SIAM J. MATH. DATA SCI.
Vol. 2, No. 3, pp. 705–739

A Maximum Principle Argument for the Uniform Convergence of Graph Laplacian Regressors*

Nicolas García Trillos† and Ryan W. Murray‡

Abstract. This paper investigates the use of methods from partial differential equations and the calculus of variations to study learning problems that are regularized using graph Laplacians. Graph Laplacians are a powerful, flexible method for capturing local and global geometry in many classes of learning problems, and the techniques developed in this paper help to broaden the methodology of studying such problems. In particular, we develop the use of maximum principle arguments to establish asymptotic consistency guarantees within the context of noise corrupted, nonparametric regression with samples living on an unknown manifold embedded in $\mathbb{R}^d$. The maximum principle arguments provide a new technical tool which informs parameter selection by giving concrete error estimates in terms of various regularization parameters. A review of learning algorithms which utilize graph Laplacians, as well as previous developments in the use of differential equation and variational techniques to study those algorithms, is given. In addition, new connections are drawn between Laplacian methods and other machine learning techniques, such as kernel regression and k-nearest neighbor methods.

Key words. empirical risk minimization, graph Laplacian, discrete to continuum, nonparametric regression

AMS subject classifications. 35J05, 49J55, 60D05, 62G08, 68R10

DOI. 10.1137/19M1245372

1. Introduction.

In this paper we present new theoretical results on the consistency of solutions to a family of variational problems that use graph Laplacian regularization for trend filtering and supervised learning with noisy labels, and determine scaling limits under which one can provably avoid overfitting as the number of data points grows. In addition, we draw new parallels between the methods studied here and other nonparametric regression methodologies found in the literature. Throughout this paper we highlight the analytical approach that we take in order to establish our high probability quantitative results, which, in particular, allows us to study the behavior of solutions to nonlinear graph-based equations. This analytical approach provides a new avenue for studying algorithms that utilize graph Laplacians; such algorithms are utilized in a wide variety of learning problems (see section 3.1 for further discussion).

Given a data set $X = \{x_1, \dots, x_n\}$ and corresponding *noisy* labels $y_1, \dots, y_n$, the idea in trend filtering is to reconstruct a trend function u taking inputs x into outputs y which closely matches the observed labels. Without further constraints, finding such a function is an ill-posed problem as there are many functions that will respect the observed data, most

*Received by the editors February 20, 2019; accepted for publication (in revised form) May 27, 2020; published electronically August 31, 2020.
https://doi.org/10.1137/19M1245372

†Department of Statistics, University of Wisconsin-Madison, Madison, WI 53711 USA (garciatrillo@wisc.edu).
‡Department of Mathematics, North Carolina State University, Raleigh, NC 27695 USA (rwmurray@ncsu.edu).

of which should be intuitively discarded as they overfit the observations y_i. To overcome this overfitting issue and to turn an ill-posed problem into a well-posed one, a popular idea used in applied mathematics [60] and statistics [66, 38] is to introduce a functional R (a "regularizer") which penalizes "irregular" functions, and to solve an optimization problem of the form

$$\min R(u) + F(u, y),$$

where $F(u, y)$ represents a loss function measuring empirical risk. In a classical statistical setting the loss function is dictated by the noise model, but, in general, the function $F(u, y)$ may simply be interpreted as a mismatch function between the observations and the regressor u. On the other hand, while the choice of regularizer is largely open ended, intuitively it should be selected to enforce "smoothness" with respect to some underlying geometry (informally, one wants to force the solution to not change too much by making small perturbations of the input).

For a generic set of points in Euclidean space, a popular choice of regularizer is the squared norm of a reproducing kernel Hilbert space (RKHS) as discussed in [4]. This approach is analogous to the use of ridge regression in classical statistics, and, in fact, can actually be rigorously cast in this way after transforming the data through a map canonically induced by the reproducing property of a kernel (see [53]).

As elegant and convenient as the use of RKHSs may be, a generic choice of kernel to be used within the above regularization procedure will typically be oblivious to the specific geometric configuration of a data set X. This is particularly problematic in settings like that of semi-supervised learning, where one hopes the underlying geometry of the data set reveals information that the limited number of available labels cannot. Motivated by this discussion, several authors, including those of [70, 57, 4, 1], consider regularizers which exploit the intrinsic geometric structure of a data set. In their set-up, a data set X comes with an additional weighted graph structure $\Gamma = (X, W)$, where W represents an $n \times n$ similarity matrix and which intuitively captures the geometry of X.

In this paper we study intrinsic regularization in the context of trend filtering and fully supervised learning. As in [57] the optimization problems we are interested in take the form

$$\min_u J(u), \qquad J(u) := \beta R_\Gamma(u) + \frac{1}{n}\sum_{i=1}^{n} F(u(x_i) - y_i), \tag{1.1}$$

where we have made the dependence of the regularizer R on the graph Γ explicit. The empirical risk that we consider here is directly linked to the log-likelihood of an *assumed* distribution $e^{-F(-s)}$ for the noise terms in an additive model of the form

$$y_i = \mu(x_i) + \xi_i, \tag{1.2}$$

where μ is some underlying ground-truth function. Notice that the *assumed* noise distribution may not be the same as the *actual* noise distribution which we will denote as $p(s)ds$. The standard choice for F is the square loss corresponding to a Gaussian model, and we will give particular attention to this choice later on. On the other hand, the regularizer R_Γ that we

will focus on in this paper is the graph *Dirichlet energy* defined by

$$R_\Gamma(u) := \frac{1}{2n}\sum_{i,j} w_{ij}|u(x_i) - u(x_j)|^2. \tag{1.3}$$

The relevance of the graph Dirichlet energy is due to its close connection to the linear in u *graph Laplacian*

$$\Delta_\Gamma u(x_i) := \sum_{j=1}^{n} w_{ij}(u(x_i) - u(x_j)), \quad x_i \in X; \tag{1.4}$$

indeed, it is straightforward to show that

$$R_\Gamma(u) = \langle \Delta_\Gamma u, u\rangle_X.$$

The graph Dirichlet energy is a special choice of intrinsic regularizer with several properties that will be discussed throughout this paper. Other sensible intrinsic regularizers are the graph p-Sobolev norms (obtained by changing the square in the above energy with a pth power) for some $p \geq 1$ (in particular, when $p > m$ where m is the intrinsic dimension of the data set; see [22]). The practical use of Laplacian regularization (or similar variants) was proposed some years ago [57], but has recently received new attention in the statistics and machine learning community [52, 59, 67]. Bayesian approaches to learning where graph Laplacians are used as covariance matrices in order to define Gaussian priors that exploit the intrinsic geometry of the data set have been considered in [70, 41, 5]. We notice that from a Bayesian perspective, the solution of problem (1.1) is the MAP estimator (maximum a posteriori estimator) for u, in a model where the unknown variable u is assumed to have a Gaussian prior distribution with mean zero and covariance matrix equal to β times the graph Laplacian, and where the observations y depend on u according to an assumed additive noise model of the form (1.2) with noise distribution $e^{-F(-s)}$.

Going back to our problem (1.1), we notice that the graph Laplacian appears explicitly in the optimality conditions satisfied by the minimizer of (1.1), namely

$$\beta\Delta_\Gamma u + f(u - y) = 0, \tag{1.5}$$

where $f = F'$. Equation (1.5) can be interpreted as an elliptic graph partial differential equation (PDE), which is linear in its highest-order term, i.e., the term that involves "derivatives" of u (in this case the graph Laplacian $\Delta_\Gamma u$), but that overall is not linear unless the loss function F is quadratic (so that its derivative is linear). The most appropriate terminology for such an equation is *semilinear equation*. This graph PDE, which emerges in a fully supervised learning setting, is nothing but one of many examples in the family of graph PDE–based machine learning methodologies. Indeed, there is a large family of machine learning methodologies for supervised and unsupervised learning that, at their core, are described by a graph PDE involving the graph Laplacian, the fractional graph Laplacian (i.e., powers of the graph Laplacian), or the p-graph Laplacian, followed by an extension step (in supervised settings) or a clustering step (e.g., k-means after the embedding step in spectral clustering). Many of

these approaches will be mentioned in section 3.1. This paper aims at studying the statistical properties of solutions to (1.5), with an emphasis on the analytical techniques that allow us to handle the nonlinear term in (1.5). More concretely, we aim at answering the following questions:

(i.) What is the behavior of the solution to (1.1) (alternatively solutions to the graph PDE (1.5)) as the number of data points goes to infinity, when the data set is assumed to be a collection of i.i.d. samples from a distribution supported on an m-dimensional curved manifold $\mathcal{M}$ (here we intuitively think of $m \ll d$, although this will not be important for our analysis), the graph is obtained from the data set by giving high weights only to points that are within Euclidean distance ε of each other, and the labels are *noisy* versions of a hidden trend function μ?

(ii.) How should β scale with n so that the solution to (1.1) converges to the Bayes regressor (i.e., the underlying label trend) in the large data limit $n \to \infty$, and, in particular, so that the regression procedure does not overfit the data?

To answer these questions we exploit properties of the graph Laplacian, and, in particular, a *maximum principle* argument at the graph level to prove that the solution of (1.5) converges *uniformly* towards the solution of an analogous homogenized partial differential equation on $\mathcal{M}$ with probability one (homogenized in x and y). We provide explicit rates of convergence with high probability (w.h.p.) in terms of the number of samples n, the parameter ε controlling data connectivity in the graph, and the parameter β controlling the strength of regularization; our rates turn out to be essentially minimax optimal (see the discussion at the end of section 3.1.2). The proposed maximum principle (see Proposition 2.2) allows us to handle any sufficiently smooth, strongly convex F. We also provide a characterization for how β must scale with n in order to recover, in the large data limit, a modified trend μ_f which depends on μ and on the function f. Indeed, unless the function F is quadratic (i.e., f is linear), in the regime $n \to \infty$, $\beta := \beta_n \to 0$ the trend μ may not be recovered, unless further assumptions on the distribution of the noisy labels are imposed. We provide uniform rates of convergence towards the modified trend. Stated in another way, we provide quantified asymptotic consistency estimates for this class of nonparametric regression algorithms. Our estimates will be decomposed into sample error (analogous to variance) and approximation estimates (analogous to bias). Our arguments to bound the sample error are essentially as difficult for the quadratic F as for the general F (and the maximum principle and monotonicity of $f = F'$ will be key to show this). For the approximation part, we will separate the analysis for the quadratic case from the general case, as a richer set of tools is available in the quadratic case.

We would like to highlight that our results are novel with respect to the existing literature in several regards that we summarize below.

- We show that the solution to (1.1) (alternatively (1.5)) converges towards the solution of an analogous limiting variational problem. This contrasts with several results in the literature that establish a connection of the graph Laplacian and the negative Laplace–Beltrami operator $\Delta_{\mathcal{M}}$ from the point of view of *pointwise convergence*, where one fixes a regular enough function $h : \mathcal{M} \to \mathbb{R}$ and establishes convergence rates for

$$\Delta_{\mathcal{M}} h - \Delta_{\Gamma} h.$$

  Some of these results include [3, 55, 37, 16, 31]. We remark that such results do not

allow one to immediately deduce convergence of solutions to optimization problems on graphs. In this paper we show how one can bootstrap the pointwise consistency estimates to obtain convergence of solutions to optimization problems on graphs that involve the graph Laplacian, using ideas from PDE theory. In particular, the regularity of the solution to a limiting PDE will be important when using the pointwise consistency of the graph Laplacian. In addition, a very careful construction of upper and lower *barrier functions* for the solution of (1.5), and the use of a maximum principle at the graph level will be crucial tools in order to handle the randomness in the data points and their noisy labels. For more details, see section 2.2.

- While there are many recent results studying the consistency of solutions to optimization problems on graphs towards solutions to variational problems and PDEs on $\mathcal{M}$ (see section 3.1.1 for a discussion), most of these results do not obtain rates of convergence. Here we use the extra structure of the graph Laplacian to obtain rates for the uniform convergence of solutions to (1.5). We emphasize that we obtain *uniform* (i.e., L^∞) rates of convergence instead of L^2 type or other type of averaged error, as is most commonly found in the literature. We will compare our work with the few exceptions to this general absence of quantitative estimates, which are usually obtained with different assumptions from ours, particularly with regard to the next point.
- In our set-up the labels y_i are assumed to be noisy versions of a hidden trend function μ so that

$$y_i = \mu(x_i) + \xi_i.$$

  We show that, provided that the regularization parameter β in (1.5) does not decay too fast to zero (and we characterize this precisely), the solution to (1.5) converges to some (modified) trend function in the limit. In particular, we show that graph Laplacian regularization is indeed capable of removing noise and prevents overfitting.
- We connect the Laplacian regularization stemming from (1.5) with other nonparametric methodologies for regression based on local averaging, which adapt to the local geometry of the underlying manifold (e.g., k-nearest neighbor (NN) regressors). This discussion is presented in section 3.1.2.
- We emphasize that our approach allows us to handle loss functions F that are different from quadratic. As mentioned earlier, from a statistical point of view this gives us some flexibility in the assumptions made on the observations. From a technical point of view, the effect that different F's have in the resulting graph PDE is the difference between having a linear PDE (in case F is quadratic) and having a nonlinear PDE (if F is not quadratic) where in principle the analysis is more difficult.

We notice that the observed data (x_i, y_i) affects (1.5) in two different ways: first, the x_i determine the graph Laplacian and in this way the "differential" operator that we study is random. Second, the noisy label observations y_i affect the right-hand side of the equation. At a high level (1.5) can be thought of as a numerical scheme for solving the limiting continuum PDE that we derive. However, the convergence analysis of such a numerical scheme is not standard in the literature, given the randomness in *both* the approximating Laplacian (in our case the graph Laplacian) and the noisy right-hand side. It is precisely the presence of the regularizer that removes noise in the limit and helps to prevent overfitting. In summary,

when going from (1.5) to the limiting PDE, there are two parts in the equation that get simultaneously homogenized: on the one hand the graph Laplacian behaves like a Laplacian operator on $\mathcal{M}$, and on the other hand, the noisy labels homogenize to some trend function.

We now outline the remainder of the work. In section 2 we give a precise description of the problem and main results. In section 3 we review related literature, comparing our results and techniques with other work. In section 4 we present the proofs of our results.

2. Problem set-up and main results.

2.1. Set-up. We consider an m-dimensional smooth, compact manifold $\mathcal{M}$ embedded in $\mathbb{R}^d$ with no boundary. A significant body of work studies how to estimate m given a family of points (see, e.g., [43]); for the purpose of this work we will treat m as a priori known. Throughout this paper we will denote by $|x - \tilde{x}|$ the Euclidean distance in $\mathbb{R}^d$ and by $d_{\mathcal{M}}(x, \tilde{x})$ the geodesic distance of two points in $\mathcal{M}$. We will denote by i_0 the injectivity radius of the manifold $\mathcal{M}$. We recall that the injectivity radius of a manifold is defined as the maximum radius for which the exponential map $\exp_x : B_m(0, i_0) \subseteq \mathcal{T}_x\mathcal{M} \to B_{\mathcal{M}}(x, i_0)$ defines a diffeomorphism for all $x \in \mathcal{M}$ (see [20]). In the remainder we use B_m to denote balls in $\mathbb{R}^m$, $B_{\mathcal{M}}$ for balls in $\mathcal{M}$ with the geodesic distance, and finally B for balls in $\mathbb{R}^d$. Finally, we will denote by $dvol_{\mathcal{M}}$ the volume form of $\mathcal{M}$ and, after rescaling as necessary, we will assume that the volume of $\mathcal{M}$ is equal to one.

In the remainder, and unless otherwise stated, we will assume that $x_1, \dots, x_n$ are i.i.d. samples from the uniform distribution on $\mathcal{M}$. We let $\mu : \mathcal{M} \to \mathbb{R}$ be an unknown trend function and assume that the labels $y_1, \dots, y_n$ are obtained by

$$y_i = \mu(x_i) + \xi_i, \tag{2.1}$$

where the ξ_i are identically distributed with distribution $p(s)ds$, are independent of each other and from the x_i, and satisfy

$$\mathbb{E}(\xi_i) = 0, \quad |\xi_i| \leq \sigma,$$

where σ is a fixed constant specifying the noise level. We have chosen to focus on the bounded noise case in order to simplify some of the technical arguments, in particular the construction and bounds of the barrier functions y^- and y^+ in the proof of Theorem 2.1. We believe that many of the arguments and results could be modified to accommodate other noise models, but we do not pursue it here.

We assume that the loss function $F : \mathbb{R} \to [0, \infty)$ is a smooth function which satisfies $F(0) = 0$ and $F''(x) > 0$ for all x; in some references this is called *strong convexity*. In particular, this implies that $f := F'$ is a strictly monotone function, and that $f'(x) > 0$ for all $x > 0$ and $f'(x) < 0$ for all $x < 0$.

All the theorems presented in section 2.2 are stated in the setting described above, but in Remark 2.6 we write precisely how they should be restated to cover a more general setting. No difficulties arise when extending these results, other than having to deal with more cumbersome notation and longer expressions.

By way of notation, $\|h\|_{C^k} = \sum_{i=1}^{k} \|D^i h\|_{C^0}$, where C^0 is the space of continuous functions equipped with the supremum norm, and where when we write the $\|D^i h\|_{C^0}$ we mean that we are summing the norm of all of the mixed derivatives of order i. For $0 < \alpha < 1$, spaces of

Hölder continuous function C^α are continuous functions equipped with the norm

$$\|h\|_{C^\alpha} = \sup_{x,y\in\mathcal{M}} \frac{|h(x)-h(y)|}{|x-y|^\alpha} + \|h\|_{C^0}.$$

The space $C^{k,\alpha}$ is given by the norm $\|h\|_{C^{k,\alpha}} = \|h\|_{C^k} + \|D^k h\|_{C^\alpha}$. For any natural number r, the function space $H^r(\mathcal{M})$ is the defined by the norm $\|h\|_{H^r} = \|h\|_{L^2(\mathcal{M})} + \|D^r h\|_{L^2(\mathcal{M})}$.

2.1.1. Graph construction and graph PDE. Given $X := \{x_i\}_{i=1}^n$ we construct a geometric graph as follows. We let $\eta : [0,\infty) \to [0,\infty)$ be a nonincreasing function which is only nonzero on $[0,1]$. We further assume η to be Lipschitz continuous and normalized so that

$$\int_{\mathbb{R}^m} \eta(|z|)\, dz = 1.$$

We define the constant

$$\tau_\eta := \int_{\mathbb{R}^m} |z_1|^2 \eta(|z|)\, dz, \tag{2.2}$$

where z_1 is the first coordinate of z. Between every two vertices $x_i, x_j \in X$ we assign the weight

$$w_{i,j} = \frac{2}{\tau_\eta \varepsilon^{m+2} n} \eta\left(\frac{|x_i - x_j|}{\varepsilon}\right). \tag{2.3}$$

Here the ε^m in the denominator is a rescaling so that the weights at each vertex sum to approximately one, while the ε^2 in the denominator is chosen so that $\frac{|u(x_i)-u(x_j)|^2}{\varepsilon^2} \sim |\nabla u|^2$. The weighted graph (X, w) is a geometric graph representing the proximity of the sample points x_i in $\mathbb{R}^d$. We have rescaled the weights for convenience (in taking limits as $n \to \infty$).

2.1.2. Limiting variational problem and PDE. At the continuum level, we first define an analogue of the graph regularizer R_Γ. The Dirichlet energy of a function $v : \mathcal{M} \to \mathbb{R}$ is defined as

$$R_{\mathcal{M}}(v) := \int_{\mathcal{M}} |\nabla v|^2 dvol_{\mathcal{M}},$$

whenever v is in the Sobolev space $H^1(\mathcal{M})$. Also, for a smooth function v we define the elliptic operator $\Delta_{\mathcal{M}}$ as

$$\Delta_{\mathcal{M}} v := -div(\nabla v),$$

i.e., the negative of the Laplace–Beltrami operator on $\mathcal{M}$. It is straightforward to show (see section 4.1) that the Euler–Lagrange equation associated to the variational problem

$$\min_v \left\{ \beta R_{\mathcal{M}}(v) + \int_{\mathcal{M}} \int_{\mathbb{R}} F(v(x) - \mu(x) - s) p(s) ds dvol_{\mathcal{M}}(x) \right\} \tag{2.4}$$

is the PDE

$$\beta \Delta_{\mathcal{M}} v + \int_{\mathbb{R}} f(v - \mu - s) p(s) ds = 0. \tag{2.5}$$

Equation (2.5) is the continuum "homogenized" analogue of the graph PDE (1.5). Existence and uniqueness as well as regularity properties of the solution to (2.5) are discussed in Theorem 2.3. We remind the reader that $p(s)ds$ is the actual distribution function of the label noise.

When passing from the graph PDE (1.5) to the PDE (2.5), we notice that there are two terms that get homogenized. On the one hand, as more feature vectors x_i are available, the graph Γ gets denser, and the graph Laplacian Δ_Γ starts behaving like $\Delta_{\mathcal{M}}$. On the other hand, as more labels y_i are acquired, we would like to obtain homogenization of the fidelity in (1.1). In the next section we present our main results relating the solutions to these two equations, i.e., (1.5) and (2.5).

2.2. Main results and discussion. Our first main result establishes probabilistic error bounds for

$$\max_{i=1,\dots,n} |u(x_i) - v(x_i)|,$$

where u is the solution to the graph PDE (1.5) (with the graph Γ as defined in section (2.1.1)), and v is the solution to (2.5). That is, we estimate the difference between the data dependent u and a homogenized function v.

Theorem 2.1 (sample error). *Suppose that u is the solution to the elliptic graph PDE (1.5) where Δ_Γ is defined in (1.4) and v is the solution to the PDE (2.5). Assume that $\mu \in C^2(\mathcal{M})$. Then for any $\delta, \zeta > 0$, with probability at least $1 - 4n\exp\big(-\frac{n\delta^2\varepsilon^{m+2}}{C(1+\varepsilon\delta)}\big) - 4n\exp(-Cn\varepsilon^m\zeta^2) - 4n\exp(-Cn\varepsilon^m)$,*

$$\max_{i=1,\dots,n} |u(x_i) - v(x_i)| \leq C\left(\frac{\varepsilon^2}{\beta} + \zeta + \beta\delta + \beta^{1/2}\varepsilon\right),$$

where the constants C depend only on μ, η, F, and $\mathcal{M}$.

One of the implications of the above result is that for β fixed, it is possible to tune ε, δ in terms of n appropriately to deduce asymptotic uniform convergence of solutions of (1.5) towards solutions of (2.5).

One of the main tools used to establish Theorem 2.1 is the following maximum principle at the graph level. As the proof is straightforward, we present it immediately.

Proposition 2.2 (maximum principle at the graph level). *Let $g : \mathbb{R} \to \mathbb{R}$ be a strictly increasing function, and let $h \in L^2(X)$ be an arbitrary function defined on the point cloud X. Suppose that the function $z : X \to \mathbb{R}$ satisfies*

$$-\beta\Delta_\Gamma z - (g(z+h) - g(h)) \geq 0.$$

Then,

$$z \leq 0,$$

i.e., the function z is nonpositive.

Proof. Notice that to prove that the function z is nonpositive, it is enough to show that $z(x_i) \leq 0$, where i is the index of the point x_i at which z is maximized. Now, we notice that

at the point x_i we have

$$\Delta_\Gamma z(x_i) = \sum_{j=1}^{n} w_{ij}(z(x_i) - z(x_j)) \geq 0.$$

It then follows that

$$-(g(z(x_i) + h(x_i)) - g(h(x_i))) \geq 0,$$

or equivalently,

$$g(z(x_i) + h(x_i)) \leq g(h(x_i)).$$

Since the function g is strictly increasing, we conclude that $z(x_i) \leq 0$, which concludes the proof. ∎

This proof is an adaptation of the proof of the maximum principle in the continuum case (see, e.g., [23, p. 344]), with the modification that we have replaced differential operators with derivatives. We also have used the fact that we have a strictly monotone lower order term, which is often not the case in the classical setting.

Theorem 2.1 is proved by showing that the difference of the functions u and v (interpreting v as its restriction to X) lies between two functions y^-, y^+ (referred to as barrier functions) which are uniformly close to zero, i.e.,

$$y^-(x_i) \leq u(x_i) - v(x_i) \leq y^+(x_i) \quad \forall i = 1, \ldots, n, \tag{2.6}$$

with

$$\|y^-\|_\infty, \|y^+\|_\infty \ll 1.$$

Up to some minor modifications, the functions y^-, y^+ take the form

$$y_i^+ \sim \frac{\varepsilon^2}{\beta}\left(\int_{\mathbb{R}} f(v_i - \mu_i - s)p(s)ds - f(v_i - \mu_i - \xi_i)\right) + \rho,$$
$$y_i^- \sim \frac{\varepsilon^2}{\beta}\left(\int_{\mathbb{R}} f(v_i - \mu_i - s)p(s)ds - f(v_i - \mu_i - \xi_i)\right) - \rho,$$

where ρ is a constant conveniently chosen so as to ensure that the functions $z^+ := (u-v) - y^+$ and $z^- := y^- - (u - v)$ satisfy the inequality required for the maximum principle at the graph level to apply with $g \equiv f$ (which then implies (2.6)). We remind the reader that $v_i = v(x_i)$ represents pointwise evaluations of the limiting PDE (2.5), $\mu_i = \mu(x_i)$ are pointwise evaluations of the limiting trend function, and ξ_i are the values of the noise. In order to guarantee that ρ can be picked to be small (so that y^-, y^+ are indeed uniformly small), we need an estimate for the difference between $\Delta_\Gamma v$ and $\Delta_\mathcal{M} v$ where v is the solution to (2.5). Such a pointwise estimate relies strongly on the regularity of the function v. The necessary regularity for the function v in turn follows from the regularity theory of solutions to elliptic PDEs (see the last part of Theorem 2.3). Apart from the pointwise estimates $\Delta_\Gamma v - \Delta_\mathcal{M} v$, the other probabilistic estimates that we need are used to control the expressions that appear when we apply the graph Laplacian to the functions y^- and y^+. After some cancellations and standard concentration inequalities, these remaining terms can be shown to be small. We

emphasize that it is precisely our convenient construction of the barrier functions y^-, y^+, and the structure of the graph Laplacian which ultimately allows us to handle the randomness in our problem, and bootstrap basic pointwise consistency results into convergence rates for solutions to optimization problems on random geometric graphs.

After establishing error estimates for the difference between u and v, we obtain estimates for the difference between v and a modified trend μ_f defined implicitly by

$$\int_{\mathbb{R}} f(\mu_f(x) - \mu(x) - s)p(s)ds = 0 \quad \forall x \in \mathcal{M}. \tag{2.7}$$

The function μ_f is the solution to (2.4) when the regularizer has been turned off (i.e., when $\beta = 0$). Notice that when the loss function F has the form $F(t) = \frac{1}{2}t^2$ the modified trend coincides with μ. This is also the case for general F when the noise distribution $p(s)ds$ is assumed to be symmetric. Moreso, if the actual noise distribution p coincides with the assumed noise distribution $e^{-F(-s)}$, then $\mu_f = \mu$. To see this notice that in that case

$$\int_{\mathbb{R}} f(-s)p(s)ds = \int_{\mathbb{R}} F'(-s)e^{-F(-s)}ds = \int_{\mathbb{R}} \frac{d}{ds}e^{-F(s)}ds = 0,$$

which implies that if we take $\mu_f = \mu$, then μ_f indeed solves (2.7).

The following result is obtained using tools from the theory of PDE.

Theorem 2.3 (regularity and approximation error). *Let μ_f be the solution to (2.7). Then, there exists a unique $v \in C^2(\mathcal{M})$ which solves the PDE (2.5). Furthermore, for β sufficiently small, this function satisfies*

$$\sup_{x\in\mathcal{M}} |v(x) - \mu_f(x)| \leq \frac{\beta\|\Delta_{\mathcal{M}}\mu\|_\infty}{c_1}, \tag{2.8}$$

where $f'(t) > c_1$ for $t \in [-\|\mu_f\|_\infty, \|\mu_f\|_\infty]$. Furthermore, assuming that $\|\mu\|_{C^2} < \infty$, then

$$\|v\|_{C^2} \leq C, \quad \|v\|_{C^3} \leq C\beta^{-1/2}, \quad \|v\|_{C^4} \leq C\beta^{-1}, \tag{2.9}$$

where here C is independent of β.

We recall that the definitions of these norms is given in section 2.1. We can combine Theorems 2.1 and 2.3 and deduce the following error estimates between our regression function u constructed by solving 1.5 and the modified trend μ_f.

Theorem 2.4. *Under the same assumptions in Theorem 2.1 and using the same notation there as well as that in Theorem 2.3, with probability greater than $1 - 4n\exp\big(-\frac{n\delta^2\varepsilon^{m+2}}{C(1+\varepsilon\delta)}\big) - 4n\exp(-Cn\varepsilon^m\zeta^2) - 4n\exp(-Cn\varepsilon^m)$, we have*

$$\max_{i=1,\dots,n} |u(x_i) - \mu_f(x_i)| \leq C\left(\beta + \frac{\varepsilon^2}{\beta} + \zeta + \beta\delta + \beta^{1/2}\varepsilon\right).$$

In particular, choosing δ of order one and $\zeta = \beta = \varepsilon$ we have that with probability larger than $1 - Cn\exp(-Cn\varepsilon^{m+2})$ we have $\max|u(x_i) - \mu_f(x_i)| \leq C\varepsilon$.

We note that as long as $\left(\frac{\log(n)}{\delta^2 n}\right)^{\frac{1}{m+2}} \ll \varepsilon \ll \beta^{\frac{1}{2}} \ll 1$, then the previous theorem gives asymptotic consistency. We emphasize that this theory provides clear asymptotic ranges of parameters for which consistency is achieved, and can be used to identify ranges of parameters where overfitting *does not* occur in the large n limit. In particular, taking δ to be of order one, and $\zeta = \beta = \varepsilon = C\left(\frac{\log(n)}{n}\right)^{1/m+2}$ for some large enough constant, we obtain, with high probability, an overall error of estimation of

$$\left(\frac{\log(n)}{n}\right)^{1/m+2}$$

which is known to be essentially minimax optimal for the recovery of the trend function. We will discuss this further in section 3.1.2.

One can cast the estimates in Theorems 2.1 and 2.3 in terms of *sample error* and *approximation error* estimates in the following sense: the first theorem gives an estimate on the random variation (in an L^∞ norm) one sees when comparing the solution of the (random) optimization problem (1.1) and the associated mean or homogenized problem (2.4). The second theorem gives an estimate between the solution of the homogenized problem (2.4) and the Bayes regressor, or in other words it provides an L^∞ estimate on the bias induced (in the homogenized limit) by the regularizing term.

We conclude this section by making a few remarks of a technical nature.

Remark 2.5. Our results can be generalized in a straightforward way to the case where the trend function μ is smooth everywhere except on a regular $m-1$ dimensional discontinuity set D_μ . In such a case we can obtain similar error bounds for the difference between the solution to the graph PDE and the solution to the continuum PDE. Such error bounds are uniform away from the discontinuity set D_μ. The reason for this is that most of our estimates are local, and even those that are not only involve averaging at the length-scale ε. It is not clear how to utilize our techniques, which are strongly grounded in the theory of elliptic PDE, to settings where the trend function is highly irregular.

Remark 2.6 (a more general statistical model). As was mentioned in section 2.1, although we state our main results assuming the data $x_1, \dots, x_n$ to be uniformly distributed in $\mathcal{M}$ and the ξ_i to be identically distributed, it is completely straightforward to extend them to a more general setting. In particular, one can suppose that

$$y_i = \mu(x_i) + \xi_i,$$

where $(x_1, \xi_1), \dots, (x_n, \xi_n)$ are samples from a joint distribution $\boldsymbol{\gamma}$ with a smooth marginal density for x denoted by γ, conditional distributions for the noise, which vary smoothly in x, of the form

$$\mathbb{P}(\xi \in ds | x) = \gamma_x(s)ds,$$

which are further assumed to be centered and to have bounded support. Then, Theorems 2.1, 2.3, and 2.4 continue to be true if we now let v be the solution to the PDE

$$\Delta_\gamma v(x) + \int_{\mathbb{R}} f(v(x) - \mu(x) - s)\gamma_x(s)ds = 0, \quad x \in \mathcal{M},$$

where

$$\Delta_\gamma v := -\frac{1}{\gamma}\mathrm{div}(\gamma^2 \nabla v),$$

and if we let μ_f be the function that satisfies

$$\int_{\mathbb{R}} f(\mu_f(x) - \mu(x) - s)\gamma_x(s)ds = 0$$

for all $x \in \mathcal{M}$.

In this paper we have also assumed the noise in labels y_i to be bounded. While our current proofs, in particular the construction of the barrier functions y^+ and y^-, do not allow us to directly drop this assumption, they do serve as a basis for future improved results. In a similar way, it is likely that the assumptions we have made on the loss function F can be relaxed further.

Remark 2.7 (constructing out of sample regressors). Since the regression function u obtained by solving (1.5) is only defined at the data points $x_1, \dots, x_n$, one needs a mechanism to extend u to the whole $\mathbb{R}^d$ in order to use it for prediction. A simple extension mechanism proposed in Chapter 5 in [70] is defined as follows: for an arbitrary $x \in \mathbb{R}^d$ we consider

$$\boldsymbol{u}(x) := \sum_{i=1}^{n} u_i \mathbf{1}_{V_i}(x),$$

where the $\{V_i\}_{i\in\mathbb{N}}$ is the Voronoi tessellation in $\mathbb{R}^d$ induced by $\{x_1, \dots, x_n\}$, that is,

$$V_i := \{x \in \mathbb{R}^d \ : \ |x - x_i| \leq |x - x_j| \quad \forall j = 1, \dots, n\}.$$

We use the above definition for points x that belong to a single Voronoi cell, and define $\boldsymbol{u}(x)$ to be the average of the $u(x_i)$ associated to the cells V_i that x belongs to. In other words, $\boldsymbol{u}$ is the 1-NN extension of u.

With the L^∞ bounds between u and μ_f that we have derived in our main theorems one can now show that $\boldsymbol{u}$ is uniformly close to μ_f when restricted to $\mathcal{M}$. Let us outline the argument. Take for simplicity a point $x \in \mathcal{M}$ which belongs only to V_i. Then, the triangle inequality implies that

$$\begin{aligned} |\boldsymbol{u}(x) - \mu_f(x)| &\leq |u_i - \mu_f(x_i)| + |\mu_f(x_i) - \mu_f(x)| \\ &\leq \sup_{j=1,\dots,n} |u_j - \mu_f(x_j)| + Lip(\mu_f)|x - x_i| \\ &\leq \sup_{j=1,\dots,n} |u_j - \mu_f(x_j)| + Lip(\mu_f)\mathrm{diam}_{\mathcal{M}}(V_i). \end{aligned}$$

We notice that the term $\sup_{j=1,\dots,n} |u_j - \mu_f(x_j)|$ is controlled in Theorem 2.4. In turn the diameter of Voroni cells constructed from random samples can be controlled following [19].

Remark 2.8. One direction of research which is worth further exploration is to study how these ideas can be used to address questions similar to the ones explored in this paper in the context of graph models Γ different from geometric graphs. An example of such a model is

the stochastic block model where points $x_1, \ldots, x_n$ have no geometric meaning and weights are determined randomly based on a probabilistic rule. We note that large n behavior of the spectra of graph Laplacians for graphs generated from a stochastic block model has been studied in [50].

3. Literature review and comparison of techniques. In this section we provide a review of relevant literature, beginning with statistical and machine learning literature related to Laplacian regularization. We then discuss recent analytical advances related to graph-based regularization. We then provide some comparison and discussion between the methods we study here and kernel and k-NN methods. Finally, we give a comparison of our analytical results with purely variational results, and demonstrate that the maximum principle method obtains more precise bounds.

3.1. Laplacians, geometry, and regularization. There is a significant literature related to the use of Laplacians in statistical problems. Here we complement (in a nonexhaustive way) the discussion started in the introduction.

The use of derivative penalties in parametric regression was advocated in the work of Wahba [66]. In that context, one seeks for a polynomial spline of specified order which interpolates observed data points, and which minimizes some derivative penalty, such as the continuum Laplacian. Around 2000, a significant body of research developed the use of Laplacians to capture intrinsic geometry in statistical tasks. Often this work was carried out with the aim of designing methods which can leverage geometry to handle settings which are not fully supervised, or where there is an active component to learning. For example, [8] investigated the use of density weighted regularizers in nonparametric regression problems. Reference [4] uses Laplacian eigenfunctions to conduct semi-supervised learning (so does [69, 70]). Reference [2] uses the Laplacian to construct dimension reduction algorithms. Reference [47] introduced spectral clustering, which uses the first nontrivial eigenvector of the Laplacian to construct meaningful clusters; theoretical consistency of these types of algorithms was studied in [65]. An algorithm linking graph cuts and spectral clustering was proposed in [40]. Reference [17] studies a method which uses Laplacian regularized regression to fit labeled points, and then k-NN regression on any unlabeled or new data points. Reference [57] studies the use of the Laplacian to construct reproducing kernels on graphs. We remark that much of this work focuses on graph-based algorithms. We also remark that much of this work was developed simultaneously with kernel methods, and the two are often mixed in these works.

Another related body of work focuses on diffusion maps and manifold learning. These techniques describe the use of Laplacian eigenfunctions as a powerful parametric family which encapsulates underlying geometry. Important early works establishing theory for this topic include [15, 16].

More recently, there has been renewed interest, especially in the statistics community, in utilizing derivative and difference-based methods on graphs in order to conduct regression tasks. In particular, *trend filtering* [59, 67] seeks to conduct regression tasks on a graph with an added regularization term based on function differences. This regularization term is permitted to take different forms in trend filtering, including powers of the Laplacian or TV norms. While the set-up is not identical to ours, in particular due to the use of higher-order

differences and ℓ^1-type regularizers, it is very much in the spirit of the model that we study here. The use of PDE-type methods to analyze these more complicated trend filtering models is the topic of future work. In a related work, [32] studies the robustness of semi-supervised methods, and proposes using sparsity-type penalties on graph derivatives similar to those from trend filtering. References [24, 34] study models for active learning which build upon the graph Laplacian framework in order to identify regions where labels should be acquired.

In addition, there has recently been a broader interest in Dirichlet (and more general derivative-based regularization terms) in order to conduct learning tasks. For example, [48] studies the use of Laplacian eigenvalues on graphs in the context of Dirichlet partitions to perform generalized clustering. Bayesian inverse methods have also utilized Laplacian regularization in [26].

3.1.1. Analysis of large sample limits of variational problems on graphs. In the past few years there has been a rapid development of a body of work borrowing ideas from the calculus of variations and PDE theory to study large sample asymptotics of optimization problems on geometric graphs closely connected to machine learning tasks. The motivation is clear: to a large extent, most of the graph-based methods for learning that are in existence can be phrased as solving either a variational problem on a graph or a graph PDE. In many instances these graph PDEs involve the graph Laplacian. Said works include the study of consistency of Cheeger and ratio graph cuts on graphs [28], consistency of graph Laplacian spectrum [61], and supervised and semi-supervised learning [11, 21, 25, 26]. In the previously listed papers, the convergence of discrete solutions to continuum counterparts is studied in the TL^p-metric introduced in [27] and later further studied in [58]. The TL^p topology can be thought of as L^p convergence after suitable matching of the ground-truth measure generating the data set X and its empirical measure. The consistency of the optimization problems is studied using variational methods, and in particular, the notion of Γ-convergence (a.k.a. epi-convergence). This is a powerful notion used to establish asymptotic convergence of minimizers of optimization problems (especially in highly nonconvex settings), but it does not offer direct ways to obtain rates of convergence.

Among the papers previously listed, our earlier paper [25] is closely connected to this paper. There we consider an optimization problem of the form

$$\min_u \frac{\beta}{n} \sum_{i,j} w_{ij} |u(x_i) - u(x_j)| + \frac{1}{n} \sum_i |u(x_i) - y_i|,$$

which is the L^1 version of the problem we study here. As is well known in the image analysis community the total variation functional (the first term in the above objective function) enforces sparsity of derivatives [12], and hence the above optimization problem seems more appropriate for the purposes of classification when binary labels are available (in our notation $y_i \in \{0, 1\}$). In that paper we study the regimes of $\beta := \beta_n$ (and how the graph connectivity ε must scale with n) so as to recover in the large n limit the Bayes classifier with probability one. No rates of convergence are provided.

The paper [56] is also related to our work. There, p-Laplacian regularization for semi-supervised learning is studied. The optimization problem takes the form

$$\min_u \frac{\beta}{n} \sum_{i,j} w_{ij} |u(x_i) - u(x_j)|^p$$

subject to

$$u(x_i) = y_i, \quad i = 1, \dots, q,$$

where q is held fixed as $n \to \infty$. The authors are able to show that when p is greater than the intrinsic dimension m, solutions to the p-Laplacian regularization problem converge *uniformly* to a continuum counterpart, as $n \to \infty$, which depends on the labels $y_1, \dots, y_q$ (in other words the labels are not forgotten in the limit). The uniform convergence is proved by bootstrapping the TL^p convergence obtained through variational methods by controlling the "oscillations" of the discrete minimizers at a certain convenient length-scale.

The paper [10] is very closely related to [56] and to this paper. In particular, it obtains analogue results to [56], but using a PDE approach rather than a calculus of variations one. A maximum principle at the graph level analogous to the one that we use in this paper is a crucial tool that is later used in conjunction with general and flexible results on consistency of viscosity solutions to elliptic PDEs. In this paper we take a PDE approach as in [10], and specifically use a maximum principle, to obtain rates for the uniform convergence of graph Laplacian regressors towards continuum counterparts. Whether results similar to the ones we present here can be obtained for regressors obtained using other regularization terms different from the graph Dirichlet energy is a question that we believe is worth exploring in the future.

Another very recent work which is related to ours is [54]. In that work the authors consider semi-supervised learning with Laplacian smoothing (with hard label constraints), but without any label noise (in particular, the labels y_i are evaluations of a $C^1(\mathcal{M})$ function). That paper establishes convergence rates which are analogous to ours, using similar technical tools (i.e. a maximum principle, but with boundary data). The semi-supervised aspect of their work is closely related to the harmonic extension problem (see [46] for additional discussion on the harmonic extension problem in machine learning). We emphasize that the noisy labels that we consider here are not covered in their analysis and are not trivial to handle: see the proof of Theorem 2.1 and the discussion in section 3.2 for further information.

3.1.2. The linear case and connections to k-NN regressors and other local averaging procedures. Here we draw a connection between the graph Laplacian regressor obtained by solving (1.5) and the classical k-NN regressor. To do this, we will focus on solutions of (1.5) when $F(t) = \frac{1}{2}t^2$. As we will see, graph Laplacian regularization with squared error loss can be interpreted as a local averaging procedure, where the "locality" is defined in terms of the intrinsic geometry of the graph, which in turn approximates the geometry of the underlying manifold $\mathcal{M}$.

Let us first briefly recall the definition of the k-NN regressor: For a $k \in \mathbb{N}$, with $k < n$, we define $N_k(x_i)$ to be the set of k nearest neighbors of x_i in the data set X. The k-NN regressor is then defined as

$$u_k(x_i) := \frac{1}{k} \sum_{x_j \in N_k(x_i)} y_j. \tag{3.1}$$

The use of local averages in nonparametric regression goes as far back as the work [62], k-NN regression being a special case of this general idea. The book [35] presents a very complete picture of many nonparametric regression techniques and dedicates a whole chapter (Chapter 6) to k-NN regression. Asymptotic properties of k-NN regressors have been a topic of investigation for several decades; see [35] and the paper [18] where L^1 convergence towards a trend function is proved in a very general setting. More recent results like [42] prove uniform convergence towards a trend in a quite general setting where, in particular, the intrinsic dimension of the underlying ground-truth may vary; as a byproduct k-NN regression is shown to adapt to the local geometry of the underlying model. The paper [13] is closely related to [42], but studies the classification problem instead.

We now show that when F is quadratic, the graph Laplacian regressor obtained by solving (1.5) can be interpreted as a local averaging procedure, where now the averaging is with respect to the heat kernel on the graph; for simplicity, we take $F(t) := \frac{1}{2}t^2$. Indeed, in this case the solution to the graph PDE (1.5) can be explicitly written as

$$u = (\beta\Delta_\Gamma + I)^{-1}y.$$

The fact that Δ_Γ is self-adjoint and positive semidefinite allows us to use the spectral theorem and write

$$u = (\beta\Delta_\Gamma + I)^{-1}y = \int_0^\infty e^{-t(\beta\Delta_\Gamma + I)}y dt.$$

Since Δ_Γ and I commute we get

$$u = \int_0^\infty e^{-t}\left(e^{-t\beta\Delta_\Gamma}y\right)dt = \int_0^\infty \frac{e^{-t/\beta}}{\beta}\left(e^{-t\Delta_\Gamma}y\right)dt, \tag{3.2}$$

where in the final step we have made a change of variables. From this formula we can conclude a couple of things. First, we notice that the function $e^{-t\Delta_\Gamma}y$ is simply the solution to the heat equation (on the graph) with initial condition y evaluated at time t and can be written as

$$e^{-t\Delta_\Gamma}y(x_i) = \frac{1}{n}\sum_{j=1}^n K_t(x_j, x_i)y_j,$$

where $K_t(x_j, x_i)$ is the heat kernel on the graph at time t. One can then show that the function $K_t(\cdot, x_i)$ is nonnegative and, moreover, that $\frac{1}{n}\sum_{j=1}^n K_t(x_j, x_i) = 1$. From this it follows that the function $e^{-t\Delta_\Gamma}y$ is obtained by computing a local average (at length-scale t) of y around each point x_i. On the other hand, since the function $\frac{1}{\beta}e^{-t/\beta}$ is a probability density on $(0, \infty)$, we conclude that the graph Laplacian regressor u is nothing but an average of averages of y over all length-scales t. The weight given to each length-scale is naturally determined by the parameter β, and in particular, if β is small, more relevance is given to more local length-scales, whereas if β is large, more relevance is given to global length-scales. Notice that the structure of (3.2) is analogous to what we would obtain if we averaged the different k-NN regressors u_k from (3.1) over the value of k to produce a regression function of the form

$$\overline{u}(x_i) := \sum_k g(k)u_k(x_i),$$

where in the above g is some probability mass function (p.m.f.) over k.

So far we have seen that the regressor u that stems from graph Laplacian regularization with squared loss is obtained by averaging over local averages of the labels y_i at different length-scales, and that these local averages use the intrinsic geometry of the graph (summarized in the graph heat kernel). Since in our setting the data is assumed to be sampled from a smooth manifold $\mathcal{M}$, it is to be expected that the graph heat kernel $e^{-t\Delta_\Gamma}$ actually behaves like the heat kernel on $\mathcal{M}$, $e^{-t\Delta_\mathcal{M}}$. Furthermore, for small values of t one has (neglecting constants) that $e^{-t\Delta_\mathcal{M}} \sim \frac{1}{t^{m/2}} e^{-d_\mathcal{M}(x,y)^2/4t}$ (see, for example, [63] or Chapter 15 in [33]), where $d_\mathcal{M}$ is the geodesic distance on the manifold $\mathcal{M}$. In particular, for small values of β the regression function u is expected to behave like

$$u(x_i) \sim \int_0^\infty \frac{e^{-t/\beta}}{\beta} \left(\frac{1}{n} \sum_{j=1}^n \frac{e^{-d_\mathcal{M}(x_i,x_j)^2/4t}}{t^{m/2}} y_j \right) dt,$$

which can be interpreted as a bandwidth average of local regression kernels, where the local regression kernels average over geodesic distances. The bottom line is that graph Laplacian regression (at least in the linear case) is very closely related to other local averaging regression procedures like k-NN regression and other methodologies that use geodesic distances to define nearest neighbors for label propagation [44, 70].

We would also like to notice that (3.2) can be written in the form

$$u(x_i) = \frac{1}{n} \sum_{j=1}^n K_n(x_i, x_j) y_j = \left(\int_0^\infty \frac{e^{-t/\beta}}{\beta} \left(e^{-t\Delta_\Gamma} y \right) dt \right) [i, j].$$

Written in this form, this construction of u can be seen as a modified *Nadaraya–Watson regression* [45, 68], also known as kernel regression. Indeed, to construct the Naradaya–Watson regressor one typically picks a kernel function K, and defines $u_{NW}(x_i) = \frac{\sum_{j=1}^n K(x_i - x_j) y_j}{\sum_{j=1}^n K(x_i - x_j)}$. Here various choices of the kernel function are permitted, such as a sharp cutoff, or the Gaussian. One can also, in principle, permit the kernel to vary in i (e.g., using the heat kernel after a fixed time t on a manifold or graph [4]), and here we see that the Laplacian regression we study in this work can be cast within that framework. This can also be seen as a version of RKHS regression, without any regularization. This method is still a topic of research; for example, recent works [51, 64] study kernel regression over graphs.

We also pause here to offer a comparison of performance guarantees between the regression procedure that we study and that of k-NN regression studied in [42]. In that work, under an identical regression model, it is shown that k-NN regression achieves, with high probability and ignoring logarithmic factors, the same $n^{-\frac{1}{m+2}}$ uniform convergence rate towards the trend μ, with a lead constant L that scales as the square of the Lipschitz constant of μ. It is also shown in that work that such a rate is nearly minimax optimal, in the sense that no estimator can achieve a better rate than $L^{m/(2+m)} n^{-1/(2+m)}$. Given the connection between k-NN and the Laplacian regression methods discussed here, it is not surprising that the two have essentially the same convergence rate, as we rigorously show in our main result, Theorem 2.4 (and the discussion below it).

It is indeed not so surprising that for small β both k-NN and graph Laplacian regularization (for quadratic F) are so similar. In both cases the whole idea is to average "enough" so as to remove the noise, and essentially any local averaging procedure should recover the underlying trend. However, as has been discussed throughout this paper, the Laplacian regularized supervised regression problem studied here belongs to a larger family of regularized graph-based algorithms for both supervised and *unsupervised* learning. One main thesis of this work is that developing tools for the fully supervised regime will facilitate the study of other, less supervised, algorithms that utilize graph Laplacians, which are currently not amenable to analysis using standard techniques for supervised methods.

3.2. PDE approach versus variational approach. In this section we illustrate the advantages of the PDE approach that we take in this paper and contrast it with a direct variational approach. In a variational approach one essentially uses the strong convexity of the functional to be minimized in (1.1) to derive convergence rates for minimizers. To illustrate this idea, we sketch the proof of the following proposition.

Proposition 3.1. *Let $F(t) = \frac{t^2}{2}$, and consider the variational problem*

$$\min_u \beta R_\Gamma(u) + \frac{1}{2n}\sum_{i=1}^n (u(x_i) - \mu(x_i))^2 := \min_u \tilde{J}(u). \tag{3.3}$$

Then the minimizer $\tilde{u}_n$ of (3.3) satisfies $\|\tilde{u} - \mu\|_{L^2(X)} \leq C\beta^{1/2}$, where here we abuse notation and let μ represent pointwise evaluation of μ at the points in the data set X. Furthermore, if u^* is the minimizer of (1.1), with the same choice of F, then $\|u^* - \tilde{u}\|_{L^2(X)} \leq C\big(\frac{\varepsilon^{1/4}}{\beta^{1/4}} + \frac{1}{(n\varepsilon^m)^{1/4}}\big)$. In turn, we have that $\|u^* - \mu\|_{L^2(X)} \leq C\big(\frac{\varepsilon^{1/4}}{\beta^{1/4}} + \beta^{1/2}\big)$. In particular, optimizing in β we find that $\|u^* - \mu\|_{L^2(X)} \leq C\big(\varepsilon^{1/6} + \frac{1}{(n\varepsilon^m)^{1/4}}\big)$.

Sketch of proof. To begin, various recent results [29, 9] allow one to show that, w.h.p., $R_\Gamma(\mu) \leq C$ as long as μ is sufficiently smooth. This, in turn, implies that $\tilde{J}(\tilde{u}) \leq C\beta$ w.h.p., which proves the first bound.

For the second bound, recalling the definition of J in (1.1) , by using the optimality of $\tilde{u}$ and u^*, summation by parts and (1.5), we compute that

$$0 \leq J(\tilde{u}) - J(u^*) = \frac{\beta}{2n}\sum_{i,j} w_{ij}\left(|\tilde{u}_i - \tilde{u}_j|^2 - |u_i^* - u_j^*|^2\right) + \frac{1}{2n}\sum_i (\tilde{u}_i - y_i)^2 - (u_i^* - y_i)^2 \tag{3.4}$$

$$= \frac{1}{2n}\sum_i \beta(\Delta_\Gamma \tilde{u})_i \tilde{u}_i - (\Delta_\Gamma u^*)_i u_i^* + (\tilde{u}_i - y_i)^2 - (u_i^* - y_i)^2 \tag{3.5}$$

$$= \frac{1}{2n}\sum_i (\mu_i - \tilde{u}_i)\tilde{u}_i - (y_i - u_i^*)u_i^* + (\tilde{u}_i - y_i)^2 - (u_i^* - y_i)^2$$

$$= \frac{1}{2n}\sum_{i=1}^n (\tilde{u}_i(\mu_i - y_i) + (u_i^* - \tilde{u}_i)y_i)\,. \tag{3.6}$$

An identical argument, but for $\tilde{J}$, yields that

$$0 \leq \tilde{J}(\tilde{u}) - \tilde{J}(u^*) = \frac{1}{2n}\sum_{i=1}^{n}(u_i^*(y_i-\mu_i) + (\tilde{u}_i - u_i^*)\mu_i)\,. \tag{3.7}$$

Adding the right-hand side of (3.7), which is positive, to the inequality (3.4) then yields

$$J(\tilde{u}) - J(u^*) \leq \frac{1}{2n}\sum_i(\tilde{u}_i - u_i^*)(\mu_i - y_i) \leq \frac{1}{2}|\langle u, \mu - y\rangle_{L^2(X)}| + \frac{1}{2}|\langle \tilde{u}, \mu - y\rangle_{L^2(X)}|. \tag{3.8}$$

At this stage, one notices that the right-hand side is composed of terms which are inner products between $\mu - y = \xi$ (which averages to zero at relatively small scales since $\mathbb{E}(\xi_i) = 0$ and the ξ_i are independent), and the functions $u, \tilde{u}$, which enjoy some degree of regularity. In the spirit of [25], we then use a local averaging procedure to bound these terms. In particular, given some ball $B_{\mathcal{M}}(x, r)$, if we let $\bar{u}$ be the average of u^* over that ball, then we may write

$$\frac{1}{n}\sum_{x_i \in B_{\mathcal{M}}(x,r)} u_i^*(\mu_i - y_i) = \frac{1}{n}\sum_{x_i \in B_{\mathcal{M}}(x,r)} (u_i^* - \bar{u})(\mu_i - y_i) + \bar{u}(\mu_i - y_i).$$

The second of these terms we may control w.h.p. using concentration estimates, as long as r is not smaller than ε, introducing some error that is in the order of $\frac{1}{\sqrt{nr^m}}$ (as we expect to have roughly nr^m terms in the sum). On the other hand, for the first term we may use the Poincaré inequality (which holds as long as ε scales appropriately with n; see [29]) to deduce that

$$\frac{1}{n}\sum_{x_i \in B(x,r)} (u_i^* - \bar{u})^2 \leq \frac{Cr}{n}\sum_{x_i \in B(x,r)}\sum_{x_j \in B(x,r)} w_{ij}(u_i^* - u_j^*)^2.$$

By then using the Cauchy–Schwarz inequality we obtain that, w.h.p.,

$$\left|\frac{1}{n}\sum_{x_i \in B(x,r)} u_i(\mu_i - y_i)\right| \leq Cr^{1/2}\left(\frac{1}{n}\sum_{x_j \in B(x,r)} w_{ij}(u_i^* - u_j^*)^2\right)^{1/2}.$$

By using a partition of unity (details of this type of argument can be found in [25]), we may then deduce that, w.h.p.,

$$\left|\frac{1}{n}\sum_{i=1}^{n} u_i^*(\mu_i - y_i)\right| \leq Cr^{1/2}(R_\Gamma(u^*))^{1/2} + C\frac{1}{\sqrt{nr^m}}.$$

An analogous bound holds for $\tilde{u}$. By noting that $J(\mu)$ is order one, we may deduce that $R_\Gamma(u^*) \leq C\beta^{-1}$. Similar logic applied to $\tilde{J}$ implies that $R_\Gamma(\tilde{u}) \leq C$. Supposing that $r \sim \varepsilon$ (the smallest permissible scaling), we then have that $J(\tilde{u}) - J(u^*) \leq C\big(\frac{\varepsilon^{1/2}}{\beta^{1/2}} + \frac{1}{\sqrt{n\varepsilon^m}}\big)$. Strong convexity of J in L^2 then implies that $\|\tilde{u} - u^*\|_2^2 \leq C\big(\frac{\varepsilon^{1/2}}{\beta^{1/2}} + \frac{1}{\sqrt{n\varepsilon^m}}\big)$, which proves the desired bound. ∎

We remark that the previous proof provides an L^2-type estimate, and one would need to use the discrete maximum principle to upgrade to uniform convergence. Such techniques are well-known in the numerical analysis community; see, e.g., [14]. The approach presented above is elegant and straightforward, and requires rather minimal assumptions on regimes for $\varepsilon := \varepsilon_n$. However, the rates that it proves are far worse than those that we prove here using the PDE approach. In particular, the Poincaré–type argument, which is used to handle the noisy labels, is overly pessimistic. In addition, the proof is elegant only when F is squared loss, but the details to make a similar argument work for more general loss functions are more complicated. One of the key ideas in this paper is that by leveraging the structure of the graph Laplacian (namely the maximum principle), one is able to provide much better theoretical guarantees.

4. Proof of main results.

In this section we provide proofs of the main results. In particular, in section 4.1 we provide a simplified proof of Theorem 2.3 for the case with quadratic loss, deferring the general argument to an appendix. In section 4.2 we use our maximum principle at the graph level in conjunction with two technical lemmas (where our probabilistic estimates are presented) to prove Theorem 2.1.

4.1. PDE estimates.

An important starting point in studying minimizers of (1.1) is to understand properties of minimizers of (2.4). The continuum problem (2.4) has a rich history: namely solutions of this variational problem are known to be highly regular. One of our goals in this work is to show that the tools used to study partial differential equations, including regularity theory, are quite helpful in analyzing regularized regression problems. However, establishing regularity estimates is a rather technical aspect of PDE theory. Hence, here we provide a simple proof for the case where the risk function is quadratic (and hence the necessary conditions are linear), and provide a more detailed proof for the nonlinear case with some extended discussion in an appendix.

Proof of Theorem 2.3 in the case with quadratic loss. In this linear case, we may express the solution of the Euler–Lagrange equation in the form

$$v = (\beta \Delta_{\mathcal{M}} + I)^{-1} \mu,$$

which in turn can be written as

$$v(x) = \int_0^\infty (e^{-t(\beta \Delta_{\mathcal{M}} + I)} \mu)(x) dt = \int_0^\infty e^{-t} (e^{-t\beta \Delta_{\mathcal{M}}} \mu)(x) dt. \tag{4.1}$$

Here we are using the spectral theorem for $\Delta_{\mathcal{M}}$. It follows that

$$v(x) - \mu(x) = \int_0^\infty e^{-t} \left(\int_{\mathcal{M}} K_{t\beta}(y, x)(\mu(y) - \mu(x)) dy \right) dt,$$

where $K_{t\beta}$ is the heat kernel on $\mathcal{M}$ (at time $t\beta$). In particular,

$$\begin{aligned} |v(x) - \mu(x)| &\leq \int_0^\infty e^{-t} \left(\int_{\mathcal{M}} K_{t\beta}(y, x) |\mu(y) - \mu(x)| dy \right) dt \\ &\leq Lip(\mu) \int_0^\infty e^{-t} \int_{\mathcal{M}} K_{t\beta}(y, x) d_{\mathcal{M}}(y, x) dy dt \leq C Lip(\mu) \beta, \end{aligned}$$

where the last inequality follows using properties of the heat kernel in $\mathcal{M}$, namely Gaussian upper bounds for the heat kernel on a smooth compact manifold (see, for example, [63] or Chapter 15 in [33]). The bottom line is that

$$\|v - \mu\|_\infty \leq C Lip(\mu)\beta,$$

where $Lip(\mu)$ is the Lipschitz constant of μ (which is finite since we have assumed that $\mu \in C^2(\mathcal{M})$; see the statements of Theorems 2.3 and 2.1).

On a heuristic level, one can argue for the remaining bounds simply using the PDE: As $\beta\Delta_{\mathcal{M}} v = v - \mu$, and as $|v - \mu| < C\beta$, we can then deduce that $|\Delta_{\mathcal{M}} v| < C$. Similarly, we formally have that $\beta\Delta^2_{\mathcal{M}} v = \Delta_{\mathcal{M}} v - \Delta_{\mathcal{M}}\mu$, and hence $|\Delta^2_{\mathcal{M}} v| \leq C\beta^{-1}$. Classical theory, as discussed in the appendix, can be used to infer a C^2 bound from a bound on $\Delta_{\mathcal{M}} v$ and a C^4 bound from a bound on $\Delta^2_{\mathcal{M}} v$. An interpolation argument gives the desired C^3 bound of order $\beta^{-1/2}$. Rigorously justifying these types of arguments requires that the PDE hold in a classical sense, which is often not clear a priori. More details on how one infers classical regularity of the solutions of variational problems is given in the appendix. ∎

We note that if the heat kernel were given by a convolution, then we could pass derivatives directly to μ and from formula (4.1) one could immediately argue that $\|v\|_{C^k} \leq C\|\mu\|_{C^k}$. It's likely that such bounds still hold in our case, but they were not necessary for our purposes and so we do not pursue them further.

4.2. Probabilistic estimates. In order to show the *sample error* bounds from Theorem 2.1 we start by making some computations based on standard concentration inequalities (see, e.g., [6]).

Proposition 4.1 (Hoeffding and Bernstein inequalities). *Suppose $U_1, \ldots, U_n$ are independent real valued random variables, with mean zero, and for which $|U_i| \leq M$ for all $i = 1, \ldots, n$. Suppose that*

$$\frac{1}{n}\sum_{i=1}^n Var(U_i) \leq \hat{\sigma}^2$$

for some $\hat{\sigma}^2$. Then,

- *(Hoeffding) For every $\delta > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n U_i\right| > \delta\right) \leq 2\exp\left(\frac{-2n\delta^2}{M^2}\right).$$

- *(Bernstein) For every $\delta > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n U_i\right| > \delta\right) \leq 2\exp\left(\frac{-n\delta^2}{2\hat{\sigma}^2 + 2M\delta/3}\right).$$

Remark 4.2. In most applications these inequalities are used to prove that the empirical average $\frac{1}{n}\sum_{i=1}^n U_i$ is small w.h.p., so that, in particular, one is typically interested in choosing $\delta \ll 1$. When the estimate on the average of variances is not significantly smaller than M^2, Bernstein's inequality does not produce any improvement over Hoeffding's.

Our first probabilistic estimates are concerned with the *pointwise* convergence of $\Delta_\Gamma h$ towards $\Delta_{\mathcal{M}} h$ for a fixed regular enough function h. Such estimates have been obtained in the literature before (see, for example, [36, 9, 10]), but here we present them again emphasizing the dependence of constants on the regularity of the function h. In particular, we will need this explicit dependence in order to quantify the error between u and v in terms of the regularity estimates obtained in Theorem 2.3 for the function v.

Proposition 4.3 (pointwise consistency of graph Laplacian). *Let $h \in C^3(\mathcal{M})$. Then, for every $\delta > 0$, with probability at least $1 - 2n\exp\big(-\frac{n\delta^2\varepsilon^{m+2}}{C(\|h\|_{C^1},\eta,\mathcal{M})(1+\varepsilon\delta)}\big)$, we have*

$$\max_{1\le i\le n} |\Delta_\Gamma h(x_i) - \Delta_{\mathcal{M}} h(x_i)| \le \delta + C(m,\eta,\|h\|_{C^3})\varepsilon,$$

where the last constant depends at most linearly on $\|h\|_{C^3}$.

We remind the reader that $\eta\big(\frac{|x-y|}{\varepsilon}\big)$ is the kernel describing the connectivity of our graph, m is the dimension of the manifold $\mathcal{M}$, and the $\|\cdot\|_{C^k}$ norm measures size of the first k derivatives of a function.

Proof. Associated to the function h we define a function

$$\Delta_\varepsilon h(x) := \frac{2}{\tau_\eta \varepsilon^{m+2}} \int_{\mathcal{M}} \eta\left(\frac{|x-\tilde{x}|}{\varepsilon}\right)(h(x) - h(\tilde{x}))\, dvol_{\mathcal{M}}(\tilde{x}), \quad x \in \mathcal{M}. \tag{4.2}$$

This function can be interpreted as a nonlocal Laplacian of h.

Fix $i \in \{1,\dots,n\}$ and denote by $U_1,\dots,U_n$ the variables

$$U_j := \frac{2}{\tau_\eta \varepsilon^{m+2}} \eta\left(\frac{|x_i - x_j|}{\varepsilon}\right)(h(x_i) - h(x_j)).$$

Notice that given x_i, we have

$$\mathbb{E}(U_j | x_1 \dots x_n) = \Delta_\varepsilon h(x_i), \quad j \neq i.$$

Also, using the bound on the support of η,

$$|U_j - \mathbb{E}(U_j | x_1,\dots,x_n)| \le \frac{4}{\tau_\eta \varepsilon^{m+1}} \|\eta\|_\infty \|\nabla h\|_\infty, \tag{4.3}$$

$$Var(U_j | x_1,\dots,x_n) \le \frac{4\|\eta\|_\infty \|\nabla h\|_\infty^2}{\tau_\eta^2 \varepsilon^{2m+2}} \int_{\mathcal{M}} \eta\left(\frac{|x_i - \tilde{x}|}{\varepsilon}\right) dvol_{\mathcal{M}}(\tilde{x}).$$

It is simple to see that for all $0 < \varepsilon < 1$ and all $x \in \mathcal{M}$ we have

$$0 < C_{\mathcal{M}}^{-1} \le \frac{1}{\varepsilon^m} \int_{\mathcal{M}} \eta\left(\frac{|x-\tilde{x}|}{\varepsilon}\right) dvol_{\mathcal{M}}(\tilde{x}) \le C_{\mathcal{M}},$$

where $C_{\mathcal{M}}$ is a positive constant. In particular, it follows that

$$\frac{1}{n}\sum_{j=1}^n Var(U_j | x_1,\dots,x_n) \le \frac{1}{\varepsilon^{m+2}} C_{\mathcal{M}} \|\eta\|_\infty \|\nabla h\|_\infty^2. \tag{4.4}$$

We notice that neither M nor σ^2 depends on $x_1, \dots, x_n$, and that the $U_j - \mathbb{E}(U_j|x_1, \dots, x_n)$ are independent random variables. We may now use Bernstein's inequality (Proposition 4.1), along with the definition of $\Delta_\Gamma h(x_i)$ and (4.3) and (4.4), to obtain

$$\mathbb{P}\left(|\Delta_\Gamma h(x_i) - \Delta_\varepsilon h(x_i)| > \delta \middle| x_1, \dots, x_n\right) \leq 2\exp\left(-\frac{n\delta^2\varepsilon^{m+2}}{C(\|h\|_{C^1}, \eta, \mathcal{M})(1+\varepsilon\delta)}\right),$$

and by the law of iterated expectation get

$$\mathbb{P}\left(|\Delta_\Gamma h(x_i) - \Delta_\varepsilon h(x_i)| > \delta\right) \leq 2\exp\left(-\frac{n\delta^2\varepsilon^{m+2}}{C(\|h\|_{C^1}, \eta, \mathcal{M})(1+\varepsilon\delta)}\right).$$

A simple union bound implies that

$$\mathbb{P}\left(\max_{i=1,\dots,n} |\Delta_\Gamma h(x_i) - \Delta_\varepsilon h(x_i)| > \delta\right) \leq 2n\exp\left(-\frac{n\delta^2\varepsilon^{m+2}}{C(\|h\|_{C^1}, \eta, \mathcal{M})(1+\varepsilon\delta)}\right). \tag{4.5}$$

Now we claim that for all $h \in C^3(\mathcal{M})$ and all $x \in \mathcal{M}$,

$$|\Delta_\varepsilon h(x) - \Delta_{\mathcal{M}} h(x)| \leq C_m Lip(\eta)\|h\|_{C^3}\varepsilon. \tag{4.6}$$

We first replace $\Delta_\varepsilon h$ with a version of it that uses the geodesic distance on $\mathcal{M}$ rather than the Euclidean distance. More precisely, we set

$$\widetilde{\Delta}_\varepsilon h(x) := \frac{2}{\tau_\eta \varepsilon^{m+2}} \int_{\mathcal{M}} \eta\left(\frac{d_{\mathcal{M}}(x,\tilde{x})}{\varepsilon}\right)(h(x) - h(\tilde{x}))\, dvol_{\mathcal{M}}(\tilde{x}),$$

where $d_{\mathcal{M}}(x,\tilde{x})$ is the geodesic distance between two points $x, \tilde{x}$ in $\mathcal{M}$. Now, as long as $|x - \tilde{x}| \leq c$ for some small enough c (that only depends on $\mathcal{M}$), we have that

$$|d_{\mathcal{M}}(x,\tilde{x}) - |x - \tilde{x}|| \leq C(\mathcal{M})|x - \tilde{x}|^3,$$

where $C(\mathcal{M})$ is a constant that depends on $\mathcal{M}$; see, for example, Proposition 2 in [29].

From this and the Lipschitz continuity of η we can conclude that if $|x - \tilde{x}| \leq 2\varepsilon$, then

$$\left|\eta\left(\frac{d_{\mathcal{M}}(x,\tilde{x})}{\varepsilon}\right) - \eta\left(\frac{|x-\tilde{x}|}{\varepsilon}\right)\right| \leq C(\mathcal{M})Lip(\eta)\varepsilon^2.$$

On the other hand, if $2\varepsilon < |x - \tilde{x}|$, we must also have $d_{\mathcal{M}}(x,\tilde{x}) > \varepsilon$. Therefore, in all cases we have

$$\left|\eta\left(\frac{d_{\mathcal{M}}(x,\tilde{x})}{\varepsilon}\right) - \eta\left(\frac{|x-\tilde{x}|}{\varepsilon}\right)\right| \leq C(\mathcal{M})Lip(\eta)\varepsilon^2 \mathbf{1}_{B_{\mathcal{M}}(x,2\varepsilon)}(\tilde{x}),$$

from which it follows that for all $x \in \mathcal{M}$,

$$\begin{aligned}
&|\Delta_\varepsilon h(x) - \widetilde{\Delta}_\varepsilon h(x)| \\
&\leq \frac{2}{\tau_\eta \varepsilon^{m+2}} \int_{\mathcal{M}\cap B_{\mathcal{M}}(x,2\varepsilon)} \left|\eta\left(\frac{|x-\tilde{x}|}{\varepsilon}\right) - \eta\left(\frac{d_{\mathcal{M}}(x,\tilde{x})}{\varepsilon}\right)\right| |h(x) - h(\tilde{x})| dvol_{\mathcal{M}}(\tilde{x}) \\
&\leq C(\mathcal{M})\frac{Lip(\eta)}{\tau_\eta}\|\nabla h\|_\infty \varepsilon.
\end{aligned}$$

Let us now compare $\widetilde{\Delta}_\varepsilon h(x)$ with $\Delta_{\mathcal{M}} h(x)$. For that purpose we use the exponential map at the point x,

$$exp_x : B_m(0,\varepsilon) \to B_{\mathcal{M}}(x,\varepsilon),$$

which takes tangent vectors v at x with norm less than ε into points $\exp_x(v)$ in $\mathcal{M}$ that are within geodesic distance ε of x. Let H be the composition $H := h \circ \exp_x(v)$, i.e., the function h written in normal coordinates around x. The regularity of h and $\mathcal{M}$ implies that H is also regular, and using a Taylor expansion around the origin we get

$$H(v) = H(0) + \langle \nabla H(0), v\rangle + \frac{1}{2}\langle D^2 H(0)v, v\rangle + r(v),$$

where the remainder r is a function that satisfies

$$|r(v)| \leq C\varepsilon^3 \quad \forall v \in B_m(0,\varepsilon).$$

The constant C depends on $\mathcal{M}$ and the third derivatives of h. We emphasize that, by the integral form of the Taylor remainder theorem, the constant in this expression scales at most linearly in the third derivatives of h. In normal coordinates we can then write

$$\begin{aligned}
\widetilde{\Delta}_\varepsilon h(x) &= \frac{2}{\tau_\eta \varepsilon^{m+2}} \int_{B_m(0,\varepsilon)} \eta\left(\frac{|v|}{\varepsilon}\right)(H(v) - H(0)) J_x(v)\, dv \\
&= \frac{2}{\tau_\eta \varepsilon^{m+2}} \int_{B_m(0,\varepsilon)} \eta\left(\frac{|v|}{\varepsilon}\right)\langle \nabla H(0), v\rangle J_x(v)\, dv \\
&+ \frac{2}{\tau_\eta \varepsilon^{m+2}} \int_{B_m(0,\varepsilon)} \eta\left(\frac{|v|}{\varepsilon}\right)\langle D^2 H(0)v, v\rangle J_x(v)\, dv \\
&+ \frac{2}{\tau_\eta \varepsilon^{m+2}} \int_{B_m(0,\varepsilon)} \eta\left(\frac{|v|}{\varepsilon}\right) r(v) J_x(v)\, dv.
\end{aligned}$$

We know that the Jacobian of the exponential map J_x satisfies

$$J_x(v) = 1 + O(|v|^2)$$

(see section 2.2 in [9]), so we can actually write

$$\begin{aligned}
\widetilde{\Delta}_\varepsilon h(x) &= \frac{2}{\tau_\eta \varepsilon^{m+2}} \int_{B_m(0,\varepsilon)} \eta\left(\frac{|v|}{\varepsilon}\right)\langle \nabla H(0), v\rangle dv \\
&+ \frac{1}{\tau_\eta \varepsilon^{m+2}} \int_{B_m(0,\varepsilon)} \eta\left(\frac{|v|}{\varepsilon}\right)\langle D^2 H(0)v, v\rangle dv + R,
\end{aligned}$$

where $|R| \leq C(\|h\|_{C^3}, m, \eta)\varepsilon$, with the constant depending at most linearly in $\|h\|_{C^3}$. We notice that the first term on the right-hand side of the above expression drops due to the radial symmetry of the kernel, and also that the second term is equal to

$$trace(D^2 H(0)) = \Delta H(0) = \Delta_{\mathcal{M}} h(x).$$

The bottom line is that, as anticipated in (4.6),

$$|\Delta_\varepsilon h(x) - \Delta_{\mathcal{M}} h(x)| \le |\Delta_\varepsilon h(x) - \widetilde{\Delta}_\varepsilon h(x)| + |\widetilde{\Delta}_\varepsilon h(x) - \Delta_{\mathcal{M}} h(x)| \le C(\|h\|_{C^3}, m, \eta)\varepsilon.$$

Combining (4.5) and (4.6) we deduce that with probability greater than $1 - 2n \exp\left(-\frac{n\delta^2\varepsilon^{m+2}}{C(\|h\|_{C^1},\eta,\mathcal{M})(1+\varepsilon\delta)}\right)$,

$$\max_{i=1,\dots,n} |\Delta_\Gamma h(x_i) - \Delta_{\mathcal{M}} h(x_i)| \le \delta + C(\|h\|_{C^3}, m, \eta)\varepsilon.$$

This completes the proof. ∎

Our next result will allow us to show that the functions y^- and y^+ mentioned in (2.6) and defined explicitly in (4.10), are uniformly small.

Lemma 4.4. *Let $h : \mathcal{M} \to \mathbb{R}$ be a smooth function. For each $i = 1, \dots, n$ let E_i be defined as*

$$E_i := \sum_{j=1}^{n} \frac{\eta_{ij}}{g_j} \int_{\mathbb{R}} (f(h(x_j) - s) - f(h(x_j) - \xi_j))p(s)ds,$$

where

$$\eta_{ij} := \frac{1}{n\varepsilon^m}\eta\left(\frac{|x_i - x_j|}{\varepsilon}\right) \quad \text{and} \quad g_i := \sum_{l=1}^{n} \eta_{il}.$$

Let $\zeta > 0$. Then with probability greater than $1 - 2n\exp\left(-cn\varepsilon^m\zeta^2\right) - 2n\exp(-cn\varepsilon^m)$,

$$|E_i| \le \zeta \quad \forall i = 1, \dots, n.$$

Proof. Fix $i = 1, \dots, n$ and let U_j be the random variables

$$U_j := \frac{n\eta_{ij}}{g_j} f(h(x_j) - \xi_j), \quad j = 1, \dots, n.$$

Conditioned on $x_1, \dots, x_n$, the variables $U_1, \dots, U_n$ are independent and satisfy

$$|U_j| \le \frac{1}{\varepsilon^m}\|\eta\|_\infty \frac{M_{h,f}}{G_{\vec{x}}},$$

where

$$M_{h,f} := \sup_{x\in\mathcal{M}} |f(h(x) \pm \sigma)|,$$

and where

$$G_{\vec{x}} := \min_{j=1,\dots,n} g_j.$$

Moreover,

$$\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} U_j \middle| x_1, \dots, x_n\right) = \frac{\eta_{ij}}{g_j}\int_{\mathbb{R}} f(h(x_j) - s)p(s)ds,$$

and

$$\frac{1}{n}\sum_{j=1}^{n} Var(U_j|x_1,\dots,x_n) \leq \frac{M_{h,f}^2}{n\varepsilon^{2m}G_{\vec{x}}^2}\sum_{j=1}^{n}\eta^2\left(\frac{|x_i-x_j|}{\varepsilon}\right) = \frac{M_{h,f}^2\|\eta\|_\infty}{\varepsilon^m G_{\vec{x}}^2}g_i \leq \frac{M_{h,f}^2\|\eta\|_\infty}{\varepsilon^m G_{\vec{x}}^2}\widetilde{G}_{\vec{x}},$$

where

$$\widetilde{G}_{\vec{x}} := \max_{i=1,\dots,n} g_i.$$

Bernstein's inequality (Proposition 4.1) then implies that

$$\begin{aligned}\mathbb{P}(|E_i| \geq \zeta|x_1,\dots,x_n) &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n}U_j - \frac{1}{n}\sum_{j=1}^{N}\mathbb{E}(U_j|x_1,\dots,x_n)\right| \geq \zeta\middle|x_1,\dots,x_n\right)\\ &\leq 2\exp\left(\frac{-n\zeta^2}{\frac{2\|\eta\|_\infty M_{h,f}^2}{\varepsilon^m G_{\vec{x}}^2}\widetilde{G}_{\vec{x}} + \frac{2\|\eta\|_\infty M_{h,f}}{3\varepsilon^m G_{\vec{x}}}\zeta}\right).\end{aligned}$$

Using a simple union bound we deduce that

$$\mathbb{P}\left(\max_{i=1,\dots,n}|E_i| \geq \zeta\middle|x_1,\dots,x_n\right) \leq 2n\exp\left(\frac{-n\varepsilon^m\zeta^2}{\frac{2\|\eta\|_\infty M_{h,f}^2}{G_{\vec{x}}^2}\widetilde{G}_{\vec{x}} + \frac{2\|\eta\|_\infty M_{h,f}}{3G_{\vec{x}}}\zeta}\right),$$

and by the law of iterated expectation we obtain

$$\mathbb{P}\left(\max_{i=1,\dots,n}|E_i| \geq \zeta\right) \leq 2n\mathbb{E}\left(\exp\left(\frac{-n\varepsilon^m\zeta^2}{\frac{2\|\eta\|_\infty M_{h,f}^2}{G_{\vec{x}}^2}\widetilde{G}_{\vec{x}} + \frac{2\|\eta\|_\infty M_{h,f}}{3G_{\vec{x}}}\zeta}\right)\right).$$

Now, the only terms in the above expression that depend on $x_1,\dots,x_n$ are $G_{\vec{x}}$ and $G_{\tilde{x}}$. These, however, can be shown to be bounded below and above by positive constants with very high probability. Indeed, we first notice that for all $\varepsilon < 1$ and all $x \in \mathcal{M}$ we have

$$0 < C_{\mathcal{M}}^{-1} \leq K_\varepsilon(x) := \frac{1}{\varepsilon^m}\int_{\mathcal{M}}\eta\left(\frac{|x-\tilde{x}|}{\varepsilon}\right)dvol_{\mathcal{M}}(\tilde{x}) \leq C_{\mathcal{M}}$$

for some positive constant $C_{\mathcal{M}}$. On the other hand, using Hoeffding's inequality we get that

$$\mathbb{P}\left(\max_{i=1,\dots,n}|g_i - K_\varepsilon(x_i)| \geq \frac{1}{2C_{\mathcal{M}}}\right) \leq 2n\exp\left(-cn\varepsilon^m\right),$$

from where it follows that except on a set with probability less than $2n\exp(-cn\varepsilon^m)$, we have

$$\frac{1}{2C_{\mathcal{M}}} \leq G_{\vec{x}} \leq \widetilde{G}_{\vec{x}} \leq 2C_{\mathcal{M}}. \tag{4.7}$$

Therefore,

$$\mathbb{P}\left(\max_{i=1,\dots,n}|E_i| \geq \zeta\right) \leq 2n\exp\left(-cn\varepsilon^m\zeta^2\right) + 2n\exp(-cn\varepsilon^m).$$

This completes the proof. ∎

We are ready to prove Theorem 2.1.

Proof of Theorem 2.1. Let us first introduce some notation. For a function $g : X \to \mathbb{R}$ we denote by g_i the value of the function at x_i, i.e., $g(x_i)$. Also, we will restrict the solution of (2.5) and its Laplacian $\Delta_{\mathcal{M}} v$ to the point cloud X, so, in particular, we will treat v and $\Delta_{\mathcal{M}} v$ as functions defined on X.

First, we notice that, by (2.5)

$$\beta \Delta_\Gamma v + \int_{\mathbb{R}} f(v - \mu - s) p(s) ds = \beta(\Delta_\Gamma v - \Delta_{\mathcal{M}} v)$$

at all points in X. Let us denote by $a : X \to \mathbb{R}$ the right-hand side of the above expression (i.e., β times the difference between $\Delta_\Gamma v$ and $\Delta_{\mathcal{M}} v$). By Proposition 4.3 and Theorem 2.3 we know that with probability at least $1 - 2n \exp\big(-\frac{n\delta^2 \varepsilon^{m+2}}{C(\mu,\eta,\mathcal{M})(1+\varepsilon\delta)}\big) =: 1 - p_{n,\delta}$ we have that

$$\max_{i=1,\dots,n} |a_i| \le \beta(\delta + C\beta^{-1/2}\varepsilon). \tag{4.8}$$

Now, let $w := u - v$. Then, by (1.5) and (2.5) ,

$$\begin{aligned} -\beta \Delta_\Gamma w &= -\beta \Delta_\Gamma u + \beta \Delta_\Gamma v \\ &= f(u - y) - \int_{\mathbb{R}} f(v - \mu - s) p(s) ds + a \\ &= f(u - \mu - \xi) - \int_{\mathbb{R}} f(v - \mu - s) p(s) ds + a \end{aligned} \tag{4.9}$$

Let us define the functions y^+ and y^- on X, respectively, by

$$\begin{aligned} y_i^+ &:= \frac{\varepsilon^2}{\beta g_i} \left(\int_{\mathbb{R}} f(v_i - \mu_i - s) p(s) ds - f(v_i - \mu_i - \xi_i) \right) + \rho, \\ y_i^- &:= \frac{\varepsilon^2}{\beta g_i} \left(\int_{\mathbb{R}} f(v_i - \mu_i - s) p(s) ds - f(v_i - \mu_i - \xi_i) \right) - \rho, \end{aligned} \tag{4.10}$$

where g is as defined in Lemma 4.4 and ρ is a constant that will be chosen later on. Indeed, we will show that with the appropriate choice of ρ, the following holds at all point in X:

$$y^- \le w \le y^+.$$

We focus on showing $w \le y^+$, the other inequality obtained in a completely analogous way. To see that $w \le y^+$, we will actually show that for an appropriate (small) value of ρ, the function

$$z := w - y^+$$

satisfies the inequality

$$-\Delta_\Gamma z - (f(z + v - \mu - \xi) - f(v - \mu - \xi)) \ge 0, \tag{4.11}$$

from where it follows, thanks to the maximum principle (Proposition 2.2), that $z \leq 0$. Let us then focus on showing (4.11). First, a direct computation shows that

$$
\begin{aligned}
(\beta \Delta_\Gamma y^+)_i = & \int_{\mathbb{R}} f(v_i - \mu_i - s) p(s) ds - f(v_i - \mu_i - \xi_i) \\
& - \sum_{j=1}^N \frac{\eta_{ij}}{g_j} \left(\int_{\mathbb{R}} f(v_j - \mu_j - s) p(s) ds - f(v_j - \mu_j - \xi_j) \right),
\end{aligned} \tag{4.12}
$$

where in the above we are using η_{ij} as defined in Lemma 4.4. It follows that

$$
\begin{aligned}
(-\beta \Delta_\Gamma z)_i &= f(u_i - \mu_i - \xi_i) - f(v_i - \mu_i - \xi_i) \\
&\quad - \sum_{j=1}^n \frac{\eta_{ij}}{g_j} \int_{\mathbb{R}} (f(v_j - \mu_j - s) - f(v_j - \mu_j - \xi_j)) p(s) ds + a_i \\
&= f(w_i + v_i - \mu_i - \xi_i) - f(v_i - \mu_i - \xi_i) \\
&\quad - \sum_{j=1}^n \frac{\eta_{ij}}{g_j} \int_{\mathbb{R}} (f(v_j - \mu_j - s) - f(v_j - \mu_j - \xi_j)) p(s) ds + a_i.
\end{aligned} \tag{4.13}
$$

Since $v - \mu$ is a bounded function (in particular, thanks to their regularity as follows from Theorem 2.3 and by the assumptions on μ), Lemma 4.4 implies that, with probability at least $1 - 2n \exp(-cn\varepsilon^m \zeta^2) - 2n \exp(-cn\varepsilon^m) =: 1 - p_{n,\zeta}$ we have

$$
\left| \sum_{j=1}^n \frac{\eta_{ij}}{g_j} \int_{\mathbb{R}} (f(v_j - \mu_j - s) - f(v_j - \mu_j - \xi_j)) p(s) ds \right| \leq \zeta \quad \forall i = 1, \dots, n.
$$

Hence, by using (4.8), with probability at least $1 - p_{n,\delta} - p_{n,\zeta}$, for all i we have

$$
(-\beta \Delta_\Gamma z)_i \geq f(w_i + v_i - \mu_i - \xi_i) - f(v_i - \mu_i - \xi_i) - \zeta - \beta(\delta + C\beta^{-1/2}\varepsilon),
$$

which can be rewritten as

$$
\begin{aligned}
& -\beta \Delta_\Gamma z - (f(z + v - \mu - \xi) - f(v - \mu - \xi)) \\
& \qquad \geq f(w + v - \mu - \xi) - f(z + v - \mu - \xi) - \zeta - \beta(\delta + C\beta^{-1/2}\varepsilon).
\end{aligned} \tag{4.14}
$$

Now, notice that ρ can be chosen in such a way that

$$
y^+ \geq 0.
$$

Indeed, since we have assumed that the noise ξ is bounded, we can conclude that $y^+ \geq -C_2 \frac{\varepsilon^2}{\beta} + \rho$ for some constant C_2, from where it follows that if ρ is chosen to be larger than $C_2 \frac{\varepsilon^2}{\beta}$, we can conclude that $y^+ \geq 0$. In particular, for such choice of ρ we have $w = z + y^+ \geq z$ and thus by the fundamental theorem of calculus,

$$
f(w + v - \mu - \xi) - f(z + v - \mu - \xi) = \int_{s_1}^{s_2} f'(s) ds \geq c(s_2 - s_1) = c y^+
$$

for some constant $c > 0$ (using the assumed strict monotonicity of f) and where

$$s_2 := w + v - \mu - \xi, \quad s_1 := z + v - \mu - \xi.$$

Plugging this back into (4.14) we deduce that (with probability at least $1 - p_{n,\delta} - p_{n,\zeta}$)

$$-\beta\Delta_\Gamma z - (f(z+v-\mu-\xi) - f(v-\mu-\xi)) \geq cy^+ - \zeta - \beta\delta - C\beta^{1/2}\varepsilon \geq c\rho - cC_2\frac{\varepsilon^2}{\beta} - \zeta - \beta\delta - C\beta^{1/2}\varepsilon).$$

Hence if we let ρ be defined according to

$$\rho := \frac{C_2\varepsilon^2}{\beta} + \frac{\zeta + \beta\delta + C\beta^{1/2}\varepsilon}{c},$$

we conclude that, with probability at least $1 - p_{n,\delta} - p_{n,\zeta}$,

$$-\Delta_\Gamma z - (f(z + v - \mu - \xi) - f(v - \mu - \xi)) \geq 0$$

as we wanted to show. Repeating this argument for y^-, and using a union bound completes the proof. ∎

Appendix A. Proof of Theorem 2.3. Here we give an outline of the proof of Theorem 2.3 in the case with general loss function. We attempt to offer some additional discussion and pointers to important inequalities that are used in this process, but do not attempt to provide all the details (see, e.g., [30] for complete details).

Proof. Given the continuum variational problem (2.4), the first question is whether a unique minimizer exists. A typical modern approach is to consider minimizing this functional over a wide class of functions (e.g., H^1, the broadest class of functions for which the Dirichlet energy is finite). Since the functional is convex and coercive, one can generally infer the existence of a solution using "soft" (i.e., nonconstructive) methods. This is done by using, e.g., weak compactness of bounded sets in H^1 along with weak lower semicontinuity of the functional. Alternatively, in the cases we're considering one can use other nonconstructive methods, such as Lax–Milgram or Browder–Minty (see Chapter 6, Theorem 3 in [23] and [49, section 10.3]), to infer the existence of minimizers. Uniqueness usually follows directly from strong convexity of the functional. Directly using these methods, we may infer the existence and uniqueness of an H^1 function minimizing (2.4).

Once one has assured the existence and uniqueness of a (H^1) solution to the problem, we would like to study finer properties of the solutions. To do this, we first notice that by taking variations in (2.4), that is, by letting $v_\varepsilon = v + \varepsilon w$, and then considering $\lim_{\varepsilon\to 0} \frac{J(v_\varepsilon)-J(v)}{\varepsilon}$, we have for any $w \in H^1$

$$\int_{\mathcal{M}} \beta\nabla w \cdot \nabla v + \left(\int_{\mathbb{R}} f(v - \mu - s)p(s)ds\right) w dvol_{\mathcal{M}}(x) = 0,$$

where v is the minimizer of (2.4). Note that if v were sufficiently regular (e.g., C^2), then we could use integration by parts in the first term and the fundamental theorem of the calculus

of variations to infer (2.5). At the moment, given only that $v \in H^1$, we simply can say that v is a *weak solution* of the PDE (2.5).

Several avenues are available at this stage to demonstrate that the optimizer v is more regular. First, we notice, in our case, that truncating v at any value above $\max \mu + \sigma$ and below $\min \mu - \sigma$ (we recall that p is supported in $[-\sigma, \sigma]$) will decrease the objective value in (2.4). This implies that

$$\|v\|_{L^\infty(\mathcal{M})} \leq \|\mu\|_{L^\infty(\mathcal{M})} + \sigma. \tag{A.1}$$

Next, various tools are available for establishing regularity of elliptic equations. For example, Theorem 2 in Chapter 6 in [23] states that any weak solution of $\Delta_{\mathcal{M}} w + g = 0$ (for an arbitrary g) will satisfy

$$\|w\|_{H^{r+2}(\mathcal{M})} \leq C(\|g\|_{H^r(\mathcal{M})} + \|w\|_{L^2(\mathcal{M})}), \tag{A.2}$$

where the inequality is only meaningful when g belongs to the Sobolev space $H^r(\mathcal{M})$ (i.e., the largest space of functions where one can make sense of rth order "weak" derivatives which are squared integrable). This is proved by using the weak elliptic equation to provide a priori bounds on difference quotients of the function w. Using a version of the chain rule in higher-order Sobolev spaces (namely that $\|f \circ v\|_{H^r} \leq \|f\|_{C^r} \|v\|_{H^r}$; see, e.g., [7] or [39]), and using the fact that μ and f are smooth, we can take $g = \int_{\mathbb{R}} f(v(\cdot) - \mu(\cdot) - s)p(s)ds$ and rewrite this estimate in the following way:

$$\|v\|_{H^{r+2}(\mathcal{M})} \leq C(\|v\|_{H^r(\mathcal{M})} + \|v\|_{L^\infty(\mathcal{M})}),$$

where here we remark that the constants in the previous line will depend on $\beta, r, \mathcal{M}, f$, and μ. Iterating this inequality then gives that $v \in H^r$ for any positive integer r.

Once one has established Sobolev regularity, we may use Morrey's inequality (see, e.g., Theorem 6 in Chapter 5 of [23]), which allows one to infer that for any $k \in \mathbb{N}$ there exists an r so that $\|v\|_{C^k} \leq C\|v\|_{H^r}$. This then implies that the minimizer of the problem (2.4) is in fact infinitely differentiable, and is a classical solution of (2.5).

Once one has a more regular solution, a variety of techniques are available to demonstrate a priori bounds on different derivatives (such as the bounds (2.8) and (2.9)). For example, the classical maximum principle (see, e.g., Theorem 2 in Chapter 6 of [23]) states that, given a C^2 function w, if $-\Delta_{\mathcal{M}} w \geq 0$ on a set $E \subset \mathcal{M}$, then w attains its maximum on the boundary of E. We may use this to prove the estimate (2.8) as follows: we note that for any point where $v \geq \mu_f$ we have

$$\begin{aligned}
\beta\Delta_{\mathcal{M}}(v - \mu_f) &= -\int_{\mathbb{R}} f(v - \mu - s)p(s)ds - \beta\Delta_{\mathcal{M}}\mu_f \\
&= \int_{\mathbb{R}} \left(f(\mu_f - \mu - s) - f(v - \mu - s)\right) p(s)ds - \beta\Delta_{\mathcal{M}}\mu_f \\
&\leq -c_1(v - \mu_f) - \beta\Delta_{\mathcal{M}}\mu_f,
\end{aligned}$$

where in the second equality we have used the definition of μ_f in (2.7), and in the inequality we have used (A.1) and the fact that on a bounded interval we have $f' \geq c_1 > 0$ for some

constant c_1 (which follows from the strict monotonicity of f). Now, suppose for the sake of contradiction that the set

$$E = \left\{ x \in \mathcal{M} : v - \mu_f > \frac{\beta \|\Delta_{\mathcal{M}} \mu_f\|_\infty}{c_1} \right\}$$

is nonempty. Notice that this is an open set given that both v and μ_f are continuous. From the above computations it follows that on E we have that

$$\beta \Delta_{\mathcal{M}} (v - \mu_f) \leq 0.$$

This implies (by the classical maximum principle) that $v - \mu_f$ when restricted to $\overline{E}$ attains its maximum on the boundary of E. However, since $\mathcal{M}$ is a manifold without boundary, ∂E takes the form $\partial E = \{x : v - \mu_f = \frac{\beta \|\Delta_{\mathcal{M}} \mu_f\|_\infty}{c_1}\}$, and we conclude that the maximum value that $v - \mu_f$ can take in $\overline{E}$ is $\frac{\beta}{c_1} \|\Delta_{\mathcal{M}} \mu_f\|_\infty$. However, this contradicts the fact that E was nonempty (where in theory $v - \mu_f$ achieves values higher than $\frac{\beta}{c_1} \|\Delta_{\mathcal{M}} \mu_f\|_\infty$). This provides the desired upper bound. The lower bound is deduced analogously. This proves (2.8) and by directly using the Euler–Lagrange equation (2.5), we then have that $\|\Delta_{\mathcal{M}} v\|_\infty \leq C$.

Now, to prove further bounds, we need a priori estimates in stronger norms. Many types of estimates are available, but we focus on two: Hölder–type estimates (due to De Giorgi, Nash, and Moser), and Schauder estimates. The classical Hölder estimates state that any H^1 solution of $\Delta_{\mathcal{M}} w + g = 0$ will satisfy (see Theorem 8.24 in [30])

$$\|w\|_{C^{0,\alpha}(\mathcal{M})} \leq C(\|w\|_{L^2(\mathcal{M})} + \|g\|_{L^\infty(\mathcal{M})})$$

for some appropriately chosen $\alpha > 0$ (here $C^{0,\alpha}$ denotes the space of α-Hölder continuous functions). We can use this to infer that $\|v\|_{C^{0,\alpha}(\mathcal{M})} < C$, with C independent of β. On the other hand, the classical Schauder estimates (see Theorem 6.6 in [30]) state that for a $C^{2,\alpha}$ solution of $\Delta_{\mathcal{M}} w + g = 0$ we have the bound

$$\|w\|_{C^{2,\alpha}(\mathcal{M})} \leq C(\|w\|_{C^0(\mathcal{M})} + \|g\|_{C^{0,\alpha}(\mathcal{M})}),$$

where here $C^{2,\alpha}$ is the space of functions with α-Hölder continuous second derivatives. By applying this to $w = v$, and noting that by the smoothness of f we have that $\|f \circ v\|_{C^\alpha} \leq \|\nabla f\|_\infty \|v\|_{C^\alpha}$, we may then apply these estimates to infer that $\|v\|_{C^{2,\alpha}(\mathcal{M})} \leq C$, independent of β. In turn, by considering $w = \Delta_{\mathcal{M}} v$, we may again apply the Schauder estimate to conclude that $\|v\|_{C^4(\mathcal{M})} \leq C\beta^{-1}$. Since we have that $\|v\|_{C^4(\mathcal{M})} \leq C\beta^{-1}$ and $\|v\|_{C^2(\mathcal{M})} \leq C$, we may use interpolation inequalities (see, e.g., Lemma 6.32 in [30]) to deduce that $\|v\|_{C^3(\mathcal{M})} \leq C\beta^{-1/2}$. These arguments then conclude the proof of Theorem 2.3. ∎

We remind the reader here that convergence of v as $\beta \to 0$ is towards μ_f, not μ. One can only guarantee convergence towards μ if one makes more specific assumptions upon the label error distribution p or on the empirical risk function F. We remark that the references in the previous proof referred to the Euclidean case, but can be extracted to the manifold case we consider here via standard localization arguments. We also emphasize that there are many other techniques and technical challenges associated with elliptic regularity (especially associated with boundary values) which were not relevant in this context; a standard reference is [30].

REFERENCES

[1] R. K. Ando and T. Zhang, *Learning on graph with Laplacian regularization*, in Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS'06, MIT Press, Cambridge, MA, 2006, pp. 25–32, http://dl.acm.org/citation.cfm?id=2976456.2976460.

[2] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Comput., 15 (2003), pp. 1373–1396.

[3] M. Belkin and P. Niyogi, *Towards a theoretical foundation for Laplacian-based manifold methods*, in Proceedings of the 18th Annual Conference on Learning Theory, COLT'05, Springer-Verlag, Berlin, Heidelberg, 2005, pp. 486–500, https://doi.org/10.1007/11503415_33.

[4] M. Belkin, P. Niyogi, and V. Sindhwani, *Manifold regularization: A geometric framework for learning from labeled and unlabeled examples*, J. Mach. Learn. Res., 7 (2006), pp. 2399–2434.

[5] A. Bertozzi, X. Luo, A. Stuart, and K. Zygalakis, *Uncertainty quantification in graph-based classification of high dimensional data*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 568–595, https://doi.org/10.1137/17M1134214.

[6] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities*, Oxford University Press, Oxford, 2013, https://doi.org/10.1093/acprof:oso/9780199535255.001.0001,

[7] G. Bourdaud, *Le calcul fonctionnel dans les espaces de Sobolev*, in Séminaire sur les Équations aux Dérivées Partielles, 1990–1991, École Polytechnic, Palaiseau, 1991, Exp. No. I.

[8] O. Bousquet, O. Chapelle, and M. Hein, *Measure based regularization*, in Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2004, pp. 1221–1228.

[9] D. Burago, S. Ivanov, and Y. Kurylev, *A graph discretization of the Laplace-Beltrami operator*, J. Spectr. Theory, 4 (2014), pp. 675–714, https://doi.org/10.4171/JST/83.

[10] J. Calder, *The game theoretic p-Laplacian and semi-supervised learning with few labels*, Nonlinearity, 32 (2019), pp. 301–330.

[11] J. Calder and D. Slepčev, *Properly-weighted graph Laplacian for semi-supervised learning*, Appl. Math. Optim., to appear.

[12] V. Caselles, A. Chambolle, and M. Novaga, *Total variation in imaging*, in Handbook of Mathematical Methods in Imaging. Vol. 1, 2, 3, Springer, New York, 2015, pp. 1455–1499.

[13] K. Chaudhuri and S. Dasgupta, *Rates of convergence for nearest neighbor classification*, in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds., Curran Associates, Red Hook, NY, 2014, pp. 3437–3445, http://papers.nips.cc/paper/5439-rates-of-convergence-for-nearest-neighbor-classification.pdf.

[14] P. G. Ciarlet and P.-A. Raviart, *Maximum principle and uniform convergence for the finite element method*, Comput. Methods Appl. Mech. Engrg., 2 (1973), pp. 17–31, https://doi.org/10.1016/0045-7825(73)90019-4.

[15] R. R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30.

[16] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 7426–7431, https://doi.org/10.1073/pnas.0500334102.

[17] O. Delalleau, Y. Bengio, and N. Le Roux, *Efficient non-parametric function induction in semi-supervised learning*, in Semi-Supervised Learning, Robert G. Cowell and Zoubin Ghahramani, eds., in Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, Barbados, 2005, Society for Artificial Intelligence and Statistics, 2005, pp. 96–103; available online at http://www.gatsby.ucl.ac.uk/aistats/.

[18] L. Devroye, L. Györfi, A. Krzyżak, and G. Lugosi, *On the strong universal consistency of nearest neighbor regression function estimates*, Ann. Statist., 22 (1994), pp. 1371–1385, https://doi.org/10.1214/aos/1176325633.

[19] L. Devroye, L. Györfi, G. Lugosi, and H. Walk, *On the measure of Voronoi cells*, J. Appl. Probab., 54 (2017), pp. 394–408, https://doi.org/10.1017/jpr.2017.7.

[20] M. P. do Carmo, *Riemannian Geometry*, Math. Theory Appl., translated from the second Portuguese edition by Francis Flaherty, Birkhäuser Boston, Boston, 1992.

[21] M. Dunlop, D. Slepčev, A. Stuart, and M. Thorpe, *Large data and zero noise limits of graph-based semi-supervised learning algorithms*, Appl. Comput. Harmon. Anal., 49 (2020), pp. 655–697.

[22] A. El Alaoui, X. Cheng, A. Ramdas, M. J. Wainwright, and M. I. Jordan, *Asymptotic behavior of ℓ_p-based Laplacian regularization in semi-supervised learning*, in Proceedings of the 29th Annual Conference on Learning Theory, 2016, pp. 879–906.

[23] L. C. Evans, *Partial Differential Equations*, 2nd ed., Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 2010, https://doi.org/10.1090/gsm/019.

[24] A. Gadde, A. Anis, and A. Ortega, *Active semi-supervised learning using sampling theory for graph signals*, in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 2014, pp. 492–501.

[25] N. García Trillos and R. Murray, *A new analytical approach to consistency and overfitting in regularized empirical risk minimization*, European J. Appl. Math., 28 (2017), pp. 886–921, https://doi.org/10.1017/S0956792517000201.

[26] N. García Trillos and D. Sanz-Alonso, *Continuum limits of posteriors in graph Bayesian inverse problems*, SIAM J. Math. Anal., 50 (2018), pp. 4020–4040, https://doi.org/10.1137/17M1138005.

[27] N. García Trillos and D. Slepčev, *Continuum limit of total variation on point clouds*, Arch. Ration. Mech. Anal., 220 (2015), pp. 1–49, https://doi.org/10.1007/s00205-015-0929-z.

[28] N. García Trillos, D. Slepčev, J. von Brecht, T. Laurent, and X. Bresson, *Consistency of Cheeger and ratio graph cuts*, J. Mach. Learn. Res., 17 (2016), 181.

[29] N. García Trillos, M. Gerlach, M. Hein, and D. Slepčev, *Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator*, Found. Comput. Math., 20 (2020), pp. 827–887.

[30] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Grundlehren Math. Wiss. 224, Springer-Verlag, Berlin, 1983.

[31] E. Giné and V. Koltchinskii, *Empirical graph Laplacian approximation of Laplace–Beltrami operators: Large sample results*, IMS Lecture Notes Monogr. Ser. 51, Institute of Mathematical Statistics, Beachwood, OH, 2006, pp. 238–259, https://doi.org/10.1214/074921706000000888.

[32] D. F. Gleich and M. W. Mahoney, *Using local spectral methods to robustify graph-based learning algorithms*, in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 2015, pp. 359–368.

[33] A. Grigor'yan, *Heat Kernel and Analysis on Manifolds*, AMS/IP Stud. Adv. Math. 47, American Mathematical Society, Providence, RI, International Press, Boston, 2009.

[34] Q. Gu, T. Zhang, J. Han, and C. H. Ding, *Selective labeling via error bound minimization*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2012, pp. 323–331.

[35] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-free Theory of Nonparametric Regression*, Springer Ser. Statist., Springer-Verlag, New York, 2002, https://doi.org/10.1007/b97848.

[36] M. Hein, J.-Y. Audibert, and U. von Luxburg, *Graph Laplacians and their convergence on random neighborhood graphs*, J. Mach. Learn. Res., 8 (2007), pp. 1325–1368.

[37] M. Hein, J.-Y. Audibert, and U. von Luxburg, *From graphs to manifolds—Weak and strong pointwise consistency of graph Laplacians*, in Learning Theory: Proceedings of the 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, 2005, Lecture Notes in Comput. Sci. 3559, Springer, Berlin, 2005, pp. 470–485.

[38] A. E. Hoerl and R. W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970), pp. 55–67.

[39] F. Isaia, *On the superposition operator between Sobolev spaces: Well-definedness, continuity, boundedness, and higher-order chain rule*, Houston J. Math., 41 (2015), pp. 1277–1294.

[40] T. Joachims, *Transductive learning via spectral graph partitioning*, in Proceedings of the 20th International Conference on Machine Learning (ICML-03), 2003, pp. 290–297.

[41] A. Kirichenko and H. van Zanten, *Estimating a smooth function on a large graph by Bayesian Laplacian regularisation*, Electron. J. Statist., 11 (2017), pp. 891–915, https://doi.org/10.1214/17-EJS1253.

[42] S. Kpotufe, *k-NN regression adapts to local intrinsic dimension*, in Advances in Neural Information Processing Systems 24, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, eds., Curran Associates, Red Hook, NY, 2011, pp. 729–737, http://papers.nips.cc/paper/4455-k-nn-regression-adapts-to-local-intrinsic-dimension.pdf.

[43] A. V. Little, M. Maggioni, and L. Rosasco, *Multiscale geometric methods for data sets* I: *Multiscale SVD, noise and curvature*, Appl. Comput. Harmon. Anal., 43 (2017), pp. 504–567, https://doi.org/https://doi.org/10.1016/j.acha.2015.09.009.

[44] A. Moscovich, A. Jaffe, and N. Boaz, *Minimax-optimal semi-supervised regression on unknown manifolds*, in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, A. Singh and J. Zhu, eds., Proceedings of Machine Learning Research 54, Fort Lauderdale, FL, PMLR, 2017, pp. 933–942, http://proceedings.mlr.press/v54/moscovich17a.html.

[45] E. A. Nadaraya, *On estimating regression*, Theory Probab. Appl., 9 (1964), pp. 141–142, https://doi.org/10.1137/1109020.

[46] B. Nadler, N. Srebro, and X. Zhou, *Semi-supervised learning with the graph Laplacian: The limit of infinite unlabelled data*, in Proceedings of the 22nd International Conference on Neural Information Processing Systems, 2009, pp. 1330–1338.

[47] A. Y. Ng, M. I. Jordan, and Y. Weiss, *On spectral clustering: Analysis and an algorithm*, in Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, MIT Press, Cambridge, MA, 2002, pp. 849–856.

[48] B. Osting, C. D. White, and É. Oudet, *Minimal Dirichlet energy partitions for graphs*, SIAM J. Sci. Comput., 36 (2014), pp. A1635–A1651, https://doi.org/10.1137/130934568.

[49] M. Renardy and R. C. Rogers, *An Introduction to Partial Differential Equations*, 2nd ed., Texts Appl. Math. 13, Springer-Verlag, New York, 2004.

[50] K. Rohe, S. Chatterjee, and B. Yu, *Spectral clustering and the high-dimensional stochastic blockmodel*, Ann. Statist., 39 (2011), pp. 1878–1915, https://doi.org/10.1214/11-AOS887.

[51] D. Romero, M. Ma, and G. B. Giannakis, *Kernel-based reconstruction of graph signals*, IEEE Trans. Signal Process., 65 (2016), pp. 764–778.

[52] L. Rosasco, S. Villa, S. Mosci, M. Santoro, and A. Verri, *Nonparametric sparsity and regularization*, J. Mach. Learn. Res., 14 (2013), pp. 1665–1714.

[53] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, 2002.

[54] Z. Shi, B. Wang, and S. J. Osher, *Error estimation of weighted nonlocal Laplacian on random point cloud*, Multiscale Model. Simul., submitted.

[55] A. Singer, *From graph to manifold Laplacian: The convergence rate*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 128–134, https://doi.org/10.1016/j.acha.2006.03.004.

[56] D. Slepčev and M. Thorpe, *Analysis of p-Laplacian regularization in semi-supervised learning*, SIAM J. Math. Anal., 51 (2019), pp. 2085–2120, https://doi.org/10.1137/17M115222X.

[57] A. J. Smola and R. Kondor, *Kernels and regularization on graphs*, in Learning Theory and Kernel Machines, Lecture Notes in Comput. Sci. 2777, Springer, Berlin, Heidelberg, 2003, pp. 144–158.

[58] M. Thorpe, S. Park, S. Kolouri, G. K. Rohde, and D. Slepčev, *A transportation L^p distance for signal analysis*, J. Math. Imaging Vision, 59 (2017), pp. 187–210, https://doi.org/10.1007/s10851-017-0726-4.

[59] R. J. Tibshirani, *Adaptive piecewise polynomial estimation via trend filtering*, Ann. Statist., 42 (2014), pp. 285–323.

[60] A. N. Tikhonov, *Regularization of incorrectly posed problems*, in Soviet Math. Dokl., 4 (1963), pp. 1624–1627.

[61] N. G. Trillos and D. Slepčev, *A variational approach to the consistency of spectral clustering*, Appl. Comput. Harmon. Anal., 45 (2018), pp. 239–281.

[62] J. W. Tukey, *Nonparametric estimation,* III. *Statistically equivalent blocks and multivariate tolerance regions–the discontinuous case*, Ann. Math. Statist., 19 (1948), pp. 30–39, https://doi.org/10.1214/aoms/1177730287.

[63] S. R. S. Varadhan, *On the behavior of the fundamental solution of the heat equation with variable coefficients*, Comm. Pure Appl. Math., 20 (1967), pp. 431–455.

[64] A. Venkitaraman, S. Chatterjee, and P. Händel, *Kernel Regression for Signals over Graphs*, preprint, https://arxiv.org/abs/1706.02191, 2017.

[65] U. von Luxburg, M. Belkin, and O. Bousquet, *Consistency of spectral clustering*, Ann. Statist., 36 (2008), pp. 555–586, https://doi.org/10.1214/009053607000000640.

[66] G. Wahba, *Spline Models for Observational Data*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 59, SIAM, Philadelphia, 1990, https://doi.org/10.1137/1.9781611970128.

[67] Y.-X. Wang, J. Sharpnack, A. J. Smola, and R. J. Tibshirani, *Trend filtering on graphs*, J. Mach. Learn. Res., 17 (2016), pp. 3651–3691.

[68] G. S. Watson, *Smooth regression analysis*, Sankhyā Ser. A, 26 (1964), pp. 359–372.

[69] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, *Learning with local and global consistency*, in Proceedings of the 16th International Conference on Neural Information Processing Systems, 2004, pp. 321–328.

[70] X. Zhu, Z. Ghahramani, and J. Lafferty, *Semi-supervised learning using Gaussian fields and harmonic functions*, in Proceedings of the Twentieth International Conference on Machine Learning, ICML'03, AAAI Press, Palo Alto, CA, 2003, pp. 912–919, http://dl.acm.org/citation.cfm?id=3041838.3041953.