# Learning Globally Smooth Functions on Manifolds

**Juan Cerviño** [1] **Luiz F.O. Chamon** [2] **Benjamin D. Haeffele** [3] **Rene Vidal** [3 4] **Alejandro Ribeiro** [1]

## Abstract

Smoothness and low dimensional structures play central roles in improving generalization and stability in learning and statistics. This work combines techniques from semi-infinite constrained learning and manifold regularization to learn representations that are globally smooth on a manifold. To do so, it shows that under typical conditions the problem of learning a Lipschitz continuous function on a manifold is equivalent to a dynamically weighted manifold regularization problem. This observation leads to a practical algorithm based on a weighted Laplacian penalty whose weights are adapted using stochastic gradient techniques. It is shown that under mild conditions, this method estimates the Lipschitz constant of the solution, learning a globally smooth solution as a byproduct. Experiments on real world data illustrate the advantages of the proposed method relative to existing alternatives. Our code is available here.

## 1. Introduction

Learning smooth functions has been shown to be advantageous in general and is of particular interest in physical systems. This is because of the general observation that close input features tend to be associated with close outputs and of the particular fact that in physical systems, Lipschitz continuity of input-output maps translates to stability and safety (Oberman & Calder, 2018; Finlay et al., 2018b; Couellan, 2021; Finlay et al., 2018a; Pauli et al., 2021; Krishnan et al., 2020; Shi et al., 2019; Lindemann et al., 2021; Arghal et al., 2021).

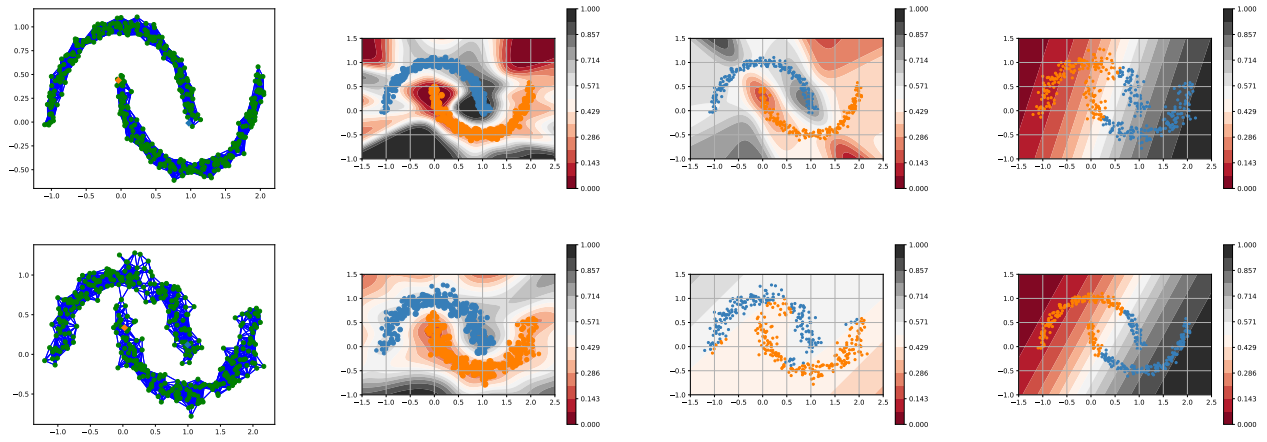
[1]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, USA [2]Excellence Cluster for Simulation Technology, University of Stuttgart, Germany [3]Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, USA [4]Innovation in Data Engineering and Science (IDEAS), University of Pennsylvania, Philadelphia, USA. Correspondence to: Juan Cerviño <jcervino@seas.upenn.edu>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. 

To learn smooth functions one can require the parameterization to be smooth. Such is the idea, e.g., of spectral normalization of weights in neural networks (Miyato et al., 2018; Zhao & Liu, 2020). Smooth parameterizations have the advantage of being globally smooth, but they may be restrictive because they impose smoothness for inputs that are not necessarily realized in the data. This drawback motivates the use of Lipschitz *penalties* in risk minimization (Oberman & Calder, 2018; Finlay et al., 2018b; Couellan, 2021; Pauli et al., 2021; Bungert et al., 2021), which offers the opposite tradeoff. Since penalties encourage but do not enforce small Lipschitz constants, we may learn functions that are smooth on average, but with no global guarantees of smoothness at every point in the support of the data (Bubeck & Sellke, 2021; Bubeck et al., 2021). Formulations that guarantee *global* smoothness can be obtained if the risk minimization problem is modified by the addition of a Lipschitz constant *constraint* (Krishnan et al., 2020; Shi et al., 2019; Lindemann et al., 2021; Arghal et al., 2021). This yields formulations that guarantee Lipschitz smoothness in all possible inputs without the drawback of enforcing smoothness outside of the input data distribution. Several empirical studies (Krishnan et al., 2020; Shi et al., 2019; Lindemann et al., 2021; Arghal et al., 2021) demonstrated the advantage of imposing global smoothness constraints on observed inputs.

In this paper we exploit the fact that data can be often modeled as points in a low-dimensional manifold. We therefore consider manifold Lipschitz constants in which function smoothness is assessed with respect to distances measured over the data manifold (Definition 1). Although this looks like a minor difference, controlling Lipschitz constants over data manifolds is quite different from controlling Lipschitz constants in the ambient space. In Figure 1, we look at a classification problem with classes arranged in two separate half moons. Constraining Lipschitz constants in the ambient space effectively assumes the underlying data is uniformly distributed in space [cf. Figure 1(d)]. Constraining Lipschitz constants in the data manifold, however, properly accounts for the data distribution [cf. Figure 1(a)].

This example also illustrates how constraining manifold Lipschitz constants is related to manifold regularization (Belkin et al., 2005; Niyogi, 2013; Li et al., 2022). The difference is that manifold regularization penalizes the average norm of the manifold gradient. This distinction is significant because



(a) Dataset (b) Manifold Lipschitz (ours) (c) Manifold Regularization (d) Ambient Regularization

Figure 1: Two moons dataset. The setting consists of a two dimensional classification problem of two classes with 1 labeled, and 200 unlabeled samples per class. The objective is to correctly classify the 200 unlabeled samples. We consider two cases (top) the estimated manifold has two connected components, and (bottom) the manifold is weakly connected (cf. Figure 1). We plot the output of a one layer neural network trained using Manifold Regularization, Manifold/Ambient Lipschitz.

regularizing is more brittle than imposing constraints. In the example in Figure 1, manifold regularization fails to separate the dataset when the moons are close [cf. Figure 1-(c), bottom]. Classification with a manifold Lipschitz constant constraint is more robust to this change in the data distribution [cf. Figure 1-(a), bottom].

The advantages of constraining Lipschitz constants on manifolds that we showcase in the synthetic example of Figure 1 manifest in the wild. To illustrate it, we empirically demonstrate this fact with two physical experiments. The first experiment entails learning a model for a differential drive ground robot in a mix of complex terrains. The second experiment consists of learning a dynamical model for a quadrotor. In both of these experiments constraining Lipschitz constants on manifolds improves upon standard risk minimization, manifold regularization, and the imposition of Lipschitz constraints in ambient space.

Global constraints on the manifold gradient yield a statistical constrained learning problem with an infinite and dense number of constraints. This is a challenging problem to approximate and solve. Here, we approach the solution of this problem in the Lagrangian dual domain and establish connections with manifold regularization that allow for the use of point cloud Laplacians. Our contributions include:

(C1) We introduce a constrained statistical risk minimization problem in which we learn a function that: (i) attains a target loss and (ii) attains the smallest possible manifold Lipschitz constant among functions that achieve this target loss (Section 2).

(C2) We introduce the Lagrangian dual problem and show that its empirical version is a statistically consistent approximation of the primal. These results do *not* require the learning parametrization to be linear (Section 3.1).

(C3) We generalize results from the manifold regularization literature to show that under regularity conditions, the evaluation of manifold Lipschitz constants can be recast in a more amenable form utilizing a weighted point cloud Laplacian (Proposition 3 in Section 3.2).

(C4) We present a dual ascent algorithm to find optimal multipliers. The function that attains the target loss and minimizes the manifold Lipschitz constant follows as a byproduct (Section 3.3).

(C5) We illustrate the merits of learning with global manifold smoothness guarantees with respect to ambient space and standard manifold regularization through two physical experiments: (i) learning model mismatches in differential drive steering over non-ideal surfaces and (ii) learning the motion dynamics of a quadrotor (Section 4).

### Related Work

This paper is at the intersection of learning with Lipschitz constant constraints (Oberman & Calder, 2018; Finlay et al., 2018b; Couellan, 2021; Pauli et al., 2021; Bungert et al., 2021; Miyato et al., 2018; Zhao & Liu, 2020; Krishnan et al., 2020; Shi et al., 2019; Lindemann et al., 2021; Arghal et al., 2021) and manifold regularization (Belkin et al., 2005;

Niyogi, 2013; Li et al., 2022; Hein et al., 2005; Belkin & Niyogi, 2005). Relative to learning with Lipschitz constraints, we offer the ability to leverage data manifolds. Since data manifolds are often obtained from unlabeled data, as in (Kejani et al., 2020; Belkin & Niyogi, 2004; Jiang et al., 2019; Kipf & Welling, 2016; Yang et al., 2016; Zhu, 2005; Lecouat et al., 2018; Ouali et al., 2020; Cabannes et al., 2021) we also use point-cloud Laplacian techniques to compute the integral of the norm of the gradient.

Relative to the literature on manifold regularization, we offer global smoothness assurances instead of an average penalty of large manifold gradients. Similar to us, (Krishnan et al., 2020) poses the problem of minimizing a Lipschitz constant. However, they utilize a softer surrogate (i.e. $p$-norm loss) which is a smoother version of the Lipschitz constant. Their approach therefore, tradeoffs numerical stability (small $p$) with accurate Lipschitz constant estimation ($p = \infty$). We do not work with surrogates and seek to minimize the maximum norm of the gradient utilizing an epigraph technique.

## 2. Globally Constraining Manifold Lipschitz Constants

We consider data pairs $(x, y)$ in which the input features $x \in \mathcal{M} \subset \mathbb{R}^D$ lie in a compact oriented Riemannian manifold $\mathcal{M}$ and the output features are real valued $y \in \mathbb{R}$. We study the regression problem of finding a function $f_\theta : \mathcal{M} \to \mathbb{R}$, parameterized by $\theta \in \Theta \subset \mathbb{R}^Q$ that minimizes the expectation of a nonnegative loss $\ell : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$, where $\ell(f_\theta(x), y)$ represents the loss of predicting output $f_\theta(x)$ when the world realizes the pair $(x, y)$. Data pairs $(x, y)$ are drawn according to an unknown probability distribution $p(x, y)$ on $\mathcal{M} \times \mathbb{R}$ which we can factor as $p(x, y) = p(x)p(y|x)$.

We are interested in learning smooth functions, i.e. functions with *controlled* variability over the manifold $\mathcal{M}$. We therefore let $\nabla_{\mathcal{M}} f_\theta(x)$ represent the manifold gradient of $f_\theta$ and introduce the following definition.

**Definition 1 (Manifold Lipschitz Constant)** *Given a Riemannian manifold $\mathcal{M}$, the function $f_\theta : \mathcal{M} \to \mathbb{R}$ is said to be $L$-Lipschitz continuous if there exists a strictly positive constant $L > 0$ such that for all pairs of points $x_1, x_2 \in \mathcal{M}$,*

$$|f_\theta(x_1) - f_\theta(x_2)| \leq L\, d_{\mathcal{M}}(x_1, x_2), \tag{1}$$

*where $d_{\mathcal{M}}(x_1, x_2)$ denotes the distance between $x_1$ and $x_2$ in the manifold $\mathcal{M}$. If the function $f_\theta$ is differentiable on the manifold, (1) is equivalent to requiring the gradient norm to be bounded by $L$,*

$$\begin{aligned}\|\nabla_{\mathcal{M}} f_\theta(x)\| \;&:=\; \lim_{\delta\to 0} \sup_{\substack{x'\in\mathcal{M}\,:\, x'\neq x,\\ d_{\mathcal{M}}(x,x')\leq\delta}} \frac{|f_\theta(x) - f_\theta(x')|}{d_{\mathcal{M}}(x, x')} \\ &\leq\; L, \quad \textit{for all } x \in \mathcal{M}. \end{aligned} \tag{2}$$

With Definition 1 in place and restricting attention to differentiable functions $f_\theta$, our stated goal of learning functions $f_\theta$ with controlled variability over the manifold $\mathcal{M}$ can be written as

$$\begin{aligned} P^* = \quad \min_{\theta\in\Theta,\rho\geq 0} \quad & \rho, \\ \text{subject to} \quad & \mathbb{E}_{p(x,y)}[\ell\left(f_\theta(x), y\right)] \leq \epsilon, \\ & \|\nabla_{\mathcal{M}} f_\theta(z)\|^2 \leq \rho, p(z)\text{-a.e.}, z \in \mathcal{M}. \end{aligned} \tag{3}$$

In this formulation, the statistical loss $\mathbb{E}_{p(x,y)}[\ell\left(f_\theta(x), y\right)]$ is required to be below a target level $\epsilon$. Of all functions $f_\theta$ that satisfy this loss requirement, Problem (3) defines as optimal those that have the smallest Lipschitz constant $L = \sqrt{\rho}$.

The goal of this paper is to develop methodologies to solve (3) when the data distribution and the manifold are unknown. To characterize the distribution, we use $N$ sample pairs $(x_n, y_n)$ drawn independently from the joint distribution $p(x, y)$. To characterize the manifold, we use $N^{'}$ samples $z_n$ drawn from the marginal distribution $p(x)$. This includes the $N$ samples $x_n$ from the (labeled) data pairs $(x_n, y_n)$, but may also include additional (unlabeled) samples $z_n, n = N + 1, \ldots, N + N^{'}$.

Observe from (3) that the problems of manifold regularization (Belkin et al., 2005; Niyogi, 2013; Kejani et al., 2020; Li et al., 2022) and Lipschitz constant control (Oberman & Calder, 2018; Finlay et al., 2018b; Couellan, 2021; Finlay et al., 2018a; Pauli et al., 2021; Krishnan et al., 2020; Shi et al., 2019; Lindemann et al., 2021; Arghal et al., 2021) are related. This connection is important to understand the merit of (3). To explain this better observe that there are three motivations for the problem formulation in (3): (i) it is often the case that if samples $x_1$ and $x_2$ are close, then the conditional distributions $p(y \mid x_1)$ and $p(y \mid x_2)$ are close as well. A function $f_\theta$ with small Lipschitz constant leverages this property, (ii) the Lipschitz constant of $f_\theta$ is *guaranteed* to be smaller than $L = \sqrt{\rho}$, this provides advantages in, e.g., physical systems where Lipschitz constant guarantees translates to stability and safety assurances, (iii) it leverages the intrinsic low-dimensional structure of the manifold $\mathcal{M}$ embedded in the ambient space. In particular, this permits taking advantage of unlabeled data.

Motivations (i) and (ii) are tropes of the Lipschitz regularization literature; e.g., (Oberman & Calder, 2018; Finlay et al., 2018b; Couellan, 2021; Finlay et al., 2018a; Pauli et al.,

2021; Krishnan et al., 2020; Shi et al., 2019; Lindemann et al., 2021; Arghal et al., 2021). Indeed, the problem formulation in (3) is inspired in similar problem formulations in which the Lipschitz constant is regularized in the ambient space,

$$\begin{aligned} \underset{\theta\in\Theta, L\geq 0}{\text{minimize}} \quad & L, \\ \text{subject to} \quad & \mathbb{E}_{p(x,y)}[\ell\left(f_\theta(x), y\right)] \leq \epsilon, \\ & |f_\theta(w) - f_\theta(z)| \leq L\,\|w - z\|, \\ & (w, z) \sim p(w) \times p(z). \end{aligned} \tag{4}$$

A key difference between (3) and (4) is that the latter uses a Lipschitz condition that does not require differentiability. A more important difference is that in (4), the Lipschitz constant is regularized in the ambient space. The distance between features $w$ and $z$ in (4) is the Euclidean distance $\|w - z\|$. This is disparate from the manifold metric $d_{\mathcal{M}}(w, z)$ that is implicit in the manifold gradient constraint in (3). Thus, the formulation in (3) improves upon (4) because it leverages the structure of the manifold $\mathcal{M}$ [cf. motivation (iii)].

Motivations (i) and (iii) are themes of the manifold regularization literature (Belkin et al., 2005; Niyogi, 2013; Kejani et al., 2020; Li et al., 2022). And, indeed, it is ready to conclude by invoking Green's first identity (see Section 3.2) that the formulation in (3) is also inspired in the manifold regularization problem,

$$\begin{aligned} \underset{\theta\in\Theta}{\text{minimize}} \quad & \mathbb{E}_{p(x,y)}[\ell\left(f_\theta(x), y\right)] \\ & + \gamma \int_{\mathcal{M}} \|\nabla_{\mathcal{M}} f_\theta(z)\|^2 p(z) dV(z). \end{aligned} \tag{5}$$

The difference between (3) and (5) is that the latter adds the manifold Lipschitz constant as a regularization penalty. This is disparate from the imposition of a manifold Lipschitz constraint in (3). The regularization in (5) *favors* solutions with small Lipschitz constant by penalizing large Lipschitz constants, while the constraint in (3) *guarantees* that the Lipschitz constant is bounded by $\sqrt{\rho}$. This is the distinction between *regularizing* a Lipschitz constant and *constraining* a Lipschitz constant. The constraint in (5) is also imposed at *all* points in the manifold, whereas the regularization in (5) is an average over the manifold. Taking an average allows for large Lipschitz constants at some specific points if this is canceled out by small Lipschitz constants in other points of the manifold (Bubeck & Sellke, 2021; Bubeck et al., 2021). Both of these observations imply that (3) improves upon (5) because it offers global smoothness guarantees that are important in, e.g., physical systems [cf. motivation (iii)].

**Remark 1 (Manifold Lipschitz formulations)** There are three arbitrary choices in (3): (a) We choose to constrain the average statistical loss $\mathbb{E}_{p(x,y)}[\ell\left(f_\theta(x), y\right)] \leq \epsilon$; (b) we choose to constrain the pointwise Lipschitz constant $\|\nabla_{\mathcal{M}} f_\theta(z)\|^2 \leq \rho$; and (c) we choose as our objective to require a target loss $\epsilon$ and minimize the Lipschitz constant $L = \sqrt{\rho}$. We can alternatively choose to constrain the pointwise loss $\ell\left(f_\theta(x), y\right) \leq \epsilon$, to constrain the average Lipschitz constant $\int_{\mathcal{M}} \|\nabla_{\mathcal{M}} f_\theta(z)\|^2 p(z) dV(z) \leq \rho$ [cf (5)], or to require a target smoothness $L = \sqrt{\rho}$ and minimize the loss $\epsilon$. All of the possible eight combinations of choices are of interest. We formulate (3) because it is the most natural intersection between the regularization of Lipschitz constants in ambient spaces [cf. (4)] and manifold regularization [cf. (5)]. The techniques we develop in this paper can be adapted to any of the other seven alternative formulations.

## 3. Learning with Global Lipschitz Constraints

Problem (3) is a constrained learning problem that we will solve in the dual domain (Chamon & Ribeiro, 2020). To that end, observe that (3) has statistical and pointwise constraints. The loss constraint $\mathbb{E}_{p(x,y)}[\ell\left(f_\theta(x), y\right)] \leq \epsilon$ is said to be statistical because it restricts the expected loss over the data distribution. The Lipschitz constant constraints $\|\nabla_{\mathcal{M}} f_\theta(z)\|^2 \leq \rho$ are said to be pointwise because they are imposed for all individual points in the manifold except perhaps for a set of zero measure. Consider then a Lagrange multiplier $\mu$ associated with the statistical constraint $\mathbb{E}_{p(x,y)}[\ell\left(f_\theta(x), y\right)] \leq \epsilon$ and a Lagrange multiplier distribution $\lambda(z)$ associated with the set of pointwise constraints $\|\nabla_{\mathcal{M}} f_\theta(z)\|^2 \leq \rho$. We define the Lagrangian $L(\theta, \mu, \lambda)$ associated with (3) as

$$\begin{aligned} L(\theta, \mu, \lambda) :=& \mu\Big(\mathbb{E}[\ell\big(f_\theta(x), y\big)] - \epsilon\Big) \\ & + \int_{\mathcal{M}} \lambda(z) \|\nabla_{\mathcal{M}} f_\theta(z)\|^2 p(z) dV(z). \end{aligned} \tag{6}$$

The dual problem associated with (3) can then be written as

$$\begin{aligned} D^* = \max_{\mu,\lambda\geq 0} \min_{\theta} & L(\theta, \mu, \lambda), \\ \text{subject to} \quad & \int_{\mathcal{M}} \lambda(z) p(z) dV(z) = 1. \end{aligned} \tag{7}$$

We point out that in (7) we remove $\rho$ from the Lagrangian by incorporating the dual variable constraint $\int_{\mathcal{M}} \lambda(x) p(x) dV(x) = 1$ (see Appendix A for details).

We henceforth use the dual problem (7) in lieu of (3). Since we are interested in situations in which we do not have access to the data distribution $p(x, y)$, we further consider empirical versions of (7). Given $N$ i.i.d. samples $(x_n, y_n)$ drawn from $p(x, y)$, define the empirical La-

grangian $\hat{L}(\theta, \hat{\mu}, \hat{\lambda})$ as

$$\hat{L}(\theta, \hat{\mu}, \hat{\lambda}) := \hat{\mu}\left(\frac{1}{N}\sum_{n=1}^{N} \ell\big(f_\theta(x_n), y_n\big) - \epsilon\right) + \frac{1}{N'}\sum_{n=1}^{N'} \hat{\lambda}(z_n)\|\nabla_{\mathcal{M}} f_\theta(z_n)\|^2. \quad (8)$$

and the empirical dual problem as

$$\hat{D}^\star = \max_{\hat{\mu}, \hat{\lambda} \geq 0} \min_{\theta} \hat{L}(\theta, \hat{\mu}, \hat{\lambda}) \quad (9)$$
$$\text{subject to } \frac{1}{N}\sum_{n=1}^{N} \hat{\lambda}(x_n) = 1.$$

To simplify notation, we assume from now on that no unlabeled samples are available. If unlabeled samples are given, the modification is straightforward (see Appendix B).

The remainder of this section provides three technical contributions:

In section 3.1 we address contribution C2: To justify the use of (9) we must show statistical consistency with respect to the primal problem (3). This is challenging for two reasons: (i) since we do not assume the use of a linear parameterization, (3) is not a linear problem in $\theta$. Thus, the primal (3) and dual (7) are not necessarily equivalent, (ii) since we are maximizing over the dual variables $\mu$ and $\lambda(z)$, we do not know if the empirical dual formulation in (9) is close to the statistical dual formulation in (7). We overcome these two challenges and show that the empirical dual problem (9) is a consistent approximation of the statistical primal problem (Proposition 1).

In section 3.2 we address contribution C3: Solving (9) requires evaluating the gradient norm sum $\sum_{n=1}^{N} \hat{\lambda}(x_n)\|\nabla_{\mathcal{M}} f_\theta(x_n)\|^2$. We will generalize results from the manifold regularization literature to show that under regularity conditions on $\lambda$, the gradient norm integral can be computed in a more amenable form utilizing a weighted point-cloud Laplacian (Proposition 3). (Contribution C3).

In section 3.3 we address contribution C4: We introduce a primal-dual algorithm to solve (9).

### 3.1. Statistical consistency of the empirical dual problem

In this section, we show that (9) is close to (3) under the following assumptions:

**Assumption 1** *The loss $\ell(\cdot, y)$ is $M$-Lipschitz continuous, $[0, B]$-valued, and convex for all $y \in \mathbb{R}$.*

**Assumption 2** *Let $\mathcal{H} = \{f_\theta \mid \theta \in \Theta\}$ with compact $\Theta \subset \mathbb{R}^Q$ be the hypothesis class and let $\bar{\mathcal{H}} = \overline{\text{conv}}(\mathcal{H})$ be the closure of its convex hull. For each $\nu > 0$ and $\varphi \in \bar{\mathcal{H}}$, there exists $\theta \in \Theta$ such that simultaneously $\sup_{z\in\mathcal{M}} |\varphi(z) - f_\theta(z)| \leq \nu$ and $\sup_{z\in\mathcal{M}} \|\nabla_{\mathcal{M}}\varphi(z) - \nabla_{\mathcal{M}} f_\theta(z)\| \leq \nu$.*

**Assumption 3** *The functions in the hypothesis class $\mathcal{H}$ satisfy $\|\nabla_{\mathcal{M}} f_{\theta_1}(z) - \nabla_{\mathcal{M}} f_{\theta_2}(z)\| \leq G|f_{\theta_1}(z) - f_{\theta_2}(z)|$, for all $\theta_1, \theta_2 \in \Theta$.*

**Assumption 4** *There exists $\zeta(N, \delta) \geq 0$, and $\hat{\zeta}(N, \delta) \geq 0$ monotonically decreasing with $N$, such that with probability $1 - \delta$ over independent draws $(x_n, y_n) \sim p$,*

$$\left|\mathbb{E}[\ell\left(f_\theta(x), y\right)] - \frac{1}{N}\sum_{n=1}^{N} \ell\big(f_\theta(x_n), y_n\big)\right| \leq \zeta(N, \delta),$$
$$\left|\mathbb{E}[f_\theta(x)] - \frac{1}{N}\sum_{n=1}^{N} f_\theta(x_n)\right| \leq \hat{\zeta}(N, \delta), \quad (10)$$

*for all $\theta \in \Theta$ and all distributions $p$.*

**Assumption 5** *There exists a feasible solution $\tilde{\theta} \in \Theta$ such that $\mathbb{E}[\ell(f_{\tilde{\theta}}(x), y)] < \epsilon - M\nu$.*

Assumption 1 holds for most losses utilized in practice. Assumption 2 is a requirement on the richness of the parametrization. In the particular case of neural networks, the covering constant $\nu$ is upper bounded by the universal approximation bound of the neural network. Assumption 3 also holds for neural networks with smooth nonlinearities – e.g., hyperbolic tangents. The *uniform convergence* property (10) is customary in learning theory to prove PAC learnability and is implied by bounds on complexity measures such as VC dimension or Rademacher complexity (Mohri et al., 2018; Vapnik, 1999; Shalev-Shwartz & Ben-David, 2014). In fact, if it has bounded Rademacher complexity, both $\zeta$ and $\hat{\zeta}$ are bounded due to Assumption 1. The following proposition provides the desired bound.

**Proposition 1** *Let $\hat{\mu}^\star, \hat{\lambda}^\star$ be solutions of the empirical dual problem (9). Under assumptions 1–5, there exists $\hat{\theta}^\star \in \operatorname{argmin}_\theta \hat{L}(\theta, \hat{\mu}^\star, \hat{\lambda}^\star)$ such that, with probability $1 - 5\delta$,*

$$|P^\star - \hat{D}^\star| \leq \mathcal{O}(\nu) + (1 + \Delta)\zeta(N, \delta) + \mathcal{O}(\hat{\zeta}(N, \delta)), \quad (11)$$

*where $\Delta = \max(\hat{\mu}^\star, \mu^\star) < \infty$ and $\mu^\star, \hat{\mu}^\star$ are solutions of (7), (9) respectively.*

Proposition 1 shows that the empirical dual problem 9 is statistically consistent. That is to say, for any draw of $N$ samples according to $p$, the difference between the empirical dual problem (9) and the statistical smooth learning problem (3) decreases as $N$ increases. This difference is bounded in terms of the richness of the parametrization ($\nu$), the difficulty of the fit requirement (as expressed by the optimal dual variables $\mu^\star, \hat{\mu}^\star$), and the number of samples ($N$).

The guarantee has a form typical for constrained learning problems (Chamon et al., 2023). Proposition 1 states that we are able to predict what is the minimum norm of the gradient that a function class can have while achieving an expected loss of at most $\epsilon$. This is important because we do not require access to the distribution $p$, only a set of $N$ samples from this distribution. On the other hand, Proposition 1 does not state that by solving the dual problem (9) we will obtain a solution of the primal problem (3). The following proposition provides a bound on the near feasibility of the solution of (9) with respect to the solution of (3).

**Proposition 2** *Let $\hat{\mu}^\star, \hat{\lambda}^\star$ be solutions of the empirical dual problem (9). Under assumptions 1–5, there exists $\hat{\theta}^\star \in \operatorname{argmin}_\theta \hat{L}(\theta, \hat{\mu}^\star, \hat{\lambda}^\star)$ such that, with probability $1 - 5\delta$,*

$$\max_{n\in[N]} \|\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n)\|^2 \leq P^* + \mathcal{O}(\nu) + (1+\Delta)\zeta(N,\delta) + \mathcal{O}(\hat{\zeta}(N,\delta)) + \hat{\mu}^* \left| \frac{1}{N}\sum_{n=1}^{N} \ell(f_{\hat{\theta}^*}(x_n), y_n) - \epsilon \right| \tag{12}$$

$$\text{and } \mathbb{E}[\ell(f_{\hat{\theta}^\star}(x), y)] \leq \epsilon + \zeta(N,\delta). \tag{13}$$

*where $\Delta = \max(\hat{\mu}^\star, \mu^\star) < \infty$ and $\mu^\star, \hat{\mu}^\star$ are solutions of (7), (9) respectively.*

Proposition 2 provides near optimality and near feasibility conditions for solutions $\hat{\theta}^*$ obtained through the empirical dual problem (9). The difference between the maximum gradient of the obtained solution $\theta$ and the optimal value $P^*$ is bounded by the number of samples $N$ as well as the empirical constraint violation. Notice that even though the optimal dual variable $\hat{\mu}$ is not known, the constraint violation can be evaluated in practice as it only requires to evaluate the obtained function over the $N$ given samples.

**Remark 2 (Interpolators)** *In practice, the number of parameters in a parametric function (e.g., Neural Network) tends to exceed the dimension of the input, which allows functions to interpolate the data, i.e., to attain zero loss on the dataset. Proposition 2 presents a connection to interpolating functions. By setting $\epsilon = 0$, if the function achieves zero error over the empirical distribution i.e. $\ell(f_{\hat{\theta}^\star}(x_n), y_n) = 0$, for all $n \in [N]$, then the dependency on $\mu^*$ disappears. This implies that within the classifiers that interpolate the data, the one with the minimum Lipschitz constant over the available samples will probably be the one with the minimum Lipschitz constant.*

### 3.2. From Manifold Gradient to Discrete Laplacian

We derive an alternative way of computing the integral of the norm of the gradient utilizing samples. To do so, we define the normalized point-cloud Laplacian according a probability distribution $\lambda$.

**Definition 2 (Point-cloud Laplacian)** *Consider a set of points $x_1, \ldots, x_N \in \mathcal{M}$, sampled according to the density $\lambda : \mathcal{M} \to \mathbb{R}$. The normalized graph Laplacian of $f_\theta$ at $z \in \mathcal{M}$ is defined as*

$$\mathbf{L}^t_{\lambda,N} f_\theta(z) = \frac{1}{N}\sum_{n=1}^{N} W(z, x_n)\big(f_\theta(z) - f_\theta(x_n)\big), \tag{14}$$

$$\text{for } W(z,x_n) = \frac{1}{t}\frac{G_t(z,x_n)}{\sqrt{\hat{w}(z)\hat{W}(x_n)}},$$

$$\text{with} \quad G_t(z,x_n) = \frac{1}{(4\pi t)^{d/2}} e^{-\frac{\|z-x_n\|^2}{4t}},$$

$$\hat{w}(z) = \frac{1}{N}\sum_{n=1}^{N} G_t(z,x_n),$$

$$\text{and} \quad \hat{W}(x_n) = \frac{1}{N-1}\sum_{m\neq n} G_t(x_m, x_n).$$

As long as the function considered is smooth enough, the following convergence result holds:

**Proposition 3 (Point-Cloud Estimate)** *Let $\Lambda$ be the set of probability densities defined on a compact $d$-dimensional differentiable manifold $\mathcal{M}$ isometrically embedded in $\mathbb{R}^D$ such that $\Lambda = \{\lambda : 0 < a \leq \lambda(z) \leq b < \infty, |\frac{\partial\lambda}{\partial x}| \leq c < \infty$ and $|\frac{\partial^2\lambda}{\partial x^2}| \leq d < \infty$ for all $z \in \mathcal{M}\}$, and let $f_\theta$, with $\theta \in \Theta$, be a family of functions with uniformly bounded derivatives up to order 3 vanishing at the boundary $\partial\mathcal{M}$. For any $\epsilon > 0, \delta > 0$, there exists $N^{'}, N$ such that*

$$P\left[\sup_{\lambda\in\Lambda,\theta\in\Theta} \left| \int \|\nabla_{\mathcal{M}} f_\theta(z)\|^2 \lambda(z) dV(z) - \frac{1}{N}\sum_{i=1}^{N} f_\theta(z_i)\mathbf{L}^t_{\lambda,N'} f_\theta(z_i)\lambda(z_i) \right| > \epsilon \right] \leq \delta \tag{15}$$

*where the point-cloud Laplacian $\mathbf{L}^t_{\lambda,N'} f_\theta$ is as in Definition 2, with $t = N^{'-\frac{1}{d+2+\alpha}}$ for any $\alpha > 0$.*

The proof of Proposition 3 relies on two steps. First, we relate the integral of the norm of the gradient over the manifold with the integral of the continuous Laplace-Beltrami operator by virtue of Green's identity. Second, we approximate the value of the Laplace-Beltrami operator by the point-cloud Laplacian. Proposition 3 connects the integral of the norm of the gradient of function $f_\theta$ with a point-cloud Laplacian operator. This result connects the dual problem (9), with the primal (3) while allowing for a more amenable way of computing the integral.

**Remark 3 (Laplacian Regularization)** *The dual problem (7) is closely related to the manifold regularization*

*problem (5). In particular, the two become equivalent by substituting $\rho = \mu^{-1}$ and utilizing a uniform distribution for $\lambda$. The key difference between the problems is given by the dual variable $\lambda$, which can be thought as a probability distribution over the manifold that penalizes regions of the manifold where the norm of the gradient of $f_\theta$ is larger. The standard procedure in Laplacian regularization is to calculate the graph-Laplacian of the set of points $\mathbf{L}$ utilizing the heat kernel and compute the integral by $\mathbf{f}^T\mathbf{L}\mathbf{f}$, where $\mathbf{f} = [f_\theta(x_1), \dots, f_\theta(x_n)]^T$. In the case of Manifold Lipschitz, the same product can be computed, but utilizing the re-weighted point-cloud laplacian.*

### 3.3. Dual Ascent Algorithm

We outline an iterative and empirical primal-dual algorithm to solve the dual problem. Upon the initialization of $\theta_0, \lambda_0$ and $\mu_0$, we set out to minimize the dual function using,

$$\theta_{k+1} = \theta_k - \eta_\theta \nabla_\theta L(\theta_k, \mu_k, \lambda_k), \tag{16}$$

where $\eta_\theta$ is a positive stepsize. Note that to compute (16), we can either utilize the gradient version of the Lagrangian (9) or the point-cloud Laplacian (15). Consequent to updating $\theta$, we update the dual variables by

$$\mu_{k+1} = \left[\mu_k + \eta_\mu \left(\frac{1}{N}\sum_{n=1}^{N} \ell(f_\theta(x_n), y_n) - \epsilon\right)\right]_+, \tag{17}$$

$$\tilde{\lambda}_{k+1}(x_n) = \lambda_k(x_n) + \eta_\lambda \|\nabla_{\mathcal{M}} f_\theta(x_n)\|^2, n = 1, \dots, N \tag{18}$$

$$\lambda_{k+1} = \underset{\lambda}{\operatorname{argmin}} \|\tilde{\lambda}_{k+1} - \lambda_{k+1}\|, \tag{19}$$

$$\text{subject to } \sum_{n=1}^{N} \lambda_{k+1}(x_n) = N,$$

where $\eta_\mu, \eta_\lambda$ are again a positive stepsize. Note that we require a convex projection over $\tilde{\lambda}$ to satisfy the normalizing constraint. In step (18), we need to estimate the norm of the gradient at data point $x_n$, which we do using neighboring data points. Intuitively, the primal-dual procedure increases the value of $\lambda(x_n)$ at points in which the norm of the gradient is larger. The role of $\mu$ is to enforce the loss $\ell$ to be smaller than $\epsilon$, by adjusting the relative importance of the loss $\ell$ over the norm of the integral (see Appendix F).

## 4. Experiments

To demonstrate the effectiveness of our method, we conduct two real world experiments with physical systems. In Section 4.1, we tackle a regression problem in which we try to predict the error made when estimating the state of a ground robot making turns in both pavement and grass. In Section 4.2 we learn the dynamics of a quadrotor when taking off and making circles in open air. In both experiments, data is acquired from real world robots and we compare our method against standard Empirical Risk Minimization (ERM), Ambient Regularization, and Manifold Regularization.

### 4.1. Ground Robot Error Prediction

In this experiment, we seek to learn the error that a model would make in predicting the dynamics of a ground robot by looking at the states throughout a trajectory (Koppel et al., 2016). The data acquisition of this experiment involves an *iRobot* Packbot equipped with high resolution camera. The setting of the data acquisition is a robot making turns on both pavement and grass. The dynamics of the system are govern by the discrete-time nonlinear state-space system

$$x_{k+1} = f(x_k, u_k) + g(u_k), \tag{20}$$

where $x_k$ is the state of the system, $u_k$ is the control input, $f(x_k, u_k)$ is the model prediction, and $g(u_k)$ is the non-modelable error of the prediction. In this setting, the robot state involves the position and the control inputs are linear and angular velocities. The dynamics of the system are modeled by $f(x_k, u_k)$ and they involve the mass of the robot, the radius of the wheel, and other known parameters of the robot. In practice, the model $f(x_k, u_k)$ is not perfect and the model mismatch is given by the difference in friction with the ground, delays in communications with the sensors, and discrepancies in the robot specifications. To make matters worse, some of the discrepancies are difficult to model or intractable to compute in practice.

We possess trajectories in the form of time series of the control signals, i.e., linear and angular velocity of the robot. For each trajectory, the average and variance of the mismatch between the real and the model-predicted states is quantified. Given a dataset of trajectories and errors made by the model, the objective is to predict the error that the model will make on a given trajectory.

In order to construct the point-cloud Laplacian, we measured the euclidean distance between samples and computed (14). In this case we added a threshold, i.e. minimum magnitude of $\mathbf{L}_{ij}$ below which the value of $\mathbf{L}_{ij}$ becomes 0. This is to disregard the effect of samples that are far away, by only considering neighborhoods of each sample. By doing this we are constructing a manifold in which samples that are sufficiently close will be forced to have similar outputs. However, if samples are not sufficiently close, the similarity between the outputs of the function will not be forced to be small (to exemplify this point we have a toy example in Appendix G.5). The manifold in this case is constructed by time series that are sufficiently similar to each other. For further details of the experiment and sample trajectories can be found in Appendix G.2.

| Method | Grass | Pavement |
|---|---|---|
| ERM | 0.42 | 0.0120 |
| Ambient Regularization | 0.31 | 0.0065 |
| Manifold Regularization | 0.38 | 0.0045 |
| Manifold Lipschitz (ours) | 0.25 | 0.0032 |

Table 1: Error prediction accuracy for the Ground Robot Experiment.

| Method | Error on Test Trajectory |
|---|---|
| ERM | 0.00666 |
| Ambient Regularization | 0.00734 |
| Manifold Regularization | 0.00625 |
| Manifold Lipschitz (ours) | 0.00036 |

Table 2: State prediction error of a quadrotor flying in circles on an unseen trajectory.

As seen on Table 1, regularization improves the accuracy over the standard ERM framework. This is related to the fact that given that the underlying predictive model is the same in all trajectories, similar trajectories will have similar errors. However, our method improves upon both regularization techniques. This is related to the idea that between experiments only the velocities change (the environment is fixed), so we can intuitively imagine a continuous transition in the error made by the model. This explains why our method (Manifold Lipschitz) by forcing the function to be smooth over the trajectory space is significantly better in both experiments. Our method also improves upon Ambient regularization because the euclidean distance is able to approximate the distance on the manifold *locally*, but it is not able to approximate it *globally*. For a better clarification on this last point see Appendix G.5. In all, in this experiment we show that our method improves the generalization capabilities of a function.

### 4.2. Quadrotor Model Mismatch

In this section introduce a state prediction problem based real world dynamics. The setting consists of a quadrotor taking off and flying in circles for 12 seconds.

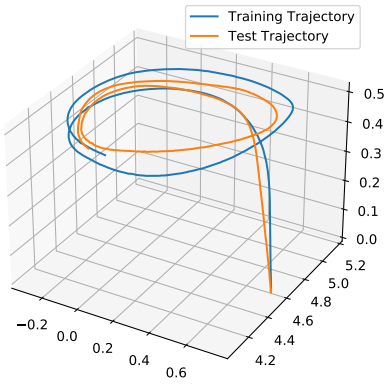



Figure 2: Quadrotor Sample Trajectories

The experimental setup is to target a speed of $0.4$m/s and to track a circular trajectory of radius $0.5$m. We consider 2 trajectories of 12 seconds each, and the starting position of the quadrotor is the same for both trajectories (the ground). For each time stamp $t$, we have measurements of position, velocity and acceleration in $\mathbb{R}^3$, i.e. $[x_t, y_t, z_t, \dot{x}_t, \dot{y}_t, \dot{z}_t, \ddot{x}_t, \ddot{y}_t, \ddot{z}_t]$.

For the learning procedure we utilize the $6000$ samples, and we seek to minimize the mean square error loss between the next state and the prediction given the current state. We train a two layer neural network with different methods as seen in Table 2. To test the accuracy of the learned function, we compute the difference between the predicted state and the next state on the test trajectory.

Analogous to experiment 4.1, to construct the point-cloud Laplacian, we compute the euclidean distances in a neighborhood of each sample. Intuitively, by looking at the trajectories 2, we can expect next states to be similiar only on samples that are sufficiently close. The euclidean distance is able to approximate the manifold *locally* but fails when it is too large. By looking at Figure 2 we can see that even though the training and testing trajectories are not the same, there is a close resemblance between the two of them. This allows a manifold model to be effective in predicting the next state. However, by looking at the results in Table 2, we see that adding regularization does not always help, as ambient regularization does not improve upon ERM. This is related to the fact that the simple euclidean distance is not necessarily a good measure of the how related two states are. Instead, one conclusion of the results shown in Table 2 is that adding regularization on the manifold space always improves upon ERM. In particular, our method give performance an order of magnitude better than standard ERM and standard Laplacian Regularization. This is a consequence of adding extra information about the problem by grouping similar states together, which reduces the impact of the noise in a particular sample. To conclude, we show that our method obtains an improvement over all the techniques considered and that by predicting the next state of a quadrotor from the current state utilizing smooth functions improves generalization for noisy measurements.

## 5. Conclusion

In this work, we presented a constraint learning method to obtain smooth functions over manifold data. We showed that under mild conditions, the problem of finding smooth

functions over a manifold can be reformulated as a weighted point-cloud Laplacian penalty over varying probability distributions whose dynamics are govern by the constraint violations. Two experiments on real world data validate the empirical advantages of obtaining functions that vary smoothly over the data.

ACKNOWLEDGMENTS

The authors would like to thank Alec Koppel, Garrett Warnell, Jonathan Fink, Ethan Stump collecting and providing the dataset utilized for experiment 4.1. The authors would also like to thank Tom Zhang and M. Ani Hsieh for collecting and providing the dataset utilized for experiment 4.2.

The work of Luiz F. O. Chamon was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy Grant EXC 2075-390740016.

## References

Arghal, R., Lei, E., and Bidokhti, S. S. Robust graph neural networks via probabilistic lipschitz constraints. *arXiv preprint arXiv:2112.07575*, 2021.

Bai, Q., Bedi, A. S., Agarwal, M., Koppel, A., and Aggarwal, V. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. *arXiv preprint arXiv:2109.06332*, 2021.

Belkin, M. and Niyogi, P. Semi-supervised learning on riemannian manifolds. *Machine learning*, 56(1):209–239, 2004.

Belkin, M. and Niyogi, P. Towards a theoretical foundation for laplacian-based manifold methods. In *International Conference on Computational Learning Theory*, pp. 486–500. Springer, 2005.

Belkin, M., Niyogi, P., and Sindhwani, V. On manifold regularization. In *International Workshop on Artificial Intelligence and Statistics*, pp. 17–24. PMLR, 2005.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Bubeck, S. and Sellke, M. A universal law of robustness via isoperimetry. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=z71OSKqTFh7.

Bubeck, S., Li, Y., and Nagaraj, D. M. A law of robustness for two-layers neural networks. In *Conference on Learning Theory*, pp. 804–820. PMLR, 2021.

Bungert, L., Raab, R., Roith, T., Schwinn, L., and Tenbrinck, D. Clip: Cheap lipschitz training of neural networks. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 307–319. Springer, 2021.

Cabannes, V., Pillaud-Vivien, L., Bach, F., and Rudi, A. Overcoming the curse of dimensionality with laplacian regularization in semi-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Castellano, A., Min, H., Bazerque, J., and Mallada, E. Reinforcement learning with almost sure constraints. *arXiv preprint arXiv:2112.05198*, 2021.

Cervino, J., Ruiz, L., and Ribeiro, A. Training stable graph neural networks through constrained learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4223–4227. IEEE, 2022.

Chamon, L. and Ribeiro, A. Probably approximately correct constrained learning. *Advances in Neural Information Processing Systems*, 33:16722–16735, 2020.

Chamon, L. F. O., Paternain, S., Calvo-Fullana, M., and Ribeiro, A. Constrained learning with non-convex losses. *IEEE Transactions on Information Theory*, 69(3):1739–1760, 2023. doi: 10.1109/TIT.2022.3187948.

Chee, K. Y., Jiahao, T. Z., and Hsieh, M. A. Knode-mpc: A knowledge-based data-driven predictive control framework for aerial robots. *IEEE Robotics and Automation Letters*, 7(2):2819–2826, 2022.

Couellan, N. The coupling effect of lipschitz regularization in neural networks. *SN Computer Science*, 2(2):1–9, 2021.

Do Carmo, M. P. and Flaherty Francis, J. *Riemannian geometry*, volume 6. Springer, 1992.

Dunson, D. B., Wu, H.-T., and Wu, N. Spectral convergence of graph laplacian and heat kernel reconstruction in $l_\infty$ from random samples. *Applied and Computational Harmonic Analysis*, 55:282–336, 2021.

Durrett, R. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Eisen, M., Zhang, C., Chamon, L. F., Lee, D. D., and Ribeiro, A. Learning optimal resource allocations in wireless systems. *IEEE Transactions on Signal Processing*, 67(10):2775–2790, 2019.

Elenter, J., NaderiAlizadeh, N., and Ribeiro, A. A lagrangian duality approach to active learning. *arXiv preprint arXiv:2202.04108*, 2022.

Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. Efficient and accurate estimation of lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Finlay, C., Calder, J., Abbasi, B., and Oberman, A. Lipschitz regularized deep neural networks generalize and are adversarially robust. *arXiv preprint arXiv:1808.09540*, 2018a.

Finlay, C., Oberman, A. M., and Abbasi, B. Improved robustness to adversarial examples using lipschitz regularization of the loss. *CoRR*, abs/1810.00953, 2018b. URL http://arxiv.org/abs/1810.00953.

Hasanbeig, M., Abate, A., and Kroening, D. Logically-constrained reinforcement learning. *arXiv preprint arXiv:1801.08099*, 2018.

Hein, M., Audibert, J.-Y., and Luxburg, U. v. From graphs to manifolds–weak and strong pointwise consistency of graph laplacians. In *International Conference on Computational Learning Theory*, pp. 470–485. Springer, 2005.

Hein, M., Audibert, J.-Y., and Luxburg, U. v. Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8(6), 2007.

Jiahao, T. Z., Hsieh, M. A., and Forgoston, E. Knowledge-based learning of nonlinear dynamics and chaos. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31 (11):111101, 2021.

Jiahao, T. Z., Chee, K. Y., and Hsieh, M. A. Online dynamics learning for predictive control with an application to aerial robots. *arXiv preprint arXiv:2207.09344*, 2022.

Jiang, B., Zhang, Z., Lin, D., Tang, J., and Luo, B. Semi-supervised learning with graph learning-convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11313–11320, 2019.

Jin, C. and Rinard, M. Manifold regularization for locally stable deep neural networks. *arXiv preprint arXiv:2003.04286*, 2020.

Kejani, M. T., Dornaika, F., and Talebi, H. Graph convolution networks with manifold regularization for semi-supervised learning. *Neural Networks*, 127:160–167, 2020.

Khoury, M. and Hadfield-Menell, D. On the geometry of adversarial examples. *arXiv preprint arXiv:1811.00525*, 2018.

Khrulkov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I., and Lempitsky, V. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Koppel, A., Fink, J., Warnell, G., Stump, E., and Ribeiro, A. Online learning for characterizing unknown environments in ground robotic vehicle models. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 626–633, 2016. doi: 10.1109/IROS.2016.7759118.

Krishnan, V., Makdah, A., AlRahman, A., and Pasqualetti, F. Lipschitz bounds and provably robust training by laplacian smoothing. *Advances in Neural Information Processing Systems*, 33:10924–10935, 2020.

Krogh, A. and Hertz, J. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

Lassance, C., Gripon, V., and Ortega, A. Laplacian networks: Bounding indicator function smoothness for neural networks robustness. *APSIPA Transactions on Signal and Information Processing*, 10, 2021.

Lecouat, B., Foo, C.-S., Zenati, H., and Chandrasekhar, V. R. Semi-supervised learning with gans: Revisiting manifold regularization. *arXiv preprint arXiv:1805.08957*, 2018.

Li, Z., Chen, Y., LeCun, Y., and Sommer, F. T. Neural manifold clustering and embedding. *arXiv preprint arXiv:2201.10000*, 2022.

Lindemann, L., Hu, H., Robey, A., Zhang, H., Dimarogonas, D., Tu, S., and Matni, N. Learning hybrid control barrier functions from data. In *Conference on Robot Learning*, pp. 1351–1370. PMLR, 2021.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M. E., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Moosavi-Dezfooli, S.-M., Fawzi, A., Uesato, J., and Frossard, P. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE/CVF Conference*

*on Computer Vision and Pattern Recognition*, pp. 9078–9086, 2019.

Niyogi, P. Manifold regularization and semi-supervised learning: Some theoretical analyses. *Journal of Machine Learning Research*, 14(5), 2013.

Oberman, A. M. and Calder, J. Lipschitz regularized deep neural networks converge and generalize. *arXiv preprint arXiv:1808.09540*, 2018.

Ouali, Y., Hudelot, C., and Tami, M. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020.

Paternain, S., Chamon, L., Calvo-Fullana, M., and Ribeiro, A. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32, 2019.

Paternain, S., Calvo-Fullana, M., Chamon, L. F., and Ribeiro, A. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 2022.

Pauli, P., Koch, A., Berberich, J., Kohler, P., and Allgöwer, F. Training robust neural networks using lipschitz bounds. *IEEE Control Systems Letters*, 6:121–126, 2021.

Robey, A., Chamon, L., Pappas, G. J., Hassani, H., and Ribeiro, A. Adversarial robustness with semi-infinite constrained learning. *Advances in Neural Information Processing Systems*, 34:6198–6215, 2021.

Rosca, M., Weber, T., Gretton, A., and Mohamed, S. A case for new neural network smoothness constraints. In Zosa Forde, J., Ruiz, F., Pradier, M. F., and Schein, A. (eds.), *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, volume 137 of *Proceedings of Machine Learning Research*, pp. 21–32. PMLR, 12 Dec 2020. URL https://proceedings.mlr.press/v137/rosca20a.html.

Rosenberg, S. *The Laplacian on a Riemannian Manifold: An Introduction to Analysis on Manifolds*. London Mathematical Society Student Texts. Cambridge University Press, 1997. doi: 10.1017/CBO9780511623783.

Ruszczynski, A. *Nonlinear optimization*. Princeton university press, 2011.

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

Shen, Z., Cervino, J., Hassani, H., and Ribeiro, A. An agnostic approach to federated learning with class imbalance. In *International Conference on Learning Representations*, 2021.

Shi, G., Shi, X., O'Connell, M., Yu, R., Azizzadenesheli, K., Anandkumar, A., Yue, Y., and Chung, S.-J. Neural lander: Stable drone landing control using learned dynamics. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 9784–9790. IEEE, 2019.

Stutz, D., Hein, M., and Schiele, B. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6976–6987, 2019.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 1999.

Wang, W. and Carreira-Perpinán, M. A. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*, 2013.

Wu, H.-T. and Wu, N. Think globally, fit locally under the manifold setup: Asymptotic analysis of locally linear embedding. *The Annals of Statistics*, 46(6B):3805–3837, 2018.

Yang, T. Advancing non-convex and constrained learning: Challenges and opportunities. *AI Matters*, 5(3):29–39, 2019.

Yang, Z., Cohen, W., and Salakhudinov, R. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pp. 40–48. PMLR, 2016.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhang, S., Huang, K., Zhu, J., and Liu, Y. Manifold adversarial training for supervised and semi-supervised learning. *Neural Networks*, 140:282–293, 2021. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2021.03.031. URL https://www.sciencedirect.com/science/article/pii/S0893608021001192.

Zhao, L. and Liu, Y. Spectral normalization for domain adaptation. *Information*, 11(2):68, 2020.

Zhu, X. J. Semi-supervised learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences*, 2005.

## A. Dual Problem Formulation

In Section 3 we introduce the optimization program in (7) as the dual of (3). Strictly speaking (7) is equivalent to the actual dual problem of (3). This follows from a ready reformulation of the dual problem as we show in the following proposition.

**Proposition 4** *The optimization program in* (7) *is equivalent to the Lagrangian dual of* (3)*.*

**Proof:** The result is true because the dual problem is linear in $\rho$. To see this, recall that $\mu$ is the dual variable associated with the statistical constraint $\mathbb{E}_{p(x,y)}[\ell\left(f_\theta(x), y\right)] \leq \epsilon$ and that $\lambda(z)$ is the Lagrange multiplier distribution associated with the set of pointwise constraints $\|\nabla_{\mathcal{M}} f_\theta(z)\|^2 \leq \rho$. The Lagrangian $\tilde{\mathcal{L}}(\rho, \theta, \mu, \lambda)$ of (3) is therefore written as

$$\tilde{\mathcal{L}}(\theta, \rho, \mu, \lambda) = \rho + \mu\bigg(\mathbb{E}[\ell\left(f_\theta(x), y\right)] - \epsilon\bigg) + \int_{\mathcal{M}} \lambda(z)\bigg(\|\nabla f_\theta(z)\|^2 - \rho\bigg) p(z) dV(z) \tag{21}$$

Reorder terms in (21) to group the two summands that involve $\rho$ to write

$$\begin{aligned}\tilde{\mathcal{L}}(\theta, \rho, \mu, \lambda) =& \rho\bigg(1 - \int \lambda(z) p(z) dV(z) dz\bigg) + \mu\bigg(\mathbb{E}_{p(x,y)}[\ell\left(f_\theta(x), y\right)] - \epsilon\bigg) \\ &+ \int \lambda(z) \|\nabla_{\mathcal{M}} f_\theta(z)\|^2 p(z) dV(z).\end{aligned} \tag{22}$$

An important observation to make is that the Lagrangian decomposes in a term that involves only $\rho$ and a term that involves only $f_\theta$. Define then

$$\begin{aligned}\tilde{\mathcal{L}}_1(\rho, \lambda) \quad &:= \rho\bigg(1 - \int \lambda(z) p(z) dV(z) dz\bigg), \\ \tilde{\mathcal{L}}_2(\theta, \mu, \lambda) \; &:= \mu\bigg(\mathbb{E}_{p(x,y)}[\ell\left(f_\theta(x), y\right)] - \epsilon\bigg) + \int \lambda(z) \|\nabla_{\mathcal{M}} f_\theta(z)\|^2 p(z) dV(z),\end{aligned} \tag{23}$$

so that we can write the Lagrangian in (22) as $\tilde{\mathcal{L}}(\theta, \rho, \mu, \lambda) = \tilde{\mathcal{L}}_1(\rho, \lambda) + \tilde{\mathcal{L}}_2(\theta, \mu, \lambda)$.

The dual problem can now be written as the maximization over multipliers of the minimum of the Lagrangian over primal variables

$$\begin{aligned}\tilde{D}^* &= \max_{\mu, \lambda} \min_{\theta, \rho} \; \tilde{\mathcal{L}}(\rho, \theta, \mu, \lambda) \\ &= \max_{\mu, \lambda} \bigg[\min_{\rho} \; \tilde{\mathcal{L}}_1(\rho, \mu, \lambda) + \min_{\theta} \; \tilde{\mathcal{L}}_2(\theta, \mu, \lambda)\bigg].\end{aligned} \tag{24}$$

where we utilized the decomposition of the Lagrangian to write the second equality.

The important observation to make is that the minimization over $\rho$ of $\tilde{\mathcal{L}}_1(\rho, \mu, \lambda)$ has an elementary solution. Indeed, as per its definition we have

$$\min_{\rho} \; \tilde{\mathcal{L}}_1(\rho, \mu, \lambda) = \min_{\rho} \; \rho\bigg(1 - \int \lambda(z) p(z) dV(z) dz\bigg). \tag{25}$$

This minimization yields $-\infty$ when $\int \lambda(z) p(z) dV(z) dz \neq 1$ and 0 when $\int \lambda(z) p(z) dV(z) dz = 1$. Since in the dual problem we are interested in the maximum over all dual variables, we know that: (i) The maximum will be attained for a dual distribution that satisfies $\int \lambda(z) p(z) dV(z) dz = 1$. (ii) When dual variables satisfy this property we know that $\min_\rho \; \rho(1 - \int \lambda(z) p(z) dV(z) dz) = 0$. It then follows that the dual problem in (24) is equivalent to

$$\begin{aligned}\tilde{D}^* = \max_{\mu, \lambda} \quad & \min_{\theta} \; \tilde{\mathcal{L}}_2(\theta, \mu, \lambda), \\ \text{subject to} \quad & \int_{\mathcal{M}} \lambda(z) p(z) dV(z) = 1.\end{aligned} \tag{26}$$

This is the problem in (7) given the definition of $\tilde{\mathcal{L}}_2(\theta, \mu, \lambda)$ in (23) which is the same as the definition of $L(\theta, \mu, \lambda)$ in (7). ■

□

Notice that the constraint $\int_{\mathcal{M}} \lambda(z)p(z)dV(z) = 1$ implies that $\lambda(z)$ is a probability distribution over the manifold $\mathcal{M}$. This is an important observation for the connections we establish to manifold regularization in Section 3.2. It also implies that even though the dual problem 7 is not an unconstrained problem, the constraint is easy to enforce as an orthogonal projection on the space of probability distributions over the manifold $\mathcal{M}$. This is a simple normalization.

## B. Empirical Dual Problem with unlabeled samples

In Section 3 we introduce the empirical dual program in (9) in the case in which unlabeled samples are unavailable. The ability to leverage unlabeled samples is an important feature of this work and ready to incorporate in (9). Indeed, if in addition to $N$ i.i.d. labeled samples $(x_n, y_n)$ with $N \in [1, N]$ drawn from $p(x, y)$ we are also given $N^{'}$ i.i.d. unlabeled samples $x_n$ with $n \in [N+1, N+N^{'}]$ drawn from the input distribution $p(x)$ we redefine (9) as

$$\hat{D}^{\star} = \max_{\hat{\mu},\hat{\lambda}\geq 0} \min_{\theta} \hat{L}(\theta, \hat{\mu}, \hat{\lambda}) := \hat{\mu}\left(\frac{1}{N}\sum_{n=1}^{N} \ell\big(f_\theta(x_n), y_n\big) - \epsilon\right) + \frac{1}{N+N^{'}} \sum_{n=1}^{N+N^{'}} \hat{\lambda}(x_n)\|\nabla_{\mathcal{M}} f_\theta(x_n)\|^2,$$
$$\text{subject to } \frac{1}{N+N^{'}} \sum_{n=1}^{N+N^{'}} \hat{\lambda}(x_n) = 1, \tag{27}$$

Results in Section 3 hold with proper modifications.

## C. Proof of Proposition 1

The proof in this appendix is a generalization of the proof in (Chamon et al., 2023). In order to show Proposition 1, we need to introduce an auxiliary problem formulation over a functional domain of functions. In this case, we take the optimization problem (3) over the convex hull of the domain of parametric functions $\phi \in \bar{\mathcal{H}}$,

$$\begin{aligned} \tilde{P}^*_\alpha = \quad & \min_{\phi\in\bar{\mathcal{H}},\rho\geq 0} \quad \rho, \\ & \text{subject to} \quad \mathbb{E}_{p(x,y)}[\ell\left(\phi(x), y\right)] \leq \epsilon - M\alpha, \\ & \qquad\qquad \|\nabla_{\mathcal{M}}\phi(z)\|^2 \leq \rho, \quad p(z)\text{-a.e.}, \quad z \in \mathcal{M}. \end{aligned} \tag{28}$$

We can now show that problem (28) with $\alpha = \nu$ is strongly dual as follows,

**Lemma 1** *Under the assumptions of Proposition 1, the functional smooth learning problem* (28) *with $\alpha = \nu$ has zero duality gap, i.e. $\tilde{P}^*_\nu = \tilde{D}^*_\nu$.*

**Proof Lemma 1:** To begin with, the non-parametric space $\bar{\mathcal{H}}$ is convex as it is the convex hull of the space of parametric functions $\mathcal{H}$, therefore the domain of the optimization problem is a convex set. Note that the objective is linear. Given that the gradient of $\phi$ is a linear function of $\phi$, and that by taking the norm we preserve convexity, the constraint $\|\nabla_{\mathcal{M}}\phi(z)\|^2 \leq \rho$ is convex. By Assumption 1, the loss $\ell$ is convex. Therefore, (3) is a semi-infinite convex problem. Moreover, since $\tilde{\theta}$ from Assumption 5 belongs to the relative interior of the feasible set, and $\mathcal{H} \subset \bar{\mathcal{H}}$, it suffices to take $\phi^\dagger(\cdot) = f(\tilde{\theta}, \cdot)$, and $\rho^\dagger > \sup_{z\in\mathcal{M}} ||\nabla_{\mathcal{M}} f_{\tilde{\theta}}(z)||^2$, and by Slater's condition as $\phi^\dagger, \rho^\dagger$ belongs to the interior of the feasible domain, problem (28) has zero duality gap, i.e. it is strongly dual.

Now we need to define the *supergradient* of function $d(\mu, \lambda)$.

**Definition 3 (Supergradient and Superdifferential)** *We say that $c \in \mathbb{R}$ is a supergradient of $d$ at $\mu$ if,*

$$d(\mu^{'}) \leq d(\mu) + c(\mu^{'} - \mu) \text{ for all } \mu^{'} \in \mathbb{R}. \tag{29}$$

*The set of all supergradients of $d$ at $\mu$ is called the superdifferential, and we denote it $\partial d(\mu)$.*

**Lemma 2 (Danskin's Theorem)** *Consider the function,*

$$F(x) = \sup_{y\in Y} f(x, y). \tag{30}$$

*where $f : \mathbb{R}^n \times Y \to \mathbb{R} \cup \{-\infty, +\infty\}$, if the following conditions are satisfied*

1. *Function $f(\cdot, y)$ is convex for all $y \in Y$.*
2. *Function $f(x, \cdot)$ is upper semicontinuous for all $x$ in a certain neighborhood of a point $x_0$.*
3. *The set $Y \subset \mathbb{R}^m$ is compact*

*Then,*

$$\partial F(x_0) = conv\left(\bigcup_{y \in \hat{Y}(x_0)} \partial_x f(x_0, y)\right) \tag{31}$$

*where $\partial_x f(x_0, y)$ denotes the subdifferential of the function $f(\cdot, y)$ at $x_0$*

**Proof of Lemma 2:** The proof can be found in (Ruszczynski, 2011)[Theorem 2.87].

**Lemma 3** *Let $\mu^*, \lambda^*$ be a solution of dual problem $D^*$ (cf. (7)). Under the conditions of Proposition 1, there exists a feasible $\theta^\dagger \in \operatorname{argmin}_\theta \mathcal{L}(\theta, \mu^*, \lambda^*)$, such that the value $D^*$ is bounded by,*

$$P^* - \nu(\mu^*_\nu M - 2\sqrt{\tilde{P}^*_\nu} - \nu) \leq D^* \leq P^* \tag{32}$$

*where $\tilde{P}^*_\nu$ and $\mu^*_\nu$ are the optimal value, and optimal dual variable of the functional version of problem (28) with constraint $\alpha = \nu$.*

**Proof:** First, we need to show that $P^*$ is finite, which is equivalent to showing that there is a feasible $\theta \in \Theta$ for problem (3), which is verified by Assumption 4.

Now, we need to verify that there exists a Lagrangian minimizer $\theta^\dagger \in \Theta^\dagger(\mu^*, \lambda^*)$ that is feasible. To do so, We begin by considering the set of Lagrange minimizers $\Theta^\dagger(\mu^*, \lambda^*)$ as follows,

$$\Theta^\dagger(\mu^*, \lambda^*) = \operatorname*{argmin}_\theta L(\theta, \mu^*, \lambda^*) \tag{33}$$

we can also define the constraint slack associated with parameters $\theta$ as follows,

$$c(\theta) = [\mathbb{E}[\ell(f_\theta(x), y)] - \epsilon]_+. \tag{34}$$

Therefore, to show that there exists a feasible solution, it is analogous to show that there is an element $\theta^\dagger$ of $\Theta^\dagger(\check{\mu}^*, \lambda^*)$ whose slack is equal to zero, i.e. $c(\theta^\dagger) = 0$. To show that this element exists, by contradiction, we can say that if no element of $\Theta(\mu^*, \lambda^*)$ is feasible, then $\mathbb{E}[\ell(f_\theta(x), y)] - \epsilon > 0$, for all $\theta \in \Theta(\mu^*, \lambda^*)$. From Lemma 2, we get that $\mathbf{0} \notin \partial\Theta^\dagger(\mu^*, \lambda^*)$, which contradicts the optimality of $\mu^*, \lambda^*$. Given that the dual variable $\mu^*$ is finite, considering the existence of a feasible solution by Assumption 2. Hence, there must be at least one element of $\Theta(\mu^*, \lambda^*)$ that is feasible.

Now we need to show that the inequality (32) holds. The upper bound is trivially verified by weak duality (Boyd et al., 2004), i.e.,

$$D^* \leq P^* \tag{35}$$

Now note that problem (28) with $\alpha = 0$ is strongly dual by Lemma 1, i.e.,

$$\tilde{P}^*_\nu = \min_\phi \operatorname*{maximize}_{\mu, \lambda \geq 0} \tilde{L}_\nu(\phi, \mu, \lambda) = \operatorname*{maximize}_{\mu, \lambda \geq 0} \min_\phi \tilde{L}_\nu(\phi, \mu, \lambda) = \tilde{D}^*_\nu \tag{36}$$

with the Lagrangian defined as,

$$\begin{aligned} \tilde{L}_\nu(\phi, \mu, \lambda) = \mu\Big(\mathbb{E}[\ell\big(\phi(x), y\big)] - (\epsilon - M\nu)\Big) + \int_{\mathcal{M}} \lambda(x) \|\nabla_{\mathcal{M}} \phi(x)\|^2 p(x) dV(x)), \\ \text{for all } \lambda \text{ that satisfy } \int_{\mathcal{M}} \lambda(x) p(x) dV(x) = 1, \end{aligned} \tag{37}$$

Coming back to the parametric dual problem (7), by optimality we know that,

$$D^* \geq \min_{\theta\in\Theta} L(\theta, \mu, \lambda), \text{for all } \mu, \lambda. \tag{38}$$

We thus utilize the optimal dual variables of the functional problem with constraints $\epsilon - M\nu$, i.e., $\mu^*_\nu, \lambda^*_\nu$, as follows,

$$\begin{aligned} D^* &\geq \min_{\theta\in\Theta} L(\theta, \mu^*_\nu, \lambda^*_\nu) && (39)\\ &\geq \min_{\phi\in\bar{\mathcal{H}}} L(\phi, \mu^*_\nu, \lambda^*_\nu) && (40)\\ &\geq \min_{\phi\in\bar{\mathcal{H}}} \tilde{L}_\nu(\phi, \mu^*_\nu, \lambda^*_\nu) - \mu^*_\nu M\nu && (41)\\ &= \tilde{P}^*_\nu - \mu^*_\nu M\nu. && (42)\end{aligned}$$

Equation (39) holds given that $\mathcal{H} \subseteq \bar{\mathcal{H}}$, and problem $P^*_\nu$ is strongly dual by Lemma 1. To complete the proof we need to show that,

$$\tilde{P}^*_\nu \geq P^* - 2\nu\sqrt{\tilde{P}^*_\nu} - \nu^2. \tag{43}$$

Denoting $\phi^*_\nu$ a solution to problem $P^*_\nu$, by Assumption 2, we know there is a parameterization $\tilde{\theta}^*_\nu$ such that simultaneously,

$$\sup_{z\in\mathcal{M}} |\phi^*_\nu(z) - f_{\tilde{\theta}^*_\nu}(z)| \leq \nu \tag{44}$$

$$\sup_{z\in\mathcal{M}} \|\nabla_{\mathcal{M}}\phi^*_\nu(z) - \nabla_{\mathcal{M}} f_{\tilde{\theta}^*_\nu}(z)\| \leq \nu \tag{45}$$

therefore,

$$\begin{aligned} \left|\mathbb{E}[\ell(\phi^*_\nu(x), y)] - \mathbb{E}[\ell(f_{\tilde{\theta}^*_\nu}(x), y)]\right| &\leq \mathbb{E}\left[|\ell(\phi^*_\nu(x), y)] - \ell(f_{\tilde{\theta}^*_\nu}(x), y)|\right] && (46)\\ &\leq M\mathbb{E}\left[|\phi^*_\nu(x) - f_{\tilde{\theta}^*_\nu}(x)|\right] \leq M\nu && (47)\end{aligned}$$

Since $\phi^*_\nu$ is feasible for $P_\nu$, this implies that $f_{\tilde{\theta}^*_\nu}$ is feasible for Problem 3. We can now define the $\rho^*_{\tilde{\theta}^*_\nu}$ as the minimum $\rho$ obtained with $f_{\tilde{\theta}^*_\nu}$ as follows,

$$\begin{aligned} P^*_\nu = \quad & \min_{\rho\geq 0} \quad \rho, && (48)\\ \text{subject to} \quad & \mathbb{E}_{p(x,y)}[\ell\left(f_{\tilde{\theta}^*_\nu}(x), y\right)] \leq \epsilon,\\ & \|\nabla_{\mathcal{M}} f_{\tilde{\theta}^*_\nu}(z)\|^2 \leq \rho, \quad p(z)\text{-a.e.}, \quad z \in \mathcal{M}.\end{aligned}$$

Returning to (39), and by optimality, $P^* \leq P^*_\nu$,

$$\begin{aligned} D^* &\geq \tilde{P}^*_\nu - \mu^*_\nu M\nu && (49)\\ &\geq \tilde{P}^*_\nu + P^* - P^*_\nu - \mu^*_\nu M\nu && (50)\end{aligned}$$

Now we need to bound the difference $\tilde{P}^*_\nu - P^*_\nu$. By compactness of $\mathcal{M}$, we know that there exist $z_1, z_2 \in \mathcal{M}$ such that $\|\nabla_{\mathcal{M}}\phi^*_\nu(z_1)\|^2 = \tilde{P}^*_\nu$ and $\|\nabla_{\mathcal{M}} f_{\tilde{\theta}^*_\nu}(z_2)\|^2 = P^*_\nu$, by optimality,

$$\begin{aligned} \tilde{P}^*_\nu - P^*_\nu &= \|\nabla_{\mathcal{M}}\phi^*_\nu(z_1)\|^2 - \|\nabla_{\mathcal{M}} f_{\tilde{\theta}^*_\nu}(z_2)\|^2 && (51)\\ &\geq \|\nabla_{\mathcal{M}}\phi^*_\nu(z_2)\|^2 - \|\nabla_{\mathcal{M}} f_{\tilde{\theta}^*_\nu}(z_2)\|^2 && (52)\end{aligned}$$

Now we can add a subtract $\nabla_{\mathcal{M}}\phi^*_\nu(z_2)$ to the second term, and use the triangle inequality twice to obtain

$$\begin{aligned} \|\nabla_{\mathcal{M}}\phi^*_\nu(z_2)\|^2 - \|\nabla_{\mathcal{M}} f_{\tilde{\theta}^*_\nu}(z_2)\|^2 &\geq \|\nabla_{\mathcal{M}}\phi^*_\nu(z_2)\|^2 - (\|\nabla_{\mathcal{M}} f_{\tilde{\theta}^*_\nu}(z_2) - \nabla_{\mathcal{M}}\phi^*_\nu(z_2)\| + \|\nabla_{\mathcal{M}}\phi^*_\nu(z_2)\|)\\ &\quad (\|\nabla_{\mathcal{M}} f_{\tilde{\theta}^*_\nu}(z_2) - \nabla_{\mathcal{M}}\phi^*_\nu(z_2)\| + \|\nabla_{\mathcal{M}}\phi^*_\nu(z_2)\|) && (53)\\ &= -\|\nabla_{\mathcal{M}} f_{\tilde{\theta}^*_\nu}(z_2) - \nabla_{\mathcal{M}}\phi^*_\nu(z_2)\|^2 - 2\|\nabla_{\mathcal{M}} f_{\tilde{\theta}^*_\nu}(z_2) - \nabla_{\mathcal{M}}\phi^*_\nu(z_2)\|\|\nabla_{\mathcal{M}}\phi^*_\nu(z_2)\| && (54)\end{aligned}$$

Now, by noting that $\|\nabla_{\mathcal{M}} f_{\tilde{\theta}^*_\nu}(z_2) - \nabla_{\mathcal{M}} \phi^*_\nu(z_2)\|^2 \leq \nu^2$ by construction of $\tilde{\theta}^*_\nu$, and that $\|\nabla_{\mathcal{M}} \phi^*_\nu(z_2)\| \leq \sqrt{\tilde{P}^*_\nu}$ by optimality, we end up obtaining,

$$\tilde{P}^*_\nu - P^*_\nu \geq -2\nu\sqrt{\tilde{P}^*_\nu} - \nu^2 \tag{55}$$

Putting (50) and (55) together, we attain the desired result.

**Proof of Proposition 1:**

This proof follows the lines of (Chamon et al., 2023)[Proposition III.4]. We begin by considering $\mu^*, \lambda^*$, and $\hat{\mu}^*, \hat{\lambda}^*$, solutions of (7), and (9), and we define the set of optimal dual minimizers as,

$$\Theta(\mu^*, \lambda^*) = \operatorname*{argmin}_{\theta \in \Theta} L(\theta, \mu^*, \lambda^*) \tag{56}$$

$$\hat{\Theta}(\hat{\mu}^*, \hat{\lambda}^*) = \operatorname*{argmin}_{\theta \in \Theta} \hat{L}(\theta, \hat{\mu}^*, \hat{\lambda}^*) \tag{57}$$

where $L$, and $\hat{L}$ are as defined in 7, and 9. We can proceed to bound the difference between the values of the dual problems as follows,

$$D^* - \hat{D}^* = \min_{\theta \in \Theta} L(\theta, \mu^*, \lambda^*) - \min_{\theta \in \Theta} \hat{L}(\theta, \hat{\mu}^*, \hat{\lambda}^*) \tag{58}$$

$$\leq \min_{\theta \in \Theta} L(\theta, \mu^*, \lambda^*) - \min_{\theta \in \Theta} \hat{L}(\theta, \mu^*, \bar{\lambda}^*) \quad \text{(by optimality)} \tag{59}$$

$$\leq L(\hat{\theta}^\dagger, \mu^*, \lambda^*) - \hat{L}(\hat{\theta}^\dagger, \mu^*, \bar{\lambda}^*) \tag{60}$$

where $[\bar{\lambda}^*(z)]_n = \lambda^*(z_n) / \sum_{n=1}^N \lambda^*(z_n)$, and we define $\hat{\theta}^\dagger \in \hat{\Theta}(\mu^*, \bar{\lambda}^*)$. By utilizing the definition of the Lagrangian, we obtain,

$$D^* - \hat{D}^* \leq |\mu^*| \left| \mathbb{E}[\ell(f_{\hat{\theta}^\dagger}(x), y)] - \frac{1}{N} \sum_{i=1}^N \ell(f_{\hat{\theta}^\dagger}(x_n), y_n) \right| \tag{61}$$

$$+ \left| \mathbb{E}_{z \sim \lambda^*}[\|\nabla_{\mathcal{M}} f_{\hat{\theta}^\dagger}(z)\|^2] - \frac{1}{N} \sum_{n=1}^N \bar{\lambda}^*(z_n) \|\nabla_{\mathcal{M}} f_{\hat{\theta}^\dagger}(z_n)\|^2 \right| \tag{62}$$

Note that the sampled dual variables converge to the continuous values as follows,

$$\lim_{N \to \infty} \frac{\lambda^*(z)}{\frac{1}{N} \sum_{n=1}^N \lambda^*(z_n)} = \lambda^*(z) \tag{63}$$

Given that $\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^N \lambda^*(z_n) = 1$.

Utilizing the same argument on the other direction of the inequality it yields,

$$D^* - \hat{D}^* \geq L(\hat{\theta}^\dagger, \hat{\mu}^*, \hat{\lambda}^*) - \hat{L}(\hat{\theta}^\dagger, \hat{\mu}^*, \hat{\lambda}^*) \tag{64}$$

$$\geq |\hat{\mu}^*| \left| \mathbb{E}[\ell(f_{\hat{\theta}^\dagger}(x), y)] - \frac{1}{N} \sum_{i=1}^N \ell(f_{\hat{\theta}^\dagger}(x_n), y_n) \right| \tag{65}$$

$$+ \left| \mathbb{E}_{z \sim \bar{\bar{\lambda}}^*}[\|\nabla_{\mathcal{M}} f_{\hat{\theta}^\dagger}(z)\|^2] - \frac{1}{N} \sum_{n=1}^N \hat{\lambda}^*(z_n) \|\nabla_{\mathcal{M}} f_{\hat{\theta}^\dagger}(z_n)\|^2 \right| \tag{66}$$

Where $\bar{\bar{\lambda}}^* = \frac{1}{N} \sum_{n=1}^N \hat{\lambda}^*(x_n) B_\alpha(x_n)$, and $B_\alpha(x_n)$ is a ball of center $x_n$, and radius $\alpha$. By tending $\alpha \to 0$, and $N \to \infty$, given that $\bar{\bar{\lambda}}^*$ is a simple function that approximates $\hat{\lambda}^\star$, by the Monotone Convergence Theorem (Durrett, 2019), $\bar{\bar{\lambda}}^*$ converges to $\hat{\lambda}^\star$. Therefore, it holds,

$$\left| D^* - \hat{D}^* \right| \leq \max \left\{ |L(\hat{\theta}^\dagger, \mu^*, \lambda^*) - \hat{L}(\hat{\theta}^\dagger, \mu^*, \bar{\lambda}^*)|, |L(\hat{\theta}^\dagger, \hat{\mu}^*, \hat{\lambda}^*) - \hat{L}(\hat{\theta}^\dagger, \hat{\mu}^*, \hat{\lambda}^*)| \right\} \tag{67}$$

Now we utilize the property of the gradients of $\nabla_{\mathcal{M}} f_\theta$ given by Assumption 2, and the fact that the set $\Theta$ is compact, and the manifold $\mathcal{M}$ is compact,

$$\left| \|\nabla_{\mathcal{M}} f_{\theta_1}(x)\|^2 - \|\nabla_{\mathcal{M}} f_{\theta_2}(x)\|^2 \right| \tag{68}$$

$$= \left| \|\nabla_{\mathcal{M}} f_{\theta_1}(x)\|^2 - \|\nabla_{\mathcal{M}} f_{\theta_2}(x) + \nabla_{\mathcal{M}} f_{\theta_1}(x) - \nabla_{\mathcal{M}} f_{\theta_1}(x)\|^2 \right| \tag{69}$$

$$= \left| \|\nabla_{\mathcal{M}} f_{\theta_1}(x)\|^2 - \|\nabla_{\mathcal{M}} f_{\theta_2}(x) - \nabla_{\mathcal{M}} f_{\theta_1}(x)\|^2 - ||\nabla_{\mathcal{M}} f_{\theta_1}(x)\|^2 + 2\|\nabla_{\mathcal{M}} f_{\theta_2}(x) - \nabla_{\mathcal{M}} f_{\theta_1}(x)\|\|\nabla_{\mathcal{M}} f_{\theta_1}(x)\| \right| \tag{70}$$

$$= \left| - \|\nabla_{\mathcal{M}} f_{\theta_2}(x) - \nabla_{\mathcal{M}} f_{\theta_1}(x)\|^2 + 2\|\nabla_{\mathcal{M}} f_{\theta_2}(x) - \nabla_{\mathcal{M}} f_{\theta_1}(x)\|\|\nabla_{\mathcal{M}} f_{\theta_1}(x)\| \right| \tag{71}$$

We can now bound $\|\nabla_{\mathcal{M}} f_{\theta_2}(x) - \nabla_{\mathcal{M}} f_{\theta_1}(x)\|$, and $2\|\nabla_{\mathcal{M}} f_{\theta_1}(x)\|$ by $2 \max_{\theta \in \{\theta_1, \theta_2\}, z \in \mathcal{M}} \|\nabla_{\mathcal{M}} f_\theta(z)\|$, and therefore, we obtain,

$$\left| \|\nabla_{\mathcal{M}} f_{\theta_1}(x)\|^2 - \|\nabla_{\mathcal{M}} f_{\theta_2}(x)\|^2 \right| \tag{72}$$

$$\leq 2 \max_{\theta \in \{\theta_1, \theta_2\}, z \in \mathcal{M}} \|\nabla_{\mathcal{M}} f_\theta(z)\| \|\nabla_{\mathcal{M}} f_{\theta_1}(x) - \nabla_{\mathcal{M}} f_{\theta_2}(x)\| \tag{73}$$

We can now use Assumption 3 to obtain,

$$\left| \|\nabla_{\mathcal{M}} f_{\theta_1}(x)\|^2 - \|\nabla_{\mathcal{M}} f_{\theta_2}(x)\|^2 \right| \tag{74}$$

$$\leq 2 \max_{\theta \in \{\theta_1, \theta_2\}, z \in \mathcal{M}} \|\nabla_{\mathcal{M}} f_\theta(z)\| G |f_{\theta_1}(x) - f_{\theta_2}(x)| \text{ (by compactness and Assumption 3.)} \tag{75}$$

To complete the proof, we leverage Talagrand's lemma (Mohri et al., 2018)[Lemma 5.7], and with probability $1 - 2\delta$,

$$\left| D^* - \hat{D}^* \right| \leq \max \left\{ |\mu^*|, |\hat{\mu}^*| \right\} \zeta(N, \delta) + 2 \max \left\{ \sqrt{P^*}, \sqrt{\tilde{P}^*} \right\} G \hat{\zeta}(N, \delta) \tag{76}$$

By leveraging Lemma 3 we attain the desired result.

## D. Proof Of Proposition 2

**Proof of Proposition 2:**

To prove that $\hat{\theta}^*$ is approximately feasible, we use the same argument that in Lemma 3. By contradiction, if there exists no $\hat{\theta}^\dagger \in \hat{\Theta}(\hat{\mu}^*, \hat{\lambda}^*)$ that is feasible, the supergradient,

$$\frac{1}{N}\sum_{n=1}^{N} \ell(f_{\hat{\theta}^\dagger}(x_n), y_n) > \epsilon \tag{77}$$

and therefore $\mathbf{0} \notin \partial\hat{\Theta}(\hat{\theta}^*, \hat{\lambda}^*)$ which contradicts the optimality of $\hat{\theta}^*$, and $\hat{\lambda}^*$. Therefore, there exists $\hat{\theta}^* \in \hat{\Theta}(\hat{\mu}^*, \hat{\lambda}^*)$ such that,

$$\mathbb{E}[\ell(f_{\hat{\theta}^\dagger}(x), y)] \leq \epsilon + \zeta(N, \delta) \tag{78}$$

with probability $1-\delta$. We can analyze the term given by the summation of the dual variables $\lambda$, noting that if there is no solution $\hat{\lambda}^*$, such that

$$\frac{1}{N}\sum_{n=1}^{N} \lambda(x_n)\|\nabla_{\mathcal{M}} f_\theta(x_n)\|^2 = \max_{n\in[N]} \|\nabla f_\theta(x_n)\|^2 \tag{79}$$

then, utilizing the same argument, $\mathbf{0} \notin \partial\hat{\Theta}(\hat{\theta}^*, \hat{\lambda}^*)$, contradicting the optimality of $\hat{\theta}^*$.

Now, we can evaluate $\hat{D}^*$ at $\hat{\theta}^*$ as follows,

$$\hat{D}^* = \mu^*\big(\sum_{n=1}^{N} \ell(f_{\hat{\theta}^*}(x_n), y_n) - \epsilon\big) + \frac{1}{N}\sum_{n=1}^{N} \hat{\lambda}^*(x_n)\|\nabla_{\mathcal{M}} f_{\hat{\theta}^*}(x_n)\|^2 \tag{80}$$

$$= \mu^*\big(\sum_{n=1}^{N} \ell(f_{\hat{\theta}^*}(x_n), y_n) - \epsilon\big) + \max_{n\in[N]} \|\nabla_{\mathcal{M}} f_{\hat{\theta}^*}(x_n)\|^2 \text{ (by optimality of } \hat{\lambda}^*) \tag{81}$$

To conclude the proof, we leverage Proposition 1, in conjunction with (81) and we get that,

$$\begin{aligned} |P^* - \max_{n\in[N]} \|\nabla_{\mathcal{M}} f_{\hat{\theta}^*}(x_n)||^2| \leq & |\hat{\mu}^*|\Big|\sum_{n=1}^{N} \ell(f_{\hat{\theta}^*}(x_n), y_n) - \epsilon\Big| \\ & + \mathcal{O}(\nu) + \max\Big\{|\mu^*|, |\hat{\mu}^*|\Big\}\zeta(N,\delta) + 2\max\Big\{\sqrt{P^*}, \sqrt{\tilde{P}^*}\Big\}G\hat{\zeta}(N,\delta) \end{aligned} \tag{82}$$

Completing the proof.

**Corollary 1** *Under the same conditions of Proposition 2, if additionally, the gradients of the parameterized functions satisfy,*

$$||\nabla_{\mathcal{M}} f_\theta(z_1) - \nabla_{\mathcal{M}} f_\theta(z_2)|| \leq L d_{\mathcal{M}}(z_1, z_2), \tag{83}$$

*and for any dataset sampled according to p the following holds,*

$$\max_{z\in\mathcal{M}} \min_{n\in[N]} d_{\mathcal{M}}(z, x_n) \leq \tilde{\zeta}(N). \tag{84}$$

*Then,*

$$\max_{z\in\mathcal{M}} \|\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(z)\|^2 \leq P^* + \mathcal{O}(\nu) + (1+\Delta)\zeta(N,\delta) + \mathcal{O}(\hat{\zeta}(N,\delta)) \tag{85}$$

$$+ 2L\sqrt{\hat{P}}\tilde{\zeta}(N) + L^2\tilde{\zeta}^2(N) + \hat{\mu}^*\left|\frac{1}{N}\sum_{n=1}^{N} \ell(f_{\hat{\theta}^*}(x_n), y_n) - \epsilon\right|$$

$$\textit{and } \mathbb{E}[\ell(f_{\hat{\theta}^\star}(x), y)] \leq \epsilon + \zeta(N,\delta). \tag{86}$$

*where* $\sqrt{\hat{P}^*} = \max_{n\in[N]} ||\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n)||$.

**Proof of Corollary 1:** Building from Proposition 2, we need to relate the samples $x_n$ to the whole manifold $\mathcal{M}$. To do so, we start from the tautological equality that holds for all $z \in \mathcal{M}$,

$$\begin{aligned}
\|\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(z)\| &= \|\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(z) + \nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n) - \nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n)\| && (87)\\
&\leq ||\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n)\| + \|\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(z) - \nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n)|| && (88)\\
&\leq ||\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n)\| + L d_{\mathcal{M}}(z, x_n) && (89)\\
&\leq ||\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(\tilde{x}_n)\| + L \min_{n\in[N]} d_{\mathcal{M}}(z, x_n) && (90)\\
&\leq \max_{x\in[N]} ||\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n)\| + L \min_{n\in[N]} d_{\mathcal{M}}(z, x_n) && (91)\\
&\leq \max_{x\in[N]} ||\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n)\| + L \max_{z\in\mathcal{M}} \min_{x_n} d_{\mathcal{M}}(z, x_n) && (92)\\
&\leq \max_{x\in[N]} ||\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n)\| + L\hat{\zeta}(N) && (93)
\end{aligned}$$

where equation (88) holds by triangle inequality, and equation (89) by assumption 83. Given that equation (89) holds for all $x_n$, it also holds for $\tilde{x}_n = \arg\min_{n\in[N]} d_{\mathcal{M}}(z, x_n)$ i.e. equation (90). Equation (91) holds given that $||\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n)\| \leq \max_{x\in[N]} ||\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n)\|$, and equation (92) holds by taking the maximum. Finally, equation (93) holds by assumption (84). We can now take squares as follows,

$$\begin{aligned}
\|\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(z)\|^2 &\leq \left(\max_{x\in[N]} ||\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n)\| + L\hat{\zeta}(N)\right)^2 && (94)\\
&\leq \max_{n\in[N]} ||\nabla_{\mathcal{M}} f_{\hat{\theta}^\star}(x_n)\|^2 + L^2\hat{\zeta}(N)^2 + 2L\sqrt{\hat{P}^*}\zeta(N). && (95)
\end{aligned}$$

were we used that the norm is non-negative, and the definition of $\sqrt{\hat{P}^*}$. By substituting equation (95) in equation (82), we complete the proof.

## E. Proof of Proposition 3

The proof of Proposition 3 utilizes the following Definitions. In this proof, we use the fact that the total volume of the smooth Manifold $\mathcal{M}$ is bounded given its compactness ,i.e., $\int_{\mathcal{M}} vol(x) = \mathbf{V}$. We also use the fact that the functions $f_\theta(z)$ are bounded up to the third derivative, and we denote $\mathbf{F}$, $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_3$ their bounds respectively. We also require to define the degree probability $d_t(z)$ and the continuous heat kearnel Laplacian $\tilde{\mathbf{L}}^t_\lambda f_\theta(z)$.

**Definition 4** *Given a probability distribution $\lambda \in \Lambda$ defined over the manifold $\mathcal{M}$, we define the degree $d_t(x)$ as,*

$$d_t(p) = \int_{\mathcal{M}} G_t(p, z)\lambda(z)vol(z) \tag{96}$$

**Definition 5** *We define the continuous heat kernel laplacian $\tilde{\mathbf{L}}^t_\lambda f_\theta(z)$ associated with density function $\lambda \in \Lambda$, function $f_\theta, \theta \in \Theta$ and point $z \in \mathcal{M}$ as,*

$$\tilde{\mathbf{L}}^t_\lambda f_\theta(z) = \frac{1}{t}\int_{\mathcal{M}} \frac{G_t(z, w)}{\sqrt{d_t(z)}\sqrt{d_t(w)}}(f_\theta(z) - f_\theta(w))\lambda(w)vol(w) \tag{97}$$

*were $d_t$ is the degree function (cf. Definition 4).*

**Lemma 4** *For a given point in the interior of the manifold, i.e., $z \in int(\mathcal{M})$, and any open set $\mathcal{B} \subset \mathcal{M}, z \in \mathcal{B}$, in the same conditions as in Proposition 3, we have that*

$$\left| \int_{\mathcal{B}} e^{-\frac{\|z-w\|^2}{4t}} \lambda(w) f_\theta(w) vol(w) - \int_{\mathcal{M}} e^{-\frac{\|z-w\|^2}{4t}} \lambda(w) f_\theta(w) vol(w) \right| \leq b\mathbf{V}\mathbf{F} e^{-\frac{r^2}{4t}}, \tag{98}$$

*where $r = \inf_{x \notin \mathcal{B}, x \in \mathcal{M}} \|z - w\|$, and $b$ in $\Lambda$.*

**Proof of Lemma 4:** The proof is similar to Lemma 4.1 in (Belkin & Niyogi, 2005). The proof follows from bounding $\lambda$ by $b$ from assumption, $\|f_\theta\|$ by $\mathbf{F}$, and the volume of $\mathcal{M} - \mathcal{B}$ by the volume of $\mathcal{M}$ given by $\mathbf{V}$, and bounding $e^{-\frac{\|p-z\|^2}{4t}}$ by $e^{-\frac{d}{4t}}$. ■

**Lemma 5** *The limit of the degree function $d_t(x)$ (defined in Definition 4) when $t \to 0$ is lower bounded for probabilities $\lambda \in \Lambda$ as in Proposition 3,i.e.,*

$$\min_{p \in \mathcal{M}, \lambda \in \Lambda} \lim_{t \to 0} d_t(x) = a. \tag{99}$$

**Proof of Lemma 5:**

Now we can write down the definition of the degree function $d_t(x)$,

$$\begin{aligned}
\min_{x \in \mathcal{M}, \lambda \in \Lambda} \lim_{t \to 0} d_t(x) &= \min_{x \in \mathcal{M}, \lambda \in \Lambda} \lim_{t \to 0} \int_{\mathcal{M}} G_t(x, z)\lambda(z)dV(z) && (100)\\
&= \min_{x \in \mathcal{M}, \lambda \in \Lambda} \lim_{t \to 0} \int_{\mathcal{M}} \frac{1}{(4\pi t)^{d/2}} e^{-\frac{\|x-z\|^2}{4t}} \lambda(z)dV(z) && (101)\\
&\geq \min_{x \in \mathcal{M}} \lim_{t \to 0} a \int_{\mathcal{M}} \frac{1}{(4\pi t)^{d/2}} e^{-\frac{\|x-z\|^2}{4t}} dV(z) && (102)\\
&\geq \min_{x \in \mathcal{M}} \lim_{t \to 0} a \int_{\mathcal{M}} \frac{1}{(4\pi t)^{d/2}} e^{-\frac{(x-z)^\intercal(\frac{1}{2t}\mathbf{I})(x-z)}{2}} dV(z) && (103)\\
&= \min_{x \in \mathcal{M}} \lim_{t \to 0} a \int_{\mathcal{B}} \frac{1}{(4\pi t)^{d/2}} e^{-\frac{(x-z)^\intercal(\frac{1}{2t}\mathbf{I})(x-z)}{2}} dV(z) && (104)\\
&= \min_{x \in \mathcal{M}} \lim_{t \to 0} a \int_{\vec{\mathcal{B}}} \frac{1}{(4\pi t)^{d/2}} e^{-\frac{(v)^\intercal(\frac{1}{2t}\mathbf{I})(v)}{2}} dV(z) && (105)\\
&= \frac{a}{(4\pi t)^{d/2}}((2\pi)^d(2t)^d)^{\frac{1}{2}} = a && (106)
\end{aligned}$$

Now, we now that $\min_{z\in\mathcal{M},\lambda\in\Lambda}\lambda(z)=a$, which means that,

$$\min_{z\in\mathcal{M},\lambda\in\Lambda}\lim_{t\to 0}\lambda(z)\leq a \tag{107}$$

Noting that the integral of $\lim_{t\to 0}\int_{\mathcal{M}}G_t(x,z)dV(z)=1$, we have,

$$\min_{z\in\mathcal{M},\lambda\in\Lambda}\lim_{t\to 0}\lambda(z)\int_{\mathcal{M}}G_t(x,z)dV(z)\leq a \tag{108}$$

$$\min_{z\in\mathcal{M},\lambda\in\Lambda}\lim_{t\to 0}\int_{\mathcal{M}}\lambda(z)G_t(x,z)dV(z)\leq a \tag{109}$$

$$\min_{z\in\mathcal{M},\lambda\in\Lambda}\lim_{t\to 0}d_t(z)\leq a \tag{110}$$

By symmetry of $G_t$, we recover the desired result. Noting that both $a\geq d_t$, and $d_t\geq a$, then the limit holds and is equal to $a$.

**Lemma 6** *For any two sufficiently close points $p,q\in\mathcal{M}$, such that $q=exp_p(v)=\mathbb{R}^k$, the relationship between the Euclidean distance and geodesic distance is given by,*

$$d^2_{\mathcal{M}}(p,q)=\|v\|^2_{\mathbb{R}^k} \tag{111}$$

**Proof of Lemma 6:** Consider a geodesic curve $\gamma$, that goes from $\gamma(0)=p$ to $\gamma(1)=q$, this geodesic can be obtained by $\gamma(t)=\exp_p(tv)$. The distance between $p,q$ can be expressed as,

$$d_{\mathcal{M}}(p,q)=\int_0^1\|\gamma^{'}(t)\|dt=\|v\|. \tag{112}$$

given that the geodesic has constant derivative, and it is given by $v$ by the definition of exponential map (see (Do Carmo & Flaherty Francis, 1992)[Proposition 3.6]). ■

**Lemma 7** *Let $\Lambda$ be the set of probability distributions defined on a compact $d$-dimensional differentiable manifold $\mathcal{M}$ isometrically embedded in $\mathbb{R}^D$ such that $\Lambda=\{\lambda:0<a\leq\lambda(z)\leq b<\infty,|\frac{\partial\lambda}{\partial x}|\leq c<\infty$ and $|\frac{\partial^2\lambda}{\partial x^2}|\leq d<\infty$ for all $z\in\mathcal{M}\}$, and let $f_\theta$, with $\theta\in\Theta$, be a family of functions with uniformly bounded derivatives up to order 3 vanishing at the boundary $\partial\mathcal{M}$. For a point $z\in\mathcal{M}$, the probability of the difference between the point-cloud Laplacian operator $\mathbf{L}^t_{\lambda,N}f_\theta(z)$ defined in definition (2) and the continuous heat kernel Laplacian is given by*

$$P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda}|\tilde{\mathbf{L}}^t_\lambda f_\theta(z)-\mathbf{L}_{\lambda,N}f_\theta(z)|\geq\epsilon\right)\leq\delta, \tag{113}$$

*where $N$ is such that $2e^{-\frac{\epsilon^2 N}{32}(\frac{at}{\mathbf{F}b})^2}+2e^{-\frac{\epsilon^2 N}{32}\left(\frac{ta^2}{2\mathbf{FG}b2}\right)^2}+2(N-1)e^{-\frac{\epsilon^2(N-1)}{32}\left(\frac{ta^2}{2\mathbf{FG}b2}\right)^2}\leq\delta$.*

**Proof of Lemma 7:** The proof is similar to section 5.1 of Theorem 5.2 from (Belkin & Niyogi, 2005). To begin with, we define a intermediate operator $\hat{\mathbf{L}}_{\lambda,n}f_\theta(z)$ as,

$$\hat{\mathbf{L}}_{\lambda,N}f_\theta(z)=\frac{1}{Nt}\sum_{i=1}^{N}\frac{G_t(z,x_i)}{\sqrt{d_t(z)}\sqrt{d_t(x_i)}}(f_\theta(z)-f_\theta(x_i)) \tag{114}$$

Now we consider equation (113), adding and subtracting the intermediate operator $\hat{\mathbf{L}}_{\lambda,n} f_\theta(z)$ as follows,

$$P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda} |\tilde{\mathbf{L}}_\lambda^t f_\theta(z) - \mathbf{L}_{\lambda,N} f_\theta(z)| \geq \epsilon\right) \tag{115}$$

$$= P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda} |\tilde{\mathbf{L}}_\lambda^t f_\theta(z) - \hat{\mathbf{L}}_{\lambda,n} f_\theta(z) + \hat{\mathbf{L}}_{\lambda,n} f_\theta(z) - \mathbf{L}_{\lambda,N} f_\theta(z)| \geq \epsilon\right) \tag{116}$$

$$= 1 - P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda} |\tilde{\mathbf{L}}_\lambda^t f_\theta(z) - \hat{\mathbf{L}}_{\lambda,n} f_\theta(z) + \hat{\mathbf{L}}_{\lambda,n} f_\theta(z) - \mathbf{L}_{\lambda,N} f_\theta(z)| \leq \epsilon\right) \tag{117}$$

$$\leq 1 - P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda} |\tilde{\mathbf{L}}_\lambda^t f_\theta(z) - \hat{\mathbf{L}}_{\lambda,n} f_\theta(z)| \leq \epsilon/2 \cap \sup_{\theta\in\Theta,z\in\mathcal{M},\lambda\in\Lambda} |\hat{\mathbf{L}}_{\lambda,n} f_\theta(z) - \mathbf{L}_{\lambda,N} f_\theta(z)| \leq \epsilon/2\right) \tag{118}$$

$$= 1 - \left(1 - P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda} |\tilde{\mathbf{L}}_\lambda^t f_\theta(z) - \hat{\mathbf{L}}_{\lambda,n} f_\theta(z)| > \epsilon/2 \cup \sup_{\theta\in\Theta,z\in\mathcal{M},\lambda\in\Lambda} |\hat{\mathbf{L}}_{\lambda,n} f_\theta(z) - \mathbf{L}_{\lambda,N} f_\theta(z)| > \epsilon/2\right)\right) \tag{119}$$

$$\leq P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda} |\tilde{\mathbf{L}}_\lambda^t f_\theta(z) - \hat{\mathbf{L}}_{\lambda,n} f_\theta(z)| > \epsilon/2\right) + P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda} |\hat{\mathbf{L}}_{\lambda,n} f_\theta(z) - \mathbf{L}_{\lambda,N} f_\theta(z)| > \epsilon/2\right) \tag{120}$$

where equation (117) holds by taking the complement, equation (118) holds given that it is a subset of the total probability, equation (119) by taking complement again, and equation (120) by the union bound.

We will now focus on the complement of the first term of (120), i.e. $P(\sup_{\theta\in\Theta,\lambda\in\Lambda} |\tilde{\mathbf{L}}_\lambda^t f_\theta(z) - \hat{\mathbf{L}}_{\lambda,N} f_\theta(z)| \leq \epsilon/2)$, as follows,

$$P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda} |\tilde{\mathbf{L}}_\lambda^t f_\theta(z) - \hat{\mathbf{L}}_{\lambda,n} f_\theta(z)| \leq \epsilon/2\right) \tag{121}$$

$$= P\Bigg(\sup_{\theta\in\Theta,\lambda\in\Lambda} \Bigg| \int_{\mathcal{M}} \frac{1}{t} \frac{G_t(z,w)}{\sqrt{d_t(z)}\sqrt{d_t(w)}} (f_\theta(z) - f_\theta(w))\lambda(w) dV(w)$$

$$- \frac{1}{tN} \sum_{i=1}^{N} \frac{G_t(z,x_i)}{\sqrt{d_t(z)}\sqrt{d_t(x_i)}} (f_\theta(z) - f_\theta(x_i))\Bigg| \leq \epsilon/2\Bigg).$$

Now we can use the fact that sampling according to $\lambda(z)dV(z)$ is equivalent to sampling $\tilde{x}_i$ according to $dV(z)$ and then multiplying by $\lambda(x_i)$, as follows,

$$P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda} |\tilde{\mathbf{L}}_\lambda^t f_\theta(z) - \hat{\mathbf{L}}_{\lambda,n} f_\theta(z)| \leq \epsilon/2\right) \tag{122}$$

$$= P\Bigg(\sup_{\theta\in\Theta,\lambda\in\Lambda} \Bigg| \int_{\mathcal{M}} \frac{1}{t} \frac{G_t(z,w)}{\sqrt{d_t(z)}\sqrt{d_t(w)}} (f_\theta(z) - f_\theta(w))\lambda(w) dV(w)$$

$$- \frac{1}{tN} \sum_{i=1}^{N} \frac{G_t(z,\tilde{x}_i)}{\sqrt{d_t(z)}\sqrt{d_t(\tilde{x}_i)}} (f_\theta(z) - f_\theta(\tilde{x}_i))\lambda(\tilde{x}_i)\Bigg| \leq \epsilon/2\Bigg). \tag{123}$$

We can now bound $|f_\theta(z)| \leq \mathbf{F}$, $a < \lambda(z) < b$, $a < d_t(z) < b$ with , $G_t \leq 1/(4\pi t)^{D/2} := \mathbf{G}$, to express (123) as,

$$P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda} |\tilde{\mathbf{L}}_\lambda^t f_\theta(z) - \hat{\mathbf{L}}_{\lambda,n} f_\theta(z)| \leq \epsilon/2\right) \geq P\left(\frac{2\mathbf{F}b}{at}\left| \int_{\mathcal{M}} G_t(z,w) dV(w) - \frac{1}{N}\sum_{i=1}^{N} G_t(z,\tilde{x}_i)\right| \leq \epsilon/2\right) \tag{124}$$

Now, using the Hoeffding's inequality we obtain,

$$P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda} |\tilde{\mathbf{L}}_\lambda^t f_\theta(z) - \hat{\mathbf{L}}_{\lambda,n} f_\theta(z)| \geq \epsilon/2\right) < 2e^{-\frac{\epsilon^2 N}{32}(\frac{at}{\mathbf{F}\mathbf{G}b})^2} \tag{125}$$

We will now focus on the second term of (120), i.e. $P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda}|\hat{\mathbf{L}}^t_{\lambda,N}f_\theta(z)-\mathbf{L}_{\lambda,N}f_\theta(z)|>\epsilon/2\right)$, as follows,

$$P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda}|\hat{\mathbf{L}}^t_{\lambda,N}f_\theta(z)-\mathbf{L}_{\lambda,N}f_\theta(z)|\leq\epsilon/2\right) \tag{126}$$

$$=P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda}\left|\frac{1}{tN}\sum_{i=1}^{N}\frac{G_t(z,x_i)}{\sqrt{d_t(z)}\sqrt{d_t(x_i)}}(f_\theta(z)-f_\theta(x_i))-\frac{1}{tN}\sum_{i=1}^{N}\frac{G_t(z,x_i)}{\sqrt{\hat{w}(z)}\sqrt{\hat{W}(x_i)}}(f_\theta(z)-f_\theta(x_i))\right|\leq\epsilon/2\right). \tag{127}$$

Now, by bounding over $f_\theta$, and $\lambda$ we obtain,

$$P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda}|\hat{\mathbf{L}}^t_{\lambda,N}f_\theta(z)-\mathbf{L}_{\lambda,N}f_\theta(z)|\leq\epsilon/2\right) \tag{128}$$

$$\leq P\left(\sup_{\lambda\in\Lambda}\frac{1}{tN}2\mathbf{FG}\sum_{i=1}^{N}\left|\frac{1}{\sqrt{d_t(z)}\sqrt{d_t(x_i)}}-\frac{1}{\sqrt{\hat{w}(z)}\sqrt{\hat{W}(x_i)}}\right|\leq\epsilon/2\right)$$

$$=P\left(\sup_{\lambda\in\Lambda}\frac{1}{tN}2\mathbf{FG}\sum_{i=1}^{N}\left|\frac{\sqrt{d_t(z)}(\sqrt{\hat{W}(x_i)}-\sqrt{d_t(x_i)})+\sqrt{\hat{W}(x_i)}(\sqrt{d_t(z)}-\sqrt{\hat{w}(z)})}{\sqrt{d_t(z)}\sqrt{d_t(x_i)}\sqrt{\hat{w}(z)}\sqrt{\hat{W}(x_i)}}\right|\leq\epsilon/2\right)$$

Now bounding over $d$, $\hat{w}$, and $\hat{W}$ we obtain,

$$P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda}|\hat{\mathbf{L}}^t_{\lambda,N}f_\theta(z)-\mathbf{L}_{\lambda,N}f_\theta(z)|\leq\epsilon/2\right) \tag{129}$$

$$=P\left(\sup_{\lambda\in\Lambda}\frac{1}{tN}\frac{2\mathbf{FG}\sqrt{b}}{a^{3/2}}\sum_{i=1}^{N}\left|(\sqrt{\hat{W}(x_i)}-\sqrt{d_t(x_i)})+(\sqrt{d_t(z)}-\sqrt{\hat{w}(z)})\right|\leq\epsilon/2\right) \tag{130}$$

Now, we can split the probability as follows,

$$P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda}|\hat{\mathbf{L}}^t_{\lambda,N}f_\theta(z)-\mathbf{L}_{\lambda,N}f_\theta(z)|\leq\epsilon/2\right) \tag{131}$$

$$\leq P\left(\sup_{\lambda\in\Lambda}\frac{1}{t}\frac{2\mathbf{FG}\sqrt{b}}{a^{3/2}}\left|\sqrt{d_t(z)}-\sqrt{\hat{w}(z)}\right|\leq\epsilon/4\right)+P\left(\sup_{\lambda\in\Lambda}\frac{1}{tN}\frac{2\mathbf{FG}\sqrt{b}}{a^{3/2}}\sum_{i=1}^{N}\left|\sqrt{d_t(x_i)}-\sqrt{\hat{W}(x_i)}\right|\leq\epsilon/4\right) \tag{132}$$

Now, we can take a closer look at $d(z)$, $w(z)$, which we can express as,

$$P\left(\sup_{\lambda\in\Lambda}\frac{1}{t}\frac{2\mathbf{FG}\sqrt{b}}{a^{3/2}}\left|\sqrt{d_t(z)}-\sqrt{\hat{w}(z)}\right|\leq\epsilon/4\right) \tag{133}$$

$$=P\left(\sup_{\lambda\in\Lambda}\frac{1}{t}\frac{2\mathbf{FG}\sqrt{b}}{a^{3/2}}\left|\sqrt{\int_{\mathcal{M}}G_t(z,w)\lambda(w)vol(w)}-\sqrt{\frac{1}{N}\sum_{n=1}^{N}G_t(z,x_n)}\right|\leq\epsilon/4\right). \tag{134}$$

We can once again use the fact that sampling according to $\lambda(z)dV(z)$ is equivalent to sampling $\tilde{x}_i$ according to $dV(z)$ and

then multiplying by $\lambda(x_i)$ and then bound $\lambda(z)$ to obtain,

$$P\left(\sup_{\lambda\in\Lambda}\frac{1}{t}\frac{2\mathbf{FG}\sqrt{b}}{a^{3/2}}\left|\sqrt{d_t(z)}-\sqrt{\hat{w}(z)}\right|\leq\epsilon/4\right) \tag{135}$$

$$=P\left(\sup_{\lambda\in\Lambda}\frac{1}{t}\frac{2\mathbf{FG}\sqrt{b}}{a^{3/2}}\left|\sqrt{\int_{\mathcal{M}}G_t(z,w)\lambda(w)dV(w)}-\sqrt{\frac{1}{N}\sum_{n=1}^{N}G_t(z,\tilde{x}_n)\lambda(\tilde{x}_n)}\right|\leq\epsilon/4\right) \tag{136}$$

$$\leq P\left(\frac{1}{t}\frac{2\mathbf{FG}\sqrt{b}}{a^{3/2}}\left|\sqrt{\int_{\mathcal{M}}G_t(z,w)dV(w)}-\sqrt{\frac{1}{N}\sum_{n=1}^{N}G_t(z,\tilde{x}_n)}\right|\leq\epsilon/4\right). \tag{137}$$

Now, given that $G_t$ is bounded by $\mathbf{G}$, and given that $\int_{\mathcal{M}}G_t(z,w)dV(w)=1>0$, we can use the Hoeffding's inequality as follows,

$$P\left(\frac{1}{t}\frac{2\mathbf{FG}\sqrt{b}}{a^{3/2}}\left|\int_{\mathcal{M}}G_t(z,w)dV(w)-\frac{1}{N}\sum_{n=1}^{N}G_t(z,\tilde{x}_n)\right|\geq\epsilon/4\right)\leq 2e^{-\frac{\epsilon^2N}{32}\left(\frac{ta^{3/2}}{2\mathbf{FG}\sqrt{b}2}\right)^2} \tag{138}$$

Returning to (132), we can repeat the steps, and by the union bound we get,

$$P\left(\frac{1}{tN}\frac{2\mathbf{FG}\sqrt{b}}{a^{3/2}}\sum_{i=1}^{N}\left|\int_{\mathcal{M}}G_t(x_i,w)dV(w)-\frac{1}{N-1}\sum_{n\neq i}G_t(x_i,\tilde{x}_n)\right|\geq\epsilon/4\right)\leq 2(N-1)e^{-\frac{\epsilon^2(N-1)}{32}\left(\frac{ta^{3/2}}{2\mathbf{FG}\sqrt{b}2}\right)^2} \tag{139}$$

Finally, we obtain that with probability,

$$P\left(\sup_{\theta\in\Theta,\lambda\in\Lambda}|\tilde{\mathbf{L}}_\lambda^t f_\theta(z)-\mathbf{L}_{\lambda,N}f_\theta(z)|\geq\epsilon\right)\leq 2e^{-\frac{\epsilon^2N}{32}(\frac{at}{\mathbf{FG}b})^2}+2e^{-\frac{\epsilon^2N}{32}\left(\frac{ta^{3/2}}{2\mathbf{FG}\sqrt{b}2}\right)^2}+2(N-1)e^{-\frac{\epsilon^2(N-1)}{32}\left(\frac{ta^{3/2}}{2\mathbf{FG}\sqrt{b}2}\right)^2} \tag{140}$$

It suffices to pick $N_0$ such that the right term is small than $\delta$. ■

**Lemma 8** *Consider $f_\theta$ with $\theta\in\Theta$, and $\lambda\in\Lambda$, for a point $z\in\mathcal{M}$, there is a uniform bound between the continuous heat kernel Laplacian (cf .definition 5) and the the Laplace-Beltrami operator $\Delta_\lambda f_\theta=\Delta f_\theta+\langle\nabla_{\mathcal{M}}f_\theta,\nabla_{\mathcal{M}}\lambda\rangle$, where $\Delta$ is the Laplace-Beltrami operator, i.e.*

$$\sup_{\theta\in\Theta,\lambda\in\Lambda}|\tilde{\mathbf{L}}_\lambda^t f_\theta(z)-\Delta_\lambda f_\theta(z)|\leq\mathcal{O}(t^{1/2}) \tag{141}$$

*with $\tilde{\mathbf{L}}_\lambda^t f_\theta(z)=\frac{1}{t}\int_{\mathcal{M}}\frac{G_t(z,w)}{\sqrt{d_t(z)}\sqrt{d_t(w)}}(f_\theta(z)-f_\theta(w))\lambda(w)dV(w)$.*

**Proof of Lemma 8:** The proof is as follows, first we will compare the heat kernel Laplacian $\tilde{\mathbf{L}}_\lambda^t f_\theta(p)$ with the heat kernel Laplacian with $\lambda$ instead of $d_t$. Next, we will consider the integral over a ball $\mathcal{B}$, as opposed to the integral over the whole manifold $\mathcal{M}$. Then, we will exploit the euclidean properties of the manifold at point $p$, which will allow us to convert an integral over a manifold into an integral over the low dimensional structure of the manifold. In this low dimensional manifold, we can compute the integrals, and obtain the desired result.

To begin with, we will introduce the continuous heat kernel laplacian, but using the probability distribution $\lambda$ as follows,

$$\tilde{\mathbf{L}}^t_\lambda f_\theta(p) = \frac{1}{t}\int_{\mathcal{M}} \frac{G_t(p,z)}{\sqrt{d_t(p)}\sqrt{d_t(z)}}(f_\theta(p) - f_\theta(z))\lambda(z)dV(z) \tag{142}$$

$$\leq \frac{1}{t}\int_{\mathcal{M}} \frac{G_t(p,z)}{\sqrt{\lambda(p)}\sqrt{\lambda(z)}}(f_\theta(p) - f_\theta(z))\lambda(z)dV(z) \tag{143}$$

$$+ |\frac{1}{t}\int_{\mathcal{M}} \frac{G_t(p,z)}{\sqrt{d_t(p)}\sqrt{d_t(z)}}(f_\theta(p) - f_\theta(z))\lambda(z)dV(z) - \tag{144}$$

$$\frac{1}{t}\int_{\mathcal{M}} \frac{G_t(p,z)}{\sqrt{\lambda(p)}\sqrt{\lambda(z)}}(f_\theta(p) - f_\theta(z))\lambda(z)dV(z)| \tag{145}$$

$$\leq \frac{1}{t}\int_{\mathcal{M}} \frac{G_t(p,z)}{\sqrt{\lambda(p)}\sqrt{\lambda(z)}}(f_\theta(p) - f_\theta(z))\lambda(z)dV(z) \tag{146}$$

$$+ \frac{1}{t}\int_{\mathcal{M}} G_t(p,z)|\frac{1}{\sqrt{d_t(p)}\sqrt{d_t(z)}} - \frac{1}{\sqrt{\lambda(p)}\sqrt{\lambda(z)}}||f_\theta(p) - f_\theta(z)|\lambda(z)dV(z) \tag{147}$$

By compactness of manifold $\mathcal{M}$, and the fact that $\lim_{t\to 0} d_t(p) = \lambda(p)$ the following holds for any point $p \in \mathcal{M}$,

$$d_t(p) = \lambda(p) + \mathcal{O}(tg(p)) \tag{148}$$

where $g$ is a smooth function depending upon higher derivatives of $\lambda$, which is bounded by $c$. Therefore, we can write

$$d_t^{-1/2}(z) = (\lambda(z) + \mathcal{O}(tg(z)))^{-1/2} = \lambda(z)^{-1/2} + \mathcal{O}(t), \tag{149}$$

where $\mathcal{O}(t) \leq c|t|$. Therefore, we can bound the right hand side of equation (147) by,

$$\frac{1}{t}\int_{\mathcal{M}} G_t(p,z)|\frac{1}{\sqrt{d_t(p)}\sqrt{d_t(z)}} - \frac{1}{\sqrt{\lambda(p)}\sqrt{\lambda(z)}}||f_\theta(p) - f_\theta(z)|\lambda(z)dV(z) \tag{150}$$

$$\leq \frac{1}{t}\int_{\mathcal{M}} G_t(p,z)ct|f_\theta(p) - f_\theta(z)|\lambda(z)dV(z) \tag{151}$$

$$\leq \int_{\mathcal{M}} G_t(p,z)c|f_\theta(p) - f_\theta(z)|\lambda(z)dV(z) \leq \mathcal{O}(t) \tag{152}$$

where the bound holds uniformly given the compactness of the manifold $\mathcal{M}$, and the upper bound on the derivative of $f_\theta$. We now return to equation (147), by virtue of Lemma 4, we can convert the integral over the manifold $\mathcal{M}$, to an integral over the ball $\mathcal{B}$ as follows,

$$\frac{1}{t}\int_{\mathcal{M}} \frac{G_t(x,y)}{\sqrt{\lambda(x)}\sqrt{\lambda(y)}}(f_\theta(x)-f_\theta(y))\lambda(y)dV(y) \tag{153}$$

$$= \frac{1}{t}\int_{\mathcal{B}} \frac{G_t(x,y)}{\sqrt{\lambda(x)}\sqrt{\lambda(y)}}(f_\theta(x)-f_\theta(y))\lambda(y)dV(y) \tag{154}$$

$$+ \frac{1}{t}\int_{\mathcal{M}} \frac{G_t(x,y)}{\sqrt{\lambda(x)}\sqrt{\lambda(y)}}(f_\theta(x)-f_\theta(y))\lambda(y)dV(y) \tag{155}$$

$$- \frac{1}{t}\int_{\mathcal{B}} \frac{G_t(x,y)}{\sqrt{\lambda(x)}\sqrt{\lambda(y)}}(f_\theta(x)-f_\theta(y))\lambda(y)dV(y) \tag{156}$$

$$\leq \frac{1}{t}\int_{\mathcal{B}} \frac{G_t(x,y)}{\sqrt{\lambda(x)}\sqrt{\lambda(y)}}(f_\theta(x)-f_\theta(y))\lambda(y)dV(y) \tag{157}$$

$$+ \left|\frac{1}{t}\int_{\mathcal{M}} \frac{G_t(x,y)}{\sqrt{\lambda(x)}\sqrt{\lambda(y)}}(f_\theta(x)-f_\theta(y))\lambda(y)dV(y) \right. \tag{158}$$

$$\left. - \frac{1}{t}\int_{\mathcal{B}} \frac{G_t(x,y)}{\sqrt{\lambda(x)}\sqrt{\lambda(y)}}(f_\theta(x)-f_\theta(y))\lambda(y)dV(y)\right| \tag{159}$$

$$\leq \frac{1}{t}\int_{\mathcal{B}} \frac{G_t(x,y)}{\sqrt{\lambda(x)}\sqrt{\lambda(y)}}(f_\theta(x)-f_\theta(y))\lambda(y)dV(y) + b\mathbf{V}\mathbf{F}e^{-\frac{r^2}{4t}} \tag{160}$$

We now consider the exponential coordinates exp around $p$, after introducing a ball $\mathcal{B}$ of radius $r$. Letting the change of variables be $\exp_p : T_p\mathcal{M} \to \mathcal{M}$. Letting $\tilde{\mathcal{B}}$ be a ball in $T_p\mathcal{M}$, then $\mathcal{B}$ is the image of the ball under the exponential map. We will use the exponential coordinates as $v = \exp_p(z)$, and $\tilde{f}_\theta(v) = f_\theta(\exp_p(v))$, and $\tilde{f}_\theta(0) = f_\theta(\exp_p(0))$.

$$\frac{1}{t}\int_{\mathcal{B}} \frac{G_t(x,y)}{\sqrt{\lambda(x)}\sqrt{\lambda(y)}}(f_\theta(x)-f_\theta(y))\lambda(y)dV(y) = \tag{161}$$

$$\frac{1}{t}\frac{1}{\sqrt{\tilde{\lambda}(0)}}\int_{\tilde{\mathcal{B}}} \frac{e^{\frac{\|\exp_p(v)-\exp_p(0)\|^2}{4t}}}{(4\pi t)^{k/2}}\sqrt{\tilde{\lambda}(v)}(\tilde{f}_\theta(0)-\tilde{f}_\theta(v))\det(g(v))dv \tag{162}$$

where $\det(g(v))$ is the determinant of the metric tensor in exponential coordinates. We can now expand $\sqrt{\tilde{\lambda}}(v)$, and $\tilde{f}_\theta(v)$ by their Taylor expansions as follows,

$$\tilde{\lambda}^{1/2}(v) \leq \tilde{\lambda}^{1/2}(0) + \frac{1}{2}\tilde{\lambda}^{-1/2}(0)v^\intercal\nabla\tilde{\lambda}(0) + \mathcal{O}(\|v\|^2) \tag{163}$$

$$\tilde{f}_\theta(v) = \tilde{f}_\theta(0) + v^\intercal\nabla\tilde{f}_\theta(0) + \frac{1}{2}v^\intercal\mathbf{H}v + \mathcal{O}(\|v\|^3) \tag{164}$$

$$\tilde{f}_\theta(0) - \tilde{f}_\theta(v) \leq -v^\intercal\nabla\tilde{f}_\theta(0) - \frac{1}{2}v^\intercal\mathbf{H}v + \mathcal{O}(\|v\|^3) \tag{165}$$

Where $\nabla\tilde{f} = (\frac{\partial\tilde{f}}{\partial x_1},\ldots,\frac{\partial\tilde{f}}{\partial x_k})^\intercal$, and $\mathbf{H}$ its corresponding hessian matrix. Now note that given that the norm of the second derivative $\lambda$ and third derivative of $f_\theta$ are uniformly bounded for all $p \in \mathcal{M}$, $\theta \in \Theta$ and $\lambda \in \Lambda$. With these two approximations in hand, we also approximate the heat kernel $e^{\frac{\|\exp_p(v)-\exp_p(0)\|^2}{4t}}$ as follows,

$$e^{-\frac{d_{\mathcal{M}}(\exp_p(v)-p)^2}{4t}} = e^{-\frac{\|v\|^2}{4t}}, \tag{166}$$

given that $y$ is close enough to $p$. We also approximate the determinant of the metric tensor in exponential coordinates $\det(g(v))$ using the fact that

$$\det(g(v)) = 1 - \frac{1}{6}z^\intercal\mathbf{R}v + \mathcal{O}(\|v\|^3) \tag{167}$$

$$\leq 1 + k_0\|v\|^2. \tag{168}$$

where $\mathbf{R}$ is the Ricci curvature tensor which is uniformly bounded given the compactness of $\mathcal{M}$. The last inequality holds uniformly for all $p \in \mathcal{M}$ given the compactness of $\mathcal{M}$. Returning to 162, putting all the pieces together we obtain,

$$\frac{1}{t}\int_{\mathcal{B}} \frac{G_t(x,y)}{\sqrt{\lambda(x)}\sqrt{\lambda(y)}}(f_\theta(x) - f_\theta(y))\lambda(y)dV(y) \tag{169}$$

$$\leq \frac{1}{t}\frac{1}{\sqrt{\tilde{\lambda}(0)}}\frac{1}{(4\pi t)^{k/2}}\int_{\tilde{\mathcal{B}}} e^{\frac{-\|v\|^2}{4t}}(\tilde{\lambda}^{1/2}(0) + \frac{1}{2}\tilde{\lambda}^{-1/2}(0)v^\intercal\nabla\tilde{\lambda}(v) + k_1\|v\|^2) \tag{170}$$

$$(v^\intercal\nabla\tilde{f}_\theta(0) + \frac{1}{2}v^t\mathbf{H}v + k_2\|v\|^3)dv \tag{171}$$

$$+ \frac{1}{t}\frac{1}{\sqrt{\tilde{\lambda}(0)}}\frac{1}{(4\pi t)^{k/2}}\int_{\tilde{\mathcal{B}}} e^{\frac{-\|v\|^2}{4t}}(\tilde{f}(0) - \tilde{f}(v))k_0\|v\|^2dv \tag{172}$$

Which can be split into 5 integrals as follows:

$$A_t = \frac{1}{t}\frac{1}{\sqrt{\tilde{\lambda}(0)}}\frac{1}{(4\pi t)^{k/2}}\int_{\tilde{\mathcal{B}}} e^{-\frac{\|v\|^2}{4t}}\tilde{\lambda}^{1/2}(0)v^\intercal\nabla\tilde{f}_\theta(0)dv \tag{173}$$

$$B_t = \frac{1}{t}\frac{1}{\sqrt{\tilde{\lambda}(0)}}\frac{1}{(4\pi t)^{k/2}}\int_{\tilde{\mathcal{B}}} e^{-\frac{\|v\|^2}{4t}}\tilde{\lambda}^{1/2}(0)\frac{1}{2}v^t\mathbf{H}vdv \tag{174}$$

$$C_t = \frac{1}{t}\frac{1}{\sqrt{\tilde{\lambda}(0)}}\frac{1}{(4\pi t)^{k/2}}\int_{\tilde{\mathcal{B}}} e^{-\frac{\|v\|^2}{4t}}\frac{1}{2}\tilde{\lambda}^{-1/2}(0)v^\intercal\nabla\tilde{\lambda}(0)\nabla\tilde{f}_\theta(0)^\intercal vdv \tag{175}$$

$$D_t = \frac{1}{t}\frac{1}{\sqrt{\tilde{\lambda}(0)}}\frac{1}{(4\pi t)^{k/2}}\int_{\tilde{\mathcal{B}}} e^{-\frac{\|v\|^2}{4t}}\frac{1}{2}\tilde{\lambda}^{-1/2}(0)v^\intercal\nabla\tilde{\lambda}(0)\frac{1}{2}v^\intercal\mathbf{H}vdv \tag{176}$$

$$E_t = \frac{1}{t}\frac{1}{\sqrt{\tilde{\lambda}(0)}}\frac{1}{(4\pi t)^{k/2}}\int_{\tilde{\mathcal{B}}} e^{\frac{-\|v\|^2}{4t}}(\tilde{f}(0) - \tilde{f}(v))k_0\|v\|^2d\mathbb{R}^k \tag{177}$$

$$\leq \frac{1}{t}\frac{1}{\sqrt{\tilde{\lambda}(0)}}\frac{1}{(4\pi t)^{k/2}}\int_{\tilde{\mathcal{B}}} e^{\frac{-\|v\|^2}{4t}}k_0\|v\|^3d\mathbb{R}^k \tag{178}$$

Where for $E_t$ we have used the fact that the first derivative has a uniform bound. Note that given the Gaussian kernel integration, the following conditions are true,

$$\int_{\tilde{B}} v_i e^{-\frac{\|v\|^2}{4t}}dv = 0 \tag{179}$$

$$\int_{\tilde{B}} v_i v_j e^{-\frac{\|v\|^2}{4t}}dv = 0 \tag{180}$$

$$\int_{\tilde{B}} v_i v_j^2 e^{-\frac{\|v\|^2}{4t}}dv = 0 \tag{181}$$

$$\frac{1}{t}\frac{1}{(4\pi t)^{k/2}}\int_{\tilde{B}} \|v\|^3 e^{-\frac{\|v\|^2}{4t}}dv = \mathcal{O}(t^{1/2}) \tag{182}$$

$$\frac{1}{t}\frac{1}{(4\pi t)^{k/2}}\int_{\tilde{B}} v_i^2 e^{-\frac{\|v\|^2}{4t}}dv = 2 \tag{183}$$

for $i \neq j$ given the zero mean Gaussian distribution with zero mean and diagonal covariance matrix. Using these conditions, we can conclude that $A_t = D_t = 0$. From $B_t$ and $C_t$, the only non-zero elements are given by the diagonal elements, therefore, we obtain,

$$A_t + B_t + C_t + D_t + E_t = -\sum_{i=1}^{k}\frac{\partial^2\tilde{f}(0)}{\partial x_i^2} - \frac{1}{\tilde{\lambda}(0)}\sum_{i=1}^{k}[\nabla\tilde{\lambda}(0)]_i[\nabla\tilde{f}_\theta(0)]_i + \mathcal{O}(t^{1/2}) \tag{184}$$

$$= -\Delta f - \frac{1}{\lambda}\langle\nabla\lambda, \nabla f_\theta\rangle_{T_p} + \mathcal{O}(t^{1/2}) \tag{185}$$

Where we have used that $\Delta_{\mathcal{M}} f(p) = \Delta_{\mathbb{R}^k} \tilde{f}(0) = -\sum_{i=1}^{k} \frac{\partial^2 \tilde{f}}{\partial x_i^2}(0)$ (see Chapter 3 of (Rosenberg, 1997)). ■

**Proof of Proposition 3:** To begin with, utilizing Green's identity, for the Laplacian $\Delta_\lambda = \Delta f_\theta + \frac{1}{\lambda}\langle \nabla \lambda, \nabla f_\theta \rangle$, given that the function vanishes at the boundary, the following holds

$$\int_{\mathcal{M}} f_\theta(z) \Delta_\lambda f_\theta(z) \lambda(z) dV(z) = \int_{\mathcal{M}} \|\nabla_{\mathcal{M}} f_\theta(z)\|^2 \lambda(z) dV(z) \tag{186}$$

An equivalent statement of the proof is that for any $\epsilon > 0$, and $\delta > 0$, there exist $N_0$ such that, for $N \geq N_0$

$$P\left( \sup_{\lambda \in \Lambda, \theta \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} f_\theta(x_i) \lambda(x_i) \mathbf{L}_{\lambda,N}^{t_N} f_\theta(x_i) - \int_{\mathcal{M}} f_\theta(z)(-\Delta_\lambda f_\theta(z)) \lambda(x) dV(z) \right| > \epsilon \right) \leq \delta \tag{187}$$

By considering the complement of the event in (187), we obtain,

$$P\left( \sup_{\lambda \in \Lambda, \theta \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} f_\theta(x_i) \lambda(x_i) \mathbf{L}_{\lambda,N}^{t_N} f_\theta(x_i) - \int_{\mathcal{M}} f_\theta(z)(-\Delta_\lambda f_\theta(z)) \lambda(x) dV(z) \right| \leq \epsilon \right) \geq 1 - \delta \tag{188}$$

We can now add and subtract $\frac{1}{N}\sum_{i=1}^{N} f_\theta(x_i)\lambda(x_i)\Delta_\lambda f_\theta(x_i)$, and split the probability as follows,

$$P\left( \sup_{\lambda \in \Lambda, \theta \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} f_\theta(x_i) \lambda(x_i) \mathbf{L}_{\lambda,N}^{t_N} f_\theta(x_i) - \int_{\mathcal{M}} f_\theta(z)(-\Delta_\lambda f_\theta(z)) \lambda(x) dV(z) \right| \leq \epsilon \right) \tag{189}$$

$$\geq P\left( \sup_{\lambda \in \Lambda, \theta \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} f_\theta(x_i) \lambda(x_i) \mathbf{L}_{\lambda,N}^{t_N} f_\theta(x_i) - \frac{1}{N} \sum_{i=1}^{N} f_\theta(x_i) \lambda(x_i)(-\Delta_\lambda f_\theta(x_i)) \right| \leq \epsilon/2 \right)$$

$$+ P\left( \sup_{\lambda \in \Lambda, \theta \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} f_\theta(x_i) \lambda(x_i)(-\Delta_\lambda f_\theta(x_i)) - \int_{\mathcal{M}} f_\theta(z)(-\Delta_\lambda f_\theta(z)) \lambda(x) dV(z) \right| \leq \epsilon/2 \right) \tag{190}$$

Where equation (190) holds given that it is a subset of the event in (189). We can now focus on the second term of (190). Given that the second derivatives of $f_\theta$ are uniformly bounded for $\theta \in \Theta$, we can bound $|\Delta f_\theta| \leq \mathbf{F}_2$, to obtain,

$$P\left( \sup_{\lambda \in \Lambda, \theta \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} f_\theta(x_i) \lambda(x_i)(-\Delta_\lambda f_\theta(x_i)) - \int_{\mathcal{M}} f_\theta(z)(-\Delta_\lambda f_\theta(z)) \lambda(x) dV(z) \right| \leq \epsilon/2 \right)$$

$$\geq P\left( \sup_{\lambda \in \Lambda, \theta \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} f_\theta(x_i) \lambda(x_i) - \int_{\mathcal{M}} f_\theta(z) \lambda(x) dV(z) \right| \leq \frac{\epsilon}{2K_\Delta} \right). \tag{191}$$

Now, for equation (191), we can use Assumption 4 with probability $\delta/2$ as follows,

$$\hat{\zeta}(N, \delta/2) \leq \frac{\epsilon}{2K_\Delta}. \tag{192}$$

Returning to (190), we can also bound the value of $|f_\theta(z)| \leq \mathbf{F}$, and $a \leq \lambda \leq b$ to obtain,

$$P\left( \sup_{\lambda \in \Lambda, \theta \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} f_\theta(x_i) \lambda(x_i) \mathbf{L}_{\lambda,N}^{t_N} f_\theta(x_i) - \frac{1}{N} \sum_{i=1}^{N} f_\theta(x_i) \lambda(x_i)(-\Delta_\lambda f_\theta(x_i)) \right| \leq \epsilon/2 \right) \tag{193}$$

$$\geq P\left( \sup_{\lambda \in \Lambda, \theta \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} \mathbf{L}_{\lambda,N}^{t_N} f_\theta(x_i) - \frac{1}{N} \sum_{i=1}^{N} (-\Delta_\lambda f_\theta(x_i)) \right| \leq \frac{\epsilon}{2b\mathbf{F}} \right) \tag{194}$$

$$\geq P\left( \sup_{\lambda \in \Lambda, \theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \left| \mathbf{L}_{\lambda,N}^{t_N} f_\theta(x_i) - (-\Delta_\lambda f_\theta(x_i)) \right| \leq \frac{\epsilon}{2b\mathbf{F}} \right) \tag{195}$$

$$\geq P\left( \bigcap_{i=1}^{N} \sup_{\lambda \in \Lambda, \theta \in \Theta} \left| \mathbf{L}_{\lambda,N}^{t_N} f_\theta(x_i) - (-\Delta_\lambda f_\theta(x_i)) \right| \leq \frac{\epsilon}{2b\mathbf{F}} \right) \tag{196}$$

$$\geq 1 - \sum_{i=1}^{N} P\left( \sup_{\lambda \in \Lambda, \theta \in \Theta} \left| \mathbf{L}_{\lambda,N}^{t_N} f_\theta(x_i) - (-\Delta_\lambda f_\theta(x_i)) \right| \geq \frac{\epsilon}{2b\mathbf{F}} \right). \tag{197}$$

Where equation (195) holds by the triangle inequality, equation (196) holds by taking the intersection of the events, and equation (197) holds by the union bound. Now we add and subtract the continuous $\tilde{\mathbf{L}}^t_\lambda f_\theta(x_i)$ (cf. definition 5) as follows,

$$P\left(\sup_{\lambda\in\Lambda,\theta\in\Theta}\left|\mathbf{L}^{t_N}_{\lambda,N}f_\theta(x_i)-(-\Delta_\lambda f_\theta(x_i))\right|\leq\frac{\epsilon}{2b\mathbf{F}}\right) \tag{198}$$

$$=P\left(\sup_{\lambda\in\Lambda,\theta\in\Theta}\left|\mathbf{L}^{t_N}_{\lambda,N}f_\theta(x_i)-\tilde{\mathbf{L}}^t_\lambda f_\theta(x_i)+\tilde{\mathbf{L}}^t_\lambda f_\theta(x_i)-(-\Delta_\lambda f_\theta(x_i))\right|\leq\frac{\epsilon}{2b\mathbf{F}}\right) \tag{199}$$

$$\geq P\left(\sup_{\lambda\in\Lambda,\theta\in\Theta}\left|\mathbf{L}^{t_N}_{\lambda,N}f_\theta(x_i)-\tilde{\mathbf{L}}^t_\lambda f_\theta(x_i)\right|\leq\frac{\epsilon}{4b\mathbf{F}}\right) \tag{200}$$

$$+P\left(\sup_{\lambda\in\Lambda,\theta\in\Theta}\left|\tilde{\mathbf{L}}^t_\lambda f_\theta(x_i)-(-\Delta_\lambda f_\theta(x_i))\right|\leq\frac{\epsilon}{4b\mathbf{F}}\right) \tag{201}$$

For term (200), we can use Lemma 8 with $\delta/4$, and $\epsilon^{'}=\epsilon/4b\mathbf{F}$ to obtain $N_0^{'}$ should be such that

$$2e^{-\frac{\epsilon^{'2}N_0^{'}}{32}(\frac{at}{\mathbf{F}b})^2}+2e^{-\frac{\epsilon^{'2}N_0^{'}}{32}\left(\frac{ta^{3/2}}{2\mathbf{FG}\sqrt{b}2}\right)^2}+2(N_0^{'}-1)e^{-\frac{\epsilon^{'2}(N_0^{'}-1)}{32}\left(\frac{ta^{3/2}}{2\mathbf{FG}\sqrt{b}2}\right)^2}\leq\delta/4N \tag{202}$$

For term (201) can be bounded utilizing Lemma 8, and therefore $N^{'}$ needs to verify,

$$t^{1/2}\leq\frac{\delta}{4N}\text{ and }N^{'-\frac{1}{d+2+\alpha}}=t. \tag{203}$$

To conclude, it suffices to pick $N$ such that condition (192) is verified, and $N^{'}$ needs to be picked such that condition (202), and (203) are meet. The total number of samples is then $N_0+N_0^{'}$. ■

## F. Dual Ascent Algorithm

In this section, we explore the two algorithms that we present to solve problem 7; the gradient based primal dual Algorithm 1 in subsection F.1, and the pointcloud laplacian based primal dual 2 in subsection F.2. The laplacian variant, presents a computational advantage, but requires conditions on the dual variables $\lambda$ that might be difficult to secure in practice. The gradient based method on the other hand, presents a less restrictive algorithm at the expense of a higher computational cost. It is important to note that the sole difference between the two relies on the gradient of the lagrangian with respect to $\theta$, that is to say, the update on the dual variables $\lambda$ and $\mu$ remains the same. In this appendix section, we provide an verbose explanation of the two procedures. Moreover, we elaborate on the estimator of the norm of the gradient F.3.

### F.1. Gradient Based Primal Dual Ascent

On this subsection we elaborate on the gradient based method to update the primal variable $\theta$. Upon showing that under certain conditions problem (3) and problem (9) are close (see Proposition 1, and 2), we will introduce a dual ascent algorithm to solve the later. The problem that we seek to solve is given by,

$$\max_{\hat{\lambda},\hat{\mu}} \hat{d}_G(\hat{\lambda},\hat{\mu}) := \min_{\theta} \quad \frac{1}{N}\sum_{n=1}^{N}\Big(\ell\big(f_\theta(x_n),y_n\big)-\mu\epsilon\Big)+\hat{\mu}^{-1}\frac{1}{N}\sum_{n=1}^{N}\hat{\lambda}(x_n)\|\nabla_{\mathcal{M}} f_\theta(x_n)\| \tag{204}$$

$$\text{subject to} \quad \frac{1}{N}\sum_{n=1}^{N}\lambda(x_n)=1 \tag{205}$$

We will procede with an iterative process that minimizes over $\theta$, and maximizes over $\lambda$ and $\mu$. In short, for each value of $\mu$, $\lambda$, we seek to minimize the Lagrangian $\hat{L}(\theta,\hat{\mu},\hat{\lambda})$ by taking gradient steps as follows,

$$\theta_{k+1} = \theta_k + \eta_\theta \nabla_\theta \hat{L}(\theta,\hat{\mu},\lambda) \tag{206}$$

$$= \theta_k + \eta_\theta \nabla_\theta \Big(\hat{\mu}\frac{1}{N}\sum_{n=1}^{N}\ell\big(f_\theta(x_n),y_n\big)+\frac{1}{N}\sum_{n=1}^{N}\hat{\lambda}(x_n)\|\nabla_{\mathcal{M}} f_\theta(x_n)\|^2\Big) \tag{207}$$

The gradient base approach computes gradients with respect to the loss function $\ell$, and with respect to the norm of the gradient. After the dual function $\hat{d}(\hat{\lambda},\hat{\mu})$ is minimized, in order to solve 7, we require to maximize the dual function $\hat{d}(\hat{\lambda},\hat{\mu})$ over both $\hat{\lambda}$, and $\hat{\mu}$, which can be done by evaluating the constraint violation as follows,

$$\lambda_n \leftarrow \lambda_n + \eta_\lambda \partial_{\hat{\lambda}_n}\hat{\mathcal{L}}(\theta,\hat{\mu},\hat{\lambda}), \text{ with } \partial_{\hat{\lambda}_n}\hat{\mathcal{L}}(\theta,\hat{\mu},\hat{\lambda}) = \|\nabla_{\mathcal{M}}\phi(x_n)\|^2 \tag{208}$$

$$\hat{\mu} \leftarrow \hat{\mu} + \eta_{\hat{\mu}}\hat{\mathcal{L}}(\theta,\hat{\mu},\hat{\lambda}), \text{ with } \partial_{\hat{\mu}}\hat{\mathcal{L}}(\theta,\hat{\mu},\hat{\lambda}) = \frac{1}{N}\sum_{n=1}^{N}\ell\big(f_\theta(x_n),y_n\big)-\epsilon, \tag{209}$$

where stepsizes $\eta_\lambda, \eta_\mu$ are positive numbers. Note that to evaluate the norm of the gradient on a particular point $\|\nabla_{\mathcal{M}}\phi(x_n)\|$ we can look at its neighboring points $\mathcal{N}(x_i)$, and estimate its norm. After updating the dual variables $\lambda$ using gradient ascent, we require $|\lambda|_1 = 1$, which can be done by either normalizing, or projecting to the simplex (Wang & Carreira-Perpinán, 2013). How to compute the norm of the gradient will be explained in the sequel in subsection F.3. The overall procedure is explained in Algorithm 1.

### F.2. point-cloud Laplacian Dual Ascent

Our algorithm will take advantage of the Laplacian formulation given in Proposition 3. That is, we will estimate the gradient of the lagrangian with respect to $\theta$ utilizing the point-cloud laplacian formulation. Formally, for a vector $\hat{\lambda}\in\mathbb{R}_+^N$, such that $\frac{1}{N}\sum_{n=1}^{N}\hat{\lambda}_n = 1$, and a constant $\hat{\mu}\in\mathbb{R}^+$, the empirical dual function associated with the Lagrangian $\hat{L}(\theta,\hat{\mu},\hat{\lambda}_n)$ of the empirical dual function associated with (7) is defined as,

$$\hat{d}(\hat{\lambda},\hat{\mu}) = \min_{\theta} \quad \frac{1}{N}\sum_{n=1}^{N}\ell\big(f_\theta(x_n),y_n\big)+\hat{\mu}^{-1}\frac{1}{N}\sum_{n=1}^{N}f_\theta(x_n)\hat{\lambda}_n\mathbf{L}^{t_N}_{\lambda p,N}f_\theta(x_n), \tag{210}$$

Note that in (210), we have omitted the term $-\hat{\mu}\epsilon$ as it is a constant term for a given $\hat{\mu}$, and we have divided over $\hat{\mu}$ which renders an equivalent problem as long as $\hat{\mu}>0$. By taking the maximum of the dual function over $\hat{\mu}$, and $\hat{\lambda}$, we recover the

**Algorithm 1** Gradient Based Smooth Learning on Data Manifold

Initialize parametric function $\theta$, define neighborhoods $\mathcal{N}(p)$ for every $p \in \mathcal{D}$
**repeat**
  **for** primal steps $k$ **do**
    Estimate maximum norm of gradient for all $x_i$: $\|\nabla_{\mathcal{M}} f_\theta(x_i)\| = \max_{z \in \mathcal{N}(x_i)} \frac{\|f_\theta(x_i) - f_\theta(z)\|}{d(x_i, z)}$)
    Update $\theta : \theta \leftarrow \theta - \eta_\theta \nabla_\theta \left( \mu \frac{1}{N} \sum_{i=1}^N \ell(f_\theta(x_i), y_i) + \frac{1}{N} \sum_{i=1}^N \lambda(x_i) \|\nabla_{\mathcal{M}} f_\theta(x_i)\|^2 \right)$
  **end for**
  Update dual variable $\mu$: $\mu \leftarrow [\mu + \eta_\mu \frac{1}{N} \sum_{i=1}^N \ell(f_\theta(x_i), y_i) - \epsilon)]_+$
  Update dual variable $\lambda(x_i)$: $\lambda(x_i) \leftarrow [\lambda(x_i) + \eta_\lambda (\max_{z \in \mathcal{N}(x_i)} \frac{\|f_\theta(x_i) - f_\theta(z)\|^2}{d(x_i, z)^2})]_+$
  Project $\lambda : \lambda = \arg\min_{0 \preccurlyeq \tilde{\lambda}} \|\tilde{\lambda} - \lambda\|$ s.t. $|\tilde{\lambda}|_1 = N$
  $e = e + 1$
**until** convergence

**Algorithm 2** point-cloud Laplacian Smooth Learning on Data Manifold

Initialize parametric function $\theta$, fix temperature $t$
**repeat**
  **for** primal steps $k$ **do**
    Update $\theta : \theta \leftarrow \theta - \eta_\theta \nabla_\theta \left( \mu \frac{1}{N} \sum_{i=1}^N \ell(f_\theta(x_i), y_i) + \frac{1}{N} \sum_{i=1}^N f_\theta(x_i) \lambda(x_i) \mathbf{L}^t_{\lambda p, N} f_\theta(x_i) \right)$
  **end for**
  Update dual variable $\mu$: $\mu \leftarrow [\mu + \eta_\mu \frac{1}{N} \sum_{i=1}^N \ell(f_\theta(x_i), y_i) - \epsilon)]_+$
  Update dual variable $\lambda(x_i)$: $\lambda(x_i) \leftarrow [\lambda(x_i) + \eta_\lambda (\max_{z \in \mathcal{N}(x_i)} \frac{\|f_\theta(x_i) - f_\theta(z)\|^2}{d(x_i, z)^2})]_+$
  Project $\lambda : \lambda = \arg\min_{0 \preccurlyeq \tilde{\lambda}} \|\tilde{\lambda} - \lambda\|$ s.t. $|\tilde{\lambda}|_1 = N$
  $e = e + 1$
**until** convergence

dual problem 9. For a given choice of dual variables, $\hat{\lambda}, \hat{\mu}$, the dual function (210) is an unconstrained problem that only depends on the parameters $\theta$. Now, the link with Manifold Regularization (5) is seen; as considering $\tilde{\lambda}(x_n) = 1/N$, and $\tilde{\mu} = \gamma$, the problems become equivalent. In order to minimize (210), we can update the parameters $\theta$ following the gradient,

$$\theta \leftarrow \theta + \eta_\theta \nabla_\theta \left( \frac{1}{N} \sum_{n=1}^N \ell\big(f_\theta(x_n), y_n\big) + \hat{\mu}^{-1} \frac{1}{N} \sum_{n=1}^N f_\theta(x_n) \hat{\lambda}_n \mathbf{L}^{t_N}_{\lambda p, N} f_\theta(x_n) \right), \tag{211}$$

where $\eta_\theta > 0$ is a step-size. Note that the updates in 211 have two parts, one that relies on the loss $\ell$ which utilizes labeled data, and another term given by $\hat{\lambda}_n \mathbf{L}^{t_N}_{\lambda p, N}$, that penalizes the Lipschitz constant of the function, and can be computed utilizing unlabeled data. A more succinct explanation of the algorithm is described in Algorithm 2.

### F.3. Gradient Norm Estimate

For both Algorithms 1, and 2, we require to compute the norm of the gradient of $f_\theta$ at sample point $x_n$ with respect to the manifold $\mathcal{M}$. Leveraging the Lipschitz constant definition 1, in order to estimate the maximum norm of the gradient, we require to evaluate the Lipschitz over the sample points in our dataset. To do so, we require to set a metric distance over the manifold $d_{\mathcal{M}}$, with which we will define the neighborhood of sample $x_n$ as the samples that are sufficiently close,i.e.

$$\mathcal{N}(x_n) = \{z \in \mathcal{M} : d_{\mathcal{M}}(z, x_n) \leq \delta\}. \tag{212}$$

Note that our definition of neighborhood 212, requires a maximum distance $\delta$. In practice, we can either choose $\delta$, or set the degree of the neighborhood,i.e. choose the $k$nearest neighbors. Upon the selection of the neighborhood, the norm of the gradient is computed as follows.

$$\|\nabla_{\mathcal{M}} f_\theta(x_i)\| = \max_{z \in \mathcal{N}(x_i)} \frac{\|f_\theta(x_i) - f_\theta(z)\|}{d(x_i, z)} \tag{213}$$

The selection of distance over the manifold is application dependent. In controls problems that involve physical systems, the metric might involve the work required from two states to reach each other. On computer vision applications, common choices of metrics can be perceptual losses (Zhang et al., 2018), or distances over embedding (Khrulkov et al., 2020).

# G. Experiments

## G.1. Baselines

In all experiments, we compared our method with empirical risk minimization (ERM) (Vapnik, 1999), ambient regularization (Krogh & Hertz, 1991) and manifold regularization (Belkin et al., 2005). Regarding memory allocation and computation, both manifold regularization, and manifold Lipschitz, require to calculate and storage the point-cloud Laplacian matrix. Moreover, our method also requires to save a real value variable for each sample. Regarding training times, our experiments show that our method takes $10\%$ more time than laplacian regularization, which over all takes twice as much time as ERM and ambient regularization. We run experiments on both NVIDIA 2080, as well as 3090 GPUs.

## G.2. Ground Robotic Vehicle Residual Learning

In this section we explain the residual learning experiment that involves a ground robot vehicle. The data acquisition of this experiment involves an *iRobot* Packbot equipped with high resolution camera. The setting of the data acquisition, is a robot making turns on both *pavement*, and *grass*. The dynamics of the system, are govern by the discrete-time nonlinear state-space system of equations

$$x_{k+1} = f(x_k, u_k) + g(u_k), \tag{214}$$

where $x_k$ is the state of the system, and $u_k$ is the action taken, $f(x_k, u_k)$ is the model prediction, and $g(u_k)$ is a non-modellable error of the prediction. In this setting, the robot state involves the position, and the action taken is given by its linear, and angular velocities. The dynamics of the system are modeled by $f(x_k, u_k)$, and they involve the mass of the robot, the radius of the wheel, among other known parameters of the robot. In practice, the model $f(x_k, u_k)$ is not perfect, and the model mismatch is given by the difference in friction with the ground, delays in communications with the sensors, and discrepancies in the robot specifications. To make matters worse, some of the discrepancies, are difficult to model, or intractable to compute in practice. Therefore, we seek to learn the model mismatch $g(u_k)$.

Figure 3 shows examples of trajectories given by time series $x_k$, each of which is associated with an average error model mismatch, and its corresponding variance,

$$\mu_i = \frac{1}{K}\sum_{k=1}^{K} \|x_{k+1} - f(x_k, u_k)\|, \tag{215}$$

$$\sigma_i = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K} \|\mu_i - (x_{k+1} - f(x_k, u_k))\|^2}. \tag{216}$$

In practice, samples from the grass dataset tend to have larger disturbance mean $\mu$, given that pavement presents a more uniform setting. For example, if we compare sample 3b to 3f, we can see that for similar trajectories, the error is larger in the case of grass. Moreover, trajectories that involve larger magnitudes in velocities, also present larger disturbance means, and variances (cf. 3a vs 3c). This is due to the fact that at high speeds, the non-modeled effects such as drift, and friction are more significant.

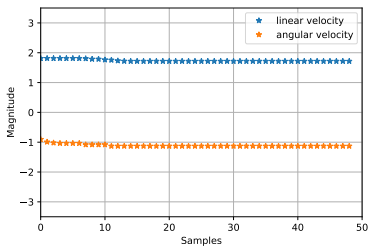
For more details on the data acquisition, please refer to the original paper (Koppel et al., 2016).

### G.2.1. Numerical Implementation

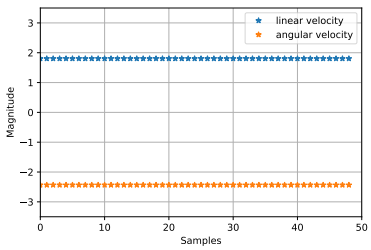
We are equipped with two datasets, one for grass with 195 samples, and one for pavement with 224 samples. We partition the datasets into 170 samples for training and 25 samples for testing, and 200 samples for training and 24 samples for testing, for grass and pavement respectively. Each sample has associated a disturbance mean, and disturbance variance vector $[\mu, \sigma_i] \in \mathbb{R}^2$.

Regarding the optimization, we utilize the mean square error as a loss, and we train a 2 layer fully connected neural network with 256 hidden dimensions and hyperbolic tangent as the non-linearity. We train for 10000 epochs, with a learning rate of $0.0015$ in the case of pavement, and $0.00015$ in the case of grass. For batch size, we utilize the whole training set. For the case of ambient Lipschitz, we utilize weight decay $w_D = 0.3$, and $w_D = 0.1$ for grass and pavement respectively.

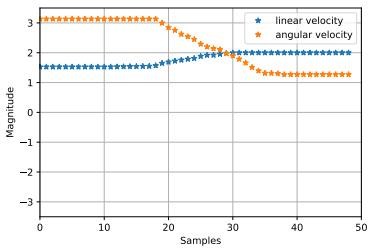
For the point-cloud laplacian, we utilize heat kernel with temperature $t = 15$, and $t = 35$ for grass and pavement respectively.



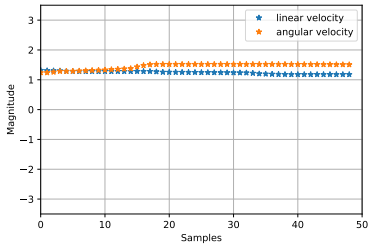

(a) Grass sample 10 with average error $\mu = 0.288$ and variance error $\sigma = 0.088$.



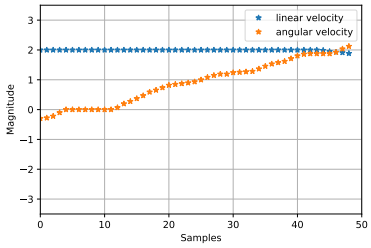

(b) Grass sample 100 with average error $\mu = 0.919$ and variance error $\sigma = 0.113$.



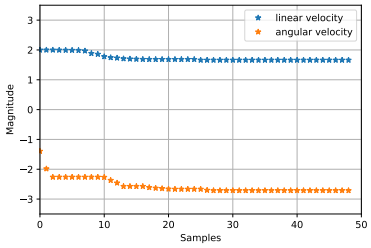

(c) Grass sample 120 with average error $\mu = 1.01$ and variance error $\sigma = 0.298$.




(d) Pavement sample 10 with average error $\mu = 0.370$ and variance error $\sigma = 0.172$.




(e) Pavement sample 100 with average error $\mu = 0.424$ and variance error $\sigma = 0.313$.




(f) Pavement sample 120 with average error $\mu = 0.652$ and variance error $\sigma = 0.747$.

Figure 3: Sample trajectories for grass, and pavement, with their corresponding average error, and variance error.

This gives us a point-cloud laplacian with one connected component. For the Laplacian regularization we utilize $1 * 10^{-3}$, and $1 * 10^{-4}$ for grass and pavement respectively.

As for the Manifold Lipschitz, we utilize $\eta_\mu = 0.01$, and $\eta_\lambda = 0.1$, and $\epsilon = 0.005$, and $0.01$ for grass and pavement respectively. We initialize $\mu = 5$, and $\lambda$ uniform.

The final results are summarized in Table 1.

### G.3. Quadrotor state prediction

In this section we present a state prediction problem based on a real world collected from a quadrotor taking off and flying in a circle for 12 seconds. The aerial robot is the open-source Crazyflie 2.1 quadrotor (`https://www.bitcraze.io/products/crazyflie-2-1/`), which has a mass of 32 g, and a mass of size 9 cm$^2$. The quadrotor communicates with a computer running on Intel $i7$ CPU, and the communication is established with the Crazyradio PA and at a nominal rate of 500 Hz ($Ts = 1/500$). To measure the position of the quadrotor a VICON is utilized. The position is obtained from the VICON, whereas the accelerations are obtained from the on-board accelerometers and gyroscope sensors. For further information on this setup, please refer to (Jiahao et al., 2022; Chee et al., 2022; Jiahao et al., 2021).

The experimental setup is to target a speed of $0.4$m/s and to track a circular trajectory of radius $0.5$m. We consider 2 trajectories of 12 seconds each (each trajectory has $6000$ time stamps), and the starting position of the quadrotor is the same for both trajectories. For each time stamp $t$, we have measurements of position, velocity and acceleration in $\mathbb{R}^3$, i.e. $[x_t, y_t, z_t, \dot{x}_t, \dot{y}_t, \dot{z}_t, \ddot{x}_t, \ddot{y}_t, \ddot{z}_t]$. The problem consists on learning the dynamical system composed by the quadrotor. We consider the dynamics given by the equation,

$$[x_{t+1}, y_{t+1}, z_{t+1}, \dot{x}_{t+1}, \dot{y}_{t+1}, \dot{z}_{t+1}] = f([x_t, y_t, z_t, \dot{x}_t, \dot{y}_t, \dot{z}_t, \ddot{x}_t, \ddot{y}_t, \ddot{z}_t]). \tag{217}$$

The learning problem consists of learning the dynamical system, i.e. we consider the dynamics given by the equation (217). For the learning procedure we utilize the $6000$ samples, and we seek to minimize the mean square error loss between the next state and the prediction given the current state (cf. equation 217). We train a two layer neural network with different methods as seen in Table 2. To test the neural network, we compute the difference between the predicted state, and the

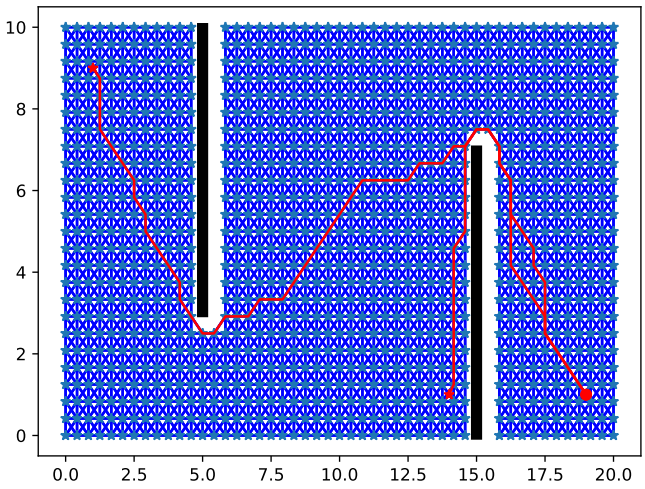



Figure 4: Point-cloud Laplacian

next state on the test trajectory. Note that even though the training, and testing trajectories are not the same, there is a resemblance between the two of them. Te begin with, we can conclude that adding regularization not always helps, as ambient regularization does not improve upon the ERM prediction method. A salient conclusion of the results shown in Table 2, is that adding regularization on the manifold space always improves upon ERM. In particular, our method is almost 3 times better than standard ERM, and more than 2 times better than standard Laplacian Regularization.

To conclude with, we show that our method obtains an improvement over all the techniques considered. This allows us to conclude that in predicting the next state of a quadrotor from the current state under noisy measurements utilizing smooth functions improves generalization.

#### G.3.1. Details on Numerical Implementation

For the state prediction problem of a quadrotor we utilized a two layer fully connected neural network with $8192$ hidden units, hyperbolic tangent as the non-linearity, and bias term. For the optimizer, we utilized a learning rate of $10^{-5}$, and the full dataset per batch. We trained until convergence in all cases with number of epochs $e = 1000$. For the ambient regularization we used weight decay $0.1$. For the construction of the Laplacian, we utilized temperature coefficient $t = 0.01$. For Laplacian regularization we utilized $\gamma = 10^{-}6$. For our method, we utilized $\mu$ dual step $0.5$, $\epsilon = 0.003$, and $\lambda$ dual step $0.1$. In all cases we trained until convergence with $e = 10000$ epochs.

### G.4. Two-Moons Dataset

In this subsection we provide the details of the experiment with the Two-moons data set utilized in 1.

To generate the data we utilized `sklearn` library, and we utilize $1$ labeled, and $200$ unlabeled samples per class (i.e. moon), and we added noise $\sigma = \{0.05, 0.1\}$. For the neural network, we utilized a two layer fully connected neural network with $64$ hidden neurons with bias term, and hyperbolic tangent as the non-linearity. For the optimizer, we utilized a learning rate of $0.9$, and no momentum. For the ambient regularization, we added a weight decay of $0.1$. For the construction of the Laplacian of Figure 1 we utilized a heat kernel temperature of $t = 0.005$, and we normalize it. For Laplacian regularization, we set $\gamma = 0.5$. For Manifold Lipschitz (our method), a $\mu$ dual step of $0.5$, and a $\lambda$ dual step of $0.1$. An ablation study over different values of temperature coefficient $t$ can be found in sections I. As seen in figure 1, Ambient Regularization fails to classify the unlabeled samples, given that ignores the distribution of samples given by the Manifold. The case in which the manifold has two connected component (cf. Figure 1a), our method works as good as Manifold Regularization, due to the fact that the Lipschitz constant will be made small in both components separately. However, when the manifold is weakly connected, Manifold Regularization fails to recognize the transition between the components, as it will penalize

large gradients across the manifold, converging to a plane that connects the two samples. Our Manifold Lipschitz method, as it requires the Lipschitz constant to be small, forces a sharp transition along the point with maximal separation.

### G.5. Navigation Controls Problem

In this section, we consider the problem of continuous navigation of an agent. The agent's objective is to reach a goal while avoiding obstacles. The state space of the agent is $\mathcal{S} = [0, 20] \times [0, 10]$, which represents the $x$ and $y$ axis respectively. The agent navigates by taking actions on the velocity $v \in \mathbb{R}^2$, and the state evolves according to the dynamics $s_{t+1} = s_t + v_t T_s$ where $T_s = 0.1s$. We construct a square grid of points in the environment that are on the free space i.e. outside of the obstacles, and utilize Dijkstra's algorithm to find the shortest path for two starting $[1, 9]^T, [14, 1]^T$ positions, and goal $[19, 1]^T$ along the grid. For those two grid trajectories, we compute the optimal actions to be taken at each point in order to follow the trajectory.

The learner is equipped with both the labeled trajectories, as well as the unlabeled point grid. To leverage the manifold structure of the data, we consider the grid of points, and we construct the point-cloud Laplacian considering adjacent points in the grid. We train a two layer neural network using the mean square error loss over the optimal set of points and actions for ERM, ERM with ambient Lipschitz regularizer, Manifold Regularization, and our method Manifold Lipschitz method. To evaluate the performance, we randomly chose $100$ starting points and compute the trajectories generated by each learned function. A trajectory is successful if it reaches the goal without colliding with the obstacles or the walls. The results are summarized in table 3, and showcase the benefit of implementing manifold lipschitz. Our method outperforms the 3 other methods due to the fact that it minimizes the gradient of the function over the domain of the data. As opposed to ERM, our method generates a smooth function outside of the labeled trajectory. Ambient Lipschitz regularization fails due to the fact that the euclidean distance ignores the real distance between samples across wall, forcing similar outputs for points that should take different actions. Manifold regularization is able to capture the similarity between points, but it fails to properly capture the sharp turns near the edges off the obstacles. The success of Manifold Lipschtiz, can be explained by its dual variables $\lambda$ shown in figure 5c. In this figure, the radius of each ball represents the value of the dual variable, which is larger close to the corners of the obstacles due to the fact that the problem requires larger gradients to make sharp turns over it. Besides, the fact that we can disentangle the loss on the labeled data, from the Lipschitz constant, allows us to overfit the data as much as we require. As measure of merit, we take $100$ random points and we compute the trajectories. A trajectory is successful if it achieves the goal without colliding. The results are shown in Table 3, and the learned functions in Figure 5.

| Method | Trajectories |
|---|---|
| ERM | 85 |
| Ambient Reg. | 66 |
| Manifold Reg. | 77 |
| Manifold Lipschitz | 94 |

Table 3: Number of successful trajectories from $100$ random starting points.

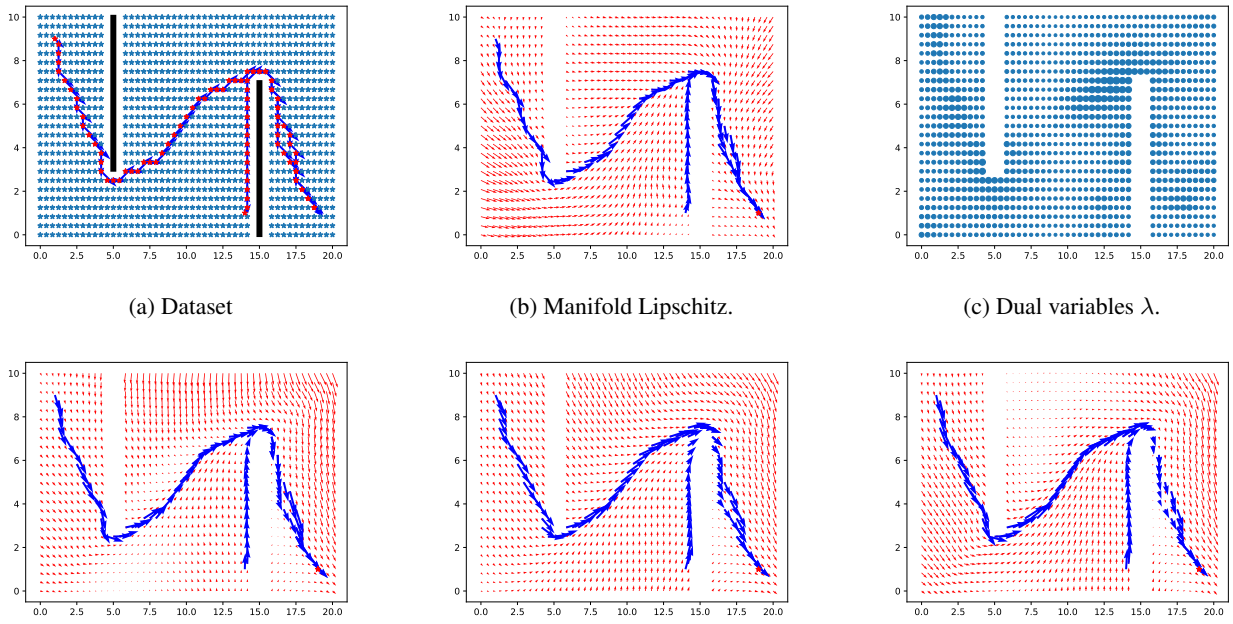
## H. Further References

Since we introduce the Lipschitz constant as a constraint to the learning problem our reformulation and solution methodologies are framed within the constrained learning paradigm (Chamon & Ribeiro, 2020; Chamon et al., 2023; Yang, 2019). Central to the solution of constrained learning problems is the use of dual formulations and dual ascent learning algorithms. These are finding increasing applicability as evidenced by their use in, e.g., adversarial robustness (Robey et al., 2021), graph neural networks (Cervino et al., 2022; Arghal et al., 2021), federated learning (Shen et al., 2021), active learning (Elenter et al., 2022), reinforcement learning (Paternain et al., 2019; 2022; Castellano et al., 2021; Bai et al., 2021; Hasanbeig et al., 2018), and wireless communications (Eisen et al., 2019).

In the context of adversarial attacks to neural networks, manifold based regularization techniques have shown a vast amount of empirical and theoretical evidence of its utility, improving its adversarial robustness(Zhang et al., 2021; Khoury & Hadfield-Menell, 2018; Ma et al., 2018; Moosavi-Dezfooli et al., 2019; Jin & Rinard, 2020; Lassance et al., 2021). Some works seek to obtain manifold attacks, which are more realistic attacks than utilizing the norm-$\infty$ ball, given the high dimensionality of the input and the low dimensional structure of the data (Stutz et al., 2019). Smooth function have also been studied in the context of robustness (Rosca et al., 2020; Bubeck & Sellke, 2021; Bubeck et al., 2021).

Our work, is based on previous results that show convergence of graph laplacians to Laplace-Beltrami operators. There exists a vast amount of work on that validates the convergence results for point-cloud operators over Manifolds (Hein et al., 2005; 2007; Dunson et al., 2021; Wu & Wu, 2018).

Regarding Lipschitz constant estimation for neural networks, (Fazlyab et al., 2019) has formulated the problem as a convex optimization problem.



(a) Dataset (b) Manifold Lipschitz. (c) Dual variables $\lambda$.

(d) ERM. (e) Ambient Regularization. (f) Manifold Regularization.

Figure 5: Figure 5a shows the training dataset, blue stars depict unlabeled point, and blue arrow the optimal action at the red star. Figure 5b shows the learned function using Manifold Learning, and 5c its associated dual variables associated. Figures 5d, 5e, 5f show the functions learned using ERM, ambient regularization, and Manifold regularization respectively.

# I. Ablation Study on Laplacian Construction

In this section we study the impact of the temperature coefficient $t$ in the construction of the Laplacian. To do so we repeat the setting of Figure 1, and we vary the value of the temperature $t$. We consider the two moons dataset problem with $1$ labeled, and $200$ unlabeled samples per class. We vary the value of the temperature coefficient, which varies the number of cross-manifold edges, and therefore makes the problem more challenging.

| Value of Heat Kernel | Connected Components | Number of Cross-Manifold Edges | Manifold Regularization | Manifold Gradient (Ours) |
|---|---|---|---|---|
| $0.0040$ | $2$ | $0$ | $100\%$ | $100\%$ |
| $0.0050$ | $1$ | $2$ | N/A | $100\%$ |
| $0.0060$ | $1$ | $3$ | N/A | $100\%$ |
| $0.0070$ | $1$ | $10$ | N/A | $100\%$ |
| $0.0080$ | $1$ | $20$ | N/A | $100\%$ |
| $0.0090$ | $1$ | $27$ | N/A | $100\%$ |
| $0.0100$ | $1$ | $42$ | N/A | $100\%$ |
| $0.0150$ | $1$ | $124$ | N/A | $100\%$ |
| $0.0175$ | $1$ | $176$ | N/A | $100\%$ |
| $0.0180$ | $1$ | $192$ | N/A | $100\%$ |
| $0.0190$ | $1$ | $217$ | N/A | $100\%$ |
| $0.0200$ | $1$ | $251$ | N/A | N/A |

Table 4: Ablation study on the temperature of the heat kernel $t$. We plot the accuracy when it achieves $100\%$, and N/A otherwise, given that an accuracy of less than $100\%$ is not representative as the method fails to capture the manifold structure of the problem.

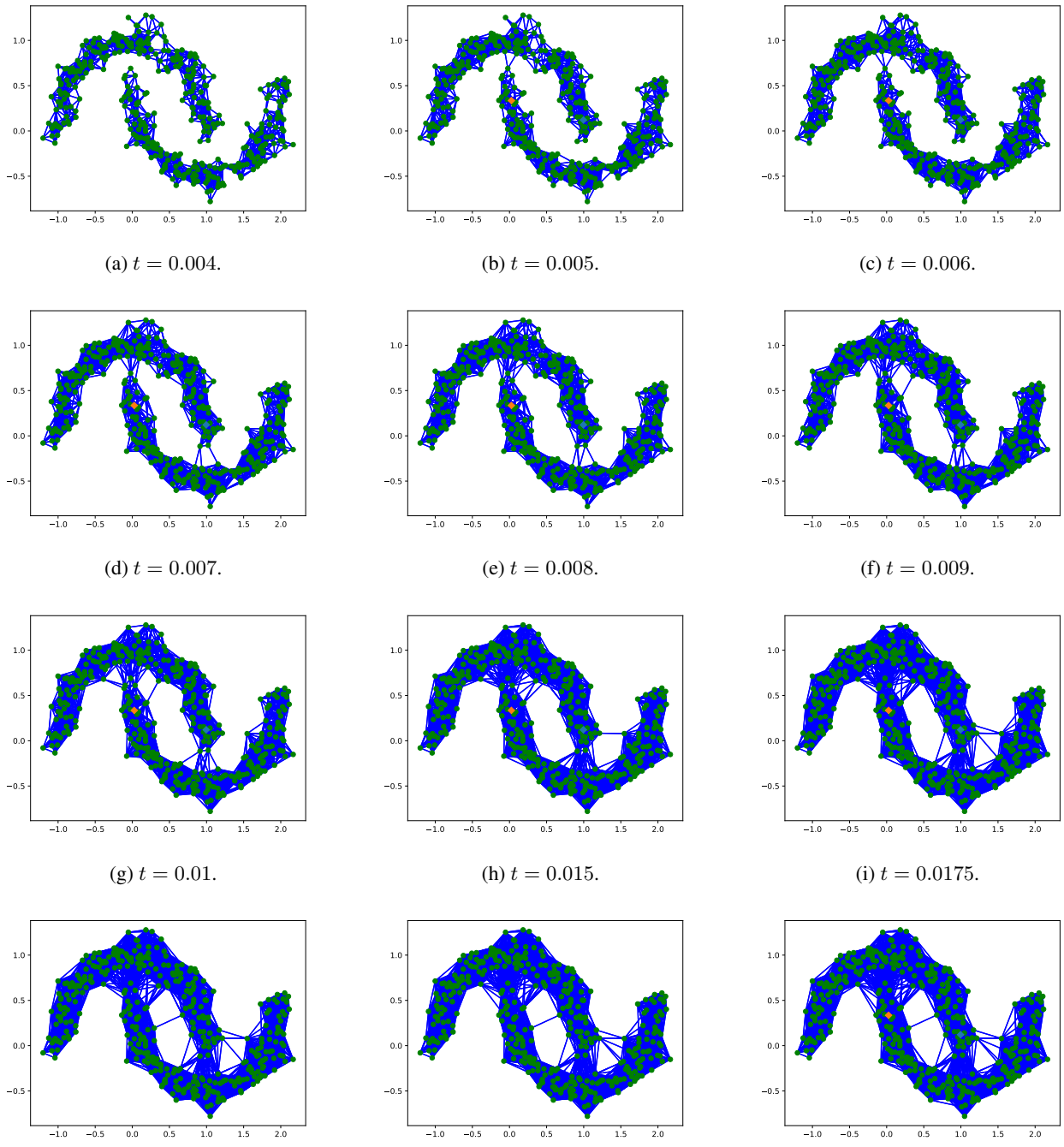
As seen in table 4, Laplacian regularization fails to achieve a perfect accuracy when the number of connected components is less than $2$. That is to say, manifold regularization achieves a perfect accuracy when each class has a connected component. However, once the components become connected, Laplacian regularization smoothness the integral of the gradient, and therefore does no properly identify the transition between components.

As can be seen in table 4, our method is more robust to non-exact manifolds. Which means that if the manifold is not calculated perfectly, and as a result we obtain $1$ connected component as opposed to $2$ separate moons, our method still works.

It is important to remark that our method still works when the connected components have cross-manifold edges in different places of the manifold. As an example, take the Laplacian with heat kernel $t = 0.007$, there are edges on both side of the manifold. Moreover, our method is able to distinguish between the two classes even with edges in the middle of the two manifold as can be seen with $t = 0.0150$, and $t = 0.0175$ (cf. 6h).

Our method brakes once the manifold structure vanishes and most of the points are interconnected, as can be seen with $t = 0.02$. In this case there are $251$ cross manifold edges, and the low-dimensional structure of the problem disappears.

In all, our proposed solution is more robust to imperfect estimation of the manifold. Even when the number of cross-edges is large, our method is able to create a partition between classes given that it finds the points with maximal separation, and allows the function to change values between them.



(a) $t = 0.004$. (b) $t = 0.005$. (c) $t = 0.006$.

(d) $t = 0.007$. (e) $t = 0.008$. (f) $t = 0.009$.

(g) $t = 0.01$. (h) $t = 0.015$. (i) $t = 0.0175$.

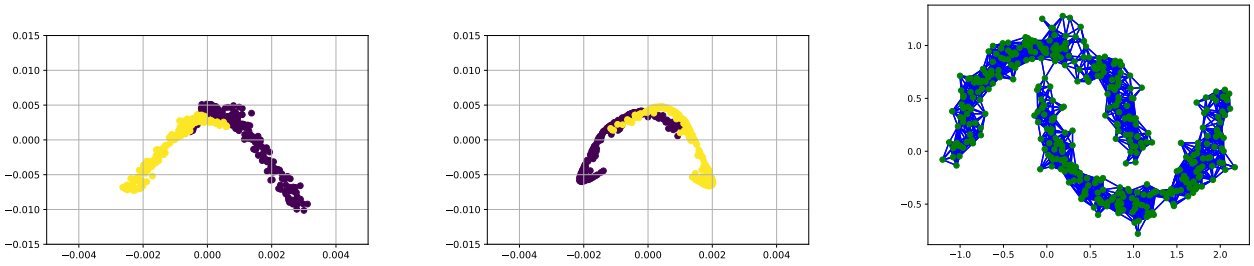
(j) $t = 0.018$. (k) $t = 0.019$. (l) $t = 0.02$.

Figure 6: Laplacian for different values of temperature coefficient $t$.

## J. Laplacian Spectral Embedding

In figure 7 we show the result of projecting the 2 moons dataset to the 2 eigenvector associated with the 2 smallest eigenvalues larger than 0 of the unweighted Laplacian $L$ in 7b, and the Laplacian multiplied by the dual variables $diag(\lambda)L$ in 7a. The dual variable $\lambda$ is obtained by utilizing our method and reaching $100\%$ accuracy.

If we compare the spectral embedding induced by the unweighted Laplacian (i.e., Laplacian eigenmap) and the re-weighted Laplacian for the two moons dataset, we notice that the latter displays better separation between the classes. It is worth noting, however, that this experiment is unfair in our favour, since the weighted Laplacian uses extra information from the primary task, namely, labeled data. Nevertheless, comparing the Laplacian spectral embeddings do provide another illustration as to the importance of considering the underlying data geometry (manifold) when imposing smoothness.




(a) Our Laplacian. (b) Unweighted Laplacian. (c) Dataset.

Figure 7: Laplacian Eigenmaps. In figures 7a,7b we project the points in the dataset to the 2 largest eigenvalues of the Laplacian $L$, and $diag(\lambda) * L$ respectively. The colors correspond to the classes predicted. In figure 1a we show the dataset.