critics: In addition to being a realistic setting, training a neural Stein critic using MNIST digits will allow us to better interpret the discrepancy. Of course, since we do not have access to the “true” score function \( p \), we cannot calculate the \( \text{MSE}_q \) for the trained critics using Equation (16). Furthermore, the computation of \( \text{MSE}_p^{(m)} \) using validation data from \( p \) via (17) becomes less accurate in high dimension, as the method would require a large amount of validation data to be an accurate representation of the population \( \text{MSE}_p \). Therefore, we introduce an additional validation metric to evaluate the fit of the neural Stein critic \( \hat{f}(\cdot, \theta) \). First, in the language of the GoF test introduced in Section V-A, we denote the quantity computed by applying the Stein operator with respect to \( q \) on neural Stein critic \( \hat{f}(\cdot, \theta) \) evaluated at a sample \( x \in \mathcal{X} \) as the “critic witness” of the sample \( x \):

\[
w(x, \theta) = T_q \hat{f}(x, \theta).
\]
Intuitively, since $f(\cdot, \theta)$ is trained on samples from $p$ (the data distribution) by maximizing the Stein discrepancy, the value of $w(x, \theta)$ represents the magnitude of the difference between distributions $p$ and $q$ at $x \in X$. Evaluating the critic witness at $n_{\text{GoF}}$ samples $x_i \sim q$, under the central limit theorem (CLT) assumption, random variables $w(x_i, \theta)$ have a (centered) normal distribution with standard deviation $\sigma(w)/\sqrt{n_{\text{GoF}}}$ when $n_{\text{GoF}}$ is large, where $\sigma(w)$ is the standard deviation of $w(x_i, \theta)$.

Note that the test statistic (43) is the mean of $w(x_i, \theta)$ computed over testing data $x_i \sim p$. As an assessment for the GoF testing power for the neural Stein critic function $f(\cdot, \theta)$ (which is expensive to compute in such a high dimension), we may compare the mean and variance of $w(x, \theta)$ applied to a testing dataset sampled from $q$ and to a “null” dataset drawn from $q$, both of size $n_{\text{GoF}}$:

$$\hat{P} = \frac{\bar{w}_p}{\sigma(w_p) + \sigma(w_q)};$$

$$\bar{w}_p = \frac{1}{n_{\text{GoF}}} \sum_{i=1}^{n_{\text{GoF}}} w(x_i, \theta),$$

$$\sigma(w_p) = \frac{1}{n_{\text{GoF}}} \sqrt{\sum_{i=1}^{n_{\text{GoF}}} (w(x_i, \theta) - \bar{w}_p)^2}. \quad (47)$$

This quantity acts to reflect the capability of the neural Stein critic to differentiate between the distributions in the GoF hypothesis testing setting described in Section V-A. In addition to the $\hat{P}$ metric from Equation (47), we may also apply the Stein discrepancy evaluated at the scaleless neural Stein critic, i.e., the $\text{SD} [\lambda f(\cdot, \theta)]$ (3), to the holdout dataset from the data distribution as an evaluation metric for the models as described in Section V-B.

3) Experimental Setup: We again train 2-hidden-layer MLP’s with Swish activation, where each hidden layer is composed of 512 hidden units, using the default Adam optimizer parameters by PyTorch. The critics observe 2,000 training samples from $p$, training on mini-batches of size 100 with a learning rate $10^{-3}$. Each model is trained for 25 epochs. We consider fixed $\lambda \in \{1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}, 1 \times 10^{0}, 2 \times 10^{0}\}$ and staging scheme $\Lambda(5 \times 10^{-1}, 1 \times 10^{-3}, 0.90)$. The frequency of updates via the staging strategy is $B_w = 20$ batches. We fit 10 critics per regularization strategy. For each critic, we compute the validation SD and the power metric (47) throughout training using $n_{\text{GoF}} = 1,000$ samples from $p$ and the same number of “null” samples from $q$.

In addition to assessing the proxy for the test power of the neural Stein critic in the GoF test via Equation (47), we examine the interpretability of the critic as a diagnostic tool for anomalous observations. By Equation (7), the scaleless optimal critic captures the difference in the score of the data distribution and model distribution. Therefore, a trained neural Stein critic can indicate which samples in a validation dataset represent the largest departure from the distribution $q$. We isolate such samples by identifying samples with high critic witness value (46). We do so by both visualizing the images in a holdout validation dataset sampled from $p$ which have a high $w(\cdot, \theta)$ value, in addition to plotting a heatmap of $w(\cdot, \theta)$ reduced using a t-SNE embedding [39] applied to the entire validation dataset.

4) Results: The power approximation using Equation (47) is plotted in Figure 7(A) for each regularization strategy, where the means and standard deviations are calculated using the ten models for each regularization scheme. While the staging strategy does not exhibit such an advantage as in the high-dimension Gaussian mixture data, staging provides a more rapid increase in the validation metric in the early training period, yielding a final model of comparable performance to the fixed-$\lambda$ strategies. In Figure 7(B), we observe that the test statistic (43) exhibits clear separation from its bootstrapped ($n_{\text{boot}} = 500$) null distribution, even in the case when the number of test samples is relatively small (in this case, $n_{\text{GoF}} = 100$). While Figure 7(B) shows this distribution for the $\Lambda(5 \times 10^{-1}, 1 \times 10^{-3}, 0.90)$ staged regularization strategy, this holds for fixed-$\lambda$ training as well. Finally, the Stein discrepancy evaluated at the scaleless neural Stein critics $\lambda f(\cdot, \theta)$ applied to the holdout dataset from $p$ is visualized throughout training for each model in Figure 7(C). This

![Figure 7](image-url)
result is similar to the result from Figure 7(A) in that the staged approach performs comparably to the best fixed-λ regularization strategies.

For a more direct understanding of how the neural Stein critics perform, consider instances of validation data sampled from ρ in which the critic witness (46) value is very large. These samples indicate the largest deviation between the model distribution q and the data distribution ρ. In the staged \( \Lambda(5 \times 10^{-1}, 1 \times 10^{-3}, 0.90) \) setting, we visualize the critic witness applied to a validation set of 6,000 samples from ρ by reducing the images to a two-dimensional embedding via t-SNE. In Figure 8(A), we observe the embedding of the validation data into this space, where the true MNIST points are highlighted in red. In Figure 8(B), the true MNIST digits are found to have a larger critic witness value than those of RBM samples in the validation dataset.

Furthermore, in the \( \Lambda(5 \times 10^{-1}, 1 \times 10^{-3}, 0.90) \) case, visualizing the images in the validation set from ρ which have the highest critic witness value in Figure 8(C) and those which have the lowest critic witness value in Figure 8(D), we find that this approach correctly identifies true digits one from MNIST as anomalous while accepting those generated by the RBM as normal. In the case of the fixed-λ regularization strategies, it seems that all methods do well to identify the true digits “1” in samples from the data distribution ρ.

VII. DISCUSSION

We have developed a novel training approach for learning neural Stein critics by starting with strong L² regularization and progressively decreasing the regularization weight over the course of training. The advantage of staged regularization is empirically observed in experiments, especially for high-dimension data. In all the experiments, it is observed that critics trained using larger regularization weights at the beginning of training enjoy a more rapid approximation of the target function, including in the task of detecting distribution differences between authentic and synthetic MNIST data. These experimental findings are consistent with the NTK lazy-training phenomenon that happens with large regularization strength at the beginning phase of training, which we theoretically prove. We apply the neural Stein method to GoF tests for which we derive a theoretical guarantee of test power.

The work can be extended in several future directions. Theoretically, within the lazy-training framework, it would be interesting to characterize the expressiveness of the NTK kernel (the assumption of Proposition 4.4). One can also explore a more advanced analysis of the neural network training dynamics, including the SGD training using mini-batches and going beyond kernel learning. A related question is to theoretically analyze the later training stage, for example, to prove the convergence guarantee with decreasing λ, and to find the theoretically optimal annealing strategy of λ in later stages. A possible way is to track the change of the NTK kernel as time evolves, which remains a challenge in studying NTK theory. For the algorithm, one can further investigate the optimal staging scheme and explore other types of regularization of the neural Stein critic than the L²-regularization. For the application to GoF testing, here we consider the typical setting for the Goodness-of-Fit test, which is a simple null hypothesis, i.e., we test whether or not the data follows a particular distribution/model; a potentially interesting direction is extending to cases where the null hypothesis is composite, i.e., comparing the data to a set of K target distributions.

An interesting strategy would be training K neural networks jointly with possible sharing parameters so as to reduce model size and computation. Finally, further applications can be conducted on other modern generative modeling approaches in the same manner as the Gaussian-Bernoulli RBM, including normalizing flow architectures.

VIII. PROOFS

A. PROOFS IN SECTIONS IV-A-IV-C

Proof of Lemma 4.3: Introduce the notation

\[ [K_0(x, x')]_{ij} = K((x, i), (x', j)), \]

where \( K(z, z') \) is a positive semi-definite (PSD) kernel defined on the space of \( Z = \mathcal{X} \times [d] \), that is,

\[ z = (x, i), \quad x \in \mathcal{X}, \quad i \in [d], \]

with \( dz \) being the induced product measure. The kernel \( K \) is PSD due to the definition (24). One can also verify that \( K \) is Hilbert-Schmidt because

\[
\int \int_{Z} K(z, z')^2 dzdz' = \sum_{i,j=1}^{d} \int_{\mathcal{X}} \int_{\mathcal{X}} |K_0(x, x')|_{ij}^2 p(x)p(x')dxdx' \]

\[
= \sum_{i,j=1}^{d} \int_{\mathcal{X}} \left( \int_{\mathcal{X}} (\partial_\theta f_i(x, \theta(0))) \partial_\theta f_j(x', \theta(0)) \right)^2 p(x)p(x')dxdx' \]

\[
\leq \left( \sum_{i=1}^{d} \int_{\mathcal{X}} \left( \|\partial_\theta f_i(x, \theta(0))\|_{E_p}^2 p(x)dx \right) \right)^2 < \infty, \quad (48)
\]

where the last step is by Cauchy-Schwarz and the integrability follows by the \( L^2 \) integrability of \( \partial_\theta f_i(x, \theta(0)) \) in \( (\mathcal{X}, p(x)dx) \) assumed in the condition of the lemma.

As a result, the spectral theorem implies that \( K(z, z') \) has discrete spectrum \( \mu_k, k = 1, 2, \cdots \) which decreases to 0, each \( \mu_k \) is associated with an eigenfunction \( v_k(z) \), and \( \{v_k\}_{k=1}^{\infty} \) form an orthonormal basis on \( Z \). This means that

\[
\sum_{j=1}^{d} \int_{\mathcal{X}} [K_0(x, x')]_{ij} v_k(x', j)p(x')dx' = \mu_k v_k(x, i),
\]

\[ \forall x \in \mathcal{X}, i \in [d]. \]

For any \( k, l = 1, 2, \ldots, \)

\[
\sum_{i=1}^{d} \int_{\mathcal{X}} v_k(x, i)v_l(x, i)p(x)dx = \delta_{kl}.
\]

At last, because the neural network has a finite width, \( \Theta \) is in a finite-dimensional Euclidean space of dimensionality \( M_{\Theta} \). By (24), the kernel \( K \) has a finite rank at most \( M_{\Theta} \). Let the rank of \( K \) be \( M \), then \( \mu_1 \geq \cdots \geq \mu_M > 0 \), and the other eigenvalues are zero. Defining \( v_k(x) \) by \( (v_k(x, i))_{i=1}^{d} \) finishes the proof. \( \Box \)
Proof of Proposition 4.4: As shown in the proof of Lemma 4.3, there are orthonormal basis $\{v_k\}_{k=1}^\infty$ of $\mathcal{X}$ with respect to $\langle \cdot , \cdot \rangle_p$, where the first $M$ many consist of eigen-functions of $K_0(x, x')$. Using the orthonormal basis, $f^*$ has the following expansion with coefficients $c_k \in \mathbb{R}$,

$$
f^* = \sum_{k=1}^\infty c_k v_k, \quad \|f^*\|_p^2 = \sum_{k=1}^\infty c_k^2 \leq \infty.
$$

(49)

The uniqueness of orthogonal decomposition gives that

$$
f_1^* = \sum_{k=1}^m c_k v_k, \quad f_2^* = \sum_{k=m+1}^\infty c_k v_k.
$$

(50)

To prove the proposition, it suffices to prove (27), and then (28) follows by that

$$
\|f_1^*\|_p \leq \|f^*\|_p
$$

which follows from (50), and that $e^{-t\delta \lambda} \leq \epsilon$ when $t \geq \log(1/\epsilon)/(\delta \lambda)$.

To prove (27): By (25), and define

$$
w(x, t) := \lambda \tilde{u}(x, t) - f^*(x),
$$

we have

$$
\partial_t w(x, t) = -\lambda \int_x K_0(x, x') \circ w(x', t) p(x') dx',
$$

(51)

and, by that $\tilde{u}(x, 0) = 0$, we have

$$
w(x, 0) = -f^*(x).
$$

Because $w(x, 0) \in L^2(p)$, the evolution equation (51) implies that

$$
w(x, t) = \sum_{k=1}^\infty b_k(t) v_k(x),
$$

$$
b_k(t) = \begin{cases} -\lambda \mu_k b_k(t), & k \leq M
0, & k > M\end{cases}, \quad b_k(0) = -c_k.
$$

Using the notation $\mu_l = 0$ for $l > M$, we have

$$
b_k(t) = -c_k e^{-t\lambda \mu_k}, \quad k = 1, 2, \ldots
$$

This gives that for any $t > 0$,

$$
-w(x, t) = \sum_{k=1}^m e^{-t\lambda \mu_k} c_k v_k(x)
$$

$$
= \sum_{k=1}^m e^{-t\lambda \mu_k} c_k v_k(x) + \sum_{k=m+1}^\infty e^{-t\lambda \mu_k} c_k v_k(x)
$$

$$
=: (1) + (2).
$$

Because $\mu_k \geq \delta$ for $k = 1, \ldots, m$,

$$
\|1\|_p^2 = \sum_{k=1}^m e^{-2 t \lambda \mu_k} c_k^2 \leq e^{-2 t \lambda \delta} \sum_{k=1}^m c_k^2 = e^{-2 t \lambda \delta} \|f_1^*\|_p^2,
$$

where the last equality is by (50). In addition,

$$
\|2\|_p^2 = \sum_{k=m+1}^\infty e^{-2 t \lambda \mu_k} c_k^2 \leq \sum_{k=m+1}^\infty c_k^2 = \|f_2^*\|_p^2.
$$

Putting together, this gives that

$$
\|w(x, t)\|_p \leq \|1\|_p + \|2\|_p \leq e^{-t \lambda \delta} \|f_1^*\|_p + \|f_2^*\|_p,
$$

which proves (27). \hfill \square

Proof of Proposition 4.5:

Proof of (i): Recall that $\dot{\theta}(t) = -\partial_\theta L_\lambda(\theta(t))$ as in (18), and then by the chain rule,

$$
\frac{d}{dt} L_\lambda(\theta(t)) = \langle \partial_\theta L_\lambda(\theta(t)), \dot{\theta}(t) \rangle_{\Theta} = -\|\partial_\theta L_\lambda(\theta(t))\|_{\Theta}^2 \leq 0,
$$

which means that $L_\lambda(\theta(t))$ is monotonically decreasing over time. Together with (11), we have that $\forall t \geq 0$,

$$
\int_0^t \|\dot{\theta}(s)\|_{\Theta}^2 ds = \int_0^t \|\partial_\theta L_\lambda(\theta(s))\|_{\Theta}^2 ds
$$

$$
= L_\lambda(\theta(0)) - L_\lambda(\theta(t))
$$

$$
= \mathcal{L}_\lambda[u(\cdot, 0)] - \mathcal{L}_\lambda[u(\cdot, t)]
$$

$$
\leq \frac{1}{2\lambda} (\|\lambda u(\cdot, 0) - f^*\|_p^2 - \|\lambda u(\cdot, t) - f^*\|_p^2)
$$

$$
\leq \frac{1}{2\lambda} (\|\lambda u(\cdot, 0) - f^*\|_p^2 - \|\lambda u(\cdot, t) - f^*\|_p^2)
$$

(52)
In the last equality, we have used the condition that \( u(x,0) = 0 \). (52) bounds the change of \( \theta(t) \) by Cauchy-Schwarz, that is,

\[
\|\theta(t) - \theta(0)\| \leq \int_0^t \|\dot{\theta}(s)\|ds \\
\leq \sqrt{t} \left( \int_0^t \|\dot{\theta}(s)\|^2ds \right)^{1/2} \\
\leq \sqrt{t} \left( \frac{1}{2\lambda} \|f(t)\| \right)^{1/2},
\]

which proves (29). Finally, by that \( \|\theta(0)\| < r/2 \), we have \( \|\theta(t)\| < r/2 + \sqrt{t/(2\lambda)} \|f(t)\| \) and then the upper bound of \( t \) in the condition of (i) ensures that \( \theta(t) \in \Theta \). Thus, for all such \( t \), \( u(x,t) = f(x,\theta(t)) \) is well-defined and \( f(x,\theta(t)) \in L^2(p) \) by Assumption 4, so that the involved integrals in the loss \( L_\lambda(\theta(t)) \) are all finite.

Proof of (ii): We first note that

\[
\theta(t) \in B_r, \quad \forall 0 \leq t \leq \frac{1}{2} \left( \frac{r}{\|f\|_p} \right)^2 \lambda. \tag{53}
\]

This is because \( r/2 + \sqrt{t/(2\lambda)} \|f\|_p \leq r \) and \( B_r \subset \Theta \), thus the condition needed for \( t \) by (i) is satisfied, and then (29) gives that \( \|\theta(t) - \theta(0)\| \leq r/2 \) for the range of \( t \) being considered. The claim (53) then follows together with the assumption that \( \|\theta(0)\| < r/2 \).

In the below, we may omit the dependence on \( t \) in notation when there is no confusion, e.g., we write \( \lambda u(x',t) - f(x') \) as \( (\Lambda u - f^*) (x') \). To prove (30), recall the evolution equations of \( u \) and \( \bar{u} \) as in (23) and (25),

\[
\begin{align*}
\partial_t u(x,t) &= -E_{x' \sim p}K_i(x,x') \circ (\Lambda u - f^*)(x') \\
\partial_t \bar{u}(x,t) &= -E_{x' \sim p}K_0(x,x') \circ (\Lambda \bar{u} - f^*)(x') \\
u(x,0) &= \bar{u}(x,0) = 0
\end{align*}
\]

which gives that \( \|u - \bar{u}\|_{t=0} = 0 \) and

\[
\partial_t (u - \bar{u}) = -E_{x' \sim p}(K_i - K_0)(x,x') \circ (\Lambda u - f^*)(x') \tag{54}
\]

and

\[
-\lambda E_{x' \sim p}K_0(x,x') \circ (u - \bar{u})(x'). \tag{55}
\]

By taking \( \langle \cdot, \cdot \rangle_p \) inner-product with \( u - \bar{u} \) on both sides, and that \( K_0 \) has the expression (24) and thus is PSD, we have that

\[
\begin{align*}
\frac{d}{dt} \|u - \bar{u}\|_p^2 \\
&\leq \langle u - \bar{u}, -E_{x' \sim p}(K_i - K_0)(x,x') \circ (\Lambda u - f^*)(x') \rangle_p \\
&\leq \|u - \bar{u}\|_p \|K_i - K_0\|_p \|\Lambda u - f^*\|_p.
\end{align*}
\]

where \( \|K_i - K_0\|_p \) stands for the operator norm of the kernel integral operator in \( L^2(p) \). We claim that

\[
\begin{align*}
\text{(Claim 1)} \quad \|\Lambda u(t,\cdot) - f^*\|_p^2 &\leq \|f^*\|_p^2, \quad \forall t \geq 0, \tag{57} \\
\text{(Claim 2)} \quad \|K_i - K_0\|_p &\leq L_1 L_2 \sqrt{\frac{2T}{\lambda}} \|f^*\|_p, \quad \forall t \leq \frac{1}{2} \left( \frac{r}{\|f\|_p} \right)^2 \lambda. \tag{58}
\end{align*}
\]

If both claims hold, then (56) continues as

\[
\begin{align*}
\frac{d}{dt} \|u - \bar{u}\|_p^2 \\
&\leq L_1 L_2 \sqrt{\frac{2T}{\lambda}} \|f^*\|_p \|u - \bar{u}\|_p \\
&\leq \frac{1}{2} \left( \frac{r}{\|f\|_p} \right)^2 \lambda.
\end{align*}
\]

We define

\[
\Delta(t) := \|\Lambda u(t,\cdot) - \lambda \bar{u}(t,\cdot)\|_p \geq 0, \quad \Delta(0) = 0,
\]

and \( \Delta(t) \) is continuous. For fixed \( t \leq \frac{1}{2} \left( \frac{r}{\|f\|_p} \right)^2 \lambda \), suppose

\[
\sup_{s \in [0,t]} \Delta(s) = \Delta(t^*), \quad \text{for some } t^* \leq t,
\]

then by (59),

\[
\frac{1}{2} \Delta(t^*)^2 \leq \lambda^2 \int_0^{t^*} \frac{d}{dt} \|u - \bar{u}(s,\cdot)\|_p^2 ds \\
\leq \lambda \int_0^{t^*} L_1 L_2 \sqrt{\frac{2s}{\lambda}} \|f^*\|_p^2 \Delta(s) ds \\
\leq L_1 L_2 \sqrt{2\lambda} \|f^*\|_p^2 \Delta(t^*) \int_0^{t^*} \sqrt{s} ds
\]

this gives that

\[
\Delta(t^*) \leq \frac{4\sqrt{\lambda}}{3} \int_0^{t^*} L_1 L_2 \sqrt{\lambda} (f^*)^2 \Delta(t^*)^{3/2},
\]

which proves (ii) by that \( \Delta(t) \leq \Delta(t^*) \) and \( t^* \leq t \).

To finish the proof, it remains to show (57) and (58).

Proof of (57): Define \( w(x,t) := \lambda u(x,t) - f^*(x) \), By (54),

\[
\partial_t w(x,t) = -\lambda E_{x' \sim p}K_i(x,x') \circ w(x'),
\]

and thus

\[
\frac{d}{dt} \|w(t,\cdot)\|_p^2 = -\lambda \langle w(x), E_{x' \sim p}K_i(x,x') \circ w(x') \rangle_p.
\]

Denoting the \( i \)-th entry of \( w_{\cdot} \) as \( w_i \), by (22), one can verify that

\[
\begin{align*}
\langle w, E_{x' \sim p}K_i(\cdot,x') \circ w(x') \rangle_p \\
&= E_{x' \sim p}E_{x' \sim p} \sum_{i,j=1}^d \langle \partial_h f_i(x,\theta(t)), \partial_j f_j(x',\theta(t)) \rangle \delta w_i(x) w_j(x') \\
&= \sum_{i,j=1}^d E_{x' \sim p} \partial_h f_i(x,\theta(t)) w_i(x) \\
&\quad \sum_{j=1}^d E_{x' \sim p} \partial_j f_j(x',\theta(t)) w_j(x') \bigg|_{\Theta} \\
&= \|E_{x' \sim p} \partial_h f(x,\theta(t)) \cdot w(x) \|_2^2, \tag{59}
\end{align*}
\]

This shows that \( \frac{d}{dt} \|w(t,\cdot)\|_p^2 \leq 0 \), that is, \( \|w(t,\cdot)\|_p^2 \) monotonically decreases over time. Thus for any \( t \geq 0 \),

\[
\|\Lambda u(t,\cdot) - f^*\|_p^2 = \|w(t,\cdot)\|_p^2 \leq \|w(0)\|_p^2 = \|f^*\|_p^2,
\]

due to that \( u(x,0) = 0 \).

Proof of (58): It suffices to show that for any \( v \in L^2(p) \),

\[
\langle v, E_{x' \sim p}K_i(\cdot,x') \circ v(x') \rangle_p \\
\leq L_1 L_2 \sqrt{2\lambda} \|f^*\|_p \|v\|_p.
\]

Denote the \( i \)-th entry of \( v \) as \( v_i \), and, similarly as in (60) and by (22) and (24), one has

\[
\begin{align*}
\langle v, E_{x' \sim p}K_i(\cdot,x') \circ v(x') \rangle_p \\
&= \|E_{x' \sim p} \partial_h f(x,\theta(t)) \cdot v(x) \|_{\Theta}^2, \\
\langle v, E_{x' \sim p}K_0(\cdot,x') \circ v(x') \rangle_p \\
&= \|E_{x' \sim p} \partial_h f(x,\theta(0)) \cdot v(x) \|_{\Theta}^2.
\end{align*}
\]

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Subtracting the two gives that

\[
\langle v, E_{x' \sim p}(K_0 - K_0)(\cdot, x') \rangle_p = \|E_{x \sim p} \partial_\theta f(x, \theta(t)) \cdot v(x)\|_\Theta^2 - \|E_{x \sim p} \partial_\theta f(x, \theta(0)) \cdot v(x)\|_\Theta^2
\]

\[
= \left( \|E_{x \sim p} \partial_\theta f(x, \theta(t)) \cdot v(x)\|_\Theta + \|E_{x \sim p} \partial_\theta f(x, \theta(0)) \cdot v(x)\|_\Theta \right)
\]

\[
\left( \|E_{x \sim p} \partial_\theta f(x, \theta(t)) \cdot v(x)\|_\Theta - \|E_{x \sim p} \partial_\theta f(x, \theta(0)) \cdot v(x)\|_\Theta \right).
\]

(62)

Because \( \theta(t) \in B_r \), cf. (53),

\[
\|E_{x \sim p} \partial_\theta f(x, \theta(t)) \cdot v(x)\| \leq E_{x \sim p} \|\partial_\theta f(x, \theta(t))\|_\Theta \|v(x)\|
\]

\[
\leq L_1 E_{x \sim p} \|v\|_p
\]

where the second inequality is by (C1). We have the same upper bound of \( \|E_{x \sim p} \partial_\theta f(x, \theta(0)) \cdot v(x)\|_\Theta \) because \( \theta(0) \in B_r \). Meanwhile, triangle inequality gives that

\[
\|E_{x \sim p} \partial_\theta f(x, \theta(t)) \cdot v(x)\| \leq E_{x \sim p} \|\partial_\theta f(x, \theta(t)) - \partial_\theta f(x, \theta(0))\|_\Theta \|v(x)\|
\]

\[
\leq L_2 \|\theta(t) - \theta(0)\| E_{x \sim p} \|v(x)\|
\]

\[
\leq L_2 \|\theta(t) - \theta(0)\|_\Theta \|v\|_p.
\]

(63)

The second-to-last line is by (C2) and both \( \theta(t) \) and \( \theta(0) \) are in \( B_r \). By (i) which is proved, the r.h.s. of (63) is upper-bounded by \( L_2 \sqrt{1/(2\lambda)} \|f^*\|_p \|v\|_p \). Putting back to (62), we have

\[
\text{(r.h.s. of (62)) } \leq 2L_1L_2 \sqrt{1/(2\lambda)} \|f^*\|_p \|v\|_p^2
\]

which proves (61).

**Proof of Theorem 4.6:** The condition on the largeness of \( \lambda \) guarantees that the range of \( t_0 \) in (31) is non-empty, and (31) ensures that \( \lambda \frac{\log(1/\epsilon)}{\epsilon} \leq t \leq \frac{2}{\lambda^2} (\frac{1}{\pi} - 2) \). For this range of \( t \), Proposition 4.5 applies to give (30). Meanwhile, Proposition 4.4 requires that \( \partial_\theta f(\cdot, \theta(0)) \) are in \( L^2(\mu) \), which follows from the uniform boundedness of \( \partial_\theta f(\cdot, \theta(0)) \) on \( X \) by Assumption 4(C1) due to that \( \theta(0) \in B_r \). The condition

\[
t \geq \frac{\log(1/\epsilon)}{\epsilon} \frac{1}{\lambda}
\]

allows Proposition 4.4 to apply to bound \( \|\lambda u(x, t) - f\|_p \) as in (28) for the range of \( t \) being considered. Putting together (30) and (28), by triangle inequality, we have

\[
\|\lambda u(x, t) - f^*\|_p \leq c\|f^*\|_p + \frac{4\sqrt{2}}{3} L_1 L_2 \sqrt{\lambda} t^{3/2} \|f^*\|_p^2
\]

and by definition \( \sqrt{\lambda} t^{3/2} = \lambda^{-1} (t_0 \log(1/\epsilon))/\delta^{3/2} \), which proves (32). \( \square \)

**B. Proofs in Section IV-D**

**Proof of Lemma 4.8:** Similarly as in the proof of Proposition 4.5(i), we have

\[
\frac{d}{dt} \hat{L}_\lambda(\hat{\theta}(t)) = \langle \partial_\theta \hat{L}_\lambda(\hat{\theta}(t)), \dot{\hat{\theta}}(t) \rangle_{\Theta^2} = -\|\partial_\theta \hat{L}_\lambda(\hat{\theta}(t))\|_{\Theta^2}^2 \leq 0,
\]

and for any \( t \geq 0 \),

\[
\int_0^t \|\dot{\hat{\theta}}(s)\|_{\Theta^2}^2 ds = \int_0^t \|\partial_\theta \hat{L}_\lambda(\hat{\theta}(s))\|_{\Theta^2}^2 ds
\]

\[
= \hat{L}_\lambda(\theta(0)) - \hat{L}_\lambda(\theta(t)).
\]

(64)

By the definition (33), because \( f(x, \theta(0)) = \hat{u}(x, 0) = 0 \), \( \hat{L}_\lambda(\theta(0)) = 0 \); Meanwhile, let \( f(x) := f(x, \theta(t)) \),

\[
\hat{L}_\lambda(\hat{\theta}(t)) = E_{x \sim p} \left( \frac{\lambda}{2} \|f(x)\|^2 - s_q(x) \cdot \hat{f}(x) \right)
\]

\[
- E_{x \sim p} \nabla_x \cdot \hat{f}(x).
\]

If the last term is with \( E_{x \sim p} \), we have that (by that \( \hat{f}(x) \) is in \( L^2(\mu) \cap F_0(\mu) \) by Assumption 3)

\[
- E_{x \sim p} \nabla_x \cdot \hat{f}(x) = E_{x \sim p} \hat{f}(x) \cdot s_p(x)
\]

\[
= E_{x \sim p} \hat{f} \cdot s_p - (E_{x \sim p} - E_{x \sim p}) \hat{f} \cdot s_p,
\]

and thus

\[
- E_{x \sim p} \nabla_x \cdot \hat{f}(x) = - E_{x \sim p} \nabla_x \cdot \hat{f}(x)
\]

\[
- (E_{x \sim p} - E_{x \sim p}) \nabla_x \cdot \hat{f}(x)
\]

\[
E_{x \sim p} \hat{f} \cdot s_p - (E_{x \sim p} - E_{x \sim p}) \hat{f} \cdot s_p
\]

\[
- (E_{x \sim p} - E_{x \sim p}) \nabla_x \cdot \hat{f}(x)
\]

\[
E_{x \sim p} \hat{f} \cdot s_p - \{ 1 \} - \{ 2 \},
\]

(66)

where

\[
\{ 1 \} := (E_{x \sim p} - E_{x \sim p}) f(x, \theta(t)) \cdot s_p(x),
\]

\[
\{ 2 \} := (E_{x \sim p} - E_{x \sim p}) \nabla_x \cdot f(x, \theta(t)).
\]

We then have, recalling that \( f^* = s_q - s_p \),

\[
- \hat{L}_\lambda(\hat{\theta}(t)) = - E_{x \sim p} \left( \frac{\lambda}{2} \|f^*\|^2 - s_q \cdot \hat{f} + s_p \cdot \hat{f} \right) + \{ 1 \} + \{ 2 \}
\]

\[
= - E_{x \sim p} \left( \frac{\lambda \|f^*\|^2}{\lambda} - \frac{\|f^*\|^2}{\lambda} \right) + \{ 1 \} + \{ 2 \}
\]

\[
\leq \frac{\|f^*\|^2}{\lambda} \{ 1 \} + \{ 1 \} + \{ 2 \}
\]

\[
= \left( E_{x \sim p} - E_{x \sim p} \right) \|f^*(x)\|^2.
\]

(67)

The difference \( (E_{x \sim p} - E_{x \sim p}) \|f^*(x)\|^2 \) is the deviation of an independent sum sample average from its expectation. By Assumption 5(C6), there is an integer \( n_0 \) possibly depending on constants \( b_2, b_4 \) s.t. \( n > n_0 \), under a good event \( E_1 \) which happens w.p. \( \geq 1 - n^{-10} \),

\[
\left( E_{x \sim p} - E_{x \sim p} \right) \|f^*(x)\|^2 \leq \sqrt{20} b_2 \sqrt{\frac{\log n}{n}}.
\]

As a result, \( \{ 3 \} \leq \frac{\sqrt{20} b_2}{\lambda} \sqrt{\frac{\log n}{n}} \). As for \( \{ 1 \} \) and \( \{ 2 \} \), we know that if \( \hat{\theta}(t) \in B_r \), then under the good event in (C5), called \( E_2 \),

\[
\max\{ \{ 1 \}, \{ 2 \} \} \leq C_F \frac{\log n}{n^\gamma} + b_F \sqrt{\frac{\log n}{n}}
\]

(68)
However, we have not shown $\hat{\theta}(t) \in B_r$ yet. We now let $n_\lambda$ such that when $n > n_\lambda$,

$$n > n_6, \quad \sqrt{20b_2} \sqrt{\log n \over n} + 4\lambda \left(C_f \over n^\gamma + b_2 \sqrt{\log n \over n}\right) < \|f^*\|^p. \quad (69)$$

Since $\|f^*\|^p > 0$ (by Assumption 2) is a fixed constant, for any $\lambda > 0$,

the l.h.s. is $\hat{O}(n^{-1/2}) + \lambda\hat{O}(n^{-\gamma})$ and thus will be less than the r.h.s. when $n$ is large enough. We also let $t_{\text{max}}$ be defined s.t. $\|f^*\|^p \sqrt{t_{\text{max}}/\lambda} = r/2$. Consider being under the intersection of good events $E_1$ and $E_2$ which happens w.p. $\geq 1 - 3n^{-10}$, we claim that, for any $t \leq t_{\text{max}}$, $\hat{\theta}(t) \in B_r$.

We prove this by contradiction. If, then by the continuity of $\hat{\theta}(t)$ over time, there must be a $0 < t_1 \leq t_{\text{max}}$ s.t. $\|\hat{\theta}(t_1)\| = r$. By that $\|\theta(0)\| < r/2$, we have

$$\|\hat{\theta}(t_1) - \theta(0)\| > r/2. \quad (70)$$

Meanwhile, by (64),(67),

$$\int_0^{t_1} \|\hat{\theta}(s)\|^2_d ds = -\hat{L}_\lambda(\hat{\theta}(t_1)) \leq \frac{\|f^*\|^2}{2\lambda} + 1 + 2 + 3$$

where $\{1, 2, 3\}$ are taking value at $t = t_1$. Under $E_1$, we already have $3 \leq \sqrt{20b_2} \sqrt{\log n \over n}$; now $\hat{\theta}(t_1) \in B_r$, applying (C5) we know that under $E_2$ the bound (68) holds. By the definition of $n_\lambda$ in (69), we have that

$$\{1, 2, 3\} < \frac{\|f^*\|^2}{2\lambda},$$

and as a result,

$$\int_0^{t_1} \|\hat{\theta}(s)\|^2_d ds < \frac{\|f^*\|^2}{\lambda}. \quad (71)$$

Then, by Cauchy-Schwarz,

$$\|\hat{\theta}(t_1) - \theta(0)\| \leq \int_0^{t_1} \|\hat{\theta}(s)\|_d ds$$

$$\leq \sqrt{t_1} \left(\int_0^{t_1} \|\hat{\theta}(s)\|^2_d ds\right)^{1/2}$$

$$\leq \sqrt{t_1} \lambda \|f^*\|^p$$

$$\leq \sqrt{t_{\text{max}}} \|f^*\|^p = r/2,$$

which means that $\|\hat{\theta}(t_1) - \theta(0)\| < r/2$ and this contradicts with (70).

Now we have shown that for $\hat{\theta}(t) \in B_r$ for any $t \leq t_{\text{max}}$. Applying the same argument to bound $\int_0^{t_{\text{max}}} \|\hat{\theta}(s)\|^2_d ds$ for any such $t$, and in particular the upper bound of $1 + 2 + 3 < \|f^*\|^2 \over 2\chi^2$ holds, one can verify that

$$\|\hat{\theta}(t) - \theta(0)\| \leq \sqrt{t \over \lambda} \|f^*\|^p$$

and this finishes the proof of the lemma. □

Proof of Proposition 4.9: Recall the evolution equation of $\hat{\mu}(x, t)$ in (25), which can be written as

$$\partial_t \hat{\mu}(x, t) = -\mathbb{E}_{x' \sim p} K_0(x, x') \circ (\hat{\lambda}(x', t) - s_q(x')).$$

Because $\|\partial_\theta f(x, \theta)\|$ is uniformly bounded by $L_1$ by Assumption 4(C1), each column of the $d$-by-$M_\theta$ matrix $\partial_\theta f(x, \theta(0))$, which can be viewed as a vector field on $X$, is uniformly bounded and thus in $L^2(p) \cap F_0(p)$, and then

$$\mathbb{E}_{x' \sim p} \partial_\theta f(x', \theta(0)) \cdot s_p(x') = -\mathbb{E}_{x' \sim p} \nabla_{x'} \cdot \partial_\theta f(x', \theta(0)).$$

As a result, we have

$$\mathbb{E}_{x' \sim p} K_0(x, x') \circ s_p(x') = -\mathbb{E}_{x' \sim p} \nabla_{x'} \cdot K_0(x, x').$$

By comparing to (38), we have

$$\partial_\lambda (\hat{\mu} - \mu)(x, t) = -\mathbb{E}_{x' \sim p} K_1(x, x') \circ (\hat{\lambda}(x', t) - s_q(x'))$$

$$+ \mathbb{E}_{x' \sim p} K_0(x, x') \circ (\hat{\lambda}(x', t) - s_q(x'))$$

$$+ \mathbb{E}_{x' \sim p} \nabla_{x'} \cdot \hat{K}_1(x, x')$$

$$- \mathbb{E}_{x' \sim p} \nabla_{x'} \cdot K_0(x, x'),$$

$$= \{1 + 2 + 3 + 4 + 5\}. \quad (71)$$

We will analyze each of the five terms respectively toward deriving a bound of $\|\lambda(\hat{\mu} - \mu)(1, t)|_d, t\|$ as has been done in the proof of Proposition 4.5(ii). We restrict to when $t \leq (\log n + 1)^2 \lambda$ by default, and considering being under the good event in Lemma 4.8 assuming that $n > n_\lambda$ (by 69), this ensures that $n$ is greater than the large $n$ threshold integer $n_6$ required by (C6), where we have $\hat{\theta}(t) \in B_r$ and

$$\|\hat{\theta}(t) - \theta(0)\| \leq \sqrt{t \over \lambda} \|f^*\|_p. \quad (72)$$

In the below, we may omit the dependence on $t$ in notation when there is no confusion, e.g., we write $\hat{\lambda}(x', t) - s_q(x')$ as $(\hat{\lambda} - s_q)(x')$.

Bound of $\|\lambda\|_d$: By definition, $\|\lambda\|_d = \|\partial_\theta f(x, \theta(0))\|, (\mathbb{E}_{x' \sim p} - \mathbb{E}_{x' \sim p}) \nabla_{x'} \cdot \partial_\theta f(x', \theta(0)) \|_d$, and thus

$$\|\lambda\|_d \leq \|\partial_\theta f(x, \theta(0))\| \|\mathbb{E}_{x' \sim p} - \mathbb{E}_{x' \sim p}\| \nabla_{x'} \cdot \partial_\theta f(x', \theta(0)) \|_d.$$ 

By (C1), $\|\partial_\theta f(x, \theta(0))\| \leq L_1$; To bound $\|\mathbb{E}_{x' \sim p} - \mathbb{E}_{x' \sim p}\| \nabla_{x'} \cdot \partial_\theta f(x', \theta(0)) \|_d$, we utilize the matrix Bernstein inequality (see, e.g., [40, Th. 6.1.1], reproduced as Lemma 1.2): The matrix $\nabla_{x'} \cdot \partial_\theta f(x', \theta(0))$ is of size $1$-by-$M_\theta$, and the operator norm is bounded by $L_3$ by Assumption 5(C3). By Lemma 1.2, let $n_7 > 2$ be an integer s.t. $n > n_7$ implies

$$(10 \log n + \log(2M_\theta))/n < (3/2)^2, \quad (73)$$

then when $n > n_7$, there is a good event $E_3$ that happens w.p. $\geq 1 - n^{-10}$ under which

$$\|\mathbb{E}_{x' \sim p} - \mathbb{E}_{x' \sim p}\| \nabla_{x'} \cdot \partial_\theta f(x', \theta(0)) \|_d \leq 4\sqrt{\Pi L_3 \log n + \log M_\theta \over n}.$$
As a result, when \( n > n_5 \) and under \( E_5 \),
\[
\|5\| \leq 4\sqrt{11L_1L_3}\sqrt{\log n + \log M_\Theta}\text{.} \tag{74}
\]

**Bound of \( \|4\| \):** By definition,
\[
\nabla x' \cdot K_1(x, x') - \nabla x' \cdot K_0(x, x')
= (\partial_\phi f(x, \hat{\theta}(t)), \nabla x'\cdot \partial_\phi f(x', \hat{\theta}(t)))_\phi
- (\partial_\phi f(x, \theta(0)), \nabla x'\cdot \partial_\phi f(x', \theta(0)))_\phi
= (\partial_\phi f(x, \hat{\theta}(t)) - \partial_\phi f(x, \theta(0)), \nabla x'\cdot \partial_\phi f(x', \hat{\theta}(t)) + \partial_\phi f(x, \theta(0)), \nabla x'\cdot \partial_\phi f(x', \theta(0)))_\phi
- \nabla x' \cdot \partial_\phi f(x', \theta(0)))_\phi.
\]

By that \( \hat{\theta}(0), \hat{\theta}(t) \in B_1 \) and \((C1)(C2)(C3)(C4)\),
\[
\|\nabla x' \cdot K_1(x, x') - \nabla x' \cdot K_0(x, x')\|
\leq (L_2L_3 + L_1L_4)\|\hat{\theta}(t) - \theta(0)\|.
\]

Together with \((72)\), we have that
\[
\|4\| \leq (L_2L_3 + L_1L_4)\sqrt{T/\lambda}\|f^*\|_p. \tag{75}
\]

**Bound of \( \|3\| \):** By definition, \( 3 = -\partial_\phi f(x, \theta(0)) - (E_{x' \sim p} - E_{x' \sim p})\partial_\phi f(x', \theta(0))\) and \( 3 = \nabla x' \cdot \partial_\phi f(x', \theta(0)) \)
\[
\|3\| \leq \|\partial_\phi f(x, \theta(0))\|
\leq \|\partial_\phi f(x, \theta(0))\|
\leq (\nabla x' \cdot \partial_\phi f(x', \theta(0)))_\phi
\leq L_1\|\phi_3 - \beta_3\|.
\]

where
\[
\phi_3 := (E_{x' \sim p} - E_{x' \sim p})\partial_\phi f(x', \theta(0)) T\lambda(\hat{\mu}(x', t) - s_q(x')).
\]

For \( \phi_3 \), we need to bound \(||(E_{x' \sim p} - E_{x' \sim p})\partial_\phi f(x', \theta(0))\|\) and \(\|\partial_\phi f(x', \theta(0))\|\) is a 1-by-\( M_\Theta \) vector, and the norm is bounded by \( L_1b_q \) uniformly on \( X' \) by \((C6)\). According to \((C6)\) and consider when \( n > n_5 \) which implies \((10\log n + \log(1 + M_\Theta))/n < \(3/2)^2, \)
there is a good event \( E_b \) that happens w.p. \( \geq 1 - n^{-10} \) under which
\[
\|\partial_\phi f(x', \theta(0))\|
< 2\sqrt{11L_1b_q}\sqrt{\log n + \log M_\Theta} \tag{76}
\]
\[
< 2\sqrt{11L_1b_q}\sqrt{\log n + \log M_\Theta}.
\]

This shows that when \( n > n_5 \) and under \( E_b \), \( \|\phi_3\| \) is bounded by the r.h.s. of \((76)\).

For \( \phi_3 \), we will bound the deviation to be \( O(n^{-1/2}) \) uniformly for all \( t \). The observation is that though \( \hat{\mu}(x, t) \) changes over \( t \), the concentration argument of the sample average only involves the kernel \( K_0(x, x') \). Specifically, by definition,
\[
\lambda(\hat{\mu}(x, t) - \lambda) = -\lambda \int_0^t \mathbb{E}_{y \sim p}[K_0(x, y)](\lambda(y, s) - f^*(y)) ds,
\]
and then
\[
\phi_3 = -\lambda \int_0^t \mathbb{E}_{y \sim p}[K_0(x, y)](\lambda(y, s) - f^*(y)) ds.
\]

Note that \( \mathbb{E}_{y \sim p}[\lambda(\hat{\mu}(y, s) - f^*(y))] \leq L_1E_{y \sim p}[\lambda(y, s) - f^*(y)] \leq L_1E_{y \sim p}[\lambda(y, s) - f^*(y)] \leq \|\lambda(y, s) - f^*(y)\|_{p}. \]

The latter can be bounded by \( f^* \) using the same argument as in \((C1)\) \((57)\) proved in the proof of \((4.5)\), that is, for any \( t \geq 0 \),
\[
\|\lambda(x, t) - f^*(y)\|_p \leq \|\lambda(x, t) - f^*(y)\|_p = \|f^*\|_p. \tag{78}
\]

Putting together, this implies that
\[
\|\phi_3\| \leq L_1\int_0^t \mathbb{E}_{y \sim p}[\lambda(y, s) - f^*(y)] ds \tag{79}
\]
\[
\leq L_1\|f^*\|_p.
\]

Next, we bound \( \|\mathbb{E}_{x' \sim p}[\lambda(y, s) - f^*(y)] \) by matrix Bernstein inequality \((1.2)\): The matrix \( \lambda(y, s) - f^*(y) \) is of size \( M_\Theta \times M_\Theta \), and its operator norm is bounded by \( L_3^2 \) by \((C1)\). When \( n > n_5 \) as defined by \((73)\), \((1.2)\) gives that w.p. \( \geq 1 - n^{-10} \),
\[
\|\mathbb{E}_{x' \sim p}[\lambda(y, s) - f^*(y)] \| \leq L_3^2\sqrt{\log n + \log M_\Theta} \tag{80}
\]

Putting together, we have that when \( n > n_5 \) and under \( E_a \cap E_b \),
\[
\|\phi_3\| \leq 2\sqrt{11L_1b_q}\|f^*\|_p \sqrt{\log n + \log M_\Theta} \tag{80}
\]

**Bound of \( \|2\| \):** By definition,
\[
\mathbb{K}_q(x, x') - \mathbb{K}_0(x, x') = \partial_\phi f(x, \hat{\theta}(t))\partial_\phi f(x', \hat{\theta}(t)) T - \partial_\phi f(x, \theta(0))\partial_\phi f(x', \theta(0)) T.
\]

Since \( \theta(0), \hat{\theta}(t) \in B_1 \), by \((C1)(C2)\), \forall x, x' \in X',
\[
\mathbb{K}_q(x, x') = \partial_\phi f(x, \hat{\theta}(t))\partial_\phi f(x', \hat{\theta}(t)) T - \partial_\phi f(x, \theta(0))\partial_\phi f(x', \theta(0)) T.
\]

where we used \((72)\) in the last inequality. This gives that
\[
\|\phi_3\| \leq 2L_1L_2\sqrt{T/\lambda}\|f^*\|_p, \tag{81}
\]

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
We will bound \( \mathbb{E}_{x', p} \| \lambda \mathbf{u}(x', t) - s_q(x') \| \) to be \( O(1) \). First, note that we have
\[
\| \lambda \mathbf{u}(\cdot, t) - s_q \|_p \leq \| \lambda \mathbf{u}(\cdot, t) - (s_q - s_p) \|_p + \| s_p \|_p \\
\leq \| f^* \|_p + \| s_p \|_p \quad \text{(by (78))} \\
\leq 2 \| f^* \|_p + b_q,
\]
where the last inequality used that \( \mathbb{E}_{x', p} \| \lambda \mathbf{u}(x, s) - f^*(x) \| \leq \| \lambda \mathbf{u}(\cdot, s) - f^* \|_p \leq \| f^* \|_p \) by (78); Meanwhile,
\[
\mathbb{E}_{x', p} \| \lambda \mathbf{u}(x', s) - f^*(x') \| ds = -\lambda \int_0^t \mathbb{E}_{x', p} \left[ \left( (E_{x'} - E_{x'})s_q(x) \right)^T \mathbf{K}_0(x, \mathbf{0}) \right] \\
\partial_\theta f(x', \theta(0)) (\lambda \mathbf{u}(x', s) - f^*(x')) ds,
\]
and, similarly,
\[
\mathbb{E}_{x', p} \| \lambda \mathbf{u}(x', s) - f^*(x') \| ds \leq (\lambda t) L_1 \| f^* \|_p.
\]
Finally,
\[
\mathbb{E}_{x', p} \| \lambda \mathbf{u}(x', s) - f^*(x') \| ds \leq (\lambda t)L_2 \| f^* \|_p b_q.
\]
To bound \( |\mathbb{E}_{x', p} | \| \lambda \mathbf{u}(x, s) - s_q(x) \| \) to be \( O(1) \) uniformly for all \( t \) similarly as in the analysis of \( \mathbb{E}_{x', p} \) above, where we leverage the expression (77) of \( \lambda \mathbf{u}(x, t) \) via \( \mathbf{K}_0(x, x') \). Based on the expression, we have
\[
\| \lambda \mathbf{u}(x, t) \|^2 = \lambda^2 \int_0^t \int_0^t \mathbb{E}_{x', p} \mathbb{E}_{x', p} (\lambda \mathbf{u}(x', s) = f^*(x'))^T [\mathbf{K}_0(x, x')^T \mathbf{K}_0(x, x')] [\lambda \mathbf{u}(x', s) - f^*(x')] ds_1 ds_2.
\]
Meanwhile, \( \| \lambda \mathbf{u}(x, t) - s_q(x) \|^2 = \| \lambda \mathbf{u}(x, t) - f^* \| + 2 \| s_q \|^2 \), and then
\[
\| \lambda \mathbf{u}(x, t) - s_q(x) \|^2 = \mathbb{E}_{x', p} \| \lambda \mathbf{u}(x', s) - s_q(x) \|^2 = \mathbb{E}_{x', p} \mathbb{E}_{x', p} \| \lambda \mathbf{u}(x', s) - f^*(x') \|^2
\]
where
\[
\mathbb{E}_{x', p} \mathbb{E}_{x', p} (\lambda \mathbf{u}(x', s) - f^*(x'))^T [\mathbf{K}_0(x, x')^T \mathbf{K}_0(x, x')] [\lambda \mathbf{u}(x', s) - f^*(x')] ds_1 ds_2
\]
and, \( \mathbb{E}_{x', p} \mathbb{E}_{x', p} (\lambda \mathbf{u}(x', s) - f^*(x'))^T [\mathbf{K}_0(x, x')^T \mathbf{K}_0(x, x')] [\lambda \mathbf{u}(x', s) - f^*(x')] ds_1 ds_2
\]
and (C1) implies that \( \| \partial_\theta f(x', \theta(0)) \|, \| \partial_\theta f(x', \theta(0)) \| \leq L_1, \)
\[
\| \partial_\theta f(x', \theta(0)) \| \leq \lambda L_1 \| f^* \|_p \| \lambda \mathbf{u}(x', s) - f^*(x') \| ds \leq \lambda L_1 \| f^* \|_p \| \lambda \mathbf{u}(x', s) - f^*(x') \| ds \leq \lambda L_1 \| f^* \|_p \| \lambda \mathbf{u}(x', s) - f^*(x') \| ds.
\]
(83)
Next, we use (95) to bound \( \|\lambda(\hat{u} - \bar{u})(\cdot, t)\|_p \). Again from (94), we have
\[
\frac{d}{dt} \frac{1}{2} \|\lambda(\hat{u} - \bar{u})(\cdot, t)\|_p^2 = \lambda^2 E_{x \sim \rho} \lambda(\hat{u} - \bar{u})(x, t) \partial_t (\hat{u} - \bar{u})(x, t)
\]
\[
\leq \lambda^2 E_{x \sim \rho} \mathbf{K}_t(x, x')(\hat{u} - \bar{u})(x', t) \leq 0.
\]
Using the same argument as in the proof of Proposition 4.5(ii) based on (59), we have that for any \( t \) being considered,
\[
\|\lambda(\hat{u} - \bar{u})(\cdot, t)\|_p \leq 2\lambda \int_0^t g(s) ds
\]
\[
\leq 2\lambda \int_0^t g(t) ds \quad (g(t) \text{ is monotone in } t)
\]
\[
= 2\lambda t g(t). \quad (95)
\]
Next, we use (95) to bound \( \|\lambda(\hat{u} - \bar{u})(\cdot, t)\|_p \). Again from (94), we have
\[
\frac{d}{dt} \frac{1}{2} \|\lambda(\hat{u} - \bar{u})(\cdot, t)\|_p^2
\]
\[
= \lambda^2 E_{x \sim \rho} \mathbf{K}_t(x, x')(\hat{u} - \bar{u})(x', t)
\]
where in the first inequality we used the upper bound \( \|\mathbf{K}_t(x, x')\| = \|\partial_t \mathbf{f}(x, \hat{\theta}(t)) \partial_t \mathbf{f}(x, \hat{\theta}(t))\| \leq L_t^1 \) by (C1). Using the argument by integrating over \( t \) again, we have
\[
\|\lambda(\hat{u} - \bar{u})(\cdot, t)\|_p \leq 2\lambda \int_0^t (2L_t^2 \lambda s + 1) g(s) ds
\]
\[
\leq 2\lambda t (2L_t^2 \lambda t + 1) g(t).
\]
((\(2L_t^2 \lambda t + 1) g(t) \) is monotone in \( t \)). Inserting the definition of \( g(t) \) in (92), we have
\[
\|\lambda(\hat{u} - \bar{u})(\cdot, t)\|_p \leq C_{2,1}(1 + 2L_t^2 \lambda t)
\]
\[
\frac{\lambda^{1/2} t^{3/2} \|\mathbf{f}^*\|_p + (\lambda t + 1) \sqrt{\log n + \log M_\Theta}}{n},
\]
which proves (40) by defining \( C_2 := 2C_{2,1} \max\{1, 2L_t^2\} \). By the definition of \( C_{2,1} \) in (93) and \( C_{2,2} \) in (84),
\[
C_2 \leq c(L_1, \ldots, L_4) (1 + b_2 + \|\mathbf{f}^*\|_p),
\]
where \( c(L_1, \ldots, L_4) \) an \( O(1) \) constant determined by \( L_1, L_2, L_3, L_4 \).

Proof of Theorem 4.10: Similarly as in the proof of Theorem 4.6, the required largeness of \( \lambda \) guarantees that the range of \( t_0 \) in (41) is non-empty. Proposition 4.4 applies same as before, and \( t_0 \geq 1 \) ensures that the bound of \( \|\lambda(\hat{u}(\cdot, t) - \mathbf{f}^*\|_p \) as in (28) holds.

The upper bound of \( t_0 \) in (41) ensures that \( t \) falls in the range by Proposition 4.9: We consider large \( n \) as required by Proposition 4.9, and, by Remark 4.4, the three thresholds \( n_\lambda, n_\delta, n_{\lambda, \delta} \) are determined by \( \lambda, \|\mathbf{f}^*\|_p, \log M_\Theta \) and \( \lambda t = t_0 \log(1/\delta) \). We also consider being under the intersection of the good events in Proposition 4.9, which happens w.p. \( \geq 1 - 7n^{-10} \), then (40) holds.

Putting together (28) and (40), by triangle inequality, we have
\[
\|\lambda(\hat{u}(x, t) - \mathbf{f}^*)\|_p \leq 2\epsilon \|\mathbf{f}^*\|_p
\]
\[
+ C_2 (1 + \lambda t) \left( \lambda^{1/2} t^{3/2} \|\mathbf{f}^*\|_p + (1 + \lambda t) \lambda t \sqrt{\frac{\log n + \log M_\Theta}{n}} \right).
\]
Then (42) follows by the equality \( \lambda t = t_0 \log(1/\delta) \) and defining
\[
\kappa_1 := (1 + t_0 \log(1/\delta)) \left( \frac{t_0 \log(1/\delta)}{\delta} \right)^{3/2},
\]
\[
\kappa_2 := (1 + t_0 \log(1/\delta))^2 t_0 \log(1/\delta). \quad \square
\]

C. Proof in Section V.A

Proof of Corollary 5.1: Applying Theorem 4.10 gives that, under \( E_{\text{tr}} \), \( \|\lambda \hat{f} - \mathbf{f}^*\|_p \) is upper bounded by the r.h.s. of (42) where setting \( n = n_{\text{tr}} \). The condition of the corollary further implies that
\[
\|\lambda \hat{f} - \mathbf{f}^*\|_p < 0.9 \|\mathbf{f}^*\|_p.
\]
Then,
\[
\langle \lambda \hat{f}, \mathbf{f}^* \rangle - \|\mathbf{f}^*\|_p = \langle \lambda \hat{f} - \mathbf{f}^*, \mathbf{f}^* \rangle
\]
Recall that \( \mathcal{F}_0(p) \) and \( \mathcal{F}_0(p) \cap L^2(p) \) (Assumption 3), (8) gives that
\[
\mathbb{E}_{x \sim p} \hat{f}(x) = (\mathbf{f}^*, \hat{f})_p > 0 \frac{1}{\lambda} \|\mathbf{f}^*\|_p^2.
\]

Next, we bound the deviation of \( \hat{T} \) and \( \hat{T}_{null} \) by concentration. By definition,
\[
\hat{T} = \mathbb{E}_{x \sim p} T_q \hat{f}(x), \quad \hat{p} := \frac{1}{n_{GoF}} \sum_{i=1}^{n_{GoF}} \delta_{x_i}.
\]
Recall that \( \hat{\theta}(t) \) is in \( B_r \) by Lemma 4.8, then by Assumption 5(C5)(C6),
\[
|T_q \hat{f}(x_i)| \leq |s_q \cdot \hat{f}(x_i)| + |\nabla \cdot \hat{f}(x_i)| \leq b_0 b_q + b(1) =: b.
\]
By the Hoeffding’s inequality, Lemma 1.1, for any \( t > 0 \), denote \( n_{GoF} \) by \( n \),
\[
\Pr[\mathbb{E}_{x \sim p} T_q \hat{f}(x) - \mathbb{E}_{x \sim p} T_q \hat{f}(x) \leq -tl|E_{tr}] \leq \exp\left\{-\frac{n t^2}{2 b^2}\right\}.
\]
By union bound, and the condition that \( 7n_{tr}^{-10} < 0.1 \min\{\alpha, \beta\} \),
\[
\Pr[\mathbb{E}_{x \sim p} T_q \hat{f}(x) - \mathbb{E}_{x \sim p} T_q \hat{f}(x) \leq -t] \leq \exp\left\{-\frac{n t^2}{2 b^2}\right\} + 7n_{tr}^{-10} < \exp\left\{-\frac{n t^2}{2 b^2}\right\} + 0.1 \beta.
\]
This gives that w.p. \( \geq 1 - \beta \),
\[
\hat{T} \geq \mathbb{E}_{x \sim p} T_q \hat{f}(x) - t_1, \quad t_1 := \sqrt{2b} \sqrt{\frac{n}{\log \frac{1}{0.92}}}
\]
Similarly,
\[
\hat{T}_{null} = \mathbb{E}_{y \sim q} T_q \hat{f}(y), \quad \hat{q} := \frac{1}{n_{GoF}} \sum_{i=1}^{n_{GoF}} \delta_{y_i}.
\]
and w.p. \( \geq 1 - \alpha \),
\[
\hat{T}_{null} \leq t_2, \quad t_2 := \sqrt{2b} \sqrt{\frac{n}{\log \frac{1}{0.92}}}
\]
We set \( t_2 \) to be the test threshold, combined with (96), we know that the target significance level and power can be achieved if \( t_1 + t_2 \leq \frac{0.1}{\lambda} \|\mathbf{f}^*\|_p^2 \). Together with that \( \sqrt{\log \frac{1}{0.92}} \leq 0.5 + \sqrt{\log \frac{1}{\alpha}} \) and similarly for \( \sqrt{\log \frac{1}{0.92}} \), this proves the corollary.

**APPENDIX A**

**ADDITIONAL TECHNICAL LEMMAS AND PROOFS**

**A. Proofs of Technical Lemmas**

**Proof of Lemma 2.1:** For any \( f \in \mathcal{F}_0(p) \), because \( pf|_{\partial X} = 0 \), we have
\[
\int_X p \nabla \cdot f = -\int_X \nabla p \cdot f
\]
Then for any \( f \in L^2(p) \cap \mathcal{F}_0(p) \),
\[
\mathbb{E}_{x \sim p} T_q f(x) = \int_X p(s_q \cdot f + \nabla \cdot f) = \sum_{X} p(s_q \cdot f - \int_X \nabla p \cdot f
\]
\[
= \int_X p(s_q - s_p) \cdot f = \mathbb{E}_{x \sim p} f \cdot \hat{f} = (\mathbf{f}^*, \hat{f})_p.
\]
Note that the integrals are well-defined because both \( s_p \) and \( s_q \) are in \( L^2(p) \). By (A.1), whenever \( \|f\|_p \leq r \),
\[
\mathbb{E}_{x \sim p} T_q f(x) \leq \|f^*\|_p \|\hat{f}\|_p \leq r \|f^*\|_p.
\]
where \( \|f^*\|_p < \infty \) because \( s_p \) and \( s_q \) are in \( L^2(p) \). This proves that SD \( (p, q) \) is upper bounded by \( r \|f^*\|_p \) and thus is finite.
To show that SD \( (p, q) \) is \( r \|f^*\|_p \), first when \( f^* = 0 \) in \( L^2(p) \) then SD \( (p, q) = 0 \) by (A.2). If \( \|f^*\|_p > 0 \), by property of the inner product, the \( f \) that is a positive multiple of \( f^* \) and has \( \|f\|_p = r \) achieves the supremum of value \( r \|f^*\|_p \). Because \( f^* \in \mathcal{F}_0(p) \), this maximizer \( f = (r/\|f^*\|_p) f^* \) is also in \( \mathcal{F}_0(p) \).

**Proof of Lemma 4.1:** By the definition of \( L_\lambda \) in (13) and (11), the latter applies to \( f = f(\cdot, \theta) \) because \( f(\cdot, \theta) \) is in \( L^2(p) \cap \mathcal{F}_0(p) \) by Assumption 3.
\[
L_\lambda(\theta) = -(\mathbf{f}^*, \mathbf{f}(\cdot, \theta))_p + \frac{\lambda}{2} \|\mathbf{f}(\cdot, \theta)\|_p^2
\]
This proves (19) according to the GD dynamic defined in (18).

**Proof of Lemma 4.2:** Inserting (19) to (21) gives that
\[
\frac{\partial}{\partial \theta} \mathbf{u}(x, t) = -\partial_{\mathbf{u}} \mathbf{f}(x, \theta(t)),
\]
\[
\mathbb{E}_{x \sim p} \partial_{\mathbf{u}} \mathbf{f}(x', \theta(t)) \cdot (\lambda \mathbf{u}(x', t) - f^*(x'))_\theta,
\]
where we used the definition (20). The lemma then follows by the definition (22) and the linearity of inner-product \( \langle \cdot, \cdot \rangle \).

**Proof of Lemma 4.7:** By (33),
\[
\partial_{\mathbf{u}} \hat{L}_\lambda(\theta) = E_{x \sim p} \partial_{\mathbf{u}} \mathbf{f}(x, \theta(t)) \cdot (\lambda \mathbf{u}(x, t) - s_q(x)) - \nabla x \cdot \partial_{\mathbf{u}} \mathbf{f}(x, \theta(t)),
\]
and combined with (34)(35), we have
\[
\partial_{\mathbf{u}} \hat{z}(x, t) = -\partial_{\mathbf{u}} \mathbf{f}(x, \theta(t)),
\]
\[
=-\left(\partial_{\mathbf{u}} \mathbf{f}(x, \theta(t)) \cdot (\lambda \mathbf{u}(x', t) - s_q(x')) - \nabla x \cdot \partial_{\mathbf{u}} \mathbf{f}(x', \theta(t))\right)_\theta.
\]

This proves the lemma by the definition of $\hat{K}_d(x, x')$ in (36).

**B. Concentration Lemmas**

The following lemma can be derived from Classical Hoeffding’s inequality.

**Lemma 1.1 (Hoeffding):** Suppose $X_1, \cdots, X_n$ are independent random variables, $|X_i| \leq L$. Then for any $t \geq 0$,

$$\Pr[\frac{1}{n}\sum_{i=1}^{n}(X_i - EX_i) \geq t], \Pr[\frac{1}{n}\sum_{i=1}^{n}(X_i - EX_i) \leq -t],$$

$$\leq \exp\{-\frac{nt^2}{2L^2}\}.$$  

The following lemma can be directly derived from [40, Th. 6.1.1].

**Lemma 1.2 (Matrix Bernstein [40]):** Let $X_i$ be a sequence of $n$ independent, random, real-valued matrices of size $d_1$-by-$d_2$. Assume that $EX_i = 0$ and $\|X_i\| \leq L$ for each $i$, and $\nu > 0$ be such that

$$\|\frac{1}{n}\sum_{i=1}^{n}EX_iX_i^T\|, \|\frac{1}{n}\sum_{i=1}^{n}EX_i^TX_i\| \leq \nu.$$  

Then, for any $t \geq 0$,

$$\Pr[\|\frac{1}{n}\sum_{i=1}^{n}X_i\| \geq t] \leq (d_1 + d_2) \exp\{-\frac{nt^2}{2(\nu + Lt/3)}\}.$$  

**APPENDIX B**

**ADDITIONAL EXPERIMENTS**

**A. Supplementary Tables**

**B. 1D Gaussian Mixture**

Consider a setting in which $q$ is a 1-dimensional Gaussian mixture of two equally-weighted, unit variance components at positions -1 and 1. Let $p$ be a similar mixture, however at positions -0.8 and 1, and the component at 1 has variance 0.25. Therefore, the score of $q$ has the form:

$$s_q(x) = \nabla q(x)/q(x) = \left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x+1)^2}{2}\right)\right)$$

$$+ \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-1)^2}{2}\right)^{-1}$$

$$\times \frac{-x+1}{\sqrt{2\pi}}\exp\left(-\frac{(x+1)^2}{2}\right)$$

$$- \frac{-x-1}{\sqrt{2\pi}}\exp\left(-\frac{(x-1)^2}{2}\right)^{-1}$$

$$+ \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-1)^2}{2}\right)^{-1}$$

$$\times \frac{-x+1}{\sqrt{2\pi}}\exp\left(-\frac{(x+1)^2}{2}\right)$$

$$\times \frac{-x+1}{\sqrt{2\pi}}\exp\left(-\frac{(x+1)^2}{2}\right)$$

which can be computed in a similar analytical manner in higher dimension given distributions of the form (45).

A neural Stein critic (MLP of 2 hidden layers with 512-node regularization in this setting) is trained using staged $\Lambda(1, 10^{-3}, 0.9)$ regularization in this setting with 1,000 training samples from $q$ with batch size 200, learning rate $10^{-3}$, and $B_w = 5$.

**C. Kernel Stein Discrepancy Goodness-of-Fit Test**

As in Section V-A, consider the setting in which we are provided with probability distribution $q$ supported on $\mathcal{X} \subset \mathbb{R}^d$ from which we can sample, and consider the scenario wherein we are provided with a finite sample of data $x_i \sim p$ of $n_{\text{GOF}}$ samples. As an alternative to learning the Stein critic function using a neural network for estimation of the test statistic in (43), one may define the function space $\mathcal{F}$ to be an RKHS defined by a kernel denoted $k(\cdot, \cdot)$. In this case, the KSD admits computation in closed form by the following relation:

$$\text{KSD} = E_{x,x' \sim p}[u_q(x, x')],$$  

where

$$u_q(x, x') = s_q(x)^T k(x, x') s_q(x') + s_q(x)^T \nabla x' k(x, x')$$

| Regularization ($\lambda$) | Mean Power | Epoch | Mean Selected $\bar{MSE}_q$ | Mean Stopping Epoch |
|---------------------------|------------|------|----------------------------|---------------------|
| 1e - 3                    | 0.814 ± 0.018 | 60   | 0.041                      | 59.5                |
| 1e - 2                    | 0.808 ± 0.027 | 50   | 0.027                      | 49.4                |
| 1e - 1                    | 0.825 ± 0.040 | 30   | 0.021                      | 27.8                |
| 1e - 0                    | 0.835 ± 0.025 | 20   | 0.025                      | 21.8                |
| $\Lambda(1e0, 5e - 2, 0.90)$ | 0.839 ± 0.044 | 30   | 0.023                      | 31.5                |
| $\Lambda(1e0, 5e - 2, 0.95)$ | 0.838 ± 0.029 | 30   | 0.020                      | 34.1                |
Given the test statistic (A.8), a “wild bootstrap” is used to approximate the distribution of the test statistic under the null hypothesis, which is outlined in [13]. Given the test set of \(n_{\text{GoF}}\) samples \(x_i \sim p\), the null hypothesis may be rejected if (A.8) computed using these \(x_i\) exceeds the \((1 - \alpha)\) quantile of the wild bootstrapped distribution under the null hypothesis (where, again, \(\alpha\) is selected to tune the Type-I error). The power is then estimated using \(n_{\text{run}}\) number of such GoF tests.

Following the approaches of [11] and [13], we use an RBF kernel to define the RKHS used for KSD in our experiments. This kernel has the following form:

\[
k(x, x') = \exp \left( -\frac{\|x - x'\|^2}{2\sigma^2} \right)
\]

where \(\sigma\) is the kernel bandwidth and \(\gamma = 1/(2\sigma^2)\). A standard heuristic we employ is the choice of the median Euclidean distance between the data considered as the bandwidth \(\sigma\) of this kernel. However, we also conduct an experiment to select a bandwidth which has the capacity to achieve higher testing power when used in the KSD GoF test. We consider bandwidths which are scaled versions of the median data distances heuristic bandwidth. That is, we compute the testing power for bandwidth factors \(\delta \in \{2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}\}\), where the scaling follows \(\gamma' = 1/(2\sigma^2)\) for \(\sigma\) equal to the median Euclidean distance between data samples. To do so, we consider the same model and data distributions as in Section VI-B.1. We then fix \(n_{\text{sample}} = 1500\), and compute the KSD test power from \(n_{\text{run}} = 400\) GoF tests using \(n_{\text{boot}} = 500\) bootstrapped samples in each test. This is conducted using \(n_{\text{replica}} = 5\) replicas to generate a mean and standard deviation of power computed for each choice of bandwidth. Finally, we compare this to the GoF testing power computed from \(n_{\text{replica}} = 5\) neural Stein critics trained using \(n_{\text{sample}} = 1500\) samples split into a 50%/50% train/test split (where 20% of the train split is

**Table II**

| Regularization (\(\lambda\)) | Mean Power | Epoch | Mean Selected MSE | Mean Stopping Epoch |
|-------------------------------|------------|-------|-------------------|---------------------|
| 2.500e - 4                    | 0.839 ±0.034 | 50   | 0.120             | 52.4                |
| 1.0000e - 3                   | 0.894 ±0.027 | 50   | 0.105             | 50.1                |
| 4.0000e - 3                   | 0.910 ±0.014 | 45   | 0.114             | 46.0                |
| 1.600e - 2                    | 0.893 ±0.025 | 30   | 0.127             | 34.1                |
| 6.400e - 2                    | 0.905 ±0.013 | 20   | 0.128             | 21.5                |
| 2.560e - 1                    | 0.878 ±0.019 | 12   | 0.132             | 12.8                |
| 1.024e - 0                    | 0.824 ±0.037 | 12   | 0.172             | 12.4                |
| \(\Lambda(5e - 1, 1e - 3, 0.80)\) | 0.928 ±0.007 | 50   | 0.085             | 49.8                |
| \(\Lambda(5e - 1, 1e - 3, 0.85)\) | 0.892 ±0.021 | 45   | 0.095             | 40.1                |

**Table III**

| Regularization (\(\lambda\)) | Mean Power | Epoch | Mean Selected MSE | Mean Stopping Epoch |
|-------------------------------|------------|-------|-------------------|---------------------|
| 2.500e - 4                    | 0.862 ±0.103 | 45   | 0.260             | 37.9                |
| 1.0000e - 3                   | 0.962 ±0.021 | 45   | 0.262             | 42.1                |
| 4.0000e - 3                   | 0.925 ±0.040 | 25   | 0.276             | 23.9                |
| 1.600e - 2                    | 0.444 ±0.111 | 10   | 0.301             | 9.3                 |
| \(\Lambda(4e - 1, 5e - 4, 0.80)\) | 0.243 ±0.050 | 4    | 0.208             | 46.2                |
| \(\Lambda(4e - 1, 5e - 4, 0.85)\) | 1.000 ±0.001 | 50   | 0.190             | 49.1                |
| \(\Lambda(4e - 1, 5e - 4, 0.90)\) | 0.995 ±0.002 | 55   | 0.202             | 54.4                |

**Fig. 9.** Comparison between the testing power of the KSD GoF test in the 10D setting described in Section VI-B.1 for a range of bandwidth choices with \(n_{\text{sample}} = 1500\). The horizontal axis indicates the scaling factor \(\delta\) of the kernel bandwidth (see Appendix B-C) and the vertical axis displays the test power. The red line and error bars display the mean and standard deviation of the testing power for each choice of kernel bandwidth, and the black line and shaded region indicate the mean and standard deviation of the testing power of the neural Stein GoF test in this setting.

\[
\mathbb{V}_{q} = \frac{1}{n_{\text{GoF}}} \sum_{i,j=1}^{n_{\text{GoF}}} u_{q}(x_{i}, x_{j}). \tag{A.8}
\]

Using this formulation, [13] defines a quadratic-time \(V\)-statistic which estimates the KSD as follows:

\[
\hat{m}_{\text{KL}}(x, x') = \nabla_{x}k(x, x')^{T}S_{p}(x') + \text{tr}((\nabla_{x}k(x, x'))). \tag{A.7}
\]
In Section VI-B.1, we outline a procedure for comparing the neural Stein test to the KSD GoF test by considering varying overall sample size \( n_{\text{sample}} \), training the neural network using half of these samples and conducting tests using the other half. This section analyzes the power and running time of the neural Stein GoF hypothesis testing procedure as outlined in Section V-A for a variety of training/testing splits.

Using the notation of Section VI-A.1 for the model and data distributions, let the model distribution \( q \) be an isotropic, two-component Gaussian mixture in \( \mathbb{R}^{50} \) with means \( \mu_1 = 0_d \) and \( \mu_2 = 0.5 \times 1_d \), and let the data distribution \( p \) have the form of (45) with covariance shift \( \rho_1 = 0.99 \) and scale \( \omega = 0.05 \). In this section, we fix \( n_{\text{sample}} = 500 \) and consider other choices of training/testing split, comparing to the KSD with “best” selected bandwidth (see Appendix B-C) for \( n_{\text{sample}} = 500 \). That is, we partition the 500 points into splits whereby the number of training samples takes on a proportion in the range of \{10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%\} of \( n_{\text{sample}} \). The testing procedure is otherwise identical to that of Section VI-B.1. The result can be found in Figure 11, which shows that the power of the neural Stein critic GoF test dominates that of both KSD test methods initially. However, the neural Stein test achieves near unit power by the time the training partition dedicated to validation. The power is then computed in each case using \( n_{\text{train}} = 400 \) GoF tests, \( n_{\text{boot}} = 500 \) bootstrapped samples in each individual test, efficient bootstrap ratio \( r_{\text{pool}} = 50 \), and \( n_{\text{replica}} = 5 \). The result can be found in Figure 11, which shows that the power of the neural Stein critic GoF test dominates that of both KSD test methods initially. However, the neural Stein test achieves near unit power by the time \( n_{\text{sample}} = 500 \), and the KSD with selected bandwidth achieves unit power by the time \( n_{\text{sample}} = 1,000 \). This increase in power compared to Figure 6(A) is due to the fact that the perturbation between the model and data distribution in this scenario is larger in this setting than in the setting outlined in Section VI-B.1.
E. Adaptive Staging

In Section III-B, we specify a method for staging the weight of regularization $\lambda$ via an exponential decay over the course of optimization. Here, we develop an adaptive procedure which stages $\lambda$ according to the monitor $\overline{\text{MSE}}_{p}^{(m)}$ (17), and compare with the non-adaptive staging in (15).

The adaptive staging method decides to stage the weight of regularization down to a smaller value when the monitored $\overline{\text{MSE}}_{p}^{(m)}$ is observed to increase. Specifically, we examine two adaptive staging approaches:

- The strategy, denoted as $\Lambda_a(\lambda_{\text{init}}, \lambda_{\text{term}}, \beta)$, begins with $\lambda = \lambda_{\text{init}}$. Over the course of training, $\overline{\text{MSE}}_{p}^{(m)}$ is computed after every $B_w = 10$ mini-batches of updates given $n_{\text{val}}$ samples from $p$ (in these examples, the validation samples are re-drawn from $p$ for each computation). If $\overline{\text{MSE}}_{p}^{(m)}$ is at any observation greater than the previous observation, the weight of regularization $\lambda$ is multiplied by the factor $\beta < 1$. This process is repeated until $\lambda = \lambda_{\text{term}}$, at which point $\lambda$ is fixed for the remainder of training. We require that $\overline{\text{MSE}}_{p}^{(m)}$ decrease at least once with each stage of $\lambda$, so that $\lambda$ may not stage down until first an improvement followed by a decrease in performance in MSE is observed.

- The staging, denoted by $\Lambda_a$, uses $\overline{\text{MSE}}_{p}$ instead of $\overline{\text{MSE}}_{p}^{(m)}$ as the monitor function to determine the staging. For simulated high dimensional Gaussian mixture data, the computation of $\overline{\text{MSE}}_{p}$ uses the prior information of the scaleless optimal critic $\hat{\mathbf{f}}^*$.

To assess the performance of the adaptive staging strategy, we conduct an experiment using the Gaussian mixture data of Section VI-A.1 in 25D with $\omega = 0.8$ and $\rho_1 = 0.5$. The neural Stein critic architecture and initialization are as in Section VI-A.2. We again use the Adam optimizer with default parameters, with learning rate $10^{-\beta}$ and batch size 200 samples. In this setting, the networks are trained using 4,000 samples from $p$ for 60 epochs each. For each staging strategy, 10 critics are trained. The $\overline{\text{MSE}}_{p}^{(m)}$ and $\overline{\text{MSE}}_{p}$ are computed each using $n_{\text{val}} = 20,000$ samples from $p$ and the $\overline{\text{MSE}}_{q}$ is computed using $n_{\text{val}} = 20,000$ samples from $q$. We compare the staging to fixed regularization strategies with $\lambda \in \{1 \times 10^{-3}, 4 \times 10^{-3}, 1.6 \times 10^{-2}\}$. The staging strategies used are $\Lambda_p(3 \times 10^{-1}, 5 \times 10^{-4}, 0.85)$, $\Lambda_p(3 \times 10^{-1}, 5 \times 10^{-4}, 0.70)$, and $\Lambda_a(3 \times 10^{-1}, 5 \times 10^{-4}, 0.70)$, where an update is made according to the nonadaptive staging strategy every $B_w = 10$ batches.

The $\overline{\text{MSE}}_{p}$ and $\overline{\text{MSE}}_{q}$ in the adaptive staging comparison experiments are shown in Figure 12(A) and (B), respectively. As in Section VI-A.3, all staging strategies outperform the fixed-$\lambda$ strategies in 25D. Furthermore, the performances of the adaptive strategies $\Lambda_a$ and $\Lambda_a$ and the non-adaptive strategy $\lambda$ are similar, with overall minimal MSE comparable across all three staging strategies. In addition, a comparison of the trajectory of $\lambda$ over the course of training via the three staged regularization strategies is displayed in Figure 12(C), indicating that the adaptive staging strategies exhibit a roughly exponential annealing of $\lambda$ over the course of training. This result justifies the rationale of the heuristic, nonadaptive staging approach proposed in Section III-B, which is computationally less expensive than the adaptive staging which relies on the repetitive computation of the monitor validation MSE.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for the helpful feedback.

REFERENCES

[1] S. Nowozin, B. Cseke, and R. Tomioka, “f-GAN: Training generative neural samplers using variational divergence minimization,” in Proc. Adv. Neural Inf. Process. Syst., vol. 29, 2016, pp. 1–9.

[2] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in Proc. Int. Conf. Mach. Learn. (PMLR), 2017, pp. 214–223.

[3] A. Hyvärinen and P. Dayan, “Estimation of non-normalized statistical models by score matching,” J. Mach. Learn. Res., vol. 6, no. 24, pp. 695–709, 2005.

[4] Y. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 1–27.
[5] I. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 1–16.

[6] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, “Your classifier is secretly an energy based model and you should treat it like one,” in Proc. Int. Conf. Learn. Represent., 2019, pp. 1–23.

[7] B. K. Sripatthanakul, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet, “On integral probability metrics, $\phi$-divergences and binary classification,” 2009, arXiv:0901.2698.

[8] J. Gorham and L. Mackey, “Measuring sample quality with Stein’s method,” in Proc. Adv. Neural Inf. Process. Syst., vol. 28, 2015, pp. 1–18.

[9] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” J. Mach. Learn. Res., vol. 13, pp. 723–773, Mar. 2012.

[10] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland, “Learning deep kernels for non-parametric two-sample tests,” in Proc. Int. Conf. Mach. Learn. (PMLR), 2020, pp. 6316–6326.

[11] Q. Liu, J. Lee, and M. Jordan, “A kernelized Stein discrepancy for goodness-of-fit tests,” in Proc. Int. Conf. Mach. Learn. (PMLR), 2016, pp. 7–15.

[12] L. Dinh, D. Krueger, and Y. Bengio, “Nice: Non-linear independent maximum Stein discrepancy estimators,” in Proc. Adv. Neural Inf. Process. Syst., Jun. 2021, pp. 1–12.

[13] C. Stein, “A bound for the error in the normal approximation to the distribution of a sum of dependent random variables,” in Proc. 6th Berkeley Symp. Math. Statist. Probab., 1972, pp. 583–603.

[14] A. Anastasiou et al., “Stein’s method meets computational statistics: A review of some recent developments,” Stat. Sci., vol. 38, no. 1, pp. 120–139, Feb. 2023.

[15] A. Anastasiou and I. M.琳, “Generalization and equilibrium in generative adversarial nets” in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 6658–6670.

[16] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland, “Learning deep kernels for non-parametric two-sample tests,” in Proc. Int. Conf. Learn. Represent., 2019, pp. 1–10.

[17] X. Cheng and Y. Xie, “Neural tangent kernel maximum mean discrepancy,” in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 17–26.

[18] A. Anastasiou et al., “Stein’s method meets computational statistics: A review of some recent developments,” Stat. Sci., vol. 38, no. 1, pp. 120–139, Feb. 2023.

[19] C. Stein, P. Diaconis, S. Holmes, and G. Reinert, “Use of exchangeable pairs in the analysis of simulations,” in Stein’s Method: Expository Lectures and Applications (Lecture Notes-Monograph Series). Beachwood, OH, USA: Institute of Mathematical Statistics, May 2004, pp. 1–25.

[20] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[21] M. J. Wainwright, High-Dimensional Statistics: A Non-Asymptotic Viewpoint, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[22] Y. LeCun, C. Cortes, and C. Burges, “Mnist handwritten digit database,” AT & T Labs, Tech. Rep., 2010.

[23] P. Ramachandran, B. Zoph, and Q. V. Le, “Searching for activation functions,” in Proc. Int. Conf. Learn. Represent., 2018, pp. 1–13.

[24] M. F. Hutchinson, “A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines,” Comput. Statist. Simul. Comput., vol. 19, no. 2, pp. 1059–1076, 1989.

[25] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, no. 11, pp. 2579–2605, 2008.

[26] J. A. Tropp, “An introduction to matrix concentration inequalities,” Found. Trends Mach. Learn., vol. 8, nos. 1–2, pp. 230–245, 2015.

Matthew Repasky received the B.S. degree in physics (computational data analysis) from the Georgia Institute of Technology in 2021, where he is currently pursuing the Ph.D. degree in machine learning with the H. Milton Stewart School of Industrial and Systems Engineering. He is interested in deep learning, spatiotemporal modeling, and reinforcement learning.

Xiujuan Cheng received the Ph.D. degree from Princeton University in 2013. She joined the Department of Mathematics, Duke University, in 2017, where she is currently an Associate Professor of mathematics. She works on theoretical and computational techniques to solve problems in high-dimensional data analysis, signal processing, and machine learning.

Yao Xie (Member, IEEE) received the Ph.D. degree in electrical engineering (mathematics) from Stanford University. She was a Research Scientist with Duke University. She is currently an Associate Professor and the Harold R. and Mary Anne Nash Early Career Professor with the H. Milton Stewart School of Industrial and Systems Engineering. She is interested in theory, and optimization, in providing the theoretical and computational techniques to solve problems in high-dimensional data analysis, signal processing, and machine learning.