Sampling Theory, Signal Processing, and Data Analysis (2023) 21:35
https://doi.org/10.1007/s43670-023-00075-5

ORIGINAL ARTICLE

Rates of convergence for regression with the graph poly-Laplacian

Nicolás García Trillos[1] · **Ryan Murray**[2] · **Matthew Thorpe**[3]

Received: 3 September 2022 / Accepted: 30 October 2023 / Published online: 27 November 2023

Abstract
In the (special) smoothing spline problem one considers a variational problem with a quadratic data fidelity penalty and Laplacian regularization. Higher order regularity can be obtained via replacing the Laplacian regulariser with a poly-Laplacian regulariser. The methodology is readily adapted to graphs and here we consider graph poly-Laplacian regularization in a fully supervised, non-parametric, noise corrupted, regression problem. In particular, given a dataset $\{x_i\}_{i=1}^n$ and a set of noisy labels $\{y_i\}_{i=1}^n \subset \mathbb{R}$ we let $u_n : \{x_i\}_{i=1}^n \to \mathbb{R}$ be the minimizer of an energy which consists of a data fidelity term and an appropriately scaled graph poly-Laplacian term. When $y_i = g(x_i) + \xi_i$, for iid noise ξ_i, and using the geometric random graph, we identify (with high probability) the rate of convergence of u_n to g in the large data limit $n \to \infty$. Furthermore, our rate is close to the known rate of convergence in the usual smoothing spline model.

Keywords Non-parametric regression on unknown domains · Supervised learning · Asymptotic consistency · Rates of convergence · PDEs on graphs · Nonlocal variational problems

Mathematics Subject Classification 49J55 · 49J45 · 62G20 · 35J20

Communicated by Rayan Saab.

✉ Matthew Thorpe
matthew.thorpe@warwick.ac.uk

Nicolás García Trillos
garciatrillo@wisc.edu

Ryan Murray
rwmurray@ncsu.edu

[1] Department of Statistics, University of Wisconsin–Madison, Madison, WI 53706, USA

[2] Department of Mathematics, North Carolina State University, Raleigh, NC 27695, USA

[3] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

1 Introduction

Given the applications to signal processing and computer science the smoothing spline problem has been attracting the interest of statisticians since the 1960s [1, 2]. The problem can be stated as [3]: given feature vectors $\{x_i\}_{i=1}^n \subset \Omega \subset \mathbb{R}^d$ and labels $\{y_i\}_{i=1}^n \subset \mathbb{R}$ minimize

$$\mathcal{E}_n^{(\text{spline})}(u) = \frac{1}{n}\sum_{i=1}^n |y_i - u(x_i)|^2 + \tau \|\nabla^s u\|^2_{\mathrm{L}^2(\Omega)} \tag{1}$$

over all $u \in \mathrm{H}^s(\Omega)$, where $\mathrm{H}^s(\Omega)$ is the Sobolev space with square integrable sth (weak) derivative on Ω. Here, one tries to find an unknown function $u\colon \Omega \to \mathbb{R}$ that is trying to match the observed labels $\{y_i\}_{i=1}^n$ at $\{x_i\}_{i=1}^n$ (through the data fidelity term $\frac{1}{n}\sum_{i=1}^n |y_i - u(x_i)|^2$) whilst being smooth in the sense of H^s regularization (by including the regularization term $\tau\|\nabla^s u\|^2_{\mathrm{L}^2(\Omega)}$). It is important to note that the regularity penalty is applied uniformly throughout the domain Ω.

Recently, the spline methodology has found use in data science and machine learning as a candidate for semi-supervised or fully-supervised learning. Zhu et al. [4] introduced the following variational problem as a method for finding missing labels. They assumed that for every pair of feature vectors x_i, x_j with $i, j \in \{1, \dots, n\}$ one has a measure of similarity $W_{i,j}$. Further assuming that there is no error in the observed labels $\{y_i\}_{i \in Z_n}$, where $Z_n \subsetneq \{1, \dots, n\}$, they proposed the variational problem: minimize

$$\mathcal{E}_n^{(\text{ZGL})}(u) = \sum_{i,j=1}^n W_{i,j}|u(x_i) - u(x_j)|^2 \tag{2}$$

subject to $u(x_i) = y_i$ for all $i \in Z_n$ over $u\colon \{x_i\}_{i=1}^n \to \mathbb{R}$. The power of this method is that the regularization will be applied more strongly when the density of data is higher, at least when the weights $W_{i,j}$ are based on the proximity of points in the Euclidean sense. Indeed, in the ε-graph setting, the continuum, $n \to \infty$, limit of (2) is, up to a multiplicative constant,

$$\mathcal{E}_\infty(u) = \int_\Omega |\nabla u(x)|^2 \rho^2(x)\, \mathrm{d}x \tag{3}$$

where ρ is the density of the data. We see that where ρ is large the minimizer of (3) should be smoother, and conversely when ρ is smaller the minimizer can fluctuate more. This is typically desirable behaviour in classification tasks since the minimizer can be expected to be approximately constant within clusters and quickly transitioning outside of clusters where the density of data is assumed to be low.

Several types of convergence results connecting (2) to (3) exist in the literature. For instance pointwise convergence of the objective functionals was established in [5, 6]. Rates of convergence between constrained minimizers of (2) to constrained minimizers of (3) appeared in [7]. Further results concern the convergence of the Laplacian

operator, without rates in [8–10] and with rates in [11–14], the game theoretic Laplacian in [15], the p-Laplacian in [16], and the ∞-Laplacian in [17]. However, none of these results consider *asymptotic consistency* in the sense that the minimizer of (2) converges to a "true function". (It would be more accurate to describe the above results as *convergence* properties of the method.) To our knowledge we are the first to consider consistency in the graph-based setting.

When there is uncertainty in the observed labels, minimising (2) with constraints is not the natural model. Instead, as in (1), we can reformulate the Zhu et al. model in the fully supervised setting with *soft* labels by

$$\mathcal{E}_{n,\tau}^{(\mathbf{y}_n)}(u) = \frac{1}{n}\sum_{i=1}^{n} |u(x_i) - y_i|^2 + \frac{\tau}{n^2\varepsilon^2}\sum_{i,j=1}^{n} W_{i,j}|u(x_i) - u(x_j)|^2 \tag{4}$$

where we include the correct scaling on the second term. The parameter τ controls the weighting between regularity and matching the data: formally, $\tau = 0$ corresponds to the hard constrained problem. In some settings the large data limits for minimizers of (4) can be inferred from the *hard* constrained problem, see [7].

To complete the generalization of the Zhu et al. model to an analogue of the spline problem (1) we discuss higher order regularization. The Dirichlet energy $\mathcal{E}_n^{(\mathrm{ZGL})}$ can be written in inner product form:

$$\mathcal{E}_n^{(\mathrm{ZGL})}(u) = \langle \Delta_n u, u\rangle_{\mathrm{L}^2(\mu_n)}$$

where Δ_n is the graph Laplacian and $\mathrm{L}^2(\mu_n)$ is the space of L^2 functions with respect to the empirical measure μ_n (defined in the following subsection). A natural method to introduce higher order regularity is to consider higher powers of the graph-Laplacian, in particular

$$\mathcal{E}_{n,\tau}^{(\mathbf{y}_n)}(u) = \frac{1}{n}\sum_{i=1}^{n}(u(x_i) - y_i)^2 + \tau\langle \Delta_n^s u, u\rangle_{\mathrm{L}^2(\mu_n)}, \tag{5}$$

in which case (4) is the special case of (5) with $s = 1$. In the context of graphs this model was introduced in [18], although the fractional Laplacian, including non-local and discrete versions, has been of interest in applied mathematics for much longer, see for example [19] and references therein. As observed in [4] and (in terms of uncertainty quantification) [20], (5) has an interpretation as a maximum a-posteriori (MAP) estimate for the Gaussian process regression method (also known as kriging). As an aside we mention works that have explored the connection between discrete and continuum problems in the Bayesian setting such as [21, 22] and [23], where the latter introduces Matérn priors on graphs and studies their continuum limits. In this paper we use the setting of [18] to recover a labeling function g as the MAP estimator given the data $\{(x_i, y_i)\}_{i=1}^n$ and using the graph poly-Laplacian to define a prior. In this Bayesian context, these works identify the higher-order regularizers as beneficial for fitting functions which are expected to have higher degree of regularity. Our focus will be on recovering the labels g *as well as* some of its higher order information at the data points $\{x_i\}_{i=1}^n$ in the form of powers of the Laplacian of g.

Theoretical analysis of splines dates back to the 1960s and we refer to [3] for an overview of more historical works and only mention a few select (more recent) references here related to large data limits. Convergence in norm of special splines under various settings have been studied in [24–32] and pointwise convergence results in [33–37]. The optimal rates of convergence for special splines were established in [38] with rate $\|u_\tau^* - g\|_{\mathrm{L}^2(\Omega)} = O(n^{-\frac{s}{2s+d}})$ where u_τ^* is the spline estimate minimizing (1). The result in [38] further established rates of convergence of derivatives, in particular showing $\|\nabla^k u_\tau^* - \nabla^k g\|_{\mathrm{L}^2(\Omega)} = O(n^{-\frac{s-k}{2s+d}})$. The general splines problem writes (1) in a more abstract framework. In particular, one seeks to find $u \in \mathcal{H}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space that can be decomposed $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, as the minimizer of

$$\mathcal{E}_n^{(\text{gen spline})}(u) = \frac{1}{n}\sum_{i=1}^{n} |y_i - L_i u|^2 + \tau \|P_1 u\|_{\mathcal{H}}^2$$

where $L_i \in \mathcal{H}^*$ (where $\mathcal{H}^*$ is the dual space of $\mathcal{H}$) and $P_1 \colon \mathcal{H} \to \mathcal{H}_1$ is the orthogonal projection. The motivation for the model is that $\mathcal{H}_1$ is an infinite dimensional space where one wants to regularise, whilst $\mathcal{H}_0$ is a finite dimensional space which doesn't need regularization. For an appropriate choice of $\mathcal{H}$ and L_i one recovers the special smoothing spline problem (as a special case of the general smoothing spline problem). In particular, $\mathcal{H}_0$ is the space of polynomials of degree at most s and $\mathcal{H}_1$ is the subspace of H^s with $\nabla^k u = 0$ for $k = 0, \dots, s-1$ and norm $\|\nabla^s u\|_{\mathrm{L}^2}$, see [39] for further details. The general smoothing spline problem has itself attracted attention with large data convergence in norm results appearing in [40–45] and weak convergence results in [39].

In this paper, using the model (5), we consider the problem of non-parametric regression of a noisy signal g observed at finitely many points that are randomly selected from an *unknown* probability distribution μ on the torus $\mathcal{T}$. More precisely, we assume we are given a set of feature vectors $\{x_i\}_{i=1}^n \subset \Omega := \mathcal{T}$ and a set of associated noisy real valued labels $\{y_i\}_{i=1}^n$ satisfying

$$y_i = g(x_i) + \xi_i, \tag{6}$$

from which we wish to recover the true signal g, also known as the label function. The random variables ξ_i are assumed to be mean zero, sub-Gaussian and independent. Our main results establish variance and bias estimates for the error of approximation of the signal g and of some of its higher order information by the solution u of a variational problem characterized by a graph PDE of the form

$$\tau \Delta_n^s u + u = y. \tag{7}$$

Up to logarithms, we establish an $O(n^{-\frac{s}{d+4s}})$ rate of convergence of minimizers of $\mathcal{E}_{n,\tau}^{(\mathbf{y}_n)}$ in (5) to g. This is comparable to the $O(n^{-\frac{s}{d+2s}})$ rate of convergence in splines, see [38]. Projecting the samples onto the first K eigenvectors of the graph Laplacian gives a better rate of $O(n^{-\frac{s}{d+2s}})$ (the minimax rate) [46], however the optimal choice

K depends on the H^s norm of g, which will usually be unknown. See Remarks 9 and 10 for further details.

Methods such as kernel ridge regression are closely related to smoothing spline models. These methods use a data fidelity (on $\{x_i\}_{i=1}^n$) plus regularization on Ω (in particular incorporating Ω's geometry *explicitly*) to attempt to recover g. In that setting, the question of how to set regularization parameters was studied in [47, 48]. Some very recent works have focused on studying the "ridgeless case", where one considers the limit as one sets the regularization parameter to zero, with both positive [49] and negative [50] results depending on the richness of the data. A related approach is to interpolate between data points using C^m penalization [51–53] or Sobolev penalization [54–56].

There are connections between the Gaussian process regression (kriging) method approach that we take here and the generalized lasso model (which includes the lasso, the fused lasso, trend filtering, and the more closely related to our work: graph fused lasso, graph trend filtering, and Kronecker trend filtering), see for example [57]. Both Gaussian process regression and generalized lasso attempt to recover an unknown function g from noisy observations of the form (6) in the fully supervised setting (i.e. for each feature vector x_i we have an observation y_i). However, in the lasso models the function g is assumed to be linear, i.e. $g(x) = \beta \cdot x$ where β is an unknown vector which parametrises g. The fused lasso, on the other hand, uses a total variation regularization in place of the graph poly-Laplacian considered here. In grid graphs this has been considered in [58–60] where the estimator is shown to be minimax (the estimator performs best amongst all other estimators in the worst case). Further results have considered chain graphs [61], and k-NN and ε-connected graphs [62] (the ε-connected graph setting is also the setting of this paper). In particular, the L^2 convergence rate of the fused lasso on an ε-connected graph is (ignoring logarithms) $O(n^{-\frac{1}{d}})$ which at least in some settings is the minimax rate [62]. Up to logarithms, and assuming a sufficiently smooth signal g, our basic L^2 convergence rates for the approximation of g coincide with these rates. This is also approximately the L^∞ convergence rate given in [63] for the case $s = 1$ in (7), which is the minimax L^∞ rate given in [64]. At this point we would like to remark that our variance estimates are meaningful and converge to zero with growing n even if the regularization parameter τ is not scaled down to zero. Our results characterize precisely the continuum limit of the solutions to the graph PDE (7), and are of relevance in case one were interested, not only on denoising, but also in enforcing additional regularization. We also remark that in our results we provide additional information about the convergence towards g, by giving convergence rates for higher-order derivatives.

Other approaches for high order regularization that do not consider Gaussian priors use instead a non-linear p-Laplacian operator for large enough p defined by

$$\Delta_n^{(p)} u(x_i) = \frac{1}{n\varepsilon^p} \sum_{j=1}^{n} W_{ij}(u(x_i) - u(x_j))|u(x_i) - u(x_j)|^{p-2}$$

(which is consistent with the definition in (8) for $p = 2$ and written $\Delta_n := \Delta_n^{(2)}$). In the graph setting, results like those in [65] establish that solutions of a p-Dirichlet

regularised problem converge with rate $n^{-\frac{1}{4}}$ to the solution of an analogue continuum non-local variational problem; although the setting differs from ours as we scale the connectivity of our graph to obtain a local limit whilst in [65] the connectivity of the graph remains fixed and the limit is to a nonlocal variational problem. Naturally, the advantage of the framework in [65] is that the dimension of the space essentially plays no role in the analysis (depending on the precise edge model one uses) and therefore it is enough to consider the problem in 1D (as the authors do). On the other hand, by not scaling down the connectivity threshold it is not possible to recover the local geometry. The same authors, in the same setting, show a rate of convergence for the associated gradient flow [66]. As was mentioned earlier when discussing generalized Lasso models (in particular in graph trend filtering), total variation is another tool used to regularise regression and classification problems. This has motivated theoretical works like [67] which study the convergence of graph total variation to a continuum weighted total variation (the same paper proposed a topology to study the convergence that didn't require regularity—in particular pointwise evaluation—of the continuum function). Total variation functionals are also widely used for clustering and segmentation such as in graph cut methods, for example ratio or Cheeger cuts [68, 69], graph modularity clustering [70, 71], and Ginzburg–Landau segmentation [72–74].

Since we have observations for all feature vectors our problem is in the fully-supervised setting. The semi-supervised setting with Laplacian regularization (closely related to (7) with $s = 1$ but with hard constraints as opposed to having a penalty term) has been considered in [7] which show an "ill-posedness result" (the labels disappear in the large data limit) if the number of labelled points scales below a critical threshold, and a "well-posedness result" (the labels remain in the continuum problem) when the number of labels scales linearly with n. Using graph p-Laplacian regularization with finite labels the authors in [15, 16] show that whether the variational problem is asymptotically well-posed depends on the choice of p and how the graph is scaled. For the fractional graph Laplacian with finite labels, it is shown in [18] that the problem is ill-posed if $s \leq \frac{d}{2}$ or the length scale on the graph is sufficiently large and conjectured that this is sharp.

We also mention that other approaches to regression on unknown manifolds include [75], where local tangent planes around points are carefully constructed to apply regression methods in the more classical functional data setting. Our approach is markedly different as it does not rely on the construction of *extrinsic* geometric objects. In particular, once a proximity graph is defined on the data cloud, all regularisers and the resulting PDEs become *intrinsic* to the graph.

Here we have reviewed a wide range of models which incorporate higher-order regularity. From a high-level, the inclusion of higher-order regularity terms will lead to improved fitting when the underlying source of labels has more regular dependence on the features, and this is quantified by concrete consistency results. Furthermore, higher-order regularizers have proved useful in smoothly propagating labels in semi-supervised and Bayesian learning settings. This work adds to this literature in the context of higher-order Laplacian regularization.

In the remainder of this section we define the graph and continuum operators that are analysed in the paper, and then state our main results.

1.1 Discrete operators

We begin by stating our basic assumptions on the data, and the graph that we use to model it:

(A1) **Assumptions on the domain Ω :** Ω is a d-dimensional torus.

The assumption that Ω is a torus is largely to simplify our arguments; we expect the same result to hold on manifolds without boundaries. For sets with boundaries (either in the manifold setting or when Ω is an open subset of $\mathbb{R}^d$) we do not expect the same results. This is because the rate of convergence of the graph Laplacian to its continuum limit deteriorates near the boundary, see for example [7, Theorem 3.2]. We leave the question of rates of convergence for regression with the graph poly-Laplacian open in this setting.

(A2) **Assumptions on the feature vectors** x_i **:** $x_i \overset{\text{iid}}{\sim} \mu$ where $\mu \in \mathcal{P}(\Omega)$, where $\mathcal{P}(\Omega)$ is the set of probability measures on Ω;

(A3) **Assumption on the density of** μ **:** μ has a density $\rho \in \mathrm{C}^\infty(\Omega)$ that is bounded from above and below by positive constants, i.e. $0 < \rho_{\min} := \min_{x\in\Omega} \rho(x) \leq \max_{x\in\Omega} \rho(x) =: \rho_{\max} < +\infty$.

(A4) **Assumptions on the graph constructed using the data** $\{x_i\}$ **:** $G_n := (\Omega_n, W_n)$ where $\Omega_n = \{x_i\}_{i=1}^n$ are the nodes and $W_n = (W_{i,j})_{i,j=1}^n$ are the edge weights defined by $W_{i,j} = \eta_\varepsilon(|x_i - x_j|)$ for $i \neq j$ and $W_{i,i} = 0$. Here $\eta_\varepsilon = \frac{1}{\varepsilon^d}\eta(\cdot/\varepsilon)$ and where $\eta : [0, +\infty) \to [0, +\infty)$ is assumed to satisfy:
(a) $\eta(t) > \frac{1}{2}$ for all $t \leq \frac{1}{2}$ and $\eta(t) = 0$ for all $t \geq 1$;
(b) η is decreasing.

(A5) **Assumptions on the labelled data :** for each $i \in \mathbb{N}$, $y_i = g(x_i) + \xi_i$, for $g \in \mathrm{H}^{2s+1+\frac{d}{2}}(\Omega)$. and $\xi_i \in \mathbb{R}$ are independent and identically distributed (iid), sub-Gaussian centred noise (where sub-Gaussian by definition means there exists $C > c > 0$ such that $\mathbb{P}\left(|\xi_j| > t\right) \leq Ce^{-ct^2}$ for all $t \geq 0$).

Our model assumes noise in the labels, but not in the feature vectors. An interesting further question would be what happens if feature vectors are also noisy. This question has been partially studied in the context of adversarial noise (where the adversary perturbs the feature vectors) in the semi-supervised setting; the results of [76] show that if the adversary cannot move feature vectors more than some maximum distance then the method is still asymptotically consistent.

Remark 1 1. The assumption that $g \in \mathrm{H}^{2s+1+\frac{d}{2}}$ is because we show $\|u_\tau^*\|_{\mathrm{H}^k(\Omega)} \lesssim \|g\|_{\mathrm{H}^k(\Omega)}$, where u_τ^* is the noiseless solution in the continuum setting (i.e. solves $\tau\Delta_\rho^s u + u = g$ where Δ_ρ is the continuum limit of the graph Laplacian, see (12) and compare to (7)). By Morrey's inequality we have that $u_\tau^* \in \mathrm{C}^{2s+1}(\Omega)$, and hence u_τ^* solves the above equation in a strong sense (see Lemma 2.14 for details).

2. The assumptions on η are technical in nature and are imposed to facilitate some very concrete steps in our analysis. Assumption (A4)(b) is slightly stronger than what is typically required in related papers, and will only be used to simplify our computations in, for example, Lemma 2.3.

The graph Laplacian Δ_n plays an important role in the regularization and is defined as follows:

$$\Delta_n := \frac{2}{n\varepsilon^2}(D_n - W_n), \qquad D_n = (D_{i,j})_{i,j=1}^n \text{ diagonal matrix with } D_{i,i} = \sum_{k=1}^n W_{i,k}. \tag{8}$$

Here we have chosen what is called the *unnormalized* graph Laplacian.

Throughout the paper we will denote the empirical measure $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. We will define an inner product with respect to a (usually probability) measure ν by

$$\langle u, v\rangle_{\mathrm{L}^2(\nu)} := \int_\Omega u(x)v(x)\,\mathrm{d}\nu(x) \qquad \text{for } u, v \text{ measurable w.r.t. } \nu,$$

and the associated L^2 norm by $\|u\|_{\mathrm{L}^2(\nu)} = \sqrt{\langle u, u\rangle_{\mathrm{L}^2(\nu)}}$. When $\nu = \mu_n$ then the norm can be written $\|u\|_{\mathrm{L}^2(\mu_n)} = \sqrt{\frac{1}{n}\sum_{i=1}^n u(x_i)^2}$.

There is a small abuse in notation in how we define Δ_n since we will also write $\Delta_n\colon \mathrm{L}^2(\mu_n) \to \mathrm{L}^2(\mu_n)$; in this case we associate $u_n \in \mathrm{L}^2(\mu_n)$ with its vector representation $(u_n(x_1), \dots, u_n(x_n))^\top$. With a secondary abuse of notation we can apply Δ_n to a continuous function on a continuum domain. That is, if $u \in \mathrm{C}^0(\Omega)$ then we can interpret $\Delta_n u = \Delta_n(u\lfloor_{\Omega_n})$ as the graph Laplacian applied to the restriction of u onto the data points $\Omega_n = \{x_i\}_{i=1}^n \subset \Omega$.

We can define the graph derivative by $\nabla_n u(x_i, x_j) = \frac{1}{\varepsilon}\sqrt{W_{ij}}(u_j - u_i)$ which is an anti-symmetric divergence (i.e. satisfies $\nabla_n u(x_i, x_j) = -\nabla_n u(x_j, x_i)$). Using the norm

$$\langle U, V\rangle_{\mathrm{L}^2(\mu_n\otimes\mu_n)} = \frac{1}{n^2}\sum_{i,j=1}^n U(x_i, x_j)V(x_i, x_j)$$

on the set of anti-symmetric divergences $U\colon \{x_i\}_{i=1}^n \times \{x_i\}_{i=1}^n \to \mathbb{R}$ we can define the graph divergence to be the negative adjoint of the graph derivative, i.e. $\mathrm{div}_n U(x_i) = \frac{2}{n\varepsilon}\sum_{j=1}^n \sqrt{W_{ij}}U(x_i, x_j)$. The normalization in the graph derivative is chosen so that the graph Laplacian can be defined as the negative of the graph divergence applied to the graph derivative: $\Delta_n = -\mathrm{div}_n \circ \nabla_n$.

Given $\mathbf{a}_n = (a_1, \dots, a_n)$, with $a_i \in \mathbb{R}$, we let

$$\mathcal{E}_{n,\tau}^{(\mathbf{a}_n)}(u_n) = \frac{1}{n}\sum_{i=1}^n |u_n(x_i) - a_i|^2 + \tau R_n^{(s)}(u_n), \tag{9}$$

where the regularization is given by

$$R_n^{(s)}(u_n) = \langle \Delta_n^s u_n, u_n\rangle_{\mathrm{L}^2(\mu_n)}, \tag{10}$$

and here s is a positive integer with Δ_n^s the s-th power of the matrix. We will mostly be concerned with the situation where $\mathbf{a}_n = \mathbf{y}_n = (y_1, \ldots, y_n)$, which gives the energy

$$\mathcal{E}_{n,\tau}^{(\mathbf{y}_n)}(u_n) = \frac{1}{n}\sum_{i=1}^{n} |u_n(x_i) - y_i|^2 + \tau R_n^{(s)}(u_n) \tag{11}$$

We will define $u_{n,\tau}^*$ to be the minimizer of (11). Note that when $s = 1$,

$$R_n^{(1)}(u_n) = \frac{1}{n^2\varepsilon^2}\sum_{i,j=1}^{n} W_{i,j}|u_n(x_i) - u_n(x_j)|^2$$

and the regularization functional is the graph Dirichlet energy. We define $R_n^{(s)}$ for non-integer powers via the eigenvector–eigenvalue expansion (however our results consider only integer powers). That is, we let $(\lambda_i^{(n)}, q_i^{(n)})$ be eigenpairs of Δ_n then, since $\{q_i^{(n)}\}_{i=1}^n$ form an orthonormal basis of $\mathrm{L}^2(\mu_n)$, we can write (for any $s \in \mathbb{R}$)

$$R_n^{(s)}(u_n) = \sum_{i=1}^{n} (\lambda_i^{(n)})^s \langle u_n, q_i^{(n)}\rangle_{\mathrm{L}^2(\mu_n)}^2.$$

1.2 Continuum operators

We now define the appropriate continuum operators and variational formulations. It is well-known that as $n \to \infty$, the operator Δ_n converges to a continuum limit Δ_ρ [7, 9–13, 15], where Δ_ρ is the differential operator defined by

$$\Delta_\rho \phi := -\frac{\sigma_\eta}{\rho}\mathrm{div}(\rho^2 \nabla \phi) \tag{12}$$

and σ_η is the constant defined by

$$\sigma_\eta := \int_{\mathbb{R}^d} \eta(|h|)|h_1|^2 \, \mathrm{d}h. \tag{13}$$

For $\tau > 0$ fixed, the continuum objective functional is defined by

$$\mathcal{E}_{\infty,\tau}^{(g)}(u) = \int_\Omega |u(x) - g(x)|^2 \rho(x) \, \mathrm{d}x + \tau R_\infty^{(s)}(u) \tag{14}$$

where

$$R_\infty^{(s)}(u) = \langle \Delta_\rho^s u, u\rangle_{\mathrm{L}^2(\mu)}. \tag{15}$$

We will define u_τ^* to be the minimizer of (14). Again, we observe that when $s = 1$ the regularization functional,

$$R_\infty^{(1)}(u) = \sigma_\eta \int_\Omega |\nabla u(x)|^2 \rho^2(x)\, \mathrm{d}x,$$

is a weighted Dirichlet energy.

We remark that, by the fact that ρ is bounded from below, we may integrate by parts to obtain

$$c R_\infty^{(s)}(u) \leq \int_\Omega |\nabla^s u(x)|^2 \, \mathrm{d}x \leq C R_\infty^{(s)}(u)$$

for some constants $C > c > 0$.

We can also define $R_\infty^{(s)}$ for non-integer powers analogously to the discrete case. More concretely, by the spectral theorem and the fact that Ω is compact, if we let (λ_i, q_i) be eigenpairs of Δ_ρ then $\{q_i\}_{i=1}^\infty$ form an orthonormal basis of $\mathrm{L}^2(\mu)$. In turn we can define

$$R_\infty^{(s)}(u) = \sum_{i=1}^\infty \lambda_i^s \langle u, q_i \rangle_{\mathrm{L}^2(\mu)}^2$$

which is well-defined for any $s \in \mathbb{R}$.

1.3 Main results

Our results are to bound the bias and variance of the estimator $u_{n,\tau}^*$, defined as the minimizer of $\mathcal{E}_{n,\tau}^{(\mathbf{y}_n)}$. Following the terminology of [47] we define the variance of the estimator by

$$\| u_{n,\tau}^* - u_\tau^* \lfloor \Omega_n \|_{\mathrm{L}^2(\mu_n)}$$

where u_τ^* is the minimizer of $\mathcal{E}_{\infty,\tau}^{(g)}$ and $\lfloor \Omega_n$ is the restriction to Ω_n, and the bias is defined to be

$$\| u_\tau^* - g \|_{\mathrm{L}^2(\mu)}.$$

The main results are the following.

1.3.1 L^2 variance estimates

We state the L^2 variance estimates in the following theorem.

Theorem 1.1 (Variance Estimates) *Let Assumptions (A1)–(A5) hold and $s \in \mathbb{N}$. We define $\mathcal{E}_{n,\tau}^{(\mathbf{y}_n)}$ by (11) and $\mathcal{E}_{\infty,\tau}^{(g)}$ by (14) where $R_n^{(s)}$ is defined by (10), $R_\infty^{(s)}$ by (15), Δ_n*

by (8) and Δ_ρ by (12). Let $u^*_{n,\tau}$ be the minimizer of $\mathcal{E}^{(\mathbf{y}_n)}_{n,\tau}$ and u^*_τ be the minimizer of $\mathcal{E}^{(g)}_{\infty,\tau}$. Then, for all $\alpha > 1$, there exists $\varepsilon_0 > 0$, $\tau_0 > 0$, $C > c > 0$ such that for all ε, n satisfying

$$\varepsilon_0 \geq \varepsilon \geq C\sqrt[d]{\frac{\log(n)}{n}} \tag{16}$$

and $\tau \in (0, \tau_0)$ we have

$$\|u^*_{n,\tau} - u^*_\tau \lfloor_{\Omega_n} \|_{L^2(\mu_n)} \leq C\left(\sqrt{\frac{\log(n)}{n\varepsilon^d}} + \frac{\varepsilon^{2s}}{\tau} + \tau\varepsilon\right)$$

with probability at least $1 - C\left(n^{-\alpha} + ne^{-cn\varepsilon^{d+4s}}\right)$.

Let $\mathbf{g}_n = (g(x_1), \ldots, g(x_n))$, and let $u^{g*}_{n,\tau}$ be the minimizer of the "noiseless" energy $\mathcal{E}^{(\mathbf{g}_n)}_{n,\tau}$. The proof of Theorem 1.1 is divided into two steps: in the first we compare $u^*_{n,\tau}$ and $u^{g*}_{n,\tau}$ (corresponding to averaging in y) which gives a quantitative bound on the effect of the noise; in the second part we compare $u^{g*}_{n,\tau}$ with u^*_τ (which corresponds to averaging in x). We do this in Sects. 2.1 and 2.2 respectively.

Remark 2 We notice that the estimates are meaningful for fixed τ when n goes to infinity, i.e. τ is not required to become smaller with growing n.

Remark 3 In addition to the bound in $\mathrm{L}^2(\mu_n)$ between $u^*_{n,\tau}$ and u^*_τ we are able to show a bound between the Laplacians $\Delta_n^{\frac{s}{2}} u^*_{n,\tau}$ and $\Delta_\rho^{\frac{s}{2}} u^*_\tau \lfloor_{\Omega_n}$ when s is even. More precisely, our results show,

$$\left\| \Delta_n^{\frac{s}{2}} u^*_{n,\tau} - \Delta_\rho^{\frac{s}{2}} u^*_\tau \lfloor_{\Omega_n} \right\|_{\mathrm{L}^2(\mu_n)} \leq C\left(\sqrt{\frac{\log(n)}{n\varepsilon^d \tau}} + \frac{\varepsilon^s}{\tau} + \varepsilon\right)$$

with the same probability as in Theorem 1.1. This inequality likely generalizes to odd s, but to prove it using the methods in this paper we would require a pointwise convergence result for the graph derivative which is beyond the scope of the paper.

Remark 4 We offer a comparison with the estimates in [65] (although note that a direct comparison is not possible as we scale $\varepsilon \to 0$ whilst [65] work in the setting where $\varepsilon > 0$ is fixed). If, as in [65], we fix $\tau > 0$, and therefore absorb it into our constants, and choose $s = 1$ then the error bound simplifies to

$$\|u^*_{n,\tau} - u^*_\tau \lfloor_{\Omega_n} \|_{\mathrm{L}^2(\mu_n)} \leq C\left(\sqrt{\frac{\log(n)}{n\varepsilon^d}} + \varepsilon\right).$$

Unfortunately, optimising over ε implies a scaling in $\varepsilon = \varepsilon_n$ of

$$\varepsilon_n \sim \left(\frac{\log(n)}{n}\right)^{\frac{1}{d+2}}$$

which is outside of the conditions of Theorem 1.1 as (16) does not hold (one needs $n\varepsilon^{d+4} \gg \log(n)$ in order to get a high probability bound). Instead we choose $\varepsilon_n \sim \left(\frac{\log(n)}{n}\right)^{\frac{1}{d+4}}$. With this choice the error scales as

$$\|u^*_{n,\tau} - u^*_\tau \lfloor \Omega_n \|_{\mathrm{L}^2(\mu_n)} \lesssim \left(\frac{\log(n)}{n}\right)^{\frac{1}{d+4}}.$$

The results in [65] show that for the p-Laplacian regularized problem a rate of convergence $n^{-\frac{1}{4}}$ when $d = 1$, $s = 1$ and $\varepsilon > 0$ is fixed independently of n, compared to our rate of convergence of $n^{-\frac{1}{5}}$ (up to logarithms).

1.3.2 L^2 bias estimates

We have the following L^2 bias estimate.

Theorem 1.2 (Bias Estimates) *Let Assumptions (A1), (A3) hold and $\tau > 0$, $s \geq 1$ and $g \in H^s(\Omega)$. We define $\mathcal{E}^{(g)}_{\infty,\tau}$ by (14) where $R^{(s)}_\infty$ is defined by (15) and Δ_ρ by (12). Let u^*_τ be the minimizer of $\mathcal{E}^{(g)}_{\infty,\tau}$, then*

$$\|u^*_\tau - g\|_{L^2(\mu)} \leq \tau \|\Delta^s_\rho g\|_{L^2(\mu)}.$$

The theorem is proved in Sect. 3.

Remark 5 We are also able to show that, for s even,

$$\left\| \Delta^{\frac{s}{2}}_\rho (g - u^*_\tau) \right\|_{\mathrm{L}^2(\mu)} \leq \sqrt{\frac{\tau}{2}} \left\| \Delta^s_\rho g \right\|_{\mathrm{L}^2(\mu)}.$$

1.3.3 L^2 error estimates

The results from Sects. 1.3.1 and 1.3.2 can be combined to bound the error between $u^*_{n,\tau}$ and g.

Theorem 1.3 *Let Assumptions (A1)–(A5) hold and let $s \in \mathbb{N}$. We define $\mathcal{E}^{(\mathbf{y}_n)}_{n,\tau}$ by (11) where $R^{(s)}_n$ is defined by (10) and Δ_n by (8). Let $u^*_{n,\tau}$ be the minimizer of $\mathcal{E}^{(\mathbf{y}_n)}_{n,\tau}$. Then for every $\alpha > 1$ there exists $\varepsilon_0 > 0$, $C > c > 0$ such that for all ε, n satisfying (16) and $\tau \in (0, \tau_0)$ we have*

$$\|u^*_{n,\tau} - g \lfloor \Omega_n \|_{L^2(\mu_n)} \leq C \left(\left(\frac{\log(n)}{n}\right)^{1/4} + \sqrt{\frac{\log(n)}{n\varepsilon^d}} + \frac{\varepsilon^{2s}}{\tau} + \tau \right)$$

with probability at least $1 - C\left(n^{-\alpha} + ne^{-cn\varepsilon^{d+4s}}\right)$.

Proof By the triangle inequality and Theorem 1.2,

$$\begin{aligned}\|u^*_{n,\tau} - g\lfloor_{\Omega_n}\|_{\mathrm{L}^2(\mu_n)} &\le \|u^*_{n,\tau} - u^*_\tau\lfloor_{\Omega_n}\|_{\mathrm{L}^2(\mu_n)} + \|u^*_\tau\lfloor_{\Omega_n} - g\lfloor_{\Omega_n}\|_{\mathrm{L}^2(\mu_n)} \\ &\le C\left(\sqrt{\frac{\log(n)}{n\varepsilon^d}} + \frac{\varepsilon^{2s}}{\tau} + \varepsilon\tau\right) + \|u^*_\tau\lfloor_{\Omega_n} - g\lfloor_{\Omega_n}\|_{\mathrm{L}^2(\mu_n)} \quad (17)\end{aligned}$$

with probability at least $1 - C(n^{-\alpha} + ne^{-cn\varepsilon^{d+4s}})$.

To simplify notation, let $\omega_i = (u^*_\tau(x_i) - g(x_i))^2$ then

$$\|u^*_\tau\lfloor_{\Omega_n} - g\lfloor_{\Omega_n}\|_{\mathrm{L}^2(\mu_n)} = \frac{1}{n}\sum_{i=1}^n (u^*_\tau(x_i) - g(x_i))^2 = \frac{1}{n}\sum_{i=1}^n \omega_i.$$

We note that $\mathbb{E}[\omega_i] = \mathbb{E}[(u^*_\tau(X) - g(X))^2] = \|u^*_\tau - g\|^2_{\mathrm{L}^2(\mu)}$ and

$$0 \le \alpha_i \le \sup_{x\in\Omega}\left(u^*_\tau(x) - g(x)\right)^2 \le 2\left(\|u^*_\tau\|^2_{\mathrm{L}^\infty} + \|g\|^2_{\mathrm{L}^\infty}\right) \le M$$

for all $\tau < \tau_0$ by Lemma 2.14 and Assumption (A5). In particular, M is independent of n and τ. Hoeffding's inequality: for all $\zeta > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \alpha_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\alpha_i] \ge \zeta\right) \le \exp\left(-\frac{2n\zeta^2}{M^2}\right),$$

with Theorem 1.1 implies, with probability at least $1 - e^{-cn\zeta^2}$,

$$\|u^*_\tau - g\lfloor_{\Omega_n}\|^2_{\mathrm{L}^2(\mu_n)} \le \|u^*_\tau - g\|^2_{\mathrm{L}^2(\mu)} + \zeta \le \tau^2\|\Delta^s_\rho g\|^2_{\mathrm{L}^2(\mu)} + \zeta.$$

We choose $\zeta^2 = \frac{\alpha\log(n)}{cn}$ so $\|u^*_\tau - g\lfloor_{\Omega_n}\|^2_{\mathrm{L}^2(\mu_n)} \le \tau^2\|\Delta^s_\rho g\|^2_{\mathrm{L}^2(\mu)} + C\sqrt{\frac{\log(n)}{n}}$ with probability at least $1 - n^{-\alpha}$. Substituting into (17) we have

$$\|u^*_{n,\tau} - g\lfloor_{\Omega_n}\|_{\mathrm{L}^2(\mu_n)} \le C\left(\sqrt{\frac{\log(n)}{n\varepsilon^d}} + \frac{\varepsilon^{2s}}{\tau} + \varepsilon\tau + \tau\|\Delta^s_\rho g\|_{\mathrm{L}^2(\mu)} + \left(\frac{\log(n)}{n}\right)^{1/4}\right)$$

with probability at least $1 - C(n^{-\alpha} + ne^{-cn\varepsilon^{d+4s}})$. □

Remark 6 In the above proof we could have obtained tighter estimates if we had used Bernstein's inequality instead of Hoeffding's inequality, since the variance of the random variables $(u^*_\tau(x_i) - g(x_i))^2$ can be easily proved to be of order τ^2. However, as we will see below, the loose estimates provided by Hoeffding's inequality are of strictly smaller order than the errors that come from Theorem 1.1 and thus it is sufficient for our purposes to use these simpler bounds.

Remark 7 Combining Remarks 3 and 5 and using a similar strategy as the one used to obtain the estimates in Corollary 1.3 we can also show, for even s,

$$\left\| \Delta_n^{\frac{s}{2}} u^*_{n,\tau} - \Delta_\rho^{\frac{s}{2}} g\lfloor_{\Omega_n} \right\|_{\mathrm{L}^2(\mu_n)} \leq C \left(\left(\frac{\log(n)}{n} \right)^{1/4} + \sqrt{\frac{\log(n)}{\tau n \varepsilon^d}} + \frac{\varepsilon^s}{\tau} + \varepsilon + \sqrt{\tau} \right)$$

with probability at least $1 - C\left(n^{-\alpha} + ne^{-cn\varepsilon^{d+4s}}\right)$.

Remark 8 When $s = 1$ the error simplifies to

$$\|u^*_{n,\tau} - g\lfloor_{\Omega_n}\|_{\mathrm{L}^2(\mu_n)} \leq C \left(\left(\frac{\log(n)}{n} \right)^{1/4} + \sqrt{\frac{\log(n)}{n\varepsilon^d}} + \frac{\varepsilon^2}{\tau} + \tau \right)$$

with probability at least $1 - C\left(n^{-\alpha} + ne^{-cn\varepsilon^{d+4}}\right)$. Choosing τ optimally with respect to ε implies $\tau = \varepsilon$ and

$$\|u^*_{n,\tau} - g\lfloor_{\Omega_n}\|_{\mathrm{L}^2(\mu_n)} \leq C \left(\left(\frac{\log(n)}{n} \right)^{1/4} + \sqrt{\frac{\log(n)}{n\varepsilon^d}} + \varepsilon \right).$$

The optimal choice of ε is $\varepsilon_n = \left(\frac{\log(n)}{n}\right)^{\frac{1}{d+2}}$, which (as in Remark 4) is outside the admissible scaling of ε_n, so we choose $\varepsilon_n \sim \left(\frac{\log(n)}{n}\right)^{\frac{1}{d+4}}$. In this regime the optimal error is then

$$\|u^*_{n,\tau} - g\lfloor_{\Omega_n}\|_{\mathrm{L}^2(\mu_n)} \leq C \left(\frac{\log(n)}{n} \right)^{\frac{1}{d+4}},$$

as the term $\left(\frac{\log(n)}{n}\right)^{1/4}$ is always of smaller order. Notice that the above error rate is approximately the minimax rate achieved for the total variation regularised problem which, in certain cases, is up to logarithms scaling as $n^{-\frac{1}{d}}$ [62], comparable to the L^∞ minimax rates and convergence of spline smoothing obtained in [38, 63, 64], which are approximately $n^{-\frac{1}{d+2}}$. This also coincides with the semi-supervised rate of convergence given in [7] when the number of labeled data is linear in n.

Remark 9 For $s \in \mathbb{N}$ we can choose $\tau = \varepsilon^s$ so that the error simplifies to

$$\|u^*_{n,\tau} - g\lfloor_{\Omega_n}\|_{\mathrm{L}^2(\mu_n)} \leq C \left(\left(\frac{\log(n)}{n} \right)^{1/4} + \sqrt{\frac{\log(n)}{n\varepsilon^d}} + \varepsilon^s \right)$$

with probability at least $1 - C(n^{-\alpha} + ne^{-cn\varepsilon^{d+4s}})$. If we choose $\varepsilon_n \sim \left(\frac{M\log(n)}{n}\right)^{\frac{1}{d+4s}}$, then the error scales as

$$\|u^*_{n,\tau} - g\lfloor_{\Omega_n}\|_{\mathrm{L}^2(\mu_n)} \leq C\left(\left(\frac{\log(n)}{n}\right)^{1/4} + \left(\frac{\log(n)}{n}\right)^{\frac{s}{d+4s}}\right) \leq C\left(\frac{\log(n)}{n}\right)^{\frac{s}{d+4s}}$$

(where C depends on M) with probability at least $1 - C\left(n^{-\alpha} + n^{1-cM}\right)$, choosing $M = \frac{1+\alpha}{c}$ we have that the bound holds with probability at least $1 - Cn^{-\alpha}$. This is close to the spline error rate, which, up to logarithms, scales as $n^{-\frac{s}{d+2s}}$; see [38].

Remark 10 The minimax rates for estimating g from noisy samples (6) is $n^{-\frac{s}{2s+d}}$ when $g \in \mathrm{H}^s$ (the rate achieved by splines). In the graph setting the minimax rate can be obtained by projecting the samples onto the first $K = K(\|g\|_{\mathrm{H}^s})$ eigenvectors of the graph Laplacian [46]. Whilst this has the advantage of better rates, one must have an a-priori estimate in the H^s norm of g in order to know K.

1.4 Outline

The rest of the paper is organized as follows. In Sect. 2 we obtain the L^2 variance estimates discussed in Sect. 1.3.1. In Sect. 3 we consider the bias of the estimation procedure given in Sect. 1.3.2.

2 L^2 variance estimates

In this section we prove the variance estimates stated precisely in Theorem 1.1. We split the proof into two main steps. First, we compare the solution $u^*_{n,\tau}$ with a discrete noiseless function $u^{g*}_{n,\tau}$. Then, we compare the function $u^{g*}_{n,\tau}$ with u^*_τ.

2.1 Removing the noise

We start by stating the main result of the section.

Proposition 2.1 *Let Assumptions (A1)–(A4) hold and $s \in \mathbb{N}$. Let $\mathcal{E}^{(\mathbf{y}_n)}_{n,\tau}$ be defined by (9), where $R^{(s)}_n$, Δ_n are defined by (10), (8) respectively, and let $u^*_{n,\tau}$, $u^{g*}_{n,\tau}$ be the minimizers of $\mathcal{E}^{(\mathbf{y}_n)}_{n,\tau}$, $\mathcal{E}^{(\mathbf{g}_n)}_{n,\tau}$ respectively. Assume that ξ_i are iid, mean zero, sub-Gaussian random variables. Then, for all $\alpha > 1$, there exists $\varepsilon_0 > 0$ and $C > 0$ such that for any ε, n satisfying (16) and $\tau > 0$ we have*

$$\left\|u^*_{n,\tau} - u^{g*}_{n,\tau}\right\|_{L^2(\mu_n)} \leq C\left(\sqrt{\frac{\log(n)}{n\varepsilon^d}} + \frac{\varepsilon^{2s}}{\tau}\right)$$

$$\left\|\Delta_n^{\frac{s}{2}} u^*_{n,\tau} - \Delta_n^{\frac{s}{2}} u^{g*}_{n,\tau}\right\|_{L^2(\mu_n)} \leq C\left(\sqrt{\frac{\log(n)}{n\varepsilon^d\tau}} + \frac{\varepsilon^s}{\tau}\right)$$

with probability at least $1 - Cn^{-\alpha}$.

The proof of the proposition will be given at the end of the section. The strategy is to compare the Euler–Lagrange equations associated with minimising $\mathcal{E}_{n,\tau}^{(\mathbf{y}_n)}$ and $\mathcal{E}_{n,\tau}^{(\mathbf{g}_n)}$. In particular, we have

$$\frac{1}{2}\nabla_{\mathrm{L}^2(\mu_n)}\mathcal{E}_{n,\tau}^{(\mathbf{y}_n)}(u_n) = \tau\Delta_n^s u_n + u_n - \mathbf{y}_n \tag{18}$$

and therefore

$$\begin{aligned}\tau\Delta_n^s u_{n,\tau}^* + u_{n,\tau}^* - \mathbf{y}_n &= 0\\ \tau\Delta_n^s u_{n,\tau}^{g*} + u_{n,\tau}^{g*} - \mathbf{g}_n &= 0.\end{aligned} \tag{19}$$

We let $w_{n,\tau}^* = u_{n,\tau}^* - u_{n,\tau}^{g*}$ then it follows that

$$\tau\Delta_n^s w_{n,\tau}^* + w_{n,\tau}^* - (\mathbf{y}_n - \mathbf{g}_n) = 0$$

and $w_{n,\tau}^*$ minimizes $\mathcal{E}_n^{(\mathbf{y}_n - \mathbf{g}_n)} = \mathcal{E}_n^{(\boldsymbol{\xi}_n)}$ where $\boldsymbol{\xi}_n = (\xi_1, \ldots, \xi_n)$. We can write

$$w_{n,\tau}^* = \left(\tau\Delta_n^s + \mathrm{Id}\right)^{-1}\boldsymbol{\xi}_n. \tag{20}$$

To obtain an estimate on $\|w_{n,\tau}^*\|_{\mathrm{L}^2(\mu_n)}$ we use an ansatz $\tilde{w}_n$ and show $\|w_{n,\tau}^* - \tilde{w}_n\|_{\mathrm{L}^2(\mu_n)} \leq C\sqrt{\frac{\log(n)}{n\varepsilon^d}}$ and $\|\tilde{w}_n\|_{\mathrm{L}^2(\mu_n)} \leq \frac{C\varepsilon^{2s}}{\tau}$ with high probability. Our choice of ansatz is to assume that the diagonal part of Δ_n dominates and therefore $\Delta_n \approx \frac{2}{n\varepsilon^2}D_n$ which leads to the choice,

$$\tilde{w}_n = \left(\tau\left(\frac{2}{n\varepsilon^2}D_n\right)^s + \mathrm{Id}\right)^{-1}\boldsymbol{\xi}_n. \tag{21}$$

This choice of ansatz is appropriate because Δ_n is increasingly sparse (for instance, if $\eta(t)$ is the indicator function on $[0, 1]$ then we have positive edge weights only when $|x_i - x_j| \leq \varepsilon$, and therefore each node has on the order of $n\varepsilon^d$ edges, hence the fraction of non-zero entries in W is $\frac{n\varepsilon^d}{n} = \varepsilon^d$) and therefore well approximated by a diagonal matrix, which with high probability will not amplify the vector ξ. We can equivalently write

$$\tilde{w}_n(x_i) = \frac{\xi_i}{\tau\left(\frac{2}{n\varepsilon^2}\sum_{k=1}^n W_{i,k}\right)^s + 1}. \tag{22}$$

The following lemmas will be useful.

Lemma 2.2 *Under Assumptions (A1)–(A4) there exists $C, C_1, C_2, c > 0$ such that, if $n\varepsilon^d \geq 1$ then*

$$C_1 \leq \frac{1}{n}\sum_{j=1}^{n} W_{i,j} \leq C_2 \quad \text{and} \quad \#\left\{j : W_{i,j} > 0\right\} \leq Cn\varepsilon^d$$

for all $i = 1, \ldots, n$ with probability at least $1 - 2ne^{-cn\varepsilon^d}$.

Proof Fix $i \in \mathbb{N}$, then $W_{i,j}$ are iid for $j \neq i$. If $M = \|\eta_\varepsilon\|_{\mathrm{L}^\infty(\mathbb{R})}$ and $\sigma^2 = \mathbb{E}\left(W_{i,j} - \mathbb{E}[W_{i,j}]\right)^2$ (where the expectation $\mathbb{E}[W_{i,j}]$ is taken over x_j) then it is straightforward to show the bounds $\sigma^2 \leq C\varepsilon^{-d}$ and $M \leq C\varepsilon^{-d}$. By Bernstein's inequality, for all $t > 0$,

$$\mathbb{P}\left(\left|\sum_{j\neq i}\left(W_{i,j} - \mathbb{E}[W_{i,j}]\right)\right| > t\right) \leq 2\exp\left(-\frac{t^2}{2n\sigma^2 + \frac{4Mt}{3}}\right) \leq 2\exp\left(-\frac{ct^2\varepsilon^d}{n+t}\right).$$

Choosing $t = \lambda n$ and restricting to $\lambda \leq 1$ we have

$$\mathbb{P}\left(\left|\sum_{j\neq i}\left(W_{i,j} - \mathbb{E}[W_{i,j}]\right)\right| > \lambda n\right) \leq 2\exp\left(-cn\varepsilon^d\lambda^2\right).$$

Hence, (recalling $W_{i,i} = 0$)

$$(n-1)\mathbb{E}[W_{i,j}] - \lambda n \leq \sum_{j=1}^{n} W_{i,j} \leq (n-1)\mathbb{E}[W_{i,j}] + \lambda n$$

with probability at least $1 - 2e^{-cn\varepsilon^d\lambda^2}$. One can show that there exists C_1, C_2 such that $C_1 \leq \mathbb{E}[W_{i,j}] \leq C_2$. For $n \geq 2$ (so that $n - 1 \geq n/2$),

$$\frac{C_1}{2} - \lambda \leq \frac{1}{n}\sum_{j=1}^{n} W_{i,j} \leq C_2 + \lambda$$

with probability at least $1 - 2e^{-cn\varepsilon^d\lambda^2}$. Choosing $\lambda = \frac{C_1}{4}$ and union bounding over $i \in \{1, \ldots, n\}$ we can conclude the first result.

The second result follows from the first by choosing $\tilde{\varepsilon} = 2\varepsilon$ and letting $\tilde{W}_{i,j} = \eta_{\tilde{\varepsilon}}(|x_i - x_j|)$. Then, $W_{i,j} > 0$ implies $\tilde{\varepsilon}^d\tilde{W}_{ij} \geq 0.5$ and therefore $\#\{j : W_{i,j} > 0\} \leq 2\tilde{\varepsilon}^d\sum_{j=1}^{n}\tilde{W}_{i,j}$. Applying the first part of the lemma we have $2\tilde{\varepsilon}^d\sum_{j=1}^{n}\tilde{W}_{i,j} \leq 2C_2n\tilde{\varepsilon}^d = 2^{d+1}C_2n\varepsilon^d$ as required. □

Lemma 2.3 *Under Assumptions (A1)–(A4) define Δ_n and D_n by (8). Then, for all $\alpha > 1$, there exists $\varepsilon_0 > 0$ and $C > 0$ such that for all ε, n satisfying (16) we have*

$$\|\Delta_n\|_{\mathrm{op}} \leq \frac{C}{\varepsilon^2}, \qquad \|D_n\|_{\mathrm{op}} \leq Cn, \qquad \text{and} \qquad \|W_n\|_{\mathrm{op}} \leq Cn$$

with probability at least $1 - Cn^{-\alpha}$.

Proof For $d \geq 3$ one can bound $\|\Delta_n\|_{\mathrm{op}} \leq C\varepsilon^{-2}$ by [18, Lemma 22]. Indeed, $\|\Delta_n\|_{\mathrm{op}} \leq C\varepsilon^{-2}$ whenever $\mathrm{d}_{\mathrm{W}^\infty}(\mu_n, \mu) < \varepsilon$, where $\mathrm{d}_{\mathrm{W}^\infty}$ is the ∞-Wasserstein distance. By [77, Theorem 1.1] this holds with probability at least $1 - Cn^{-\alpha}$. The same argument is used in (23) below as one step in the proof for $d = 2$.

For $d = 2$ a small modification is required to remove the additional logarithmic factors that are present in the scaling of $\mathrm{d}_{\mathrm{W}^\infty}(\mu_n, \mu)$, i.e. one has $\mathrm{d}_{\mathrm{W}^\infty}(\mu_n, \mu) \sim \frac{(\log(n))^{\frac{3}{4}}}{\sqrt{n}}$ and therefore requires $\varepsilon \geq C\frac{(\log(n))^{\frac{3}{4}}}{\sqrt{n}}$. However, this can be avoided by comparing μ_n to a smooth approximation of μ.

In [78, Lemma 3.1] the authors show, in the Euclidean setting, that if $\varepsilon = \varepsilon_n \to 0$ satisfies $\frac{\log n}{n\varepsilon_n^d} \to 0$ then there exists an absolutely continuous probability measure $\tilde{\mu}_n \in \mathcal{P}(\Omega)$ such that

$$\frac{\mathrm{d}_{\mathrm{W}^\infty}(\mu_n, \tilde{\mu}_n)}{\varepsilon_n} \to 0 \qquad \text{and} \qquad \|\rho - \tilde{\rho}_n\|_{\mathrm{L}^\infty(\mu)} \to 0$$

where $\tilde{\rho}_n$ is the density of $\tilde{\mu}_n$. As in [14, Proposition 2.10] the proof can be modified to give a non-asymptotic quantitative high probability bound. In particular, there exists constants C, ε_0 and θ_0 such that if $n^{-\frac{1}{d}} \leq \varepsilon \leq \varepsilon_0$ and $\theta \leq \theta_0$ then

$$\mathrm{d}_{\mathrm{W}^\infty}(\mu_n, \tilde{\mu}_n) \leq \varepsilon \qquad \text{and} \qquad \|\rho - \tilde{\rho}_n\|_{\mathrm{L}^\infty(\mu)} \leq C(\theta + \varepsilon)$$

with probability at least $1 - 2ne^{-cn\theta^2\varepsilon^d}$. For the rest of the proof we fix $\theta = \theta_0$ and absorb it into the constants.

Note that if we define $\bar{\eta} = \eta((|\cdot| - 1)_+)$ and $T_n \colon \Omega \to \Omega$ satisfies $\|T_n - \mathrm{Id}\|_{\mathrm{L}^\infty(\Omega)} \leq \varepsilon$ then

$$\begin{aligned}
\eta\left(\frac{|x - T_n(y)|}{\varepsilon}\right) &\leq \eta\left(\frac{(|x - y| - |T_n(y) - y|)_+}{\varepsilon}\right) \\
&\leq \eta\left(\left(\left|\frac{x - y}{\varepsilon}\right| - 1\right)_+\right) \\
&= \bar{\eta}\left(\frac{|x - y|}{\varepsilon}\right).
\end{aligned}$$

We choose T_n to be a transport map satisfying $T_{n\#}\tilde{\mu}_n = \mu_n$ and $\|T_n - \mathrm{Id}\|_{\mathrm{L}^\infty(\tilde{\mu}_n)} = \mathrm{d}_{\mathrm{W}^\infty}(\mu_n, \tilde{\mu}_n)$ (since $\tilde{\mu}_n$ has a Lebesgue density then we may apply standard optimal transport results, in particular Brenier's theorem, to infer existence of such a map).

Let $\lambda_{\max}$ be the largest eigenvalue of Δ_n then, as in the proof of [18, Lemma 22],

$$\begin{aligned}
\lambda_{\max} &= \sup_{\|u\|_{\mathrm{L}^2(\mu_n)}=1} \langle u, \Delta_n u\rangle_{\mathrm{L}^2(\mu_n)} \\
&\leq \frac{4}{n^2\varepsilon^{d+2}} \sup_{\|u\|_{\mathrm{L}^2(\mu_n)}=1} \sum_{i,j=1}^{n} \eta\left(\frac{|x_i - x_j|}{\varepsilon}\right) u(x_i)^2 \\
&= \frac{4}{n\varepsilon^{d+2}} \sup_{\|u\|_{\mathrm{L}^2(\mu_n)}=1} \sum_{i=1}^{n} u(x_i)^2 \int_\Omega \eta\left(\frac{|x_i - T_n(y)|}{\varepsilon}\right) \tilde{\rho}_n(y)\,\mathrm{d}y \\
&\leq \frac{4}{n\varepsilon^{d+2}} \sup_{\|u\|_{\mathrm{L}^2(\mu_n)}=1} \sum_{i=1}^{n} u(x_i)^2 \int_\Omega \bar{\eta}\left(\frac{|x_i - y|}{\varepsilon}\right) \tilde{\rho}_n(y)\,\mathrm{d}y \\
&\leq \frac{4}{n\varepsilon^{d+2}} \sup_{\|u\|_{\mathrm{L}^2(\mu_n)}=1} \sum_{i=1}^{n} u(x_i)^2 \int_\Omega \bar{\eta}\left(\frac{|x_i - y|}{\varepsilon}\right) (\rho(y) + C)\,\mathrm{d}y \\
&\leq \frac{C}{\varepsilon^2}
\end{aligned} \tag{23}$$

since $\int_{\mathbb{R}^d} \bar{\eta}(|z|)\,\mathrm{d}z < +\infty$.

Although the bound holds for probability at least $1 - Cne^{-cn\varepsilon^d}$ when $d = 2$ we can assume that the C in (16) is sufficiently large so that $\frac{n\varepsilon^d}{\log(n)} \geq \frac{\alpha+1}{c}$. After some elementary algebra one has that $1 - Cne^{-cn\varepsilon^d} \geq 1 - Cn^{-\alpha}$.

For any $v \in \mathrm{L}^2(\mu_n)$ we have, by Lemma 2.2 with probability at least $1 - 2ne^{-cn\varepsilon^d}$,

$$\|D_n v\|^2_{\mathrm{L}^2(\mu_n)} = \frac{1}{n}\sum_{i=1}^{n}\left(v(x_i)\sum_{j=1}^{n} W_{i,j}\right)^2 \leq C_2^2 n \sum_{i=1}^{n} v(x_i)^2 = C_2^2 n^2 \|v\|^2_{\mathrm{L}^2(\mu_n)}$$

which implies $\|D_n\|_{\mathrm{op}} \leq C_2 n$. The operator norm bound on W_n follows from the bounds on the operator norms of Δ_n, D_n and the triangle inequality. Choosing C in Equation (16) sufficiently large we can assume that $1 - 2ne^{-cn\varepsilon^d} \geq 1 - Cn^{-\alpha}$. □

In fact, [18, Lemma 22], suggests the operator bound on Δ_n is sharp (up to a constant), that is, there exists $c > 0$ such that $\|\Delta_n\|_{\mathrm{op}} \geq \frac{c}{\varepsilon^2}$.

Lemma 2.4 *Let Assumptions (A1)–(A4) hold and $s \geq 1$, $k \in \mathbb{N}$. Let ξ_i be iid, mean zero, sub-Gaussian random variables. Define $\tilde{w}_n$ by (22) and D_n by (8). Then, for any $\alpha > 1$ there exists $\varepsilon_0 > 0$ and $C > 0$ such that for all ε, n satisfying (16) and $\tau > 0$ we have*

$$\|W_n (D_n)^{k-1} \tilde{w}_n\|_{L^2(\mu_n)} \leq \frac{Cn^k\varepsilon^{2s}}{\tau}\sqrt{\frac{\log(n)}{n\varepsilon^d}}$$

with probability at least $1 - Cn^{-\alpha}$.

Proof Let us condition on a graph G_n that satisfies the two inequalities:

(i) $C_1 \leq \frac{1}{n}\sum_{j=1}^n W_{i,j} \leq C_2$, for all $i = 1, 2, \ldots, n$,
(ii) $\#\left\{j : W_{i,j} > 0\right\} \leq Cn\varepsilon^d$ for all $i = 1, 2, \ldots, n$.

Fix $i \in \{1, 2, \ldots, n\}$ and define

$$q_j = \frac{\tau W_{i,j}}{\varepsilon^{2s} n^{k-1}} \left[D_n^{k-1}\tilde{w}_n\right]_j.$$

Conditioned on G_n we have that q_j are zero mean and independent random variables. Moreover, since

$$q_j = \frac{\tau W_{i,j}\left(\sum_{\ell=1}^n W_{j,\ell}\right)^{k-1}\xi_j}{\varepsilon^{2s} n^{k-1}\left(\tau\left(\frac{2}{n\varepsilon^2}\sum_{\ell=1}^n W_{j,\ell}\right)^s + 1\right)}$$

then we have $|q_j| \leq \frac{C|\xi_j|}{\varepsilon^d}$ so q_j is sub-Gaussian and $\|q_j\|_{\psi_2} \leq \frac{C\|\xi_j\|_{\psi_2}}{\varepsilon^d} \lesssim \frac{1}{\varepsilon^d}$ where $\|\cdot\|_{\psi_2}$ is the Birnbaum–Orlicz norm defined by

$$\|Q\|_{\psi_2} := \inf\left\{c \geq 0 : \mathbb{E}e^{\frac{Q^2}{c^2}} \leq 2\right\} \quad \text{for a random variable } Q.$$

By Hoeffding's inequality, for any $t > 0$

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{j=1}^n q_j\right| > t|G_n\right) &\leq \mathbb{P}\left(\left|\sum_{j\,:\,W_{i,j}>0} q_j\right| > t|G_n\right)\\ &\leq 2\exp\left(-\frac{ct^2}{\sum_{j\,:\,W_{i,j}>0}\|q_j\|_{\psi_2}^2}\right)\\ &\leq 2e^{-\frac{ct^2\varepsilon^d}{n}}.\end{aligned}$$

We choose $t = \lambda\sqrt{\frac{n\log(n)}{\varepsilon^d}}$ so

$$\frac{\tau}{\varepsilon^{2s}n^{k-1}}\left|\left[W_n D_n^{k-1}\tilde{w}_n\right]_i\right| = \left|\sum_{j=1}^n q_j\right| \leq \lambda\sqrt{\frac{n\log(n)}{\varepsilon^d}}$$

with probability at least $1 - 2n^{-c\lambda^2}$, conditioned on G_n. Union bounding and selecting $\lambda = \sqrt{\frac{\alpha+1}{c}}$, we then get that the above bound holds for all $i \in \{1, 2, \ldots, n\}$ with probability at least $n^{1-c\lambda^2} = n^{-\alpha}$. Hence, after absorbing α into the constant C,

$$\left\|W_n D_n^{k-1}\tilde{w}_n\right\|_{\mathrm{L}^2(\mu_n)} \leq \left\|W_n D_n^{k-1}\tilde{w}_n\right\|_{\mathrm{L}^\infty(\mu_n)} \leq \frac{Cn^k\varepsilon^{2s}}{\tau}\sqrt{\frac{\log(n)}{n\varepsilon^d}}$$

conditioned on G_n with probability at least $1 - 2n^{-\alpha}$. Since, by Lemma 2.2, the probability of G_n satisfying conditions (i) and (ii) is at least $1 - 2e^{-cn\varepsilon^d}$, and by choosing C sufficiently large we have that $2e^{-cn\varepsilon^d} \leq Cn^{-\alpha}$ we can conclude the lemma. □

To control $\|\tilde{w}_n - w^*_{n,\tau}\|_{\mathrm{L}^2(\mu_n)}$ we take advantage of the convexity of our objective functional $\mathcal{E}^{(\boldsymbol{\xi}_n)}_{n,\tau}$, where we recall that $\boldsymbol{\xi}_n = \mathbf{y}_n - \mathbf{g}_n$. In particular, one can easily show that $\mathcal{E}^{(\boldsymbol{\xi}_n)}_{n,\tau}$ satisfies

$$\begin{aligned}&\left\langle \nabla_{\mathrm{L}^2(\mu_n)}\mathcal{E}^{(\boldsymbol{\xi}_n)}_{n,\tau}(v_n) - \nabla_{\mathrm{L}^2(\mu_n)}\mathcal{E}^{(\boldsymbol{\xi}_n)}_{n,\tau}(u_n), v_n - u_n\right\rangle_{\mathrm{L}^2(\mu_n)}\\ &\quad = 2\tau \left\| \Delta_n^{\frac{s}{2}}(v_n - u_n)\right\|^2_{\mathrm{L}^2(\mu_n)} + 2\|u_n - v_n\|^2_{\mathrm{L}^2(\mu_n)}\end{aligned}$$

for any $u_n, v_n \in \mathrm{L}^2(\mu_n)$. Hence,

$$\begin{aligned}\|u_n - v_n\|_{\mathrm{L}^2(\mu_n)} &\leq \frac{1}{2}\left\|\nabla_{\mathrm{L}^2(\mu_n)}\mathcal{E}^{(\boldsymbol{\xi}_n)}_n(v_n) - \nabla_{\mathrm{L}^2(\mu_n)}\mathcal{E}^{(\boldsymbol{\xi}_n)}_n(u_n)\right\|_{\mathrm{L}^2(\mu_n)}\\ \left\|\Delta_n^{\frac{s}{2}}(u_n - v_n)\right\|_{\mathrm{L}^2(\mu_n)} &\leq \frac{1}{2\sqrt{\tau}}\left\|\nabla_{\mathrm{L}^2(\mu_n)}\mathcal{E}^{(\boldsymbol{\xi}_n)}_n(v_n) - \nabla_{\mathrm{L}^2(\mu_n)}\mathcal{E}^{(\boldsymbol{\xi}_n)}_n(u_n)\right\|_{\mathrm{L}^2(\mu_n)}.\end{aligned}$$

Applying this bound to $u_n = w^*_{n,\tau}$ and $v_n = \tilde{w}_n$, and using the optimality of $w^*_{n,\tau}$, we have

$$\|w^*_{n,\tau} - \tilde{w}_n\|_{\mathrm{L}^2(\mu_n)} \leq \frac{1}{2}\left\|\nabla_{\mathrm{L}^2(\mu_n)}\mathcal{E}^{(\boldsymbol{\xi}_n)}_n(\tilde{w}_n)\right\|_{\mathrm{L}^2(\mu_n)} \tag{24}$$

$$\left\|\Delta_n^{\frac{s}{2}}(w^*_{n,\tau} - \tilde{w}_n)\right\|_{\mathrm{L}^2(\mu_n)} \leq \frac{1}{2\sqrt{\tau}}\left\|\nabla_{\mathrm{L}^2(\mu_n)}\mathcal{E}^{(\boldsymbol{\xi}_n)}_n(\tilde{w}_n)\right\|_{\mathrm{L}^2(\mu_n)}. \tag{25}$$

The next lemma will bound these gradients in order to prove L^2 convergence rates.

Lemma 2.5 *Let Assumptions (A1)–(A4) hold and $s \in \mathbb{N}$. Let ξ_i be iid, mean zero, sub-Gaussian, random variables. Define $\tilde{w}_n$ by (22) and $w^*_{n,\tau}$ by (20) where Δ_n is given by (8). Then, for any $\alpha > 1$, there exists $C > 0$ such that if ε, n satisfy (16) and $\tau > 0$ we have*

$$\begin{aligned}\|\tilde{w}_n - w^*_{n,\tau}\|_{L^2(\mu_n)} &\leq C\sqrt{\frac{\log(n)}{n\varepsilon^d}}\\ \left\|\Delta_n^{\frac{s}{2}}\tilde{w}_n - \Delta_n^{\frac{s}{2}}w^*_{n,\tau}\right\|_{L^2(\mu_n)} &\leq C\sqrt{\frac{\log(n)}{n\varepsilon^d\tau}}\end{aligned}$$

with probability at least $1 - Cn^{-\alpha}$.

Proof By the definition of $\tilde{w}_n$, namely Equation (21) and the Fréchet derivative of $\mathcal{E}_{n,\tau}^{(\boldsymbol{\xi}_n)}$, see (18) we have

$$\begin{aligned}\frac{1}{2}\nabla_{\mathrm{L}^2(\mu_n)}\mathcal{E}_{n,\tau}^{(\boldsymbol{\xi}_n)}(\tilde{w}_n) &= \left(\tau\Delta_n^s + \mathrm{Id}\right)\tilde{w}_n - \boldsymbol{\xi}_n \\ &= \frac{2^s\tau}{n^s\varepsilon^{2s}}\left((D_n - W_n)^s - D_n^s\right)\tilde{w}_n \\ &= \frac{2^s\tau}{n^s\varepsilon^{2s}}\left(\left(\sum_{\chi\in\{0,1\}^s}\prod_{i=1}^s D_n^{\chi_i}(-W_n)^{1-\chi_i}\right) - D_n^s\right)\tilde{w}_n.\end{aligned}$$

Using the bounds from Lemmas 2.3 and 2.4 and their associated probability estimates, along with the fact that we have cancelled the D_n^s term, we then may bound

$$\frac{1}{2}\left\|\nabla_{\mathrm{L}^2(\mu_n)}\mathcal{E}_{n,\tau}^{(\boldsymbol{\xi}_n)}(\tilde{w}_n)\right\|_{\mathrm{L}^2(\mu_n)} \leq \frac{2^s\tau}{n^s\varepsilon^{2s}}\sum_{\chi\in\{0,1\}^s}\frac{Cn^s\varepsilon^{2s}}{\tau}\sqrt{\frac{\log(n)}{n\varepsilon^d}} \leq C\sqrt{\frac{\log(n)}{n\varepsilon^d}}.$$

This concludes the proof. □

Our final lemma before proving Proposition 2.1 is to bound $\|\tilde{w}_n\|^2_{\mathrm{L}^2(\mu_n)}$.

Lemma 2.6 *Let Assumptions (A1)–(A4) hold and $s \in \mathbb{N}$. Let ξ_i be iid, mean zero, sub-Gaussian, random variables. Define $\tilde{w}_n$ by (22). Then, for any $\alpha > 1$, there exists $C > 0$ such that for all ε, n satisfying $n\varepsilon^d \geq 1$ and $\tau > 0$ we have*

$$\|\tilde{w}_n\|_{L^2(\mu_n)} \leq \frac{C\varepsilon^{2s}}{\tau}$$

with probability at least $1 - n^{-\alpha}$.

Proof By application of Lemma 2.2 we have

$$\begin{aligned}\|\tilde{w}_n\|^2_{\mathrm{L}^2(\mu_n)} &= \frac{1}{n}\sum_{i=1}^n \frac{\xi_i^2}{\left(\tau\left(\frac{2}{n\varepsilon^2}\sum_{k=1}^n W_{i,k}\right)^s + 1\right)^2} \\ &\leq \frac{\varepsilon^{4s}}{4^s C_1^{2s}\tau^2 n}\sum_{i=1}^n \xi_i^2.\end{aligned}$$

Applying a Chernoff bound we have, for all $s, t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^n \xi_i^2 \geq t\right) \leq \frac{\mathbb{E}\left[e^{s\sum_{i=1}^n \xi_i^2}\right]}{e^{st}} = \frac{\prod_{i=1}^n \mathbb{E}\left[e^{s\xi_i^2}\right]}{e^{st}}.$$

Choosing $s = \|\xi_i\|_{\Psi_2}^{-2}$ and $t = An$ we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i^2 \geq A\right) \leq 2^n e^{-An\|\xi_i\|_{\Psi_2}^{-2}}.$$

Now we choose A sufficiently large so that $\frac{A}{\|\xi_i\|_{\Psi_2}^2} \geq \alpha + \log 2$ and hence

$$2^n e^{-An\|\xi_i\|_{\Psi_2}^{-2}} \leq e^{-n\alpha} \leq n^{-\alpha}.$$

In particular, $\|\tilde{w}_n\|_{\mathrm{L}^2(\mu_n)} \leq \frac{C\varepsilon^{2s}}{\tau}$ with probability at least $1 - n^{-\alpha}$ as required. □

The proof of Proposition 2.1 now follows from Lemmas 2.3, 2.5 and 2.6 since

$$\|u_{n,\tau}^* - u_{n,\tau}^{g*}\|_{\mathrm{L}^2(\mu_n)} \leq \|w_{n,\tau}^* - \tilde{w}_n\|_{\mathrm{L}^2(\mu_n)} + \|\tilde{w}_n\|_{\mathrm{L}^2(\mu_n)} \leq C\left(\sqrt{\frac{\log(n)}{n\varepsilon^d}} + \frac{\varepsilon^{2s}}{\tau}\right)$$

and

$$\begin{aligned}\left\|\Delta_n^{\frac{s}{2}} u_{n,\tau}^* - \Delta_n^{\frac{s}{2}} u_{n,\tau}^{g*}\right\|_{\mathrm{L}^2(\mu_n)} &\leq \left\|\Delta_n^{\frac{s}{2}} w_{n,\tau}^* - \Delta_n^{\frac{s}{2}}\tilde{w}_n\right\|_{\mathrm{L}^2(\mu_n)} + \left\|\Delta_n^{\frac{s}{2}}\tilde{w}_n\right\|_{\mathrm{L}^2(\mu_n)} \\ &\leq \left\|\Delta_n^{\frac{s}{2}} w_{n,\tau}^* - \Delta_n^{\frac{s}{2}}\tilde{w}_n\right\|_{\mathrm{L}^2(\mu_n)} + \|\Delta_n\|_{\mathrm{op}}^{\frac{s}{2}}\|\tilde{w}_n\|_{\mathrm{L}^2(\mu_n)} \\ &\leq C\left(\sqrt{\frac{\log(n)}{n\varepsilon^d\tau}} + \frac{\varepsilon^s}{\tau}\right)\end{aligned}$$

with probability at least $1 - Cn^{-\alpha}$.

2.2 Discrete-to-continuum in the noiseless case

In this subsection we prove the following estimates which relate the functions $u_{n,\tau}^{g*}$ (the minimizer of $\mathcal{E}_{n,\tau}^{(\mathbf{g}_n)}$ defined in (9) with $\mathbf{a}_n = (g(x_1), \ldots, g(x_n))$) with the function u_τ^* (the minimizer of $\mathcal{E}_{\infty,\tau}^{(g)}$ defined in (14)).

As in (19) we can write the Euler–Lagrange equations associated with minimizing $\mathcal{E}_{\infty,\tau}^{(g)}$ by

$$\tau\Delta_\rho^s u_\tau^* + u_\tau^* - g = 0. \tag{26}$$

Our main result for this section is the following proposition.

Proposition 2.7 *Let Assumptions (A1)–(A5) hold and $s \in \mathbb{N}$. Define Δ_n, Δ_ρ and σ_η by (8), (12) and (13) respectively. Let $u_{n,\tau}^{g*}$ and u_τ^* satisfy (19) and (26) respectively. Then, for any $\alpha > 1$ and $\tau_0 > 0$ there exists constants $\varepsilon_0 > 0$, $C > c > 0$ such that, for any ε, n satisfying (16) and $\tau \in (0, \tau_0)$ we have*

$$\left\|u_{n,\tau}^{g*} - u_\tau^*\lfloor_{\Omega_n}\right\|_{\mathrm{L}^2(\mu_n)} \leq C\tau\varepsilon \qquad \left\|\Delta_n^{\frac{s}{2}} u_{n,\tau}^{g*} - \Delta_\rho^{\frac{s}{2}} u_\tau^*\lfloor_{\Omega_n}\right\|_{\mathrm{L}^2(\mu_n)} \leq C\varepsilon$$

with probability at least $1 - Cn^{-\alpha} - Cne^{-cn\varepsilon^{d+4s}}$.

The proof of the proposition is given in Sect. 2.2.3. The proof of Theorem 1.1 follows immediately from the triangle inequality and Propositions 2.1 and 2.7. One of the main ingredients for proving Proposition 2.7 is the following result which is of interest on its own.

Theorem 2.8 *Let Assumptions Assumptions (A1)–(A4) hold and* $s \in \mathbb{N}$. *Define* Δ_n, Δ_ρ *and* σ_η *by (8), (12) and (13) respectively. Then, for any* $\alpha > 1$, *there exists* $C > c > 0$ *and* $\varepsilon_0 > 0$ *such that for any* $u \in \mathrm{C}^{2s+1}$ *and* ε, n *satisfying (16),*

$$\left\| \left(\Delta_n^s - \Delta_\rho^s \right) u \right\|_{\mathrm{L}^2(\mu_n)} \leq C\varepsilon \left(\|u\|_{\mathrm{C}^{2s+1}(\Omega)} + 1 \right)$$

with probability at least $1 - Cn^{-\alpha} - Cne^{-cn\varepsilon^{d+4s}}$.

We notice that when $s = 1$ it is well known that the graph Laplacian is pointwise consistent and the rate at which it converges, e.g. [12]. Theorem 2.8 generalizes this result, and states that with high probability $\Delta_n^s u \to \Delta_\rho^s u$ in an L^2 sense for all $s \in \mathbb{N}$ where u is sufficiently smooth and $\varepsilon = \varepsilon_n$ satisfies a lower bound. The proof of Theorem 2.8 is given in Sect. 2.2.2.

Before presenting a rigorous proof of Proposition 2.7, let us present a heuristic argument. First, we write

$$\begin{aligned}
\Delta_n^s u(x) - \Delta_\rho^s u(x) &= \Delta_n^{s-1} \left(\Delta_n - \Delta_\rho \right) v^{(0)}(x) + \left(\Delta_n^{s-1} - \Delta_\rho^{s-1} \right) v^{(1)}(x) \\
&= \Delta_n^{s-1} \left(\Delta_n - \Delta_\rho \right) v^{(0)}(x) + \Delta_n^{s-2} \left(\Delta_n - \Delta_\rho \right) v^{(1)}(x) \\
&\quad + \left(\Delta_n^{s-2} - \Delta_\rho^{s-2} \right) v^{(2)}(x) \\
&= \cdots \\
&= \sum_{k=1}^{s} \Delta_n^{s-k} \left(\Delta_n - \Delta_\rho \right) v^{(k-1)}(x)
\end{aligned}$$

where $v^{(k)} = \Delta_\rho^k u$. We keep track of higher order errors in the pointwise consistency of the graph Laplacian, following the method in [15] to estimate, when $v \in \mathrm{C}^r$,

$$\left(\Delta_n - \Delta_\rho \right) v(x) = \varepsilon E_1(x) + \varepsilon^2 E_2(x) + \cdots \varepsilon^{r-3} E_{r-3}(x) + \varepsilon^{r-2} E_{r-2}(x) \quad (27)$$

where $E_i \in \mathrm{C}^{r-i-2}$. Now, heuristically one expects (with high probability) $\|\Delta_n^j E_i\|_{\mathrm{L}^2(\mu_n)} \lesssim \|E_i\|_{\mathrm{C}^{2j}(\Omega)}$ (when $j \leq \frac{r-i-2}{2}$) and we recall a worse case (high probability) operator norm bound $\|\Delta_n^j\|_{\mathrm{op}} \leq C\varepsilon^{-2j}$, see Lemma 2.3. Letting $u = u_\tau^*$, and assuming $g \in \mathrm{C}^1(\Omega)$, we can immediately infer that $u \in \mathrm{C}^{2s+1}(\Omega)$ from (26) (as a standard elliptic regularity result). We choose $v = v^{(k-1)}$ in (27) and note that $r = 2(s-k) + 3$. Now, (with high probability)

$$\begin{aligned}\left\|\Delta_n^{s-k} E_i\right\|_{\mathrm{L}^2(\Omega)} &= \left\|\Delta_n^{\frac{i-1}{2}} \Delta_n^{s-k-\frac{i-1}{2}} E_i\right\|_{\mathrm{L}^2(\Omega)} \\ &\leq \|\Delta_n\|_{\mathrm{op}}^{\frac{i-1}{2}} \left\|\Delta_n^{s-k-\frac{i-1}{2}} E_i\right\|_{\mathrm{L}^2(\Omega)} \\ &\leq C\varepsilon^{1-i} \|E_i\|_{\mathrm{C}^{2(s-k)-i+1}(\Omega)}.\end{aligned}$$

So, (with high probability)

$$\begin{aligned}\left\|\Delta_n^{s-k}\left(\Delta_n - \Delta_\rho\right) v^{(k-1)}\right\|_{\mathrm{L}^2(\mu_n)} &\leq \sum_{i=1}^{2(s-k)+1} \varepsilon^i \left\|\Delta_n^{s-k} E_i\right\|_{\mathrm{L}^2(\mu_n)} \\ &\leq C\varepsilon \sum_{i=1}^{2(s-k)+1} \|E_i\|_{\mathrm{C}^{2(s-k)-i+1}(\Omega)} \\ &\leq C\varepsilon.\end{aligned}$$

Thus, $\left\|\Delta_n^s u - \Delta_\rho^s u\right\|_{\mathrm{L}^2(\mu_n)} = O(\varepsilon)$ (note that C in the above inequality depends on u, in the proof we will show that this dependence is in terms of $\|u\|_{\mathrm{C}^{2s+1}(\Omega)}$, i.e. $\left\|\Delta_n^s u - \Delta_\rho^s u\right\|_{\mathrm{L}^2(\mu_n)} \leq C\varepsilon\left(\|u\|_{\mathrm{C}^{2s+1}(\Omega)} + 1\right)$).

The above discussion is clearly formal and we spend the remainder of the section making the proof rigorous. We do this in two stages. The first step gives operator bounds on Δ_n for smooth functions, i.e. quantifying $\|\Delta_n^j E_i\|_{\mathrm{L}^\infty(\mu_n)} \lesssim \|E_i\|_{\mathrm{C}^{2j}(\Omega)}$. The second step derives (27) from which we can prove Theorem 2.8 when combined with the first step.

2.2.1 Operator bounds on powers of the graph Laplacian

The aim of this subsection is to prove the following proposition.

Proposition 2.9 *Let Assumptions (A1)–(A4) hold and $m \in \mathbb{N}$. Define Δ_n by (8). Then, for all $\alpha > 1$, there exists $C > c > 0$ and $\varepsilon_0 > 0$ such that for any ε, n satisfying (16) and for all $v \in \mathrm{C}^{2m}(\Omega)$ we have*

$$\|\Delta_n^m v\|_{\mathrm{L}^2(\mu_n)} \leq C(\|v\|_{\mathrm{C}^{2m}(\Omega)} + 1)$$

with probability at least $1 - Cne^{-cn\varepsilon^{d+4m-2}} - Cn^{-\alpha}$.

Let us define the *non-local continuum Laplacian* by

$$\Delta_\varepsilon v(x) = \frac{2}{\varepsilon^2} \int_\Omega \eta_\varepsilon(|x-y|)\,(v(x) - v(y))\,\rho(y)\,\mathrm{d}y. \tag{28}$$

We prove the proposition in two steps. In the first step we show $\|\Delta_\varepsilon^m v\|_{\mathrm{L}^2(\Omega)} \leq C\|v\|_{\mathrm{C}^{2m}(\Omega)}$. In the second step we bound the difference $\|\Delta_n^m v - \Delta_\varepsilon^m v\|_{\mathrm{L}^2(\mu_n)}$. Initially

we consider the case when $m = 1$, which is just the difference of $\Delta_n v(x)$ to its expected value $\Delta_\varepsilon v(x) = \mathbb{E}[\Delta_n v(x)]$. We can then bootstrap this to $m > 1$. Putting the two steps together proves Proposition 2.9.

Lemma 2.10 *Let Assumptions (A1), (A3) and (A4) hold, and $k \in \mathbb{N}$. Define Δ_ε by (28). Then, there exists $C > 0$, $\varepsilon_0 > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$ and for all $v \in \mathrm{C}^{k+2}(\Omega)$ we have*

$$\|\Delta_\varepsilon v\|_{\mathrm{C}^k(\Omega)} \leq C\|v\|_{\mathrm{C}^{k+2}(\Omega)}. \tag{29}$$

Furthermore, if $v \in \mathrm{C}^{2k}(\Omega)$ then

$$\left\|\Delta_\varepsilon^k v\right\|_{\mathrm{C}^0(\Omega)} \leq C\|v\|_{\mathrm{C}^{2k}(\Omega)}. \tag{30}$$

Proof We can write, for ε sufficiently small, where ∇ above is the gradient in $\mathbb{R}^d$, and D^2 the matrix of second derivatives of a function on $\mathbb{R}^d$,

$$\begin{aligned}\Delta_\varepsilon v(x) &= \frac{2}{\varepsilon^2}\int_{B(x,\varepsilon)} \eta_\varepsilon(|x-y|)(v(x)-v(y))\rho(y)\,\mathrm{d}y \\ &= -\frac{2}{\varepsilon^2}\int_{\mathbb{R}^d}\eta(|z|)\left(\varepsilon\nabla v(x)\cdot z + \varepsilon^2\int_0^1\int_0^t D^2 v(x+\varepsilon s z)[z,z]\,\mathrm{d}s\,\mathrm{d}t\right) \\ &\quad\times\left(\rho(x)+\varepsilon\int_0^1\nabla\rho(x+\varepsilon s z)\cdot z\,\mathrm{d}s\right)\mathrm{d}z,\end{aligned}$$

by Taylor's theorem and a change of variables. Using the reflective symmetry of η we have $\int_{\mathbb{R}^d}\eta(|z|)z\,\mathrm{d}z = 0$ and hence,

$$\begin{aligned}\Delta_\varepsilon v(x) &= -2\nabla v(x)\cdot\int_{\mathbb{R}^d}\eta(|z|)z\int_0^1\nabla\rho(x+\varepsilon s z)\cdot z\,\mathrm{d}s\,\mathrm{d}z \\ &\quad - 2\rho(x)\int_{\mathbb{R}^d}\eta(|z|)\int_0^1\int_0^t D^2 v(x+\varepsilon s z)[z,z]\,\mathrm{d}s\,\mathrm{d}t\,\mathrm{d}z \\ &\quad - 2\varepsilon\int_{\mathbb{R}^d}\eta(|z|)\left(\int_0^1\int_0^t D^2 v(x+\varepsilon s z)[z,z]\,\mathrm{d}s\,\mathrm{d}t\right)\left(\int_0^1\nabla\rho(x+\varepsilon s z)\cdot z\,\mathrm{d}s\right)\mathrm{d}z.\end{aligned}$$

If $v \in \mathrm{C}^{k+2}(\Omega)$ and $\rho \in \mathrm{C}^{k+1}(\Omega)$ then $\Delta_\varepsilon v \in \mathrm{C}^k(\Omega)$ and moreover

$$\begin{aligned}\|\Delta_\varepsilon v\|_{\mathrm{C}^k(\Omega)} &\leq C\left(\|v\|_{\mathrm{C}^{k+1}(\Omega)}\|\rho\|_{\mathrm{C}^{k+1}(\Omega)} + \|v\|_{\mathrm{C}^{k+2}(\Omega)}\|\rho\|_{\mathrm{C}^k(\Omega)} + \varepsilon\|v\|_{\mathrm{C}^{k+2}(\Omega)}\|\rho\|_{\mathrm{C}^{k+1}(\Omega)}\right) \\ &\leq C\|v\|_{\mathrm{C}^{k+2}(\Omega)}.\end{aligned}$$

This proves the first part of the lemma. Iterating the estimate (29) implies (30). □

Now we turn to Step 2 and bounding the difference $\Delta_n - \Delta_\varepsilon$.

Lemma 2.11 *Let Assumptions (A1)–(A4) hold. Define Δ_n by (8) and Δ_ε by (28). For any $\varepsilon_0 > 0$ there exists $C > c > 0$ such that for any $\varepsilon \in (0, \varepsilon_0)$, $p > 0$, $n \in \mathbb{N}$ and $w \in \mathrm{C}^1(\Omega)$ we have*

$$\sup_{x \in \Omega_n} |(\Delta_n - \Delta_\varepsilon)\, w(x)| \leq \varepsilon^p \|w\|_{\mathrm{C}^1(\Omega)}$$

with probability at least $1 - Cne^{-cn\varepsilon^{d+2p+2}}$.

Proof Fix $w \in \mathrm{C}^1(\Omega)$, $x \in \Omega_n$ and let $\Xi_i = \frac{2}{\varepsilon^2}\eta_\varepsilon(|x - y|)\,(w(x) - w(y))$. So,

$$\frac{1}{n}\sum_{i=1}^{n} \Xi_i = \Delta_n w(x) \qquad \text{and} \qquad \mathbb{E}[\Xi_i] = \Delta_\varepsilon w(x). \tag{31}$$

Note that

$$|\Xi_i - \mathbb{E}[\Xi_i]| \leq \frac{C\|w\|_{\mathrm{C}^1(\Omega)}}{\varepsilon^{d+1}} \qquad \text{and} \qquad \mathbb{E}\,[\Xi_i - \mathbb{E}[\Xi_i]]^2 \leq \frac{C\|w\|^2_{\mathrm{C}^1(\Omega)}}{\varepsilon^{d+2}}.$$

By Bernstein's inequality for any $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} (\Xi_i - \mathbb{E}[\Xi_i]) \geq t\right) \leq \exp\left(-\frac{ct^2\varepsilon^{d+2}}{n\|w\|^2_{\mathrm{C}^1(\Omega)} + t\varepsilon\|w\|_{\mathrm{C}^1(\Omega)}}\right).$$

Choosing $t = n\varepsilon^p\|w\|_{\mathrm{C}^1(\Omega)}$ implies

$$\mathbb{P}\left(\sum_{i=1}^{n} (\Xi_i - \mathbb{E}[\Xi_i]) \geq n\varepsilon^p\|w\|_{\mathrm{C}^1(\Omega)}\right) \leq \exp\left(-\frac{cn\varepsilon^{d+2p+2}}{1+\varepsilon^{p+1}}\right) \leq \exp\left(-cn\varepsilon^{d+2p+2}\right).$$

Symmetrising the argument we have

$$\left|\sum_{i=1}^{n} (\Xi_i - \mathbb{E}[\Xi_i])\right| \leq n\varepsilon^p\|w\|_{\mathrm{C}^1(\Omega)}$$

with probability at least $1 - 2e^{-cn\varepsilon^{d+2p+2}}$. Substituting in (31) and union bounding over all $x \in \Omega_n$ we have proved the lemma. □

Using the above lemma we can provide a bound on $\Delta_n^m - \Delta_\varepsilon^m$.

Lemma 2.12 *Assume Assumptions (A1)–(A4) hold and $m \in \mathbb{N}$. Define Δ_n by (8) and Δ_ε by (28). Then, for all $\alpha > 1$, there exists $C > c > 0$ and $\varepsilon_0 > 0$ such that for any ε, n satisfying (16) and $v \in \mathrm{C}^{2m-1}(\Omega)$ we have*

$$\left\|\Delta_n^m v - \Delta_\varepsilon^m v\right\|_{\mathrm{L}^2(\mu_n)} \leq C\|v\|_{\mathrm{C}^{2m-1}(\Omega)}$$

with probability at least $1 - Cne^{-cn\varepsilon^{d+4m-2}} - Cn^{-\alpha}$.

Proof We can write

$$\begin{aligned}
\left\| \Delta_n^m v - \Delta_\varepsilon^m v \right\|_{\mathrm{L}^2(\mu_n)} &\leq \sum_{i=0}^{m-1} \left\| \Delta_n^{m-i} \Delta_\varepsilon^i v - \Delta_n^{m-i-1} \Delta_\varepsilon^{i+1} v \right\|_{\mathrm{L}^2(\mu_n)} \\
&\leq \sum_{i=0}^{m-1} \|\Delta_n\|_{\mathrm{op}}^{m-i-1} \left\| (\Delta_n - \Delta_\varepsilon) \Delta_\varepsilon^i v \right\|_{\mathrm{L}^2(\mu_n)} \\
&\leq C \sum_{i=0}^{m-1} \|\Delta_\varepsilon^i v\|_{\mathrm{C}^1(\Omega)} \\
&\leq C \sum_{i=0}^{m-1} \|v\|_{\mathrm{C}^{2i+2}(\Omega)} \\
&\leq C \|v\|_{\mathrm{C}^{2m}(\Omega)}
\end{aligned}$$

by Lemmas 2.3, 2.10 and 2.11 with probability at least $1 - Cne^{-cn\varepsilon^{d+4m-2}} - Cn^{-\alpha}$. □

2.2.2 Proof of Theorem 2.8

Now, we note that

$$\begin{aligned}
\Delta_n^s u(x) - \Delta_\rho^s u(x) &= \Delta_n^{s-1} \left(\Delta_n - \Delta_\rho\right) v^{(0)}(x) + \left(\Delta_n^{s-1} - \Delta_\rho^{s-1}\right) v^{(1)}(x) \\
&= \Delta_n^{s-1} \left(\Delta_n - \Delta_\rho\right) v^{(0)}(x) + \Delta_n^{s-2} \left(\Delta_n - \Delta_\rho\right) v^{(1)}(x) \\
&\quad + \left(\Delta_n^{s-2} - \Delta_\rho^{s-2}\right) v^{(2)}(x) \\
&= \cdots \\
&= \sum_{k=1}^{s} \Delta_n^{s-k} \left(\Delta_n - \Delta_\rho\right) v^{(k-1)}(x)
\end{aligned}$$

where $v^{(i)} = \Delta_\rho^i u$.

The idea is now to use pointwise convergence but to keep track of higher order terms than the estimates that appear in [12, 15]. For example, [15] shows that if $f \in \mathrm{C}^3$ then

$$\left|\Delta_n f(x) - \Delta_\rho f(x)\right| \leq C \|f\|_{\mathrm{C}^3} \vartheta, \tag{32}$$

where $\vartheta \geq \varepsilon$, with probability at least $1 - Cne^{-cn\varepsilon^{d+2}\vartheta^2}$. Directly applying the operator bounds we have

$$\left\| \Delta_n^s u - \Delta_\rho^s u \right\|_{\mathrm{L}^2(\mu_n)} \leq \sum_{k=1}^{s} \|\Delta_n\|_{\mathrm{op}}^{s-k} \left\| \left(\Delta_n - \Delta_\rho\right) v^{(k-1)} \right\|_{\mathrm{L}^2(\Omega)}$$

$$\leq C \sum_{k=1}^{s} \varepsilon^{-2(s-k)} \|v^{(k-1)}\|_{\mathrm{C}^3(\Omega)} \vartheta_k$$

$$\leq C \|u\|_{\mathrm{C}^{2s+1}} \sum_{k=1}^{s} \varepsilon^{-2(s-k)} \vartheta_k.$$

If we could choose $\vartheta_k = \varepsilon^{1+2(s-k)}$ then the proof is immediate; however the pointwise convergence result (32) requires $\vartheta \geq \varepsilon$ which rules out this choice. However, we will show that this gives the right answer, in particular, that the convergence is within ε with probability at least $1 - Cne^{-cn\varepsilon^{d+4s}}$. The rest of the section is devoted to removing the assumption that $\vartheta_k \geq \varepsilon$.

Proof of Theorem 2.8 Let us fix k and write $v = v^{(k-1)}$. Then, assuming $u \in \mathrm{C}^{2s+1}(\Omega)$ we have $v \in \mathrm{C}^{2(s-k)+3}(\Omega)$ and so, for y sufficiently close to x,

$$v(y) = v(x) + \sum_{j=1}^{2(s-k+1)} \sum_{i^{(j)} \in \{1,\ldots,d\}^j} a^{(j)}_{i^{(j)}} \prod_{\ell=1}^{j} \left(y_{i^{(j)}_\ell} - x_{i^{(j)}_\ell} \right) + O\left(\left| y_{i^{(j)}_\ell} - x_{i^{(j)}_\ell} \right|^{2(s-k)+3} \right)$$

where

$$a^{(j)}_{i^{(j)}} = \frac{1}{j!} \frac{\partial^j v}{\partial x_{i^{(j)}_1} \cdots \partial x_{i^{(j)}_j}}(x),$$

$i^{(j)} = (i^{(j)}_1, \ldots, i^{(j)}_j) \in \{1, \ldots d\}^j$ and the big-O notation is understood as meaning that there exists a bounded function, say $\Xi_{i^{(j)}_\ell}$ depending on $x_{i^{(j)}_\ell}$ and $y_{i^{(j)}_\ell}$ such that $O\left(\left| y_{i^{(j)}_\ell} - x_{i^{(j)}_\ell} \right|^{2(s-k)+3} \right) = \Xi_{i^{(j)}_\ell}(x_{i^{(j)}_\ell}, y_{i^{(j)}_\ell}) |y_{i^{(j)}_\ell} - x_{i^{(j)}_\ell}|^{2(s-k)+3}$. Now we can write

$$\begin{aligned}
\Delta_n v(x) &= \frac{2}{n\varepsilon^2} \sum_{y \in \Omega_n} W_{xy}(v(x) - v(y)) \\
&= -\frac{2}{n\varepsilon^2} \sum_{y \in \Omega_n} W_{xy} \left[\sum_{j=1}^{2(s-k+1)} \sum_{i^{(j)} \in \{1,\ldots d\}^j} a^{(j)}_{i^{(j)}} \prod_{\ell=1}^{j} \left(y_{i^{(j)}_\ell} - x_{i^{(j)}_\ell} \right) \right] \\
&\quad + O\left(\frac{\varepsilon^{2(s-k)+1}}{n} \sum_{y \in \Omega_n} W_{xy} \right).
\end{aligned}$$

By Lemma 2.2, $\frac{1}{n}\sum_y W_{xy} \leq C$ for all $x \in \Omega_n$ with probability at least $1 - 2ne^{-cn\varepsilon^d}$, hence we can write (with probability at least $1 - 2ne^{-cn\varepsilon^d}$)

$$\Delta_n v(x) = -\sum_{j=1}^{2(s-k+1)} \sum_{i^{(j)} \in \{1,\dots d\}^j} a^{(j)}_{i^{(j)}} I^{(j)}_{i^{(j)}} + O(\varepsilon^{2(s-k)+1})$$

where

$$I^{(j)}_{i^{(j)}} = \sum_{y \in \Omega_n} \Psi^{(j)}_{i^{(j)}}, \qquad \Psi^{(j)}_{i^{(j)}}(y) = \frac{2}{n\varepsilon^2} W_{xy} \prod_{\ell=1}^{j} \left(y^{(j)}_{i^{(j)}_\ell} - x^{(j)}_{i^{(j)}_\ell} \right).$$

Note that $\|\Psi^{(j)}_{i^{(j)}}\|_{\mathrm{L}^\infty} \leq \frac{C\varepsilon^{j-2-d}}{n}$ and $\mathbb{E}[\Psi^{(j)}_{i^{(j)}}(Y)^2] \leq \frac{C\varepsilon^{2(j-2)-d}}{n^2}$, which follows from

$$\begin{aligned}
|\Psi^{(j)}_{i^{(j)}}(y)| &\lesssim \frac{1}{n\varepsilon^2} \underbrace{W_{xy}}_{\lesssim \frac{1}{\varepsilon^d}} \prod_{\ell=1}^{j} \left(\underbrace{y^{(j)}_{i^{(j)}_\ell} - x^{(j)}_{i^{(j)}_\ell}}_{\lesssim \varepsilon} \right) \lesssim \frac{1}{n\varepsilon^{2+d-j}} \\
\mathbb{E}[\Psi^{(j)}_{i^{(j)}}(Y)^2] &\lesssim \frac{1}{n^2\varepsilon^4} \mathbb{E}\left[\underbrace{W^2_{xy}}_{\lesssim \frac{1}{\varepsilon^{2d}}} \prod_{\ell=1}^{j} \underbrace{\left(y^{(j)}_{i^{(j)}_\ell} - x^{(j)}_{i^{(j)}_\ell} \right)^2}_{\varepsilon^2} \right] \\
&\lesssim \frac{1}{n^2\varepsilon^{4+2d-2j}} \underbrace{\int \int_{|x-y| \leq \varepsilon} \mathrm{d}x\, \mathrm{d}y}_{\lesssim \varepsilon^d} \\
&\lesssim \frac{1}{n^2\varepsilon^{4+d-2j}}.
\end{aligned}$$

Hence, by Bernstein's inequality

$$\begin{aligned}
I^{(j)}_{i^{(j)}} &= \frac{2}{\varepsilon^2} \int_\Omega \eta_\varepsilon(|x-y)|) \left[\prod_{\ell=1}^{j} \left(y_{i^{(j)}_\ell} - x_{i^{(j)}_\ell} \right) \right] \rho(y)\, \mathrm{d}y + O(\varepsilon^{j-2}\vartheta) \\
&= 2\varepsilon^{j-2} \int_{\mathbb{R}^d} \eta(|z|) \left[\prod_{\ell=1}^{j} z_{i^{(j)}_\ell} \right] \rho(x + \varepsilon z)\, \mathrm{d}z + O(\varepsilon^{j-2}\vartheta)
\end{aligned}$$

with probability at least $1 - 2ne^{-cn\varepsilon^d\vartheta^2}$ for all $x \in \Omega_n$. After union bounding we may assume that the above estimate holds for all $x \in \Omega_n$, for all $k = 1, \dots, s$, for all $j = 1, \dots, k$, and for all $i^{(j)} \in \{1, \dots, d\}^j$ with probability at least $1 - Cne^{-cn\varepsilon^d\vartheta^2}$. We

choose $\vartheta = \varepsilon^{2(s-k)+3-j}$ and so, since $\vartheta \geq \varepsilon^{2s}$, the following holds with probability at least $1 - Cne^{-cn\varepsilon^{d+4s}}$.

Now we approximate

$$\rho(x+\varepsilon z) = \sum_{m=0}^{2(s-k+1)-j} \varepsilon^m \sum_{p^{(m)}\in\{1,\dots,d\}^m} b^{(m)}_{p^{(m)}} \prod_{q=1}^{m} z_{p_q^{(m)}} + O(\varepsilon^{2(s-k)-j+3})$$

where

$$b^{(m)}_{p^{(m)}} = \frac{1}{m!} \frac{\partial^m \rho}{\partial x_{p_1^{(m)}} \cdots \partial x_{p_m^{(m)}}}(x).$$

Hence,

$$I^{(j)}_{i^{(j)}} = 2 \sum_{m=0}^{2(s-k+1)-j} \sum_{p^{(m)}\in\{1,\dots,d\}^m} \varepsilon^{m+j-2} b^{(m)}_{p^{(m)}} \int_{\mathbb{R}^d} \eta(|z|) \left[\prod_{\ell=1}^{j} z_{i_\ell^{(j)}}\right] \left[\prod_{q=1}^{m} z_{p_q^{(m)}}\right] \mathrm{d}z + O(\varepsilon^{(2(s-k)+1}).$$

Let

$$F(j,m) = \sum_{i^{(j)}\in\{1,\dots,d\}^j} \sum_{p^{(m)}\in\{1,\dots,d\}^m} a^{(j)}_{i^{(j)}} b^{(m)}_{p^{(m)}} C(i^{(j)}, p^{(m)})$$

and

$$C(i^{(j)}, p^{(m)}) = -2 \int_{\mathbb{R}^d} \eta(|z|) \left[\prod_{\ell=1}^{j} z_{i_\ell^{(j)}}\right] \left[\prod_{q=1}^{m} z_{p_q^{(m)}}\right] \mathrm{d}z$$

so that

$$\Delta_n v(x) = \sum_{j=1}^{2(s-k+1)} \sum_{m=0}^{2(s-k+1)-j} \varepsilon^{m+j-2} F(j,m) + O(\varepsilon^{(2(s-k)+1}).$$

We now look at the following terms: (i) $j=1, m=0$; (ii) $j=1, m=1$; and (iii) $j=2, m=0$ (the terms which are potentially of order ε^{-1} and ε^0). For (i),

$$C(i,\emptyset) = -2 \int_{\mathbb{R}^d} \eta(|z|) z_i \,\mathrm{d}z = 0.$$

For (ii),

$$C(i,p) = -2 \int_{\mathbb{R}^d} \eta(|z|) z_i z_p \,\mathrm{d}x = \begin{cases} 0 & \text{if } i \neq p \\ -2\sigma_\eta & \text{if } i = p. \end{cases}$$

For (iii),

$$C((i_1,i_2),\emptyset) = -2\int_{\mathbb{R}^d} \eta(|z|) z_{i_1} z_{i_2}\, \mathrm{d}z = \begin{cases} 0 & \text{if } i_1 \neq i_2 \\ -2\sigma_\eta & \text{if } i_1 = i_2. \end{cases}$$

So $F(1,0) = 0$,

$$F(1,1) = -2\sigma_\eta \sum_{i=1}^{d} a_i^{(1)} b_i^{(1)} = -2\sigma_\eta \nabla v(x)\cdot\nabla\rho(x),$$

and

$$F(2,0) = -2\sigma_\eta \sum_{i=1}^{d} a_{i,i}^{(2)} b^{(0)} = -\sigma_\eta \rho(x)\mathrm{trace}(D^2 v(x)).$$

As $F(1,0)\varepsilon^{-1} + F(1,1) + F(2,0) = -\frac{\sigma_\eta}{\rho(x)}\mathrm{div}(\rho^2\nabla v)(x) = \Delta_\rho v(x)$ then we have (adding back the k dependence on v)

$$\begin{aligned}\Delta_n v^{(k-1)}(x) - \Delta_\rho v^{(k-1)}(x) = &\sum_{m=2}^{2(s-k)+1} \varepsilon^{m-1} F(1,m) + \sum_{m=1}^{2(s-k)} \varepsilon^m F(2,m) \\ &+ \sum_{j=3}^{2(s-k+1)} \sum_{m=0}^{2(s-k+1)-j} \varepsilon^{m+j-2} F(j,m) + O(\varepsilon^{2(s-k)+1}).\end{aligned}$$

In particular, if we let $F_{j,m}^{(k)}(x) = F(j,m)$ and define $E^{(k)}(x)$ to satisfy $O(\varepsilon^{2(s-k)+1}) = \varepsilon^{2(s-k)+1} E^{(k)}(x)$ then

$$\begin{aligned}\left\|\left(\Delta_n^s - \Delta_\rho^s\right) u\right\|_{\mathrm{L}^2(\mu_n)} &\leq \sum_{k=1}^{s} \left\|\Delta_n^{s-k}\left(\Delta_n - \Delta_\rho\right) v^{(k-1)}\right\|_{\mathrm{L}^2(\mu_n)} \\ &\leq \sum_{k=1}^{s}\sum_{m=2}^{2(s-k)+1} \varepsilon^{m-1} \left\|\Delta_n^{s-k} F_{1,m}^{(k)}\right\|_{\mathrm{L}^2(\mu_n)} \\ &\quad + \sum_{k=1}^{s}\sum_{m=1}^{2(s-k)} \varepsilon^{m} \left\|\Delta_n^{s-k} F_{2,m}^{(k)}\right\|_{\mathrm{L}^2(\mu_n)} \\ &\quad + \sum_{k=1}^{s}\sum_{j=3}^{2(s-k+1)}\sum_{m=0}^{2(s-k+1)-j} \varepsilon^{m+j-2} \left\|\Delta_n^{s-k} F_{j,m}^{(k)}\right\|_{\mathrm{L}^2(\mu_n)} \\ &\quad + \sum_{k=1}^{s} \varepsilon^{2(s-k)+1} \left\|\Delta_n^{s-k} E^{(k)}\right\|_{\mathrm{L}^2(\mu_n)}.\end{aligned}$$

By Lemma 2.3 (with probability at least $1 - Cn^{-\alpha}$) we have

$$\varepsilon^{2(s-k)+1} \left\| \Delta_n^{s-k} E^{(k)} \right\|_{\mathrm{L}^2(\mu_n)} \leq \varepsilon^{2(s-k)+1} \|\Delta_n\|_{\mathrm{op}}^{s-k} \|E^{(k)}\|_{\mathrm{L}^2(\mu_n)} \leq \varepsilon \|E^{(k)}\|_{\mathrm{L}^2(\mu_n)}.$$

We also have $F_{j,m}^{(k)} \in \mathrm{C}^{2(s-k)+3-j}$ and $\|F_{j,m}^{(k)}\|_{\mathrm{C}^{2(s-k)+3-j}(\Omega)} \leq C\|u\|_{\mathrm{C}^{2s+1}(\Omega)}$, therefore we have for $j \geq 3$

$$\begin{aligned} \varepsilon^{m+j-2} \left\| \Delta_n^{s-k} F_{j,m}^{(k)} \right\|_{\mathrm{L}^2(\mu_n)} &\leq \varepsilon^{m+j-2} \|\Delta_n\|_{\mathrm{op}}^{\frac{j-3}{2}} \left\| \Delta_n^{\frac{2(s-k)+3-j}{2}} F_{j,m}^{(k)} \right\|_{\mathrm{L}^2(\mu_n)} \\ &\leq C\varepsilon^{m+1} \left(\left\| F_{j,m}^{(k)} \right\|_{\mathrm{C}^{2(s-k)+3-j}(\Omega)} + 1 \right) \end{aligned}$$

with probability at least $1 - Cn^{-\alpha} - Cne^{-cn\varepsilon^{d+4m-2}} \geq 1 - Cn^{-\alpha} - Cne^{-cn\varepsilon^{d+2s}}$ by Lemma 2.3 and Proposition 2.9. When $j = 1, 2$ we have, directly from Proposition 2.9,

$$\left\| \Delta_n^{s-k} F_{j,m}^{(k)} \right\|_{\mathrm{L}^2(\mu_n)} \leq \left\| F_{j,m}^{(k)} \right\|_{\mathrm{C}^{2(s-k)}(\Omega)} \leq \left\| F_{j,m}^{(k)} \right\|_{\mathrm{C}^{2(s-k)+3-j}(\Omega)} \leq C\|u\|_{\mathrm{C}^{2s+1}(\Omega)}$$

with probability at least $1 - Cne^{-cn\varepsilon^{d+2s}}$. Hence,

$$\left\| \left(\Delta_n^s - \Delta_\rho^s\right) u \right\|_{\mathrm{L}^2(\mu_n)} \leq C\varepsilon \left(\|u\|_{\mathrm{C}^{2s+1}(\Omega)} + 1 \right)$$

with probability at least $1 - Cn^{-\alpha} - Cne^{-cn\varepsilon^{d+4s}}$. □

Remark 11 In our proofs we avoid attempting to establish pointwise consistency results for the difference $\Delta_n^s - \Delta_\rho^s$ (for arbitrary $s \in \mathbb{N}$) when acting on smooth enough functions, and instead by careful manipulation of the equations, we rely only on the existing pointwise consistency results for the case $s = 1$ [11, 15].

2.2.3 Proof of Proposition 2.7

We start with two preliminary lemmas which will be used in the proof of Proposition 2.7.

Lemma 2.13 *Let $\tau > 0$, $s > 0$, Δ_n be defined by (8) and Δ_ρ defined by (12) where σ_η is defined by (13). Assume w_n and w solve*

$$\begin{aligned} \tau \Delta_n^s w_n + w_n &= h_n \\ \tau \Delta_\rho^s w + w &= h \end{aligned}$$

for $h_n \in \mathrm{L}^2(\mu_n)$ and $h \in \mathrm{L}^2(\mu)$. Then,

$$\begin{aligned} \|w_n\|_{\mathrm{L}^2(\mu_n)} &\leq \|h_n\|_{\mathrm{L}^2(\mu_n)} \\ \|w\|_{\mathrm{L}^2(\mu)} &\leq \|h\|_{\mathrm{L}^2(\mu)}. \end{aligned}$$

Proof Let $\{q_i^{(n)}\}_{i=1}^n$ be an eigenbasis of Δ_n with non-negative eigenvalues $\{\lambda_i^{(n)}\}_{i=1}^n$. Then w_n solving $\tau \Delta_n^s w_n + w_n = h_n$ implies

$$\left(\tau[\lambda_i^{(n)}]^s + 1\right) \langle w_n, q_i^{(n)} \rangle_{\mathrm{L}^2(\mu_n)} = \langle h_n, q_i^{(n)} \rangle_{\mathrm{L}^2(\mu_n)}.$$

So,

$$\begin{aligned} \|w_n\|^2_{\mathrm{L}^2(\mu_n)} &= \sum_{i=1}^n \left| \left\langle w_n, q_i^{(n)} \right\rangle_{\mathrm{L}^2(\mu_n)} \right|^2 \\ &= \sum_{i=1}^n \left| \frac{\left\langle h_n, q_i^{(n)} \right\rangle_{\mathrm{L}^2(\mu_n)}}{1 + \tau \left[\lambda_i^{(n)}\right]^s} \right|^2 \\ &\leq \sum_{i=1}^n \left| \left\langle h_n, q_i^{(n)} \right\rangle_{\mathrm{L}^2(\mu_n)} \right|^2 \\ &= \|h_n\|^2_{\mathrm{L}^2(\mu_n)}. \end{aligned}$$

The proof for $\|w\|_{\mathrm{L}^2(\mu)} \leq \|h\|_{\mathrm{L}^2(\mu)}$ is analogous. □

Lemma 2.14 *Assume Assumptions (A3) and (A5) hold and $s > 0$. Define Δ_ρ by (12) where σ_η is defined by (13). Let u_τ^* be the solution to (26). Then, for all $\tau_0 > 0$ there exists C such that*

$$\sup_{\tau \in (0, \tau_0)} \|u_\tau^*\|_{\mathrm{C}^{2s+1}(\Omega)} \leq C.$$

Proof Let $\{(\lambda_i, q_i)\}_{i=1}^\infty$ be eigenpairs of Δ_ρ. Define $\mathcal{H}^k(\Omega) = \left\{u \in \mathrm{L}^2(\mu) \colon \sum_{i=1}^\infty \lambda_i^k \langle u, q_i \rangle^2_{\mathrm{L}^2(\mu)} < +\infty\right\}$ with the norm $\|u\|^2_{\mathcal{H}^k(\Omega)} = \sum_{i=1}^\infty \lambda_i^k \langle u, q_i \rangle^2_{\mathrm{L}^2(\mu)}$. And let $\mathrm{H}^k(\Omega)$ be the usual Sobolev space with square integrable kth (weak) derivative. By [18, Lemma 17] $\mathcal{H}^k(\Omega) \subseteq \mathrm{H}^k(\Omega)$ and there exists $C > c > 0$ (depending only on the choice of k) such that

$$C\|u\|_{\mathcal{H}^k(\Omega)} \geq \|u\|_{\mathrm{H}^k(\Omega)} \geq c\|u\|_{\mathcal{H}^k(\Omega)} \qquad \text{for all } u \in \mathcal{H}^k(\Omega).$$

As in the proof of Lemma 2.13 we take advanatge of the fact that $\langle u_\tau^*, q_i \rangle_{\mathrm{L}^2(\mu)} = \frac{\langle g, q_i \rangle_{\mathrm{L}^2(\mu)}}{1 + \tau \lambda_i^s}$ to infer

$$\begin{aligned} \|u_\tau^*\|^2_{\mathcal{H}^k(\Omega)} &= \sum_{i=1}^\infty \lambda_i^k \langle u_\tau^*, q_i \rangle^2_{\mathrm{L}^2(\mu)} \\ &= \sum_{i=1}^\infty \lambda_i^k \left| \frac{\langle g, q_i \rangle_{\mathrm{L}^2(\mu)}}{1 + \tau \lambda_i^s} \right|^2 \end{aligned}$$

$$\begin{aligned}
&\leq \sum_{i=1}^{\infty} \lambda_i^k \langle q, q_i \rangle^2_{\mathrm{L}^2(\mu)} \\
&= \|g\|^2_{\mathcal{H}^k(\Omega)} \\
&\leq C^2 \|g\|^2_{\mathrm{H}^k(\Omega)}.
\end{aligned}$$

Hence $\|u_\tau^*\|_{\mathrm{H}^k(\Omega)} \leq \frac{C}{c}\|g\|_{\mathrm{H}^k(\Omega)}$. By choosing k sufficiently large and employing Morrey's inequality we can find $\bar{C}$ such that $\|u\|_{\mathrm{C}^{2s+1}(\Omega)} \leq \bar{C}\|u\|_{\mathrm{H}^k(\Omega)}$ for all $u \in \mathrm{H}^k(\Omega)$. In particular, $\|u_\tau^*\|_{\mathrm{C}^{2s+1}(\Omega)} \leq \frac{C\bar{C}}{c}\|g\|_{\mathrm{H}^k(\Omega)}$ which proves the lemma. □

Remark 12 The constant C in the previous lemma depends on the choice of k and g. In particular, we use equivalence of norms between the spaces $\mathcal{H}^k$ and H^k, Morrey's inequality to embed C^{2s+1} into H^k, and the H^k norm of g. We note, however, that the constant c depends on g only through $\|g\|_{\mathrm{H}^k(\Omega)}$, and this dependence is linear.

We can now prove Proposition 2.7.

Proof of Proposition 2.7 We have

$$\tau \Delta_n^s u_\tau^* + u_\tau^* - g = \tau \left(\Delta_n^s - \Delta_\rho^s\right) u_\tau^*$$

on Ω_n. So, letting $w = u_\tau^* \lfloor_{\Omega_n} - u_{n,\tau}^{g*}$ we can bound

$$\tau \Delta_n^s w + w = \tau \left(\Delta_n^s - \Delta_\rho^s\right) u_\tau^*.$$

By Lemma 2.13 and Theorem 2.8

$$\|w\|_{\mathrm{L}^2(\mu_n)} \leq \tau \left\|\left(\Delta_n^s - \Delta_\rho^s\right) u_\tau^*\right\|_{\mathrm{L}^2(\mu_n)} \leq C\tau\varepsilon \left(\|u_\tau^*\|_{\mathrm{C}^{2s+1}(\Omega)} + 1\right)$$

with probability at least $1 - Cn^{-\alpha} - Cne^{-cn\varepsilon^{d+4s}}$. By Lemma 2.14 $\|u_\tau^*\|_{\mathrm{C}^{2s+1}(\Omega)}$ can be bounded for all $\tau \in (0, \tau_0)$. This completes the proof of the first inequality.

We can derive the second inequality from the first inequality as follows

$$\begin{aligned}
\left\|\Delta_n^{\frac{s}{2}} u_{n,\tau}^{g*} - \Delta_\rho^{\frac{s}{2}} u_\tau^* \lfloor_{\Omega_n}\right\|^2_{\mathrm{L}^2(\mu_n)} &\leq 2\left\|\Delta_n^{\frac{s}{2}} u_{n,\tau}^{g*} - \Delta_n^{\frac{s}{2}} u_\tau^*\right\|^2_{\mathrm{L}^2(\mu_n)} + 2\left\|\Delta_n^{\frac{s}{2}} u_\tau^* - \Delta_\rho^{\frac{s}{2}} u_\tau^* \lfloor_{\Omega_n}\right\|^2_{\mathrm{L}^2(\mu_n)} \\
&\leq 2\left\|\Delta_n^s \left(u_{n,\tau}^{g*} - u_\tau^* \lfloor_{\Omega_n}\right)\right\|_{\mathrm{L}^2(\mu_n)} \left\|u_{n,\tau}^{g*} - u_\tau^* \lfloor_{\Omega_n}\right\|_{\mathrm{L}^2(\mu_n)} \\
&\quad + C\varepsilon^2 \left(\|u_\tau^*\|_{\mathrm{C}^{s+1}(\Omega)} + 1\right)^2
\end{aligned}$$

with probability at least $1 - Cn^{-\alpha} - Cne^{-cn\varepsilon^{d+4s}}$ where we have used Theorem 2.8 on the second term and the computation

$$\left\|\Delta_n^{\frac{s}{2}} u_{n,\tau}^{g*} - \Delta_n^{\frac{s}{2}} u_\tau^*\right\|^2_{\mathrm{L}^2(\mu_n)} = \left\langle \Delta_n^{\frac{s}{2}} (u_{n,\tau}^{g*} - u_\tau^*), \Delta_n^{\frac{s}{2}} (u_{n,\tau}^{g*} - u_\tau^*)\right\rangle_{\mathrm{L}^2(\mu_n)}$$

$$= \left\langle \Delta_n^s (u_{n,\tau}^{g*} - u_\tau^*), u_{n,\tau}^{g*} - u_\tau^* \right\rangle_{\mathrm{L}^2(\mu_n)}$$
$$\leq \left\| \Delta_n^s (u_{n,\tau}^{g*} - u_\tau^*) \right\|_{\mathrm{L}^2(\mu_n)} \left\| u_{n,\tau}^{g*} - u_\tau^* \right\|_{\mathrm{L}^2(\mu_n)}$$

on the first term. Comparing the Euler–Lagrange equations we have

$$\tau \Delta_n^s \left(u_{n,\tau}^{g*} - u_\tau^* \lfloor \Omega_n \right) + \left(u_{n,\tau}^{g*} - u_\tau^* \right) = \tau \left(\Delta_\rho^s - \Delta_n^s \right) u_\tau^*.$$

By Theorem 2.8 we can derive the bound

$$\begin{aligned} \left\| \Delta_n^s \left(u_{n,\tau}^{g*} - u_\tau^* \lfloor \Omega_n \right) \right\|_{\mathrm{L}^2(\mu_n)} &\leq \frac{1}{\tau} \left\| u_{n,\tau}^{g*} - u_\tau^* \lfloor \Omega_n \right\|_{\mathrm{L}^2(\mu_n)} + \left\| \left(\Delta_\rho^s - \Delta_n^s \right) u_\tau^* \right\|_{\mathrm{L}^2(\mu_n)} \\ &\leq C\varepsilon \left(1 + \| u_\tau^* \|_{\mathrm{C}^{2s+1}(\Omega)} \right) \end{aligned}$$

with probability at least $1 - Cn^{-\alpha} - Cne^{-cn\varepsilon^{d+4s}}$. Therefore,

$$\left\| \Delta_n^{\frac{s}{2}} u_{n,\tau}^{g*} - \Delta_\rho^{\frac{s}{2}} u_\tau^* \lfloor \Omega_n \right\|_{\mathrm{L}^2(\mu_n)}^2 \leq C\varepsilon^2 \left(\tau + (1+\tau) \| u_\tau^* \|_{\mathrm{C}^{2s+1}(\Omega)}^2 \right)$$

with probability at least $1 - Cn^{-\alpha} - Cne^{-cn\varepsilon^{d+4s}}$. If $\tau \leq \tau_0$ then we can bound by $C\varepsilon^2$ as required. □

Putting together Propositions 2.1 and 2.7 proves Theorem 1.1 and Remark 3.

3 L^2 bias estimates

Recalling that the Fréchet derivative of $\mathcal{E}_{\infty,\tau}^{(g)}$ is

$$\frac{1}{2} \nabla_{\mathrm{L}^2(\mu)} \mathcal{E}_{\infty,\tau}^{(g)}(u) = \tau \Delta_\rho^s u + u - g$$

then one can easily check that the following subgradient equality holds

$$\left\langle \nabla_{\mathrm{L}^2(\mu)} \mathcal{E}_{\infty,\tau}^{(g)}(w), w - v \right\rangle_{\mathrm{L}^2(\mu)} - \| w - v \|_{\mathrm{L}^2(\mu)}^2 - \tau \left\| \Delta_\rho^{\frac{s}{2}} (w - v) \right\|_{\mathrm{L}^2(\mu)}^2 = \mathcal{E}_{\infty,\tau}^{(g)}(w) - \mathcal{E}_{\infty,\tau}^{(g)}(v) \tag{33}$$

for any $v, w \in \mathrm{H}^s(\Omega)$. Since $\nabla_{\mathrm{L}^2(\mu)} \mathcal{E}_{\infty,\tau}^{(g)}(u_\tau^*) = 0$ and g is sufficiently regular then

$$\begin{aligned} \| u_\tau^* - g \|_{\mathrm{L}^2(\mu)}^2 + \tau \left\| \Delta_\rho^{\frac{s}{2}} (u_\tau^* - g) \right\|_{\mathrm{L}^2(\mu)}^2 &= \mathcal{E}_{\infty,\tau}^{(g)}(g) - \mathcal{E}_{\infty,\tau}^{(g)}(u_\tau^*) \\ &= \left\langle \nabla_{\mathrm{L}^2(\mu)} \mathcal{E}_{\infty,\tau}^{(g)}(g), g - u_\tau^* \right\rangle_{\mathrm{L}^2(\mu)} - \| g - u_\tau^* \|_{\mathrm{L}^2(\mu)}^2 \\ &\quad - \tau \left\| \Delta_\rho^{\frac{s}{2}} (g - u_\tau^*) \right\|_{\mathrm{L}^2(\mu)}^2 \end{aligned}$$

where for the first equality we let $w = u_\tau^*$, $v = g$ in (33), and in the second equality we let $w = g$, $v = u_\tau^*$ in (33). Hence

$$\begin{aligned}\|u_\tau^* - g\|_{\mathrm{L}^2(\mu)}^2 + \tau \left\|\Delta_\rho^{\frac{s}{2}}(u_\tau^* - g)\right\|_{\mathrm{L}^2(\mu)}^2 &= \frac{1}{2}\left\langle \nabla_{\mathrm{L}^2(\mu)}\mathcal{E}_{\infty,\tau}^{(g)}(g), g - u_\tau^*\right\rangle_{\mathrm{L}^2(\mu)} \\ &\leq \frac{1}{2}\left\|\nabla_{\mathrm{L}^2(\mu)}\mathcal{E}_{\infty,\tau}^{(g)}(g)\right\|_{\mathrm{L}^2(\mu)} \|g - u_\tau^*\|_{\mathrm{L}^2(\mu)}.\end{aligned}$$

It follows that

$$\left\|u_\tau^* - g\right\|_{\mathrm{L}^2(\mu)} \leq \frac{1}{2}\left\|\nabla_{\mathrm{L}^2(\mu)}\mathcal{E}_{\infty,\tau}^{(g)}(g)\right\|_{\mathrm{L}^2(\mu)} = \tau \left\|\Delta_\rho^s g\right\|_{\mathrm{L}^2(\mu)}$$

and

$$\left\|\Delta_\rho^{\frac{s}{2}}(u_\tau^* - g)\right\|_{\mathrm{L}^2(\mu)}^2 \leq \frac{1}{2}\left\|\nabla_{\mathrm{L}^2(\mu)}\mathcal{E}_{\infty,\tau}^{(g)}(g)\right\|_{\mathrm{L}^2(\mu)} \left\|\Delta_\rho^s g\right\|_{\mathrm{L}^2(\mu)} = \frac{\tau}{2}\left\|\Delta_\rho^s g\right\|_{\mathrm{L}^2(\mu)}^2$$

which proves Theorem 1.2 and Remark 5.

Acknowledgements RM was supported by the NSF Grant DMS-2307971 and a Simons Foundation Grant MP-TSM-00002904. NGT was supported by the NSF grants DMS-2005797 and DMS-2236447. MT was supported by the European Research Council under the European Union's Horizon 2020 research and innovation programme Grant agreement No 777826 (NoMADS).

Data availability No data was created or analyzed in this study, as such data sharing is not applicable.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Schoenberg, I.: Spline functions and the problem of graduation. Proc. Natl. Acad. Sci. U.S.A. **52**(4), 947–950 (1964)
2. Schoenberg, I.: On interpolation by spline functions and its minimum properties. Int. Ser. Numer. Anal. **5**, 109–129 (1964)
3. Wahba, G.: Spline Models for Observational Data, vol. 59. SIAM, Philadelphia (1990)
4. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using Gaussian fields and harmonic functions. In: Proceedings of the 20th International Conference on Machine Learning, pp. 912–919 (2003)

5. Nadler, B., Srebro, N., Zhou, X.: Statistical consistency of semi-supervised learning: The limit of infinite unlabelled data. In: Advances in Neural Information Processing Systems, pp. 1330–1338 (2009)
6. El Alaoui, A., Cheng, X., Ramdas, A., Wainwright, M.J., Jordan, M.I.: Asymptotic behavior of ℓ_p-based Laplacian regularization in semi-supervised learning. In: Conference on Learning Theory, pp. 879–906 (2016)
7. Calder, J., Slepčev, D., Thorpe, M.: Rates of convergence for Laplacian semi-supervised learning with low labelling rates. Res. Math. Sci. **10**(1) (2023). arXiv:2006.02765
8. von Luxburg, U., Belkin, M., Bousquet, O.: Consistency of spectral clustering. Ann. Stat. **36**(2), 555–586 (2008)
9. Belkin, M., Niyogi, P.: Convergence of Laplacian eigenmaps. In: Advances in Neural Information Processing Systems, p. 129 (2007)
10. García Trillos, N., Slepčev, D.: A variational approach to the consistency of spectral clustering. Appl. Comput. Harmon. Anal. **45**(2), 239–281 (2018)
11. Hein, M., Audibert, J.-Y., Luxburg, U.: From graphs to manifolds—weak and strong pointwise consistency of graph Laplacians. In: Learning Theory. Lecture Notes in Computer Science, vol. 3559, pp. 470–485. Springer, Berlin (2005)
12. Singer, A.: From graph to manifold Laplacian: The convergence rate. Appl. Comput. Harmon. Anal. **21**(1), 128–134 (2006)
13. García Trillos, N., Gerlach, M., Hein, M., Slepčev, D.: Error estimates for spectral convergence of the graph Laplacian on random geometric graphs towards the Laplace–Beltrami operator. Found. Comput. Math. (2019). https://doi.org/10.1007/s10208-019-09436-w
14. Calder, J., García Trillos, N.: Improved spectral convergence rates for graph Laplacians on ε-graphs and k-nn graphs. Appl. Comput. Harmon. Anal. **60**, 123–175 (2022)
15. Calder, J.: The game theoretic p-Laplacian and semi-supervised learning with few labels. Nonlinearity **32**(1), 301 (2018)
16. Slepčev, D., Thorpe, M.: Analysis of p-Laplacian regularization in semi-supervised learning. SIAM J. Math. Anal. **51**(3), 2085–2120 (2019)
17. Calder, J.: Consistency of Lipschitz learning with infinite unlabeled data and finite labeled data. SIAM J. Math. Data Sci. **1**(4), 780–812 (2019)
18. Dunlop, M.M., Slepčev, D., Stuart, A.M., Thorpe, M.: Large data and zero noise limits of graph-based semi-supervised learning algorithms. Appl. Comput. Harmon. Anal. **49**(2), 655–697 (2020)
19. Ciaurri, Ó., Roncal, L., Stinga, P.R., Torrea, J.L., Varona, J.L.: Fractional discrete Laplacian versus discretized fractional Laplacian. Preprint arXiv:1507.04986 (2015)
20. Bertozzi, A.L., Luo, X., Stuart, A.M., Zygalakis, K.C.: Uncertainty quantification in graph-based classification of high dimensional data. SIAM/ASA J. Uncertain. Quantification **6**(2), 568–595 (2018)
21. García Trillos, N., Sanz-Alonso, D.: Continuum limits of posteriors in graph Bayesian inverse problems. SIAM J. Math. Anal. **50**(4), 4020–4040 (2018)
22. García Trillos, N., Kaplan, Z., Samakhoana, T., Sanz-Alonso, D.: On the consistency of graph-based Bayesian semi-supervised learning and the scalability of sampling algorithms. J. Mach. Learn. Res. **21**(28), 1–47 (2020)
23. Sanz-Alonso, D., Yang, R.: The SPDE approach to Matérn fields: Graph representations. Stat. Sci. **37**(4), 519–540 (2022)
24. Bissantz, N., Hohage, T., Munk, A.: Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. Inverse Probl. **20**(6), 1773–1789 (2004)
25. Bissantz, N., Hohage, T., Munk, A., Ruymgaart, F.: Convergence rates of general regularization methods for statistical inverse problems and applications. SIAM J. Numer. Anal. **45**(6), 2610–2636 (2007)
26. Claeskens, G., Krivobokova, T., Opsomer, J.D.: Asymptotic properties of penalized spline estimators. Biometrika **96**(3), 529–544 (2009)
27. Hall, P., Opsomer, J.D.: Theory for penalised spline regression. Biometrika **92**(1), 105–118 (2005)
28. Kauermann, G., Krivobokova, T., Fahrmeir, L.: Some asymptotic results on generalized penalized spline smoothing. J. R. Stat. Soc. B: Stat. Methodol. **71**(2), 487–503 (2009)
29. Lai, M.-J., Wang, L.: Bivariate penalized splines for regression. Stat. Sin. **23**, 1399–1417 (2013)
30. Lukas, M.A.: Robust generalized cross-validation for choosing the regularization parameter. Inverse Probl. **22**(5), 1883–1902 (2006)
31. Wang, X., Shen, J., Ruppert, D.: On the asymptotics of penalized spline smoothing. Electron. J. Stat. **5**, 1–17 (2011)

32. Arcangeli, R., Ycart, B.: Almost sure convergence of smoothing D^m-splines for noisy data. Numerische Mathematik **66**(1), 281–294 (1993)
33. Li, Y., Ruppert, D.: On the asymptotics of penalized splines. Biometrika **95**(2), 415–436 (2008)
34. Shen, J., Wang, X.: Estimation of monotone functions via P-splines: A constrained dynamical optimization approach. SIAM J. Optim. **49**(2), 646–671 (2011)
35. Xiao, L., Li, Y., Apanasovich, T.V., Ruppert, D.: Local asymptotics of P-splines. Preprint arXiv:1201.0708 (2012)
36. Yoshida, T., Naito, K.: Asymptotics for penalized additive B-spline regression. J. Jpn. Stat. Soc. **42**(1), 81–107 (2012)
37. Yoshida, T., Naito, K.: Asymptotics for penalised splines in generalised additive models. J. Nonparametr. Stat. **26**(2), 269–289 (2014)
38. Stone, C.J.: Optimal global rates of convergence for nonparametric regression. Ann. Stat. **10**(4), 1040–1053 (1982)
39. Thorpe, M., Johansen, A.M.: Pointwise convergence in probability of general smoothing splines. Ann. Inst. Stat. Math. **70**(4), 717–744 (2017)
40. Wahba, G.: A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. Ann. Stat. **13**(4), 1378–1402 (1985)
41. Kimeldorf, G.S., Wahba, G.: A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann. Math. Stat. **41**(2), 495–502 (1970)
42. Cox, D.D.: Approximation of method of regularization estimators. Ann. Stat. **16**(2), 694–712 (1988)
43. Nychka, D.W., Cox, D.D.: Convergence rates for regularized solutions of integral equations from discrete noisy data. Ann. Stat. **17**(2), 556–572 (1989)
44. Carroll, R.J., Van Rooij, A.C.M., Ruymgaart, F.H.: Theoretical aspects of ill-posed problems in statistics. Acta Applicandae Mathematica **24**(2), 113–140 (1991)
45. Mair, B.A., Ruymgaart, F.H.: Statistical inverse estimation in Hilbert scales. SIAM J. Appl. Math. **56**(5), 1424–1444 (1996)
46. Green, A., Balakrishnan, S., Tibshirani, R.J.: Minimax optimal regression over Sobolev spaces via Laplacian eigenmaps on neighborhood graphs. Preprint arXiv:2111.07394 (2021)
47. Cucker, F., Smale, S.: Best choices for regularization parameters in learning theory: On the bias-variance problem. Found. Comput. Math. **2**(4), 413–428 (2002)
48. Caponnetto, A., De Vito, E.: Optimal rates for the regularized least-squares algorithm. Found. Comput. Math. **7**(3), 331–368 (2007)
49. Liang, T., Rakhlin, A.: Just interpolate: Kernel "ridgeless" regression can generalize. Ann. Stat. **48**(3), 1329–1347 (2020)
50. Rakhlin, A., Zhai, X.: Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon. In: Conference on Learning Theory, PMLR, pp. 2595–2623 (2019)
51. Fefferman, C., Klartag, B.: Fitting a C^m-smooth function to data I. Ann. Math. **169**(1), 315–346 (2009)
52. Fefferman, C., Klartag, B.: Fitting a C^m-smooth function to data II. Revista Matemática Iberoamericana **25**(1), 49–273 (2009)
53. Fefferman, C.: Fitting a C^m-smooth function to data III. Ann. Math. **170**(1), 427–441 (2009)
54. Fefferman, C., Israel, A., Luli, G.K.: Fitting a Sobolev function to data I. Revista Matemática Iberoamericana **32**(1), 275–376 (2016)
55. Fefferman, C., Israel, A., Luli, G.K.: Fitting a Sobolev function to data II. Revista Matemática Iberoamericana **32**(2), 649–750 (2016)
56. Fefferman, C., Israel, A., Luli, G.K.: Fitting a Sobolev function to data III. Revista Matemática Iberoamericana **32**(3), 1039–1126 (2016)
57. Tibshirani, R.J., Taylor, J.: The solution path of the generalized lasso. Ann. Stat. **39**(3), 1335–1371 (2011)
58. Hütter, J.-C., Rigollet, P.: Optimal rates for total variation denoising. In: Conference on Learning Theory, pp. 1115–1146 (2016)
59. Sadhanala, V., Wang, Y.-X., Tibshirani, R.J.: Total variation classes beyond 1D: Minimax rates, and the limitations of linear smoothers. In: Advances in Neural Information Processing Systems, pp. 3513–3521 (2016)
60. Sadhanala, V., Wang, Y.-X., Sharpnack, J.L., Tibshirani, R.J.: Higher-order total variation classes on grids: Minimax theory and trend filtering methods. In: Advances in Neural Information Processing Systems, pp. 5800–5810 (2017)

61. Padilla, O.H.M., Sharpnack, J., Scott, J.G., Tibshirani, R.J.: The DFS fused lasso: Linear-time denoising over general graphs. J. Mach. Learn. Res. **18**(176), 1–36 (2018)
62. Padilla, O.H.M., Sharpnack, J., Chen, Y., Witten, D.M.: Adaptive non-parametric regression with the K-NN fused lasso. Preprint arXiv:1807.11641 (2018)
63. García Trillos, N., Murray, R.: A maximum principle argument for the uniform convergence of graph Laplacian regressors. SIAM J. Math. Data Sci. **2**(3), 705–739 (2020)
64. Kpotufe, S.: k-NN regression adapts to local intrinsic dimension. In: Advances in Neural Information Processing Systems, pp. 729–737 (2011)
65. Hafiene, Y., Fadili, J., Elmoataz, A.: Continuum limits of nonlocal p-Laplacian variational problems on graphs. SIAM J. Imaging Sci. **12**(4), 1772–1807 (2019)
66. Hafiene, Y., Fadili, J., Elmoataz, A.: Nonlocal p-Laplacian evolution problems on graphs. SIAM J. Numer. Anal. **56**(2), 1064–1090 (2018)
67. García Trillos, N., Slepčev, D.: Continuum limit of total variation on point clouds. Arch. Ration. Mech. Anal. **220**(1), 193–241 (2016)
68. Szlam, A., Bresson, X.: Total variation, Cheeger cuts. In: Proceedings of the 27th International Conference on Machine Learning, pp. 1039–1046 (2010)
69. García Trillos, N., Murray, R., Thorpe, M.: From graph cuts to isoperimetric inequalities: Convergence rates of Cheeger cuts on data clouds. Arch. Ration. Mech. Anal. **244**(3), 541–598 (2022)
70. Hu, H., Laurent, T., Porter, M., Bertozzi, A.: A method based on total variation for network modularity optimization using the MBO scheme. SIAM J. Appl. Math. **73** (2013). https://doi.org/10.1137/130917387
71. Davis, E., Sethuraman, S.: Consistency of modularity clustering on random geometric graphs. Ann. Appl. Probab. **28**(4), 2003–2062 (2018)
72. Cristoferi, R., Thorpe, M.: Large data limit for a phase transition model with the p-Laplacian on point clouds. Eur. J. Appl. Math. **31**(2), 185–231 (2020)
73. Thorpe, M., Theil, F.: Asymptotic analysis of the Ginzburg–Landau functional on point clouds. Proc. R. Soc. Edinb. Sect. A Math. **149**(2), 387–427 (2019)
74. Gennip, Y., Bertozzi, A.L.: Γ-convergence of graph Ginzburg–Landau functionals. Adv. Differ. Equ. **17**(11–12), 1115–1180 (2012)
75. Lin, Z., Yao, F.: Functional regression on manifold with contamination. Biometrika **108**(1), 167–181 (2021)
76. Thorpe, M., Wang, B.: Robust certification for Laplace learning on geometric graphs. In: Mathematical and Scientific Machine Learning, PMLR, pp. 896–920 (2022)
77. García Trillos, N., Slepčev, D.: On the rate of convergence of empirical measures in ∞-transportation distance. Can. J. Math. **67**(6), 1358–1383 (2015)
78. Caroccia, M., Chambolle, A., Slepčev, D.: Mumford–Shah functionals on graphs and their asymptotics. Nonlinearity **33**(8), 3846 (2020)