PDEs on Graphs

Leon Bungert, Eloi Martinet

Winter semester 2025/26

1 Graphs

Definition 1.1 ((Weighted) graphs)**.**

- A graph is a tuple $G = (V, E)$, consisting of a finite set of vertices V and a set of edges $E \subset \{(x, y) : x, y \in V, x \neq y\}$. G is called undirected if $(x, y) \in E$ holds if and only if $(y, x) \in E$ and directed otherwise.
- A weighted graph is a tuple $G = (V, w)$, consisting of a finite set of vertices V and a function $w : V \times V \to [0, \infty)$. G is called undirected if $w(x, y) = w(y, x)$ for all $x, y \in V$ and directed otherwise.

Remark 1.1. A few remarks are in order:

- A graph is a special case of a weighted graph that arises for weight functions $w : V \times V \to \{0, 1\}$ and by defining $E = \{(x, y) \in V \times V : w(x, y) = 1\}$.
- Sometimes we denote weight functions by w_{xy} instead of $w(x, y)$.
- Some authors define the edge set of unweighted graphs as subset of $\{\{x, y\} : x, y \in V, x \neq y\}$ which has only half as many elements as our tuple-based definition.
- In this definition of graphs, the edges just encode binary relations between vertices but the edges themselves are no relevant objects. In particular, the space of functions on the graph vertices is a finite-dimensional one. This is in stark contrast to so-called metric graphs, where edges are subintervals of $\mathbb{R}$ with specified lengths and one typically considers infinite-dimensional spaces of functions on the union of all these intervals with coupling constraints on the vertices.

Next we define what it means for a graph to be connected.

Definition 1.2 (Connectedness)**.** A graph $G = (V, E)$ or a weighted graph $G = (V, w)$ is called connected if for every $x, y \in V$ there exists a number $k \in \mathbb{N}$ and points $x_1, \dots, x_k \in V$ with $x_1 = x$, $x_k = y$, and for all $i \in \{1, \dots, k-1\}$ it holds $(x_i, x_{i+1}) \in E$ or $w(x_i, x_{i+1}) > 0$, respectively, for weighted graphs.

We continue with some examples of how sets of vertices can be converted into a graph.

Example 1.1 (Fully-connected graph)**.** Let V be an arbitrary finite set and $E = V \times V$. Then $G = (V, E)$ is a graph which is fully-connected, meaning that for all $x, y \in V$ we have $(x, y) \in E$.

Example 1.2 (Erdős–Rényi graph)**.** Let V be an arbitrary finite set of $n \in \mathbb{N}$ elements and $p \in [0, 1]$. By including (x, y) and (y, x) into E independently with probability p, one obtains a so-called $G(n, p)$-graph. Obviously, for $p = 0$ the graph does not have any edge, whereas for $p = 1$ one obtains a fully-connected graph almost surely. Less trivially, if $p > \frac{\log n}{n}$ then a $G(n, p)$-graph is connected almost surely. Erdős–Rényi graphs defined like this are undirected.

If the vertices are subset of a metric space, one can use the metric to construct more sparsely connected graphs.

Example 1.3 (ε-ball graph)**.** Let $V \subset M$ be a finite set of vertices contained in a metric space (M, d), and let $\varepsilon > 0$. By setting $E = \{(x, y) \in V \times V \,:\, d(x, y) < \varepsilon\}$ one obtains a so-called ε-ball graph. For $\varepsilon \to \infty$ this tends to a fully-connected graph. Note that ε-ball graphs are automatically undirected.

Example 1.4 (k-nearest neighbor graph)**.** Let $V \subset M$ be a finite set of vertices contained in a metric space (M, d), and let $k \in \mathbb{N}$. By setting $E = \{(x, y) \in V \times V \,:\, y \text{ is among the } k\text{-nearest neighbors of } x\}$ one obtains a so-called directed k-nearest neighbor graph. Symmetrized versions also exist. For $k \to \infty$ these graphs become fully-connected.

While all these graph constructions require the set of vertices to be given, a common model assumption is that the vertices are in fact random sample from some probability distribution, as outlined in the next example.

Example 1.5 (Random geometric graphs)**.** Let $\Omega \subset \mathbb{R}^d$ be an open set, equipped with a probability measure $\mu \in \mathcal{P}(\Omega)$. Let $V = \{x_i\}_{i=1,\dots,n}$ be *i.i.d.* random samples from μ, meaning that x_i for $i = 1, \dots, n$ are independent random variable with law μ and hence satisfying $\mathbb{P}(x_i \in A) = \mu(A)$ for any Borel subset $A \subset \Omega$. Equipping V with an ε-ball or k-nearest neighbor structure, one obtains a so-called random geometric graph.

We now prove that random geometric ε-ball graphs are connected with high probability if $\varepsilon \geq C \left(\frac{\log n}{n}\right)^{\frac{1}{d}}$. We give the proof for graphs samples from the hypercube but it generalizes to more general domains.

Proposition 1.1 (Connectedness of a random geometric ε-ball graph)**.** *Let $\Omega = [0, 1]^d$ be the hypercube, let $\mu \in \mathcal{P}(\Omega)$ be a probability measure which has the density ρ with respect to the d-dimensional Lebesgue measure, and assume that there exists a constant $c_\rho > 0$ such that $\rho \geq c_\rho$ almost everywhere in Ω.*

Then there exist constants $C_1, C_2 > 0$ depending only on d and c_ρ such that the associated random geometric ε-ball graph $G_{n,\varepsilon}$ is connected with probability at least $1 - C_1 n \exp(-C_2 n \varepsilon^d)$.

Proof. We cover Ω by non-overlapping boxes $\{B_i\}_{i=1,\dots,M}$ where $M = \left\lceil 2^d d^{\frac{d}{2}} \varepsilon^{-d} \right\rceil$ of side length $h \leq \frac{\varepsilon}{2\sqrt{d}}$. The maximal distance between two points in neighboring boxes is at most ε. Hence, if all boxes contain a point in $G_{n,\varepsilon}$, the graph is connected. Conversely, if the graph is not connected (we denote this event by N), there has to be an empty box. Using a union bound and that the graph points are *i.i.d.*, we get

$$\begin{aligned} \mathbb{P}(N) &\leq \mathbb{P}\left(\bigcup_{i=1}^{M} \{B_i \cap G_{n,\varepsilon} = \emptyset\}\right) \\ &\leq \sum_{i=1}^{M} \mathbb{P}\left(B_i \cap G_{n,\varepsilon} = \emptyset\right) = \sum_{i=1}^{M} \mathbb{P}\left(x_1 \notin B_i\right)^n . \end{aligned}$$

It holds $\mathbb{P}(x_1 \notin B_i) = 1 - \int_{B_i} \rho \,\mathrm{d}x \leq 1 - |B_i|\, c_\rho = 1 - C_2 \varepsilon^d$, where C_2 depends on d and c_ρ. There are two cases to consider:

If $n\varepsilon^d \geq 1$ we can use the elementary inequality $1 - t \leq \exp(-t)$ for all $t \in \mathbb{R}$ to get

$$\mathbb{P}(N) \leq M(1 - C_2 \varepsilon^d)^n \leq C_1 \varepsilon^{-d} \exp(-C_2 n \varepsilon^d) \leq C_1 n \exp(-C_2 n \varepsilon^d).$$

If $n\varepsilon^d \leq 1$ we have the trivial estimate

$$\mathbb{P}(N) \leq 1 \leq C_1 n \exp(-C_2) \leq C_1 n \exp(-C_2 n\varepsilon^d)$$

if C_1 is increased to be at least $\exp(C_2)$. $\square$

Exercise 1.1. Use the Borel–Cantelli lemma to show that for $\varepsilon \geq C\left(\frac{\log n}{n}\right)^{\frac{1}{d}}$ with a sufficiently large constant $C > 0$, the graph $G_{n,\varepsilon}$ is connected almost surely as $n \to \infty$.

Remark 1.2. The length scale restriction from Exercise 1.1 is often referred to as the connectivity length scale of a random geometric graph. We shall later work with stronger length scale restrictions which in particular imply the almost sure connectedness of the considered graphs.

Remark 1.3. With a similar argument to the one used in the proof of Proposition 1.1 one can prove that a random geometric k-nearest neighbor graph is connected with high probability if $k \geq C \log n$ for a sufficiently large constant $C > 0$. For this one tessellates the domain into squares of side length of order $h \lesssim \left(\frac{k}{n}\right)^{\frac{1}{d}}$ such that the expected number of graph points in an h-neighborhood of any graph point is of order k.

2 Unsupervised learning

In this section we will study unsupervised learning methods involving graphs. Unsupervised learning refers to the situation where one works with unlabeled data and tries to extract meaningful information from it. As two prototypical examples we will consider clustering, i.e., the task of subdividing data sets into a fixed number of semantically meaningful components, and ranking, where one assigns an importance score to each data point based on its relation to the remaining data.

2.1 Spectral clustering

The general clustering task is to subdivide a data set into a fixed number of components such that the similarity is high between data in the same component and low between data in different components.

The most elementary approach to clustering is the k-means algorithm which measures similarity based on the pairwise Euclidean distance of data points in $\mathbb{R}^d$. We will instead consider spectral clustering–a more sophisticated approach which is able to cluster more complex data sets and is based on the use of graphs.

To set the scene we let $G = (V, w)$ be a weighted and undirected graph. The fact that the graph is undirected will actually be important for the method. We will first introduce some notation: For a vertex $x \in V$ we define its degree by

$$\deg(x) = \sum_{y \in V} w_{xy}.$$

For a subset of the vertices $A \subset V$ we define its volume by

$$\mathrm{vol}(A) = \sum_{x \in A} \deg(x).$$

Finally, we define the perimeter of a subset $A \subset V$ as the sum of all weights that need to be cut for separating A from its complement $A^c = V \setminus A$:

$$\mathrm{Per}(A) = \sum_{\substack{x \in A \\ y \in A^c}} w_{xy}.$$

Note that for an unweighted graph where $w_{xy} \in \{0, 1\}$ this is precisely the number of edges separating A from its complement. Finally, we define the function $\mathbb{1}_A : V \to V$ via

$$\mathbb{1}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \in A^c, \end{cases}$$

and we set $\mathbb{1} = \mathbb{1}_V$.

Next we will start deriving the spectral clustering problem and we will limit ourselves to dividing the graph into two clusters. Later we will comment on generalizations to more clusters.

Clustering the graph into two clusters amounts to finding a subset $A \subset V$ such that the vertices in A form one cluster and the vertices in A^c form the second one. Ideally, we would like the clusters to be interconnected as little as possible. Hence, the naive approach to clustering would be to search for a subset A which solves

$$\min_{A \subset V} \mathrm{Per}(A).$$

However, this problem has the trivial solution $A = \emptyset$. Even if one were to exclude the empty set the solution could be given by just a single vertex with just one adjacent edge.

To enforce a more balanced problem, let us consider the problem of minimizing the so-called normalized cut

$$\min_{A \subset V} \underbrace{\frac{\mathrm{Per}(A)}{\mathrm{vol}(A)} + \frac{\mathrm{Per}(A^c)}{\mathrm{vol}(A^c)}}_{=\mathrm{NCut(A)}}. \tag{2.1}$$

This new objective function enforces that neither A nor A^c are too small since otherwise this objective function attains large values.

Remark 2.1. Note that using $\mathrm{Per}(A^c) = \mathrm{Per}(A)$ and $\mathrm{vol}(A) + \mathrm{vol}(A^c) = \mathrm{vol}(V)$ we have

$$\mathrm{NCut}(A) = \frac{\mathrm{Per}(A)}{\mathrm{vol}(A)} + \frac{\mathrm{Per}(A^c)}{\mathrm{vol}(A^c)} = \mathrm{Per}(A)\frac{\mathrm{vol}(A^c) + \mathrm{vol}(A)}{\mathrm{vol}(A)\,\mathrm{vol}(A^c)} = \mathrm{Per}(A)\frac{\mathrm{vol}(V)}{\mathrm{vol}(A)\,\mathrm{vol}(A^c)}$$

and hence problem (2.1) is equivalent to minimizing

$$\min_{A \subset V} \frac{\mathrm{Per}(A)}{\mathrm{vol}(A)\,\mathrm{vol}(A^c)}.$$

Remark 2.2. Without loss of generality assume $\mathrm{vol}(A) \leq \mathrm{vol}(A^c)$. Using that

$$\mathrm{Per}(A^c) = \mathrm{Per}(A) = \sum_{\substack{x \in A \\ y \in A^c}} w_{xy} \leq \sum_{\substack{x \in A \\ y \in V}} w_{xy} = \sum_{x \in A} \deg(x) = \mathrm{vol}(A)$$

we get that

$$\mathrm{NCut}(A) \leq 1 + \frac{\mathrm{vol}(A)}{\mathrm{vol}(A^c)} \leq 2 \quad \forall A \subset V.$$

2.1.1 Relaxation

Note that (2.1) is a challenging combinatorial optimization problem, meaning that to arrive to a solution one would have to try all possible subsets of the vertex set V. If the latter has n elements these are 2^n different choices. Already for $n = 100$ data points these are about 10^{30} many possibilities.

Hence, in order to obtain a feasible algorithm, we have to relax the optimization problem. To do this, let us associate with a subset $A \subset V$ the following function $u_A : V \to \mathbb{R}$, defined as

$$u_A(x) = \begin{cases} \sqrt{\dfrac{\mathrm{vol}(A^c)}{\mathrm{vol}(A)}}, & x \in A, \\ -\sqrt{\dfrac{\mathrm{vol}(A)}{\mathrm{vol}(A^c)}}, & x \in A^c. \end{cases} \tag{2.2}$$

By definition, we have

$$|u_A(x) - u_A(y)|^2 = \begin{cases} 0, & x, y \in A, \\ 0, & x, y \in A^c, \\ \dfrac{\mathrm{vol}(A^c)}{\mathrm{vol}(A)} + \dfrac{\mathrm{vol}(A)}{\mathrm{vol}(A^c)} + 2, & x \in A,\ y \in A^c \text{ or } x \in A^c,\ y \in A. \end{cases}$$

We make a few observations.

Quadratic form Our first observation is that $\mathrm{NCut}(A)$ can be rewritten as a quadratic form involving u_A as follows:

$$\begin{aligned} \frac{1}{2} \sum_{x,y \in V} w_{xy} |u_A(x) - u_A(y)|^2 &= \sum_{\substack{x \in A \\ y \in A^c}} w_{x,y} \left(\frac{\mathrm{vol}(A^c)}{\mathrm{vol}(A)} + \frac{\mathrm{vol}(A)}{\mathrm{vol}(A^c)} + 2 \right) \\ &= \mathrm{Per}(A) \left(\frac{\mathrm{vol}(A^c) + \mathrm{vol}(A)}{\mathrm{vol}(A)} + \frac{\mathrm{vol}(A) + \mathrm{vol}(A^c)}{\mathrm{vol}(A^c)} \right) \\ &= \mathrm{vol}(V) \left(\frac{\mathrm{Per}(A)}{\mathrm{vol}(A)} + \frac{\mathrm{Per}(A^c)}{\mathrm{vol}(A^c)} \right) \\ &= \mathrm{vol}(V)\, \mathrm{NCut}(A). \end{aligned}$$

Normalization We observe that

$$\sum_{x \in V} \left(\sqrt{\deg(x)} u_A(x) \right)^2 = \frac{\mathrm{vol}(A^c)}{\mathrm{vol}(A)} \sum_{x \in A} \deg(x) + \frac{\mathrm{vol}(A)}{\mathrm{vol}(A^c)} \sum_{x \in A^c} \deg(x) = \mathrm{vol}(A^c) + \mathrm{vol}(A) = \mathrm{vol}(V).$$

Orthogonality Finally, we observe the following orthogonality:

$$\begin{aligned} \sum_{x \in V} \left(\sqrt{\deg(x)} u_A(x) \right) \left(\sqrt{\deg(x)} \mathbb{1}(x) \right) &= \sqrt{\frac{\mathrm{vol}(A^c)}{\mathrm{vol}(A)}} \sum_{x \in A} \deg(x) - \sqrt{\frac{\mathrm{vol}(A)}{\mathrm{vol}(A^c)}} \sum_{x \in A^c} \deg(x) \\ &= \sqrt{\mathrm{vol}(A^c)\, \mathrm{vol}(A)} - \sqrt{\mathrm{vol}(A)\, \mathrm{vol}(A^c)} = 0. \end{aligned}$$

Hilbert space structure We equip the (finite-dimensional!) vector space $\ell^2(V) = \{u : V \to \mathbb{R}\}$ with a Hilbert space structure, by defining an inner product

$$\langle u, v\rangle = \sum_{x\in V} u(x)v(x).$$

Note that $\ell^2(V)$ is isometric to $\mathbb{R}^n$ equipped with the Euclidean inner product where $n = \#V$.

Let us define the linear operators $D, L : \ell^2(V) \to \ell^2(V)$ via

$$Du(x) = \deg(x)u(x),$$
$$Lu(x) = \sum_{y\in V} w_{xy}(u(y) - u(x)).$$

D is called the degree operator and L the graph Laplacian. Note that D has a natural square root $D^{1/2} : \mathcal{H} \to \mathcal{H}$, given by

$$D^{1/2}u(x) = \sqrt{\deg(x)}u(x).$$

Exercise 2.1. Prove that

$$\langle -Lu, u\rangle = \frac{1}{2}\sum_{x,y\in V} w_{xy} \left|u(x) - u(y)\right|^2 \tag{2.3}$$

holds for all $u \in \ell^2(V)$.

Exercise 2.2. Prove that D and L are self-adjoint, meaning that $\langle Du, v\rangle = \langle u, Dv\rangle$ and $\langle Lu, v\rangle = \langle u, Lv\rangle$ holds for all $u, v \in \ell^2(V)$. Prove also that $-L$ is positive semi-definite.

Relaxation Using the above identification and Equation (2.3) we can equivalently rewrite (2.1) as

$$\min\left\{\langle -Lu_A, u_A\rangle \,:\, A \subset V,\ \left\|D^{1/2}u_A\right\|^2 = \mathrm{vol}(V),\ \langle D^{1/2}u_A, D^{1/2}\mathbb{1}\rangle = 0\right\}.$$

Without loss of generality we can assume that $\deg(x) > 0$ for all $x \in V$. Otherwise there would be an isolated vertex without any neighbor which can just be removed from the graph. Let us make the substitution $v_A = D^{1/2}u_A$ to obtain using Exercise 2.2

$$\min\left\{\langle -L_{sym}v_A, v_A\rangle \,:\, A \subset V,\ \|v_A\|^2 = \mathrm{vol}(V),\ \left\langle v_A, D^{1/2}\mathbb{1}\right\rangle = 0\right\}, \tag{2.4}$$

where the normalized graph Laplacian $L_{sym} = D^{-1/2}LD^{-1/2}$ is given by

$$L_{sym}u(x) = \frac{1}{\sqrt{\deg(x)}}\sum_{y\in V} w_{xy}\frac{u(y)}{\sqrt{\deg(y)}} - u(x).$$

Note that since $\deg(x) > 0$ by assumption, we get that $D^{-1/2}$ is invertible and has the inverse $D^{-1/2} : \ell^2(V) \to \ell^2(V)$, defined via $D^{-1/2}u(x) = \frac{u(x)}{\sqrt{\deg(x)}}$ for $x \in V$.

Exercise 2.3. Prove this formula for the normalized graph Laplacian.

Remark 2.3. The normalized graph Laplacian L_{sym} has the subscript which stands for symmetric. This is to set it apart from another normalized graph Laplacian, the so-called random walk graph Laplacian, which is not symmetric or self-adjoint.

We have now equivalently rewritten the clustering problem (2.1) as (2.4) which is still a combinatorial optimization problem. However, we can relax it by dropping the assumption that $v_A = D^{1/2}u_A$ where u_A is given by (2.2). Instead, we optimize over arbitrary functions $v \in \ell^2(V)$ which leads to

$$\min\left\{\langle -L_{sym}v, v\rangle \,:\, v \in \ell^2(V),\ \|v\|^2 = \mathrm{vol}(V),\ \left\langle v, D^{1/2}\mathbb{1}\right\rangle = 0\right\}, \tag{2.5}$$

where it holds that the infimal value in (2.6) is smaller or equal than the infimal value in (2.4). Finally, we note that problem (2.5) is equivalent to

$$\min\left\{\frac{\langle -L_{sym}v, v\rangle}{\|v\|^2} \,:\, v \in \ell^2(V)\setminus\{0\},\ \left\langle v, D^{1/2}\mathbb{1}\right\rangle = 0\right\}, \tag{2.6}$$

in the sense that any solution of (2.5) solves (2.6) and, vice versa, any solution of (2.6), when normalized such that its square norm equals $\mathrm{vol}(V)$, solves (2.5). We shall see that the solutions of (2.6) are exactly the eigenvectors of $-L_{sym}$ corresponding to its second eigenvalue. Note that, analogously to Equation (2.3), also the normalized graph Laplacian encodes a quadratic form.

Exercise 2.4. Prove that

$$\langle -L_{sym}u, u\rangle = \frac{1}{2}\sum_{x,y\in V} w_{xy}\left|\frac{u(x)}{\sqrt{\deg(x)}} - \frac{u(y)}{\sqrt{\deg(y)}}\right|^2$$

holds for all $u \in \ell^2(V)$.

Proposition 2.1 (Connected components)**.** *The negative normalizes Laplacian $-L_{sym}$ has non-negative eigenvalues. Furthermore, the dimension $M \in \mathbb{N}$ of the eigenspace corresponding to the eigenvalue $\lambda_1 = 0$ equals the number of connected components $\{V_i\}_{i=1,\dots,M}$ of the graph and the eigenspace is spanned by eigenvectors of the form $v_1 = D^{1/2}\sum_{i=1}^M c_i \mathbb{1}_{V_i}$, where $c_i \in \mathbb{R}$ for $i = 1,\dots,M$, meaning that $-L_{sym}v_1 = 0$.*

Proof. The result is a simple consequence of (2.4). □

Proposition 2.2 (Solution of spectral clustering)**.** *If G is connected, then the minimum in* (2.6) *is attained by any eigenvector v_2 of $-L_{sym}$ associated to its second eigenvalue $\lambda_2 > 0$, meaning that $-L_{sym}v_2 = \lambda_2 v_2$.*

Proof. Since $-L_{sym}$ is a self-adjoint operator on a finite-dimensional space, the standard spectral theorem from linear algebra implies the existence of an orthonormal basis of eigenvectors $\{v_i\}_{i=1}^n$ where $n = \#V$. We assume these eigenvectors correspond to ordered eigenvalues $0 = \lambda_1 < \lambda_2 \leq \dots \leq \lambda_n$. Hence, any $v \in \ell^2(V)\setminus\{0\}$ can be written as

$$v = \sum_{i=1}^n \alpha_i v_i,$$

where $\alpha_i = \langle v, v_i \rangle$. Since G is connected we know by Proposition 2.1 that the eigenspace corresponding to the eigenvalue $\lambda_1 = 0$ is spanned by the function $D^{1/2}\mathbb{1}$. Hence, any $v \in \ell^2(V) \setminus \{0\}$ with $\langle v, D^{1/2}\mathbb{1} \rangle = 0$ can be written as

$$v = \sum_{i=2}^{n} \alpha_i v_i \quad \text{where} \quad \sum_{i=2}^{n} \alpha_i^2 = \|v\|^2 > 0.$$

Using this decomposition, we have

$$\frac{\langle -L_{sym} v, v \rangle}{\|v\|^2} = \frac{\left\langle \sum_{i=2}^{n} \lambda_i \alpha_i v_i, \sum_{j=2}^{n} \alpha_j v_j \right\rangle}{\sum_{i=2}^{n} \alpha_i^2} = \frac{\sum_{i=2}^{n} \lambda_i \alpha_i^2}{\sum_{i=2}^{n} \alpha_i^2}.$$

The right hand side is a convex combination of the eigenvalues $\{\lambda_i\}_{i=2,\dots,n}$ and which is minimized for $\alpha_2 > 0$ and $\alpha_i = 0$ for $i \geq 3$. Hence, a solution of (2.6) is given by v_2. Furthermore, since $\frac{\langle -L_{sym} v, v \rangle}{\|v\|^2}$ has the same value for every other choice of non-zero eigenvector in the same eigenspace, and since the whole eigenspace is orthogonal to $D^{1/2}\mathbb{1}$, we can conclude the proof. □

Remark 2.4. In practice, one typically works with solutions of the generalized eigenvalue problem $-Lu = \lambda Du$ which is equivalent to the one of the symmetric graph Laplacian L_{sym}. To see this, we perform the re-substitution $u = D^{-1/2} v$ where v solves $-L_{sym} v = \lambda v$. Then we have

$$-Lu = -LD^{-1/2} v = -D^{1/2} L_{sym} v = -\lambda D^{1/2} v = -\lambda Du.$$

Furthermore, the fact that eigenvectors of $-L_{sym}$ corresponding to different eigenvalues are orthogonal, i.e., $\langle v, \tilde{v} \rangle = 0$, translates to

$$0 = \langle v, \tilde{v} \rangle = \langle D^{1/2} u, D^{1/2} \tilde{u} \rangle = \langle u, D\tilde{u} \rangle.$$

Hence, for clustering of connected graphs we solve

$$-Lu = \lambda_2 Du \quad \text{where} \quad \langle u, D\mathbb{1} \rangle = 0.$$

Sometimes this is also written as

$$-L_{rw} u = \lambda_2 u \quad \text{where} \quad \langle u, D\mathbb{1} \rangle = 0.$$

Here $L_{rw} = D^{-1} L$ is the so-called random walk graph Laplacian which is given by

$$L_{rw} u(x) = \frac{1}{\deg(x)} \sum_{y \in V} w_{xy} u(y) - u(x)$$

and is not self-adjoint.

2.1.2 The Cheeger inequality

It is obvious that we cannot just relax the problem (2.4) into (2.6) without loosing anything. A natural question is hence whether we can quantify how much the minimal values of (2.1) and (2.6) differ. The answer to this question is provided by the Cheeger inequality.

Actually, the Cheeger inequality works with the conductance instead of the normalized cut of the graph. The conductance of a set $A \subset V$ is defined as

$$\phi(A) = \frac{\mathrm{Per}(A)}{\min(\mathrm{vol}(A), \mathrm{vol}(A^c))}$$

and the Cheeger constant of the graph is defined as

$$\mathrm{Cheeg}(G) = \min_{A \subset V} \phi(A).$$

Remark 2.5. It is easy to check that $\phi(A) \leq 1$ for all $A \subset V$ and hence $\mathrm{Cheeg}(G) \leq 1$.

In fact, the conductance of a set is in a way equivalent to its normalized cut in the following sense.

Lemma 2.1. *For every subset $A \subset V$ it holds*

$$\phi(A) \leq \mathrm{NCut}(A) \leq 2\phi(A).$$

Exercise 2.5. Prove Lemma 2.1.

The Cheeger inequality now relates the Cheeger constant with the second eigenvalue of the normalized Laplacian. Thanks to Lemma 2.1 this implies the same relation up to constants for the normalized cut.

Theorem 2.1 (Cheeger's inequality)**.** *Let $G = (V, w)$ be a graph with $\deg(x) > 0$ for all $x \in V$ and let $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ denote the eigenvalues of $-L_{sym}$. Then it holds*

$$\frac{\lambda_2}{2} \leq \mathrm{Cheeg}(G) \leq \sqrt{2\lambda_2}.$$

Proof. We can assume that G is connected, otherwise the result is trivial since $\lambda_2 = 0 = \mathrm{Cheeg}(G)$.

The first inequality is easy to show and the **proof idea** is to take a set A which attains the Cheeger constant, i.e, $\mathrm{Cheeg}(G) = \frac{\mathrm{Per}(A)}{\min(\mathrm{vol}(A),\mathrm{vol}(A^c))}$, and show that the indicator function of that set (suitably centered) has a Rayleigh quotient bounded by two times the conductance of A.

We plan to apply Proposition 2.2 which states that λ_2 is given by the minimal value of the optimization problem (2.6). First note, that by making the substitution $v = D^{1/2}u$ we see that (2.6) is equivalent to

$$\min\left\{\frac{\langle -Lu, u\rangle}{\left\|D^{1/2}u\right\|^2} \,:\, u \in \ell^2(V) \setminus \{0\},\ \langle u, D\mathbb{1}\rangle = 0\right\}, \tag{2.7}$$

Without loss of generality we can assume $\mathrm{vol}(A) \leq \mathrm{vol}(A^c)$ and define the function

$$v(x) = \mathbb{1}_A(x) - \frac{\mathrm{vol}(A)}{\mathrm{vol}(V)} = \begin{cases} 1 - \dfrac{\mathrm{vol}(A)}{\mathrm{vol}(V)}, & x \in A, \\ -\dfrac{\mathrm{vol}(A)}{\mathrm{vol}(V)}, & x \in A^c. \end{cases}$$

First note that it holds

$$\langle v, D\mathbb{1}\rangle = \sum_{x\in V} d(x)\left(\mathbb{1}_A(x) - \frac{\mathrm{vol}(A)}{\mathrm{vol}(V)}\right) = \mathrm{vol}(A) - \mathrm{vol}(A) = 0,$$

so v is feasible for (2.7) and we have

$$\lambda_2 \leq \frac{\langle -Lv, v\rangle}{\left\|D^{1/2}v\right\|^2}.$$

Hence, we have to upper-bound this quotient. For the numerator it holds

$$\langle -Lv, v\rangle = \frac{1}{2}\sum_{x,y\in V} w_{xy}\left|v(x)-v(y)\right|^2 = \sum_{x\in A}\sum_{y\in A^c} w_{xy} = \mathrm{Per}(A).$$

For the denominator it holds

$$\begin{aligned}
\left\|D^{1/2}u\right\|^2 &= \sum_{x\in V} d(x)\left(\mathbb{1}_A(x) - 2\frac{\mathrm{vol}(A)}{\mathrm{vol}(V)}\mathbb{1}_A(x) + \left(\frac{\mathrm{vol}(A)}{\mathrm{vol}(V)}\right)^2\right) \\
&= \mathrm{vol}(A) - \frac{\mathrm{vol}(A)^2}{\mathrm{vol}(V)} \\
&= \mathrm{vol}(A)\left(1 - \frac{\mathrm{vol}(A)}{\mathrm{vol}(V)}\right) \geq \frac{\mathrm{vol}(A)}{2}
\end{aligned}$$

using that $\mathrm{vol}(A) \leq \frac{\mathrm{vol}(V)}{2}$. Combining these three estimates we can conclude the proof of the first inequality:

$$\lambda_2 \leq 2\frac{\mathrm{Per}(A)}{\mathrm{vol}(A)} = 2\,\mathrm{Cheeg}(G).$$

The second inequality is much harder to prove but **proof idea** is very intuitive: We let u be a minimizer of (2.7), meaning that

$$\lambda_2 = \frac{\langle -Lu, u\rangle}{\left\|D^{1/2}u\right\|^2} \quad \text{and} \quad \langle u, D\mathbb{1}\rangle = 0. \tag{2.8}$$

From this eigenvector u we aim to construct a subset A which satisfies $\phi(A) \leq \sqrt{2\lambda_2}$ which would prove the second inequality. For this we consider all level set of the function u and take the one with minimal conductance.

Let us enumerate the vertices $V = \{x_i\}_{i=1}^n$ such that

$$u(x_1) \geq u(x_2) \geq \ldots u(x_n).$$

We define the subsets $A_0 = \emptyset$ and $A_i = \{x_1, \ldots, x_i\} \subset V$ for $i = 1, \ldots, n$, and we define

$$\alpha = \min_{i=1}^{n} \phi(A_i)$$

which obviously satisfies $\alpha \geq \mathrm{Cheeg}(G)$. So our goal is to prove that $\frac{\alpha^2}{2} \leq \lambda_2$. Let $r \in \{1, \ldots, n\}$ denote the largest index such that $\mathrm{vol}(A_r) \leq \mathrm{vol}(V)/2$ (without loss of generality we can assume that $\operatorname{vol}(A_1) \leq \operatorname{vol}(V)/2$ and consider the function $v = u - u(x_r) \in \ell^2(V)$. We aim to estimate λ_2 from below and according to (2.8) this involves estimating the denominator $\left\|D^{1/2}u\right\|^2$ from above. Note that we have

$$\begin{aligned}
\left\|D^{1/2}v\right\|^2 &= \sum_{x\in V} \deg(x)(u(x) - u(x_r))^2 \\
&= \sum_{x\in V} \deg(x)u(x)^2 - 2u(x_r) \underbrace{\sum_{x\in V} \deg(x)u(x)}_{=\langle u, D\mathbb{1}\rangle = 0} + \underbrace{\operatorname{vol}(V)u(x_r)^2}_{\geq 0} \\
&\geq \sum_{x\in V} \deg(x)u(x)^2 = \left\|D^{1/2}u\right\|^2.
\end{aligned}$$

Next we introduce the notation $v = v^+ - v^-$ where $v^\pm(x) = \max(\pm v(x), 0)$ for $x \in V$ and analogously $a^\pm = \max(\pm a, 0)$ for real numbers $a \in \mathbb{R}$. Using (2.3) and (2.8) we have

$$\begin{aligned}
\lambda_2 &= \frac{\langle -Lu, u\rangle}{\left\|D^{1/2}u\right\|^2} \geq \frac{\langle -Lu, u\rangle}{\left\|D^{1/2}v\right\|^2} \\
&= \frac{1}{2}\frac{\sum_{x,y\in V} w_{xy}\left|u(x) - u(y)\right|^2}{\sum_{x\in V} \deg(x)v(x)^2} \\
&= \frac{1}{2}\frac{\sum_{x,y\in V} w_{xy}\left|v(x) - v(y)\right|^2}{\sum_{x\in V} \deg(x)v(x)^2} \\
&\geq \frac{1}{2}\frac{\sum_{x,y\in V} w_{xy}\left(\left|v^+(x) - v^+(y)\right|^2 + \left|v^-(x) - v^-(y)\right|^2\right)}{\sum_{x\in V} \deg(x)\left(v^+(x)^2 + v^-(x)^2\right)}.
\end{aligned}$$

In the last inequality we used that for any $a, b \in \mathbb{R}$ it holds $a^2 = (a^+)^2 + (a^-)^2$ as well as

$$\begin{aligned}
(a-b)^2 &= (a^+ - a^- - b^+ + b^-)^2 = (a^+ - b^+ - (a^- b^-))^2 \\
&= (a^+ - b^+)^2 + (a^- - b^-)^2 - 2(a^+ - b^+)(a^- - b^-) \\
&= (a^+ - b^+)^2 + (a^- - b^-)^2 - 2\left(\underbrace{a^+a^-}_{=0} + \underbrace{b^+b^-}_{=0} - a^+b^- - a^-b^+\right) \\
&= (a^+ - b^+)^2 + (a^- - b^-)^2 + 2a^+b^- + 2a^-b^+ \\
&\geq (a^+ - b^+)^2 + (a^- - b^-)^2.
\end{aligned}$$

Next we use the following elementary inequalities

$$\begin{aligned}
\frac{a+b}{c+d} &\geq \min\left\{\frac{a}{c}, \frac{b}{d}\right\}, \quad a, b \geq 0, c, d > 0, \\
(a+b)^2 &\leq 2(a^2 + b^2), \quad a, b \in \mathbb{R}.
\end{aligned}$$

Assuming without loss of generality that the minimum is attained by the first quotient we get

$$
\begin{aligned}
\lambda_2 &\geq \frac{1}{2} \frac{\sum_{x,y\in V} w_{xy} \left|v^+(x) - v^+(y)\right|^2}{\sum_{x\in V} \deg(x) v^+(x)^2} \\
&= \frac{1}{2} \frac{\left[\sum_{x,y\in V} w_{xy} \left|v^+(x) - v^+(y)\right|^2\right] \left[\sum_{x,y\in V} w_{xy} \left|v^+(x) + v^+(y)\right|^2\right]}{\left[\sum_{x\in V} \deg(x) v^+(x)^2\right] \left[\sum_{x,y\in V} w_{xy} \left|v^+(x) + v^+(y)\right|^2\right]} = \frac{1}{2} \frac{N}{D}.
\end{aligned}
$$

We will estimate both parts of the quotient separately, beginning with the denominator D. Using the second elementary inequality and the symmetry of the weights we get

$$
\begin{aligned}
D &= \left[\sum_{x\in V} \deg(x) v^+(x)^2\right] \left[\sum_{x,y\in V} w_{xy} \left|v^+(x) + v^+(y)\right|^2\right] \\
&\leq 2 \left[\sum_{x\in V} \deg(x) v^+(x)^2\right] \left[\sum_{x,y\in V} w_{xy} \left(v^+(x)^2 + v^+(y)^2\right)\right] = 4 \left[\sum_{x\in V} \deg(x) v^+(x)^2\right]^2.
\end{aligned}
$$

We continue with a monster estimate of the numerator N for which we use the Cauchy–Schwarz inequality for sums, a few index shifts, the definition of α, the fact that $v^+(x_k) = 0$ for $k \geq r$, and the fact that $\mathrm{vol}(A_{k+1}) - \mathrm{vol}(A_k) = \sum_{i=1}^{k+1} \deg(x_i) - \sum_{i=1}^{k} \deg(x_i) = \deg(x_{k+1})$

to get

$$
\begin{aligned}
N &= \left[\sum_{x,y\in V} w_{xy}\left|v^+(x)-v^+(y)\right|^2\right]\left[\sum_{x,y\in V} w_{xy}\left|v^+(x)+v^+(y)\right|^2\right] \\
&= 4\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_{x_ix_j}\left|v^+(x_i)-v^+(x_j)\right|^2\right]\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_{x_ix_j}\left|v^+(x_i)+v^+(x_j)\right|^2\right] \\
&\geq 4\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_{x_ix_j}\left(v^+(x_i)^2-v^+(x_j)^2\right)\right]^2 \\
&= 4\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_{x_ix_j}\sum_{k=i}^{j-1}\left(v^+(x_k)^2-v^+(x_{k+1})^2\right)\right]^2 \\
&= 4\left[\sum_{i=1}^{n-1}\sum_{k=i}^{n-1}\sum_{j=k+1}^{n} w_{x_ix_j}\left(v^+(x_k)^2-v^+(x_{k+1})^2\right)\right]^2 \\
&= 4\left[\sum_{k=1}^{n-1}\sum_{i=1}^{k}\sum_{j=k+1}^{n} w_{x_ix_j}\left(v^+(x_k)^2-v^+(x_{k+1})^2\right)\right]^2 \\
&= 4\left[\sum_{k=1}^{n-1}\operatorname{Per}(A_k)\left(v^+(x_k)^2-v^+(x_{k+1})^2\right)\right]^2 \\
&\geq 4\alpha^2\left[\sum_{k=1}^{n-1}\min(\operatorname{vol}(A_k),\operatorname{vol}(A_k^c))\left(v^+(x_k)^2-v^+(x_{k+1})^2\right)\right]^2 \\
&= 4\alpha^2\left[\sum_{k=1}^{r-1}\operatorname{vol}(A_k)\left(v^+(x_k)^2-v^+(x_{k+1})^2\right)\right]^2 \\
&= 4\alpha^2\left[\sum_{k=0}^{r-2}\operatorname{vol}(A_{k+1})v^+(x_{k+1})^2-\sum_{k=1}^{r-1}\operatorname{vol}(A_k)v^+(x_{k+1})^2\right]^2 \\
&= 4\alpha^2\left[\sum_{k=0}^{r-2}\left(\operatorname{vol}(A_{k+1})-\operatorname{vol}(A_k)\right)v^+(x_{k+1})^2\right]^2 \\
&= 4\alpha^2\left[\sum_{k=0}^{r-2}\deg(x_{k+1})v^+(x_{k+1})^2\right]^2 \\
&= 4\alpha^2\left[\sum_{x\in V}\deg(x)v^+(x)^2\right]^2.
\end{aligned}
$$

Combining the estimates for D and N, we obtain

$$
\lambda_2 \geq \frac{1}{2}\frac{N}{D} \geq \frac{\alpha^2}{2} \geq \frac{\operatorname{Cheeg}(G)^2}{2}
$$

as desired. This concludes the proof. □

Corollary 2.1. *Under the conditions of Theorem 2.1 it holds*

$$\frac{\lambda_2}{2} \leq \min_{A\subset V} \mathrm{NCut}(A) \leq \sqrt{8\lambda_2}.$$

Proof. The result is a combination of Lemma 2.1 and Theorem 2.1. □

Remark 2.6. Note that the Cheeger inequality gives some a-priori bounds for the Cheeger constant $\mathrm{Cheeg}(G)$ and the second eigenvalue λ_2, namely

$$\lambda_2 \leq 8 \quad \text{and hence} \quad \mathrm{Cheeg}(G) \leq 4.$$

These are not sharp since by definition we have $\mathrm{Cheeg}(G) \leq 1$ and hence $\lambda_2 \leq 2$ which is a sharp upper bound.

2.1.3 A random walk perspective

In this section we will briefly discuss an interpretation of graph clustering using random walks. A random walk $(X_k)_{k\in\mathbb{N}_0} \subset V^{\mathbb{N}}$ on a graph, starting at some $X_0 \in V$ is a Markov chain defined through the transition probabilities

$$\mathbb{P}(X_k = y \mid X_{k-1} = x) = \frac{w_{xy}}{\sum_{y\in V} w_{xy}} = \frac{w_{xy}}{\deg(x)}.$$

Let $u \in \ell^2(V)$ be a function on the graph. Then we obtain that

$$\begin{aligned}\mathbb{E}(u(X_k) \mid X_{k-1} = x) &= \sum_{y\in V} \mathbb{P}(X_k = y \mid X_{k-1} = x)u(y) = \frac{1}{\deg(x)}\sum_{y\in V} w_{xy}u(y)\\ &= L_{rw}u(x) + u(x).\end{aligned}$$

If u is an eigenvector of the random walk Laplacian L_{rw} we have $-L_{rw}u = \lambda u$ and hence

$$\mathbb{E}(u(X_k) \mid X_{k-1} = x) = (1-\lambda)u(x).$$

Using the tower formula for conditional expectations we can iterate this an obtain

$$\begin{aligned}\mathbb{E}(u(X_k) \mid X_{k-2} = x) &= \mathbb{E}\big(\mathbb{E}(u(X_k) \mid X_{k-1}) \mid X_{k-2} = x\big) = \mathbb{E}\big((1-\lambda)u(X_{k-1}) \mid X_{k-2} = x\big)\\ &= (1-\lambda)\mathbb{E}(u(X_{k-1}) \mid X_{k-2} = x) = (1-\lambda)^2 u(x)\end{aligned}$$

and recursively

$$\mathbb{E}(u(X_k) \mid X_0 = x) = (1-\lambda)^k u(x) \tag{2.9}$$

If $\lambda = 0$ this formula is not very interesting and the corresponding eigenvector is just $u = \mathbb{1}$. It becomes more interesting if we assume that $\lambda = \lambda_2 \in (0, 2)$ and $u = u_2$ the corresponding eigenvector. Let us identify the clusters as $A = \{u_2(x) > 0\}$ and $A^c = \{u_2(x) \leq 0\}$. Then we can interpret (2.9) as follows: First of all we note that as $k \to \infty$ we have $\mathbb{E}(u(X_k) \mid X_0 = x) \to 0$ which can be interpreted as a “mixing” behavior of the random walk. Second, the smaller λ_2 is the more likely is the random walk to remain within the cluster it started in, larger it is, the faster does the mixing of the random walk take place.

There is also a nice relation between one step of a random walk and the normalized cut NCut. To see this, let us fix a set $A \subset V$ and define

$$\mathbb{P}(A \mid A^c) = \mathbb{P}(X_1 \in A \mid X_0 \in A^c).$$

Proposition 2.3. *Let X_0 be drawn according to the stationary distribution $\pi = \frac{\deg}{\operatorname{vol}(V)}$. Then it holds*

$$\operatorname{NCut}(A) = P(A \mid A^c) + P(A^c \mid A).$$

Proof. For two arbitrary subsets $A, B \subset V$ we have by definition of the conditional probability:

$$\mathbb{P}(X_1 \in B \mid X_0 \in A) = \frac{P(X_1 \in B,\, X_0 \in A)}{\mathbb{P}(X_0 \in A)}.$$

The numerator is given by

$$\begin{aligned}
\mathbb{P}(X_1 \in B,\, X_0 \in A) &= \sum_{\substack{x \in A \\ y \in B}} \mathbb{P}(X_0 = x,\, X_1 = y) = \sum_{\substack{x \in A \\ y \in B}} \mathbb{P}(X_0 = x)\mathbb{P}(X_1 = y \mid X_0 = x) \\
&= \sum_{\substack{x \in A \\ y \in B}} \pi(x) \frac{w_{xy}}{\deg(x)} = \frac{\sum_{\substack{x \in A \\ y \in B}} w_{xy}}{\operatorname{vol}(V)}.
\end{aligned}$$

Similarly, we compute

$$\mathbb{P}(X_0 \in A) = \sum_{x \in A} \mathbb{P}(X_0 = x) = \sum_{x \in A} \pi(x) = \frac{\operatorname{vol}(A)}{\operatorname{vol}(V)}$$

and hence we have

$$\mathbb{P}(X_1 \in B \mid X_0 \in A) = \frac{\sum_{\substack{x \in A \\ y \in B}} w_{xy}}{\operatorname{vol}(A)}.$$

Using this for $B = A^c$ and vice versa yields the result. $\square$

The interpretation of this proposition is clear: The smaller the normalized cut of A, the smaller is the probability that a random walk starting in A jumps to A^c in one step (or vice versa).

2.1.4 Clustering with more than two clusters

Here we describe the brief idea how clustering with multiple clustering works. In this case we are looking for a disjoint partition $A_1, \dots, A_k \subset V$ which minimizes the normalized cut

$$\operatorname{NCut}(\{A_i\}_{i=1}^k) = \sum_{i=1}^k \frac{\operatorname{Per}(A_i)}{\operatorname{vol}(A_i)}.$$

Similarly as in the case $k = 2$, one can rewrite the normalized cut in terms of suitable functions $u_{A_i} = \operatorname{vol}(A_i)^{-1/2} \mathbb{1}_{A_i}$ for $i = 1, \dots, k$. It holds

$$\min_{A_1, \dots, A_k} \operatorname{NCut}(\{A_i\}_{i=1}^k) = \min \left\{ \sum_{i=1}^k \langle -L u_{A_i}, u_{A_i} \rangle \,:\, A_1, \dots, A_k \subset V \text{ disjoint}, \langle u_{A_i}, D u_{A_j} \rangle = \delta_{ij} \right\},$$

where δ_{ij} is the Kronecker delta that equals one if $i = j$ and zero otherwise. The natural relaxation for this problem is

$$\min \left\{ \sum_{i=1}^{k} \langle -Lu_i, u_i \rangle \,:\, u_1, \dots, u_k \in \ell^2(V),\, \langle u_i, Du_j \rangle = \delta_{ij} \right\},$$

As in Proposition 2.2 it can be shown that the minimum is given by $\lambda_1 + \dots \lambda_k$ and is attained by the first k generalized eigenvectors solving $-Lu_i = \lambda D u_i$.

Definition 2.1 (Spectral embedding)**.** Let $G = (V, E)$, $k \leq \#V$, and let $(u_1, \dots, u_k)$ denote the first k generalized eigenvectors of the graph Laplacian. Then the following subset of $\mathbb{R}^k$

$$\Phi = \{(u_1(x), \dots, u_k(x)) \in \mathbb{R}^k \,:\, x \in V\}$$

is called the spectral embedding of G (into $\mathbb{R}^k$).

Note that the spectral embedding Φ is a set of n points in $\mathbb{R}^k$ where $n = \#V$.

Spectral clustering for more than two clusters is performed by applying any clustering algorithm (e.g, k-means) to the spectral embedding.

2.1.5 Nonlinear spectral clustering

In fact, the idea of relaxing the normalized cut of a set A into an objective function of the form $\frac{1}{2} \sum_{x,y \in V} w_{xy} \left| u_A(x) - u_A(y) \right|^2$ is not limited to having a quadratic structure. Indeed, the same argument works for any exponent $1 \leq p < \infty$.

Note that we can rewrite the spectral clustering problem (2.5) as (2.7), given by

$$\min \left\{ \frac{\frac{1}{2} \sum_{x,y \in V} w_{xy} \left| u(x) - u(y) \right|^2}{\sum_{x \in V} \deg(x) \left| u(x) \right|^2} \,:\, u \in \ell^2(V) \setminus \{0\}, \langle u, D\mathbb{1} \rangle = 0 \right\}.$$

It turns out we can get rid of the orthogonality constraint by equivalently rewriting the problem as

$$\min \left\{ \frac{\frac{1}{2} \sum_{x,y \in V} w_{xy} \left| u(x) - u(y) \right|^2}{\min_{c \in \mathbb{R}} \sum_{x \in V} \deg(x) \left| u(x) - c \right|^2} \,:\, u \in \ell^2(V) \setminus \{0\} \right\}. \tag{2.10}$$

Now a straightforward generalization of that problem is to replace the exponent 2 by $p \in [1, \infty)$ everywhere:

$$\min \left\{ \frac{\frac{1}{2} \sum_{x,y \in V} w_{xy} \left| u(x) - u(y) \right|^p}{\min_{c \in \mathbb{R}} \sum_{x \in V} \deg(x) \left| u(x) - c \right|^p} \,:\, u \in \ell^2(V) \setminus \{0\} \right\}. \tag{2.11}$$

For convenience we define the Rayleigh quotient

$$R^{(p)}(u) = \frac{\frac{1}{2} \sum_{x,y \in V} w_{xy} \left| u(x) - u(y) \right|^p}{\min_{c \in \mathbb{R}} \sum_{x \in V} \deg(x) \left| u(x) - c \right|^p}, \qquad u \in \ell^2(V).$$

It turns out that this problem is still a relaxation of the normalized cut minimization as the following proposition states:

Proposition 2.4. *Let $p \in (1,\infty)$, $A \subset V$ be a subset and define*

$$u_A^{(p)}(x) = \begin{cases} \left(\dfrac{1}{\operatorname{vol} A}\right)^{\frac{1}{p-1}}, & x \in A \\ -\left(\dfrac{1}{\operatorname{vol} A^c}\right)^{\frac{1}{p-1}}, & x \in A^c. \end{cases}$$

Then it holds that

$$R^{(2)}(u_A^{(2)}) = \mathrm{NCut}(A)$$
$$\lim_{p\to 1} R^{(p)}(u_A^{(p)}) = \phi(A).$$

Exercise 2.6. Prove this.

Hence, p-spectral clustering based on the minimization of (2.11) in the limit $p \to 1$ is a relaxation of the Cheeger constant. Amazingly, one can also prove a version of Theorem 2.1 for p-spectral clustering which becomes sharp as $p \to 1$.

Theorem 2.2 (p-Cheeger inequality)**.** *Let $G = (V,w)$ be a graph with $\deg(x) > 0$ for all $x \in V$ and let*

$$\lambda_2^{(p)} := \min\left\{ \frac{\frac{1}{2}\sum_{x,y\in V} w_{xy}\,|u(x)-u(y)|^p}{\min_{c\in\mathbb{R}}\sum_{x\in V}\deg(x)\,|u(x)-c|^p} \,:\, u \in \ell^2(V)\setminus\{0\}\right\}.$$

Then it holds

$$\frac{\lambda_2^{(p)}}{2^{p-1}} \le \mathrm{Cheeg}(G) \le p\sqrt[p]{\frac{\lambda_2^{(p)}}{2^{p-1}}}$$

and in particular

$$\lim_{p\to 1}\lambda_2^{(p)} = \mathrm{Cheeg}(G).$$

Proof. The proof goes along the lines of Theorem 2.1 but it is beyond the scope of these lecture notes to present it here. It can be found in [Amg03], see also [BH09b]. □

Remark 2.7. It can also be proved (see [BH09a, Theorem 4.4]) that thresholding a solution u^* of (2.11) via $A_t = \{x \in V : u^*(x) > t\}$ the value $\min_{t\in\mathbb{R}}\phi(A_t)$ converges to $\mathrm{Cheeg}(G)$ as $p \to 1$.

Remark 2.8. Finally we would like to remark that (2.11) is equivalent to the *nonlinear* eigenvalue problem $-L^{(p)}(u) = \lambda\phi^{(p)}(u)$ where

$$L^{(p)}(u)(x) = \frac{1}{\deg(x)}\sum_{y\in V} w_{xy}\phi^{(p)}(u(y)-u(x))$$

is the random walk graph p-Laplacian operator and $\phi^{(p)}(t) = |t|^{p-2}\,t$ for $t \in \mathbb{R}$.

2.2 The PageRank algorithm

In this section we shall discuss the PageRank algorithm which was delevoped in [Pag99] and is the basis of Google's ranking of websites. It is an unsupervised method since it works with unlabeled data—namely websites and links between them—represented as directed graph. For more generality we will actually work with weighted directed graph $G = (V, w)$ where the weights satisfy $w_{xy} = w_{yx}$, in general, and it holds

$$w_{xy} = \begin{cases} > 0 & \text{if there exists a link from page } x \text{ to page } y, \\ 0 & \text{otherwise.} \end{cases}$$

The simplest choice would be $w_{xy} = 1$ if x links to y, but with general weights one could model, for instance, how prominently placed the link is or how many links there are.

The **main idea** of PageRank is the following: Take a random walk of K steps on the internet by randomly clicking on a link on the current website, and define the rank of a website x as

$$\operatorname{rank}(x) = \lim_{K\to\infty} \frac{\text{number of times } x \text{ is visited}}{K}.$$

The issue with this approach is that such a random surfer would very soon get stuck since many websites do not have links that let one leave the website. Therefore, the random surfer sitting at x acts a follows:

- With probability $\alpha \in [0, 1)$ the surfer clicks a random link from x to y with probability $\frac{w_{xy}}{\sum_{z\in V} w_{xz}}$.
- With probability $1-\alpha \in (0, 1]$ the surfer decides to visit a random website y on the internet, following a so-called teleportation distribution $(v(y))_{y\in V}$, meaning $\sum_{y\in V} v(y) = 1$ and $v(y) \geq 0$ for all $y \in V$.

Remark 2.9 (The teleportation distribution). Here we discuss three ways of choosing the teleportation distribution:

- (Uniform): The simplest choice is $v(y) = \frac{1}{n}$ for $y \in V$ where $n = \#V$ is the number of websites. In this model all websites are equally likely.
- (Localized): Once can fix some website $x_0 \in V$ and define $v(y) = \delta_{x_0,y}$ such that the random surfer is always teleported to the same website.
- (Popularity): It can be chosen based on a ranking of the most popular websites (like `instagram.com`, `amazon.com`, etc.) assuming that this is where surfers go frequently.

With $(X_k)_{k\in\mathbb{N}_0}$ we denote the position of the random surfer following the above strategy after k steps.[1] If we denote by

$$u_k(x) = \mathbb{P}(X_k = x) \tag{2.12}$$

the probability that the random surfer is at page $x \in V$ after k steps, the PageRank vector is defined as

$$u(x) := \lim_{k\to\infty} u_k(x), \tag{2.13}$$

[1] Note that for what follows the initial condition of the random surfer is irrelevant.

provided the limit exists. To characterize the limit, we first derive a recursive formula for $u_k(x)$. For this we define the probabilities

$$P(x,y) = \mathbb{P}(\text{surfer clicks on a link from } y \text{ to } x) = \frac{w_{yx}}{\sum_{z\in V} w_{yz}}.$$

Proposition 2.5. *It holds for every $x \in V$ that*

$$u_{k+1}(x) = (1-\alpha)v(x) + \alpha \sum_{y\in V} P(x,y)u_k(y).$$

Proof. To prove the identity we use the law of total probability:

$$\begin{aligned}
u_{k+1}(x) = \mathbb{P}(X_{k+1} = x) &= \sum_{y\in V} \mathbb{P}(X_{k+1} = x \mid X_k = y)\mathbb{P}(X_k = y) \\
&= \sum_{y\in V} \left[(1-\alpha)v(x) + \alpha P(x,y)\right] u_k(y) \\
&= (1-\alpha)v(x) \underbrace{\sum_{y\in V} u_k(y)}_{=1} + \alpha \sum_{y\in V} P(x,y)u_k(y) \\
&= (1-\alpha)v(x) + \alpha \sum_{y\in V} P(x,y)u_k(y),
\end{aligned}$$

which concludes the proof. □

From Proposition 2.5 we see that if the PageRank vector u exists, it has to satisfy the equation

$$u = (1-\alpha)v + \alpha P u \tag{2.14}$$

where the operator $P : \ell^2(V) \to \ell^2(V)$ is defined as

$$Pu(x) := \sum_{y\in V} P(x,y)u(y), \quad x \in V.$$

For what follows it is convenient to work with the 1-norm, defined as

$$\|u\|_1 := \sum_{x\in V} |u(x)|, \quad u \in \ell^2(V).$$

Lemma 2.2. *It holds that $\|Pu\|_1 \le \|u\|_1$ for all $u \in \ell^2(V)$.*

Exercise 2.7. Prove Lemma 2.2.

2.2.1 Existence and uniqueness

We are now ready to prove a well-posedness result.

Theorem 2.3 (Existence and uniqueness of the PageRank vector)**.** *If v is a probability distribution over V and $\alpha \in [0,1)$, then there exists a unique $u \in \ell^2(V)$ solving* (2.14)*. Furthermore, u is a probability distribution over V.*

Proof. Equation (2.14) is equivalent to the linear problem $Au = v$ where

$$A := (1-\alpha)^{-1}(1-\alpha P)$$

is a linear operator from $\ell^2(V)$ to itself. Since $\ell^2(V)$ is finite-dimensional, (2.14) possesses a unique solution if and only if $\ker A = \{0\}$. Hence, let $u \in \ell^2(V)$ with $Au = 0$. Since $\alpha < 1$ this is equivalent to $(1-\alpha P)u = 0$ and hence $\alpha P u = u$. Taking the 1-norm yields

$$\|u\|_1 = \alpha \|Pu\|_1 \leq \alpha \|u\|_1$$

which (taking into account $\alpha < 1$) is a contradiction unless $u = 0$. This shows $\ker A = \{0\}$, as desired.

To prove that u is a probability distribution we sum and use (2.14) and the fact that v is a probability distribution over V to obtain

$$\begin{aligned}
\sum_{x\in V} u(x) &= (1-\alpha)\sum_{x\in V} v(x) + \alpha \sum_{x\in V}\sum_{y\in V} P(x,y)u(y) \\
&= 1-\alpha+\alpha\sum_{y\in V} u(y) \underbrace{\sum_{x\in V} P(x,y)}_{=1} \\
&= 1-\alpha+\alpha\sum_{y\in V} u(y).
\end{aligned}$$

Reordering and using $\alpha < 1$ implies $\sum_{x\in V} u(x) = 1$. Similarly, we compute

$$\begin{aligned}
\sum_{x\in V} |u(x)| &= \sum_{x\in V} \left|(1-\alpha)v(x) + \sum_{y\in V} P(x,y)u(y)\right| \\
&\leq \sum_{x\in V} \left[(1-\alpha)v(x) + \sum_{y\in V} P(x,y)\,|u(y)|\right] \\
&\leq 1-\alpha+\alpha\sum_{y\in V} |u(y)| = (1-\alpha)\sum_{x\in V} u(x) + \alpha \sum_{y\in V} |u(y)|\,.
\end{aligned}$$

Reordering and using $\alpha < 1$ implies $\sum_{x\in V} |u(x)| \leq \sum_{x\in V} u(x)$ which implies that $u(x) \geq 0$ for all $x \in V$. Hence, u is a probability distribution over V. □

Remark 2.10 (Eigenvalue problem). As a matter of fact, the PageRank problem (2.14) can be written as an eigenvalue problem. For this we define the linear teleportation operator $T : \ell^2(V) \to \ell^2(V)$ via

$$T_v u(x) = v(x)\sum_{y\in V} u(y) = v(x)$$

we see that (2.14) is equivalent to the eigenvalue problem

$$P_\alpha u = u$$

where $P_\alpha := (1-\alpha)T_v + \alpha P$.

Exercise 2.8. Prove that 1 is the largest eigenvalue of P_α.

Remark 2.11. Perhaps surprisingly, one can relate the PageRank problem (2.14) (or equivalently the eigenvalue problem $P_\alpha u = u$) to the random walk graph Laplacian $L_{rw} = D^{-1}L$. Indeed, u solves (2.14) if and only if it solves

$$u - \frac{\alpha}{1-\alpha} L_{rw}^* u = v,$$

where L_{rw}^* is the adjoint of the random walk Laplacian.

Exercise 2.9. Prove the statements in Remark 2.11.

2.2.2 Convergence

Next we study how fast the iteration from Proposition 2.5 convergence to the PageRank vector u solving (2.14). It turns out the convergence is exponentially fast in terms of k.

Theorem 2.4 (Convergence)**.** *If v is a probability distribution over V and $\alpha \in [0,1)$, then it holds*

$$\|u_k - u\|_1 \leq \alpha^k \|u_0 - u\|_1 ,$$

where $u \in \ell^2(V)$ is the unique solution of (2.14).

Proof. Subtracting the formulas for u_k (cf. Proposition 2.5) and u we get, using also Lemma 2.2, that

$$u_k - u = \alpha P(u_{k-1} - u)$$

and hence

$$\|u_k - u\|_1 = \alpha \|P(u_{k-1} - u)\|_1 \leq \alpha \|u_{k-1} - u\|_1$$

and inductively

$$\|u_k - u\|_1 \leq \alpha^k \|u_0 - u\|_1$$

□

Remark 2.12. From Theorem 2.4 we see that the choice of $\alpha \in [0,1)$ determines the speed of convergence and hence α should not be chosen too close to 1.

Remark 2.13. The PageRank iteration from Proposition 2.5 is equivalent to the power iteration $u_{k+1} = P_\alpha u_k$ for the linear operator P_α which is known to converge to an eigenvector corresponding to the largest eigenvalue. Note that since $\|P_\alpha u\|_1 = 1$ if u is a probability distribution, there is no normalization required.

2.3 The t-SNE embedding

The last unsupervised method which we shall discuss in these notes is the so-called t-distributed stochastic neighbor embedding (t-SNE) which, similarly to the spectral embedding from 2.1 aims to embed a given dataset into a low-dimensional space. t-SNE is very popular for data visualization and correspondingly the typical embedding spaces are $\mathbb{R}^2$ or $\mathbb{R}^3$.

To set the scene, let $G = (V, w)$ denote a weighted and directed graph, satisfying $\deg(x) > 0$ for all $x \in V$. Enumerating the vertices of G as $V = \{x_1, \dots, x_n\}$, we define its weight and degree matrix $W, D \in \mathbb{R}^{n\times n}$ via

$$W_{ij} := \begin{cases} w(x_i, x_j) & \text{if } i, j \in \{1, \dots, n\},\ i \neq j \\ 0 & \text{otherwise,} \end{cases}$$

$$D_{ij} := \delta_{ij} \deg(x_i) = \delta_{ij} \sum_{k=1}^{n} W_{ik} \quad i, j = 1, \dots, n,\ i \neq j.$$

Next, we define a symmetrized and normalized version of the weight matrix as

$$P = \frac{1}{2n}\left(D^{-1}W + W^T D^{-1}\right) \in \mathbb{R}^{n\times n} \tag{2.15}$$

We emphasize that P is a discrete probability distribution since $P_{ij} \geq 0$ for all i, j and $\sum_{i,j=1}^{n} P_{ij} = 1$.

Exercise 2.10. Prove this.

The idea of t-SNE is to find a representation of the data as $\{y_1, \dots, y_n\} \subset \mathbb{R}^k$ (with $k = 2$ or 3) such that points x_i, x_j with a high similarity as encoded through a large value of P_{ij} are mapped to points y_i, y_j which are close in the Euclidean sense. For this we define a similarity matrix $Q \in \mathbb{R}^{n\times n}$ via

$$Q_{ij} := \begin{cases} \dfrac{\left(1 + |y_i - y_j|^2\right)^{-1}}{\sum_{k\neq l}\left(1 + |y_k - y_l|^2\right)^{-1}}, & \text{if } i, j \in \{1, \dots, n\},\ i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

We see that Q_{ij} is large if $|y_i - y_j|$ is small and vice versa. Note also that Q, just like P, is a discrete probability distribution. The simple idea of t-SNE is to determine points y_i such that Q and P are as close as possible, measured through their Kullback–Leibler (KL) divergence (a.k.a. cross-entropy in machine learning)

$$\mathrm{KL}(P, Q) := \sum_{i,j=1}^{n} P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right)$$

with the conventions that $0 \log\left(\frac{0}{q}\right) = 0$ for all $q \geq 0$ and that $p \log\left(\frac{p}{0}\right) = \infty$ for all $p > 0$.

We would like to emphasize that because of the first convention and the fact that $P_{ii} = 0$, the KL divergence is well-defined even though $Q_{ii} = 0$.

Note also that $\mathrm{KL}(P, Q) \geq 0$ for all probability distributions P and Q, and that $\mathrm{KL}(P, P) = 0$.

Exercise 2.11. Prove that for two discrete probability distributions $(p_i)_{i=1,\dots,n}, (q_i)_{i=1,\dots,n}$ it holds

$$\sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) \geq 0$$

with equality if and only if $p = q$.

Hint: Use the inequality $\log x \leq x - 1$ *and the same conventions for the function* $x \log x$ *as above.*

We note that the KL distance is not symmetric in P and Q but the choice of ordering is intentional since it enforces that the similarity of the embedded points Q_{ij} is very close to the similarity of the original data P_{ij} if the latter is large. For very dissimilar points, i.e., $P_{ij} \approx 0$ the similarity in embedding space does not matter much.

In t-SNE one now performs a gradient descent of $\mathrm{KL}(P, Q)$ with respect to the variables $y_1, \dots, y_n$ on which Q depends. Thanks to the decomposition

$$\mathrm{KL}(P, Q) = \sum_{i,j=1}^{n} P_{ij} \log P_{ij} - \sum_{i,j=1}^{n} P_{ij} \log Q_{ij},$$

where the first term does not depend on Q, and using that $P_{ii} = 0$, it suffices to minimize the energy

$$E : \mathbb{R}^{kn} \to \mathbb{R}, \quad E(y_1, \dots, y_n) := -\sum_{i \neq j} P_{ij} \log Q_{ij}, \tag{2.16}$$

where, for now, we suppress the dependency of Q on the y variables. Using the definition of Q, we can express the energy as

$$E(y_1, \dots, y_n) = \sum_{i \neq j} P_{ij} \log \left(1 + |y_i - y_j|^2\right) + \log \left(\sum_{k \neq l} \left(1 + |y_k - y_l|^2\right)^{-1} \right). \tag{2.17}$$

We notice that this energy is the sum of an attraction and a repulsion term. Indeed, for points with P_{ij} large, the first term is minimized by choosing y_i close to y_j. The second term, however, encourages that nearby points spread out.

The t-SNE method is then given by the gradient descent of energy E. Starting with some initial guess $y_1^{(0)}, \dots, y_n^{(0)} \in \mathbb{R}^k$ the points are updated via

$$y_i^{(k+1)} = y_i^{(k)} - h \nabla_{y_i} E(y_1^{(k)}, \dots, y_n^{(k)}) \qquad \text{for } i = 1, \dots, n, \ k \in \mathbb{N}_0, \tag{2.18}$$

where $h > 0$ is a step size. It remains to compute the gradient of E to obtain an explicit algorithm.

Proposition 2.6. *The gradient of the energy* $E : \mathbb{R}^{kn} \to \mathbb{R}$*, defined in* (2.17)*, with respect to the variable* y_i *for* $i = 1, \dots, n$ *is given by*

$$\nabla_{y_i} E(y_1, \dots, y_n) = 4Z \sum_{j \neq i} (P_{ij} - Q_{ij}) \, Q_{ij} (y_i - y_j),$$

where we abbreviate $Z := \sum_{k \neq l} \left(1 + |y_k - y_l|^2\right)$.

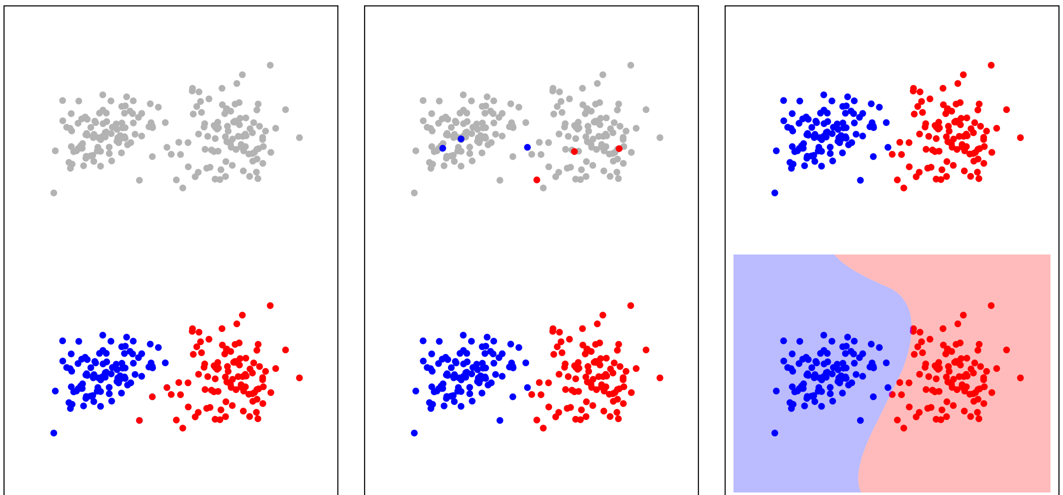
Remark 2.14. From Proposition 2.6 we see that the negative gradient $-\nabla_{y_i} E$ pulls the point y_i towards its neighbors y_j for which $P_{ij} \geq Q_{ij}$. All other points have a repulsive force.

Exercise 2.12. Prove Proposition 2.6.

3 Semi-supervised learning

In this section we will study graph-based semi-supervised learning methods. Semi-supervised learning refers to the situation where one has a large but finite dataset where only a very small subset carries labels. The task is to propagate these labels to the whole data set. This is in stark contrast to supervised learning where typically the whole dataset is labeled and one would like to assign labels to new previously unseen data. The latter is typically achieved by fitting a parametrized function, e.g., a neural network. See Figure 1 for an illustration of these different paradigms.

Semi-supervised learning is typically used whenever abundant data is available but labels are expensive or hard to get, e.g., for tumor classification in medical images.



(a) Unsupervised learning (b) Semi-supervised learning (c) Supervised learning

Figure 1: Different learning paradigms, using increasing amounts of labeled data.

Mathematically speaking, the setup for semi-supervised learning involves a dataset V and a labeled set $\Gamma \subset V$ with labels $g : \Gamma \to Y$, where Y denotes the set of possible labels. The task is to find a function $u : V \to Y$ which extends the labels, meaning that $u = g$ on Γ. Obviously, this extension problem has potentially infinitely many solutions and the goal is to construct one which is meaningful.

The simplest semi-supervised learning algorithm is the nearest-neighbor classifier. Given a dataset V and a labeled set $\Gamma \subset V$ with labels $g : \Gamma \to \mathbb{R}$, the nearest-neighbor classifier assigns the following labels

$$u(x) = g(x_0) \qquad \text{where} \qquad x_0 \in \arg\min_{y \in V} |y - x| \,, \tag{3.1}$$

together with a rule to break ties. The issue with this approach is that it does not at all take the distribution of unlabeled data points into account, see Figure 2.

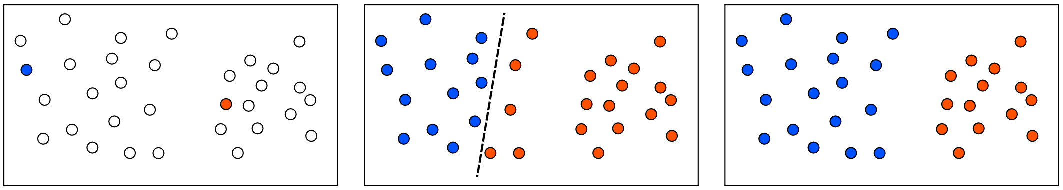


Figure 2: **Left:** Data with labels, **middle:** nearest-neighbor classifier with separating hyperplane, **right:** desired classifier.

3.1 The graph Laplace equation

We let $\mathcal{G} = (V, w)$ be a weighted undirected graph. We also fix a set of labeled vertices $\Gamma \subset V$ and a label function $g : \Gamma \to \mathbb{R}$.

Example 3.1 (Geometric weights)**.** For $V \subset \mathbb{R}^d$, a non-increasing function $\eta : [0, \infty) \to [0, \infty)$ and a bandwidth $\varepsilon > 0$ we can define $w_{xy} := \eta\left(\frac{|x-y|}{\varepsilon}\right)$. A popular choice for η is $\eta(t) = \exp(-\sigma t)1_{[0,1]}(t)$ for $\sigma \geq 0$ or $\eta(t) = \frac{1}{t}1_{[0,1]}(t)$.

Assumption 3.1 (Semi-supervised smoothness assumption)**.** Similar data points should get similar labels.

To extend the labels from Γ to V we try to determine a function $u : V \to \mathbb{R}$ which coincides with g on Γ and enforces Assumption 3.1 on $V \setminus \Gamma$. For this we let $\ell^2(V)$ denote the Hilbert space of all functions $u : V \to \mathbb{R}$ equipped with the inner product $\langle u, v\rangle_{\ell^2(V)} := \sum_{x\in V} u(x)v(x)$ and define the convex set of admissible functions

$$\mathcal{A} := \{u \in \ell^2(V) \,:\, u = g \text{ on } \Gamma\}.$$

We consider the following optimization problem

$$\min_{u\in\mathcal{A}} \mathcal{E}(u), \tag{3.2}$$

where we define the graph Dirichlet energy of $u \in \ell^2(V)$ as

$$\mathcal{E}(u) := \frac{1}{2}\sum_{x,y\in V} w_{xy}\,|u(x) - u(y)|^2\,. \tag{3.3}$$

Proposition 3.1. *Problem* (3.2) *admits a solution.*

Proof. Since $\mathcal{E}$ is a continuous function of u, we plan to apply the Bolzano–Weierstraß theorem. For this, however, we need to restrict the minimization to a compact set. Note that $\mathcal{A}$ is not compact. To this end, we note that truncation does not increase the Dirichlet energy, i.e., for $u \in \ell^2(V)$ the function $u_{a,b}(x) := \min(\max(u(x), b), a)$ satisfies $\mathcal{E}(u_{a,b}) \leq \mathcal{E}(u)$. This is because

$$|u_{a,b}(x) - u_{a,b}(y)| \leq |u(x) - u(y)| \qquad \forall x, y \in V.$$

Hence, setting $a = \min_\Gamma g$ and $b = \max_\Gamma g$, we can introduce the compact and non-empty set

$$\mathcal{B} = \{u \in \mathcal{A} \,:\, a \leq u(x) \leq b \;\forall x \in V\} \subset \mathcal{A}$$

and note that (3.2) is equivalent to $\min_{u\in\mathcal{B}}\mathcal{E}(u)$ which, by Bolzano–Weierstraß, possesses a solution. $\square$

For proving uniqueness we need the extra assumption that the graph is connected to Γ, meaning that for all $x \in V$ there exists $y \in \Gamma$ as well as $x_1 = x, \dots, x_m = y \in V$ with $w_{x_i x_{i+1}} > 0$ for all $i = 1, \dots, m-1$.

Under this assumption we can prove uniqueness directly using strong convexity of (3.3). However, we will pursue a different strategy based on the maximum principle. For this we first derive a necessary optimality condition for (3.2). If $u \in \mathcal{A}$ is a minimizer and $v \in \ell^2(V)$ satisfies $v = 0$ on Γ then $u + tv \in \mathcal{A}$ for all $t \geq 0$ and we get

$$\begin{aligned}
0 = \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(u+tv) &= \sum_{x,y\in V} w_{xy}(u(x)-u(y))(v(x)-v(y)) \\
&= \sum_{x,y\in V} w_{xy}(u(x)-u(y))v(x) - \sum_{x,y\in V} w_{yx}(u(y)-u(x))v(x) \\
&= 2\sum_{x,y\in V} w_{xy}(u(x)-u(y))v(x),
\end{aligned}$$

using the symmetry of the weights. Since v was arbitrary, a necessary condition of optimality for (3.2) is the graph Laplace equation

$$\begin{cases} Lu = 0 & \text{in } V \setminus \Gamma, \\ u = g & \text{in } \Gamma, \end{cases} \tag{3.4}$$

where the graph Laplace operator $L : \ell^2(V) \to \ell^2(V)$ is defined as

$$Lu(x) = \sum_{y\in V} w_{xy}\left(u(y) - u(x)\right), \qquad u \in \ell^2(V). \tag{3.5}$$

Note that $Lu(x) = 0$ for $x \in V$ is equivalent to the mean-value property

$$u(x) = \frac{1}{d(x)} \sum_{y\in V} w_{xy} u(y),$$

where $d(x) = \sum_{y\in V} w_{xy}$ is the degree of x.

Now we can prove the main result of this section, namely the maximum principle for subsolutions of the graph Laplace equation.

Theorem 3.1 (Maximum principle)**.** *Let $u \in \ell^2(V)$ satisfy $Lu(x) \geq 0$ for all $x \in V \setminus \Gamma$. If $G = (V, w)$ is connected to Γ, it holds*

$$\max_{x\in V} u(x) = \max_{x\in\Gamma} u(x).$$

Proof. Let $x_0 \in V$ be such that $\max_{x\in V} u(x) = u(x_0)$ and assume that $x_0 \notin \Gamma$. Using this together with $Lu(x_0) \geq 0$ we get

$$u(x_0) \leq \frac{1}{d(x_0)} \sum_{y\in V} w_{x_0 y} u(y) \leq \frac{1}{d(x_0)} \sum_{y\in V} w_{x_0 y} u(x_0) = u(x_0)$$

and it follows $\sum_{y\in V} w_{x_0 y}\left(u(y)-u(x_0)\right) = 0$. Since by definition of x_0 we have that $w_{x_0 y}\left(u(y)-u(x_0)\right) \leq 0$ for all $y \in V$, it follows that $w_{x_0 y}\left(u(y)-u(x_0)\right) = 0$ for all $y \in V$. This implies that $u(x_0) = u(y)$ for all $y \in V$ with $w_{x_0 y} > 0$. By picking a path $x_0, x_1, \dots, x_n$ from x_0 to some $x_n \in \Gamma$ we obtain that $u(x_0) = u(x_1) = \max_{x\in V} u(x)$. We can hence repeat the argument for x_1 and get $u(x_2) = u(x_1)$ and inductively $\max_{x\in V} u(x) = u(x_0) = \dots = u(x_n) \leq \max_{x\in\Gamma} u(x)$. Since $\max_{x\in\Gamma} u(x) \leq \max_{x\in V} u(x)$ holds trivially, we can conclude the proof. □

Corollary 3.1. *If $G = (V, w)$ is connected to Γ, then* (3.2) *and* (3.4) *possess a unique solution.*

Exercise 3.1. Give two proofs of Corollary 3.1, one using Theorem 3.1, and one using just the properties of the graph Dirichlet energy $\mathcal{E}$.

We can also obtain a maximum principle that does not require connectedness of the graph but instead requires strict subsolutions.

Lemma 3.1. *Let $u \in \ell^2(V)$ satisfy $Lu(x) > 0$ for all $x \in V \setminus \Gamma$. Then it holds*

$$\max_{x\in V} u(x) = \max_{x\in\Gamma} u(x).$$

Proof. If $x_0 \in V \setminus \Gamma$ is such that $u(x_0) = \max_{x\in V} u(x)$ then it follows

$$0 < Lu(x_0) = \sum_{y\in V} w_{x_0 y}\left(u(y)-u(x_0)\right) \leq 0$$

which is a contradiction and hence $x_0 \in \Gamma$. □

3.2 A random walk perspective on Laplace learning

We will see that the problem Equation (3.4) admits a nice reformulation in term of stopped random walk. The intuitive idea is the following: Imagine starting from a vertex x, and walk randomly from a vertex to one of its neighbors, until you reach a vertex $y \in \Gamma$. You write down the value of $g(y_1)$, and you start walking again from x until you hit another vertex $y_2 \in \Gamma$; you note $g(y_2)$ and repeat the same process indefinitely. Then $u(x)$ is the average value of $g(y_1), ..., g(y_n), ...$

This intuition is put rigorously in the following theorem:

Theorem 3.2. *Let $G = (V, w)$ a weighted graph and $\Gamma \subset V$ a non empty set such that the graph is connected to Γ. Let $x \in V$ and $(X_k)_{0\leq k}$ be a random walk starting at x, with transition probability*

$$\mathbb{P}(X_{k+1} = y | X_k = x) := \frac{w_{xy}}{\deg(x)}.$$

The solution of Equation (3.4) *satisfies:*

$$u(x) = \mathbb{E}\left[g(X_\tau) | X_0 = x\right] \tag{3.6}$$

where

$$\tau := \inf\{k \geq 0 : X_k \in \Gamma\}.$$

Remark 3.1. The random variable τ is the *hitting time* of the random walk, i.e. the first time at which the random walk reaches Γ.

In order to prove the theorem, we need first to make sure that the random variable X_τ is well-defined (this would not be the case if $\tau = \infty$):

Lemma 3.2. *τ is finite a.s.*

Proof. Let $x \in V$ and $b(x) \in \Gamma$ such that there exists a path from x to b_x. Moreover, we can assume that this path is simple and of length m_x. We denote this path as

$$x = x_0 \to x_1 \to \cdots \to x_{m_x} = b_x.$$

This the path is simple, we have that $m_x \leq |V| =: M$, the number of vertices of the graph. Define

$$\varepsilon_x := \mathbb{P}(X_1 = x_1, \dots, X_{m_x} = b_x | X_0 = x).$$

We have

$$\begin{aligned}
\varepsilon_x &= \mathbb{P}(X_1 = x_1, \dots, X_{m_x} = b_x | X_0 = x) \\
&= \mathbb{P}(X_2 = x_2, \dots, X_{m_x} = b_x | X_1 = x_1, X_0 = x)\mathbb{P}(X_1 = x_1 | X_0 = x) \\
&= \mathbb{P}(X_2 = x_2, \dots, X_{m_x} = b_x | X_1 = x_1)\mathbb{P}(X_1 = x_1 | X_0 = x) \\
&= \dots \\
&= \prod_{i=0}^{m_x - 1} \mathbb{P}(X_{i+1} = x_{i+1} | X_i = x_i) \\
&= \prod_{i=0}^{m_x - 1} \frac{w_{x_i, x_{i+1}}}{\deg(x_i)}.
\end{aligned}$$

Now, let $\alpha := \min_{x,y \in V}\{\frac{w_{xy}}{\deg(x)} : w_{xy} > 0\}$. This implies that $\varepsilon_x \geq \alpha^{m_x} \geq \alpha^M > 0$, uniformly for all $x \in V$. Hence the probability of the walk to stop in less than M steps is lower bounded by the probability to take the particular previous path:

$$\mathbb{P}(\tau \leq M) \geq \varepsilon_x \geq \alpha^M \quad \Longrightarrow \quad \mathbb{P}(\tau > M) \leq 1 - \alpha^M.$$

In the same way, we have that for all $n \in N$, $\mathbb{P}(\tau > (n+1)M | \tau > nM) \leq 1 - \alpha^M$, which leads to

$$\begin{aligned}
\mathbb{P}(\tau > (n+1)M) &= \mathbb{P}(\tau > (n+1)M | \tau > nM)\mathbb{P}(\tau > nM) + \underbrace{\mathbb{P}(\tau > (n+1)M | \tau \leq nM)}_{=0}\mathbb{P}(\tau \leq nM) \\
&\leq (1 - \alpha^M)\mathbb{P}(\tau > nM) \\
&\leq \dots \\
&\leq (1 - \alpha^M)^n.
\end{aligned}$$

But $\{z = \infty\} = \bigcap_{n \geq 0}\{\tau > nM\}$ which is the intersection of a decreasing sequence of events. Hence:

$$\mathbb{P}(\tau = \infty) = \lim_{n \to \infty} \mathbb{P}(\tau > nM) = 0,$$

which proves the lemma. $\square$

Proof of the Theorem. Thanks to the previous lemma, the random variable X_τ is well defined. Let $k \in \mathbb{N}$ and $x \in V \setminus \Gamma$. We have:

$$\mathbb{E}[u(X_{k+1})|X_k = x] = \sum_{y\in V} \mathbb{P}(X_{k+1} = y|X_k = x)u(y) = \frac{1}{\deg(x)} \sum_{y\in V} w_{xy}u(y) = u(x)$$

where the last equality comes from the fact that u is harmonic on $V \setminus \Gamma$. Hence, $u(X_k) = \mathbb{E}[u(X_{k+1})|X_k]$. By recursively applying this formula from time 0 to τ, we have:

$$\begin{aligned} u(x) &= \mathbb{E}[u(X_1)|X_0 = x] && (3.7)\\ &= \mathbb{E}[\mathbb{E}[u(X_2)|X_1]|X_0 = x] && (3.8)\\ &= \dots && (3.9)\\ &= \mathbb{E}[\dots\mathbb{E}[u(X_\tau)|X_{\tau-1}]\dots|X_0 = x] && (3.10)\\ &= \mathbb{E}[\dots\mathbb{E}[g(X_\tau)|X_{\tau-1}]\dots|X_0 = x] && (3.11) \end{aligned}$$

since $X_\tau \in \Gamma$. Now, using repeatedly the Tower Formula $\mathbb{E}[\mathbb{E}[X|Y]|Z] = \mathbb{E}[X|Z]$ from the last line, we can collapse back to :

$$\begin{aligned} u(x) = &= \mathbb{E}[\dots\mathbb{E}[\mathbb{E}[g(X_\tau)|X_{\tau-1}]|X_{\tau-2}]\dots|X_0 = x]\\ &= \mathbb{E}[\dots\mathbb{E}[g(X_\tau)|X_{\tau-2}]\dots|X_0 = x]\\ &= \dots\\ &= \mathbb{E}[g(X_\tau)|X_0 = x]. \end{aligned}$$

□

3.3 Random geometric graphs

In this section we aim to prove consistency of the graph Laplace equation with a partial differential equation. This requires that the graph G and its associated graph Laplacian (3.5) approximate a given Euclidean domain $\Omega \subset \mathbb{R}^d$ and a certain partial differential operator sufficiently well.

To ensure that this is the case, we therefore need a model assumption on the graph, or equivalently on the data that is used to construct it.

Assumption 3.2 (Manifold assumption)**.** The data points $\{x_i\}_{i=1,\dots,n}$ are *i.i.d.* random samples from a probability distribution on a manifold.

For the purpose of this lecture we will work with a more restrictive assumption for the rest of this section which, however, already requires most techniques and tools to deal with the general case of Assumption 3.2:

We assume that $\Omega \subset \mathbb{R}^d$ is a domain with smooth boundary, $V_n = \{x_i\}_{i=1,\dots,n} \subset \overline{\Omega}$ is a *i.i.d.* sample from a probability distribution which has density $\rho \in C^2(\overline{\Omega})$ with respect to the Lebesgue measure restricted to Ω and satisfies $c_\rho \leq \rho \leq C_\rho$ on Ω. Remember that this means $\mathbb{P}(x_i \in A) = \int_A \rho(x)\,\mathrm{d}x$ for all $i = 1,\dots,n$. Furthermore, as in Example 3.1 we consider weights of the form

$$w_{xy}^{\varepsilon,n} = \frac{2}{\sigma_\eta n\varepsilon^2}\eta_\varepsilon(|x-y|) \tag{3.12}$$

with a non-increasing and bounded function $\eta : [0, \infty) \to [0, \infty)$ that satisfies $\operatorname{supp} \eta \subset [0, 1]$. Here we also used the notations $\eta_\varepsilon(t) = \varepsilon^{-d}\eta(t/\varepsilon)$ as well as

$$\sigma_\eta = \int_{\mathbb{R}^d} \eta(|z|) \, |z_1|^2 \, \mathrm{d}z < \infty. \tag{3.13}$$

3.4 Continuum limit

In the setting of a random geometric graph we can now prove our main result, the discrete to continuum convergence of (suitably normalized) solutions to the graph Laplace equation to the solution of a boundary value problem involving a linear elliptic operator of Laplacian type. For this we define the graph Laplace operator

$$L_{n,\varepsilon} u(x) = \frac{2}{\sigma_\eta n \varepsilon^2} \sum_{y \in V_n} \eta_\varepsilon\left(|x - y|\right) \left(u(y) - u(x)\right), \qquad x \in V,\ u \in \ell^2(V_n), \tag{3.14}$$

which arises by using the weights (3.12), and the following linear differential operator

$$\Delta_\rho u := \rho^{-1} \operatorname{div}(\rho^2 \nabla u) = \rho^{-1} \sum_{i=1}^{d} \partial_i \left(\rho \partial_i u\right), \qquad x \in \Omega,\ u \in C^2(\Omega).$$

We will see that the differential operator Δ_ρ arises as a limit of the graph Laplacian $L_{n,\varepsilon}$ for large number of data points $n \in \mathbb{N}$ and small $\varepsilon > 0$.

To state our theorem, for $\varepsilon > 0$ we define $\partial_\varepsilon\Omega := \{x \in \Omega : \operatorname{dist}(x, \Omega^c) \le \varepsilon\}$ and $\Omega_\varepsilon = \Omega \setminus \partial_\varepsilon\Omega$.

Theorem 3.3 (Continuum limit)**.** *Let $0 < \varepsilon \le 1$, $n \in \mathbb{N}$, and $g \in C^3(\overline{\Omega})$. Define $\Gamma_n := V_n \cap \partial_\varepsilon\Omega$, let $u_{n,\varepsilon} \in \ell^2(V_n)$ be a solution of*

$$\begin{cases} L_{n,\varepsilon} u_{n,\varepsilon}(x) = 0, & x \in V_n \setminus \Gamma_n, \\ u_{n,\varepsilon}(x) = g(x), & x \in \Gamma_n, \end{cases} \tag{3.15}$$

and $u \in C^3(\overline{\Omega})$ be the unique solution of

$$\begin{cases} \Delta_\rho u(x) = 0, & x \in \Omega, \\ u(x) = g(x), & x \in \partial\Omega. \end{cases} \tag{3.16}$$

There exist constants $C_1, C_2 > 0$ such that for any $0 < \lambda \le 1$ the event that

$$\max_{x \in V_n} |u_{n,\varepsilon}(x) - u(x)| \le C_1 \left(\|u\|_{C^3(\overline{\Omega})} + 1 \right) (\lambda + \varepsilon)$$

has probability at least

$$1 - 4 \exp\left(-C_2 n \varepsilon^{d+2} \lambda^2 + \log n \right).$$

Laplace learning is asymptotically well-posed for sufficiently dense graphs and sufficiently large label sets.

Remark 3.2. The best error we can get from Theorem 3.3 is $O(\varepsilon)$ for the choice $\lambda = \varepsilon$. If $\varepsilon = \varepsilon_n$ satisfies

$$\varepsilon_n \gg \left(\frac{\log n}{n}\right)^{\frac{1}{d+4}},$$

where we remark that the right hand side is larger than the connectivity threshold, the convergence from Theorem 3.3 holds true almost surely as $n \to \infty$ by the Borel–Cantelli lemma.

For proving this theorem, we require a consistency statement for the graph Laplace operator, meaning that $L_{n,\varepsilon}u(x) \approx \Delta_\rho u(x)$ for $x \in V_n$ and a sufficiently regular function u. To show this, we shall pass through a nonlocal operator that arises as expectation of the graph Laplacian. It is given by

$$\mathcal{L}_\varepsilon u(x) := \frac{2}{\sigma_\eta \varepsilon^2} \int_\Omega \eta_\varepsilon(|x-y|)(u(y)-u(x))\rho(y)\,\mathrm{d}y. \tag{3.17}$$

For relating it to the graph Laplacian we will require results on concentration of measure.

3.4.1 Concentration of measure

Concentration of measure deals with the question of quantifying the probability that a random variable is close to its expected value. The simplest such concentration inequality is Markov's inequality.

Proposition 3.2 (Markov's inequality)**.** *Let S be a non-negative random variable and $t > 0$. Then it holds*

$$\mathbb{P}\left[S \geq t\right] \leq \frac{\mathbb{E}\left[S\right]}{t}$$

Proof. The statement follows from:

$$\mathbb{E}\left[S\right] = \mathbb{E}\left[S 1_{S \geq t}\right] + \mathbb{E}\left[S 1_{S<t}\right] \geq t\,\mathbb{P}\left[S \geq t\right].$$

□

We will be particularly interested in the case of sums or averages of *i.i.d.* random variables, i.e., random variables of the form $S_n = \frac{1}{n}\sum_{i=1}^n X_i$. The central limit theorem tells us that if the X_i's are *i.i.d.* with expectation μ and variance σ^2, then $\sqrt{n}(S_n - \mu)$ converges in distribution to a $\mathcal{N}(0, \sigma^2)$-distributed random variable. In particular, we expect to get the Gaussian bounds of the form

$$\mathbb{P}\left[|S_n - \mu| \geq t\right] \leq C \exp\left(-\frac{nt^2}{2\sigma^2}\right) \qquad \forall t > 0.$$

Note that such a bound would be much sharper than Markov's inequality which just gives an algebraic decay in t. Our goal will be to prove Gaussian bounds for S_n under some extra condition on the random variables X_i which essentially requires them to be almost surely bounded.

Exercise 3.2. Show that if $Z \sim \mathcal{N}(0,1)$ then for all $t > 0$,

$$\mathbb{P}(Z \geq t) \leq \frac{e^{-t^2/2}}{t\sqrt{2\pi}}.$$

We start with the Chernoff bounding technique which involves the moment generating function of a random variable.

Definition 3.1 (Moment generating function)**.** We define the moment generating function M_X of a random variable X as

$$M_X(\lambda) := \mathbb{E}\left[\exp(\lambda X)\right], \qquad \lambda \in \mathbb{R},$$

if the value exists.

Using the moment generating function we can always produce an exponential tail bound, the so-called Chernoff bound.

Proposition 3.3 (Chernoff bounds)**.** *For a random variable X and any $\lambda > 0$ it holds that*

$$\mathbb{P}\left[X \geq t\right] \leq M_X(\lambda)\exp(-t\lambda).$$

Proof. We can use Markov's inequality from Proposition 3.2 to compute

$$\begin{aligned}\mathbb{P}\left[X \geq t\right] &= \mathbb{P}\left[\lambda X \geq \lambda t\right] = \mathbb{P}\left[\exp(\lambda X) \geq \exp(\lambda t)\right] \leq \mathbb{E}\left[\exp(\lambda X)\right]\exp(-\lambda t)\\ &= M_X(\lambda)\exp(-\lambda t).\end{aligned}$$

□

Corollary 3.2. *Let X_i, $i = 1,\dots,n$ be independent random variables. Then it holds*

$$\mathbb{P}\left[\sum_{i=1}^n (X_i - \mathbb{E}\left[X_i\right]) \geq t\right] \leq \prod_{i=1}^n M_{X_i - \mathbb{E}[X_i]}(\lambda)\exp(-\lambda t)$$

Proof. Applying Proposition 3.3 to $X := \sum_{i=1}^n (X_i - \mathbb{E}\left[X_i\right])$ It suffices to compute the moment-generating function of X. Using independence we have

$$\begin{aligned}M_X(\lambda) &= \mathbb{E}\left[\exp\left(\sum_{i=1}^n (X_i - \mathbb{E}\left[X_i\right])\right)\right] = \mathbb{E}\left[\prod_{i=1}^n \exp\left(X_i - \mathbb{E}\left[X_i\right]\right)\right]\\ &= \prod_{i=1}^n \mathbb{E}\left[\exp\left(X_i - \mathbb{E}\left[X_i\right]\right)\right] = \prod_{i=1}^n M_{X_i - \mathbb{E}[X_i]}(\lambda)\end{aligned}$$

which concludes the proof. □

Example 3.2. Chernoff bounds are most famously used for Bernoulli random variables. If X_i, $i = 1,\dots,n$ are independent Bernoulli random variables which take the value 1 with probability $p \in [0,1]$ (and 0 with probability $1-p$) the Chernoff bounds can be written as

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq (1+\delta)np\right] \leq \exp\left(-\frac{np\delta^2}{2\left(1+\frac{\delta}{3}\right)}\right) \qquad \forall \delta > 0$$

which can be proved by using Corollary 3.2, computing the moment generating function and optimizing over $\lambda > 0$.

In general, it is impossible to compute the moment-generating function and we have to resort to upper-bounding it.

The simplest way of upper-bounding it gives rise to the Hoeffding inequality which we just state here but do not prove since we do not need it later.

Theorem 3.4 (Hoeffding's inequality). *Let X_i, $i = 1, \dots, n$ be* i.i.d. *random variables with expectation $\mu = \mathbb{E}[X_i]$ and assume there exists $b > 0$ such that $|X_i - \mu| \leq b$ almost surely. Then it holds for $S_n = \frac{1}{n}\sum_{i=1}^n X_i$ that*

$$\mathbb{P}[S_n - \mu \geq t] \leq \exp\left(-\frac{nt^2}{2b^2}\right).$$

We first remark that the variance σ^2 can not be larger than b^2 due to the bound $|X_i - \mu| \leq b$. Moreover, Hoeffding's inequality is sharp if $b^2 \approx \sigma^2$ since then we get the Gaussian bound that we expect from the central limit theorem. This is the case, e.g., for uniform random variables on an interval. If the variance is significantly smaller, we expect to get σ^2 in place of b^2. This can essentially be achieved in Bernstein's inequality which takes the following form.

Theorem 3.5 (Bernstein's inequality). *Let X_i, $i = 1, \dots, n$ be* i.i.d. *random variables with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}[X_i] = \sigma^2$. If there exists $b \geq 0$ such that $|X_i - \mu| \leq b$ for all $i = 1, \dots, n$ almost surely, it holds for all $t > 0$ that*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq t\right) \leq \exp\left(-\frac{nt^2}{2\left(\sigma^2 + \frac{bt}{3}\right)}\right).$$

Remark 3.3. We note the different parameter regimes of Theorem 3.5. If $bt \leq \sigma^2$ (the small deviations regime) then we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq t\right) \leq \exp\left(-\frac{3nt^2}{8\sigma^2}\right)$$

which are the Gaussian bounds (up to constants) which we expect from the central limit theorem. On the other hand, if $bt \geq \sigma^2$ (the large deviations regime) then we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq t\right) \leq \exp\left(-\frac{3nt}{8b}\right)$$

which is merely an exponential bound.

To prove Bernstein's inequality we have to establish an upper-bound for the moment-generating function and prove some auxiliary lemmas.

Lemma 3.3 (Bernstein's lemma). *For a random variable X with expectation $\mathbb{E}[X] = \mu$ and variance $\mathbb{V}[X] = \sigma^2$ and assume that there exists a constant $b > 0$ such that $|X - \mu| \leq b$ almost surely. Then it holds*

$$M_{X-\mu}(\lambda) \leq \exp\left(\frac{\sigma^2}{b^2}\left(\exp(\lambda b) - 1 - \lambda b\right)\right).$$

Proof. Using the Taylor series of the exponential function we get for all $x \in \mathbb{R}$ with $|x| \leq b$:

$$\begin{aligned}\exp(\lambda x) &= \sum_{k=0}^{\infty} \frac{(\lambda x)^k}{k!} = 1 + \lambda x + x^2 \sum_{k=2}^{\infty} \frac{\lambda^k x^{k-2}}{k!} \leq 1 + \lambda x + x^2 \sum_{k=2}^{\infty} \frac{\lambda^k b^{k-2}}{k!} \\ &= 1 + \lambda x + \frac{x^2}{b^2} \left(\exp(\lambda b) - 1 - \lambda b\right).\end{aligned}$$

By taking expectations we get

$$\begin{aligned}M_{X-\mu}(\lambda) &= \mathbb{E}\left[\exp(\lambda X)\right] \leq 1 + \lambda \mathbb{E}\left[X - \mu\right] + \mathbb{E}\left[\frac{(X-\mu)^2}{b^2}\right] \left(\exp(\lambda b) - 1 - \lambda b\right) \\ &= 1 + \frac{\sigma^2}{b^2} \left(\exp(\lambda b) - 1 - \lambda b\right) \leq \exp\left(\frac{\sigma^2}{b^2} \left(\exp(\lambda b) - 1 - \lambda b\right)\right),\end{aligned}$$

where we used the elementary inequality $1 + x \leq \exp(x)$ for $x \in \mathbb{R}$. □

For what follows we use the function $h(\delta) := (1+\delta)\log(1+\delta) - \delta$, defined for all $\delta > -1$.

Lemma 3.4. *For any number $\delta > 0$ we have*

$$\max_{x \geq 0} \left\{\delta x - (\exp(x) - 1 - x)\right\} = h(\delta).$$

Proof. Defining the function $f(x) = \delta x - (\exp(x) - 1 - x)$ we see that $f'(x) = \delta - \exp(x) + 1 = 0$ if and only if $x = \log(1+\delta) > 0$ and furthermore we have $f''(x) = -\exp(x) < 0$ so that x is a global maximum. The maximal value is then given by $f(\log(1+\delta)) = h(\delta)$. □

Lemma 3.5. *For any $\delta > 0$ we have*

$$h(\delta) \geq \frac{\delta^2}{2\left(1 + \frac{\delta}{3}\right)}.$$

Proof. Let $\delta \geq 0$. The idea is to compare the derivatives of the function h and the function $f(\delta) = \frac{\delta^2}{2(1+\delta/3)}$ which appears on the right hand side. We note that $h(0) = h'(0) = f(0) = f'(0) = 0$ and furthermore

$$h''(\delta) = \frac{1}{1+\delta} \geq \frac{1}{(1+\delta/3)^3} = f''(\delta) \qquad \forall \delta > 0$$

where we used that $(a+b)^3 = a^3 + 3ab^2 + 3a^b + b^3$. Using the fundamental theorem of calculus thus allows us to show that $h'(\delta) \geq f'(\delta)$ for all $\delta > 0$. Applying once more give $h(\delta) \geq f(\delta)$. □

Now we are ready to prove Bernstein's inequality.

Proof of Theorem 3.5. Using Corollary 3.2 and applying Lemma 3.3 to $X = X_i$ we have

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^{n} (X_i - \mu) \geq t\right] &\leq \prod_{i=1}^{n} M_{X_i - \mu}(\lambda) \exp(-\lambda t) \\ &\leq \exp\left(\frac{n\sigma^2}{b^2} \left(\exp(\lambda b) - 1 - \lambda b\right) - \lambda t\right) \\ &= \exp\left(-\frac{n\sigma^2}{b^2} \left(\frac{bt}{n\sigma^2} \lambda b - \left(\exp(\lambda b) - 1 - \lambda b\right)\right)\right).\end{aligned}$$

Minimizing the left hand side with respect to $\lambda > 0$ and using Lemmas 3.4 and 3.5 we obtain

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu) \geq t\right] &\leq \exp\left(-\frac{n\sigma^2}{b^2} h\left(\frac{bt}{n\sigma^2}\right)\right) \\ &\leq \exp\left(-\frac{n\sigma^2}{b^2}\frac{\left(\frac{bt}{n\sigma^2}\right)^2}{2\left(1+\frac{1}{3}\frac{bt}{n\sigma^2}\right)}\right) \\ &= \exp\left(-\frac{t^2}{2n\left(\sigma^2+\frac{bt}{3n}\right)}\right).\end{aligned}$$

Finally, we conclude the proof by replacing $t > 0$ by $nt > 0$ we can conclude the proof. $\square$

Exercise 3.3. Prove the Chernoff bounds from Example 3.2, using again Lemmas 3.4 and 3.5.

3.4.2 Consistency

Now we turn to the important consistency results which are necessary to prove Theorem 3.3. For this we first prove that with high probability the graph Laplacian evaluated on a Lipschitz-continuous function is close to the nonlocal operator $\mathcal{L}_\varepsilon$. This result requires Bernstein's inequality. As a next step, we will prove that the nonlocal operator $\mathcal{L}_\varepsilon$ evaluated on a C^3-function is close to the weighted Laplace operator Δ_ρ.

Lemma 3.6 (Discrete to nonlocal consistency)**.** *There exists a constant $C > 0$ such that for $u \in \mathrm{Lip}(\Omega)$, $0 < \lambda \leq \varepsilon^{-1}$, $0 < \varepsilon \leq 1$, and $n \in \mathbb{N} \setminus \{1\}$ the event that*

$$\max_{x \in V_n} |L_{n,\varepsilon} u(x) - \mathcal{L}_\varepsilon u(x)| \leq \mathrm{Lip}(u)\lambda$$

has probability at least

$$1 - 2\exp\left(-Cn\varepsilon^{d+2}\lambda^2 + \log n\right).$$

Proof. We fix $x \in \Omega$ and shall apply Bernstein's inequality to the i.i.d. random variables

$$Y_i := \frac{2}{\sigma_\eta \varepsilon^2}\eta_\varepsilon(|x_i - x|)(u(x_i) - u(x))$$

which are such that $L_{n,\varepsilon}u(x) = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\mathbb{E}(Y_i) = \mathcal{L}_\varepsilon u(x)$. Furthermore, we estimate

$$\begin{aligned}\mathbb{V}(Y_i) \leq \mathbb{E}(Y_i^2) &= \frac{4}{\sigma_\eta^2\varepsilon^4}\int_\Omega \eta_\varepsilon(|y-x|)^2(u(y)-u(x))^2\rho(y)\,\mathrm{d}y \\ &\leq \frac{4\,\mathrm{Lip}(u)^2}{\sigma_\eta^2\varepsilon^2}\int_{\Omega\cap B(x,\varepsilon)} \eta_\varepsilon(|y-x|)^2 \leq \frac{C\,\mathrm{Lip}(u)^2}{\sigma_\eta^2\varepsilon^{2+d}},\end{aligned}$$

since by assumption, $\eta_\varepsilon(|x-y|) \leq C\frac{1_{|x-y|<\varepsilon}}{\varepsilon^d}$. Remark that as often in analysis, C denotes any constant. Finally, using $\varepsilon \leq 1$ we compute

$$|Y_i - \mathbb{E}(Y_i)| \leq |Y_i| + |\mathbb{E}(Y_i)| \leq \frac{C\,\mathrm{Lip}(u)}{\sigma_\eta\varepsilon^2}\varepsilon^{1-d} + \frac{C\,\mathrm{Lip}(u)}{\sigma_\eta\varepsilon^2}\varepsilon \leq \frac{C\,\mathrm{Lip}(u)}{\sigma_\eta\varepsilon^{1+d}} = b,$$

where C depends on η and d and changed in the inequality. Theorem 3.5 then implies

$$\mathbb{P}\left(|L_{n,\varepsilon}u(x)-\mathcal{L}_\varepsilon u(x)|\geq t\right)\leq 2\exp\left(-\frac{nt^2}{2\left(\frac{C\operatorname{Lip}(u)^2}{\sigma_\eta\varepsilon^{2+d}}+\frac{C\operatorname{Lip}(u)t}{3\sigma_\eta\varepsilon^{1+d}}\right)}\right).$$

Choosing $t=\operatorname{Lip}(u)\lambda$ for $0<\lambda\leq\varepsilon^{-1}$ we get

$$\mathbb{P}\left(|L_{n,\varepsilon}u(x)-\mathcal{L}_\varepsilon u(x)|\geq \operatorname{Lip}(u)\lambda\right)\leq 2\exp\left(-Cn\varepsilon^{d+2}\lambda^2\right).$$

where we redefined the constant $C>0$. Conditioning on $x_i=x$ for $i=1,\dots,n$, using the previous result for the remaining $n-1$ i.i.d. random variables, and using a union bound one obtains

$$\begin{aligned}
&\mathbb{P}\left(\max_{x\in V_n}|L_{n,\varepsilon}u(x)-\mathcal{L}_\varepsilon u(x)|\leq \operatorname{Lip}(u)\lambda\right)\\
&=\mathbb{P}\left(\bigcap_{i=1}^n\{|L_{n,\varepsilon}u(x)-\mathcal{L}_\varepsilon u(x)|\leq \operatorname{Lip}(u)\lambda\}\right)\\
&=1-\mathbb{P}\left(\bigcup_{i=1}^n\{|L_{n,\varepsilon}u(x)-\mathcal{L}_\varepsilon u(x)|\geq \operatorname{Lip}(u)\lambda\}\right)\\
&\geq 1-\sum_{i=1}^n\mathbb{P}\left(|L_{n,\varepsilon}u(x_i)-\mathcal{L}_\varepsilon u(x_i)|\geq \operatorname{Lip}(u)\lambda\right)\\
&\geq 1-\sum_{i=1}^n\int_\Omega\mathbb{P}\left(|L_{n,\varepsilon}u(x_i)-\mathcal{L}_\varepsilon u(x_i)|\geq \operatorname{Lip}(u)\lambda\mid x_i=x\right)\rho(x)\,\mathrm{d}x\\
&\geq 1-2n\exp\left(-C(n-1)\varepsilon^{d+2}\lambda^2\right)\\
&\geq 1-2\exp\left(-Cn\varepsilon^{d+2}\lambda^2+\log n\right),
\end{aligned}$$

where we used $n-1\geq n/2$ for $n\geq 2$ and the constant $C>0$ changed its value. ☐

Exercise 3.4. Note that Lemma 3.6 does not hold uniformly in u (which is no problem for what we treat in this lecture). Prove that there exist constants $C_1,C_2,C_3>0$ such that for all $t\in(0,\varepsilon^{-1})$,

$$\mathbb{P}\left(\forall u\in C^3,\max_{x\in V_n}|L_{n,\varepsilon}u(x)-\mathcal{L}_\varepsilon u(x)|\leq C_1\|u\|_{C^3}t\right)\leq 1-C_2\exp(-C_3n\varepsilon^{d+2}t^2+\log n).$$

For this Taylor-expand u before applying Bernstein.

Exercise 3.5. Prove a version of Lemma 3.6 by using Hoeffding's inequality from Theorem 3.4 instead of Bernstein's. How does the result change?

Lemma 3.7 (Nonlocal to local consistency)**.** *There exists a constant $C>0$ depending on ρ such that for every $u\in C^3(\overline{\Omega})$ it holds*

$$\max_{x\in\Omega_\varepsilon}|\mathcal{L}_\varepsilon u(x)-\Delta_\rho u(x)|\leq C\,\|u\|_{C^3(\overline{\Omega})}\,\varepsilon.$$

Proof. To prove the statement one utilizes the following Taylor expansions in $\mathcal{L}_\varepsilon u(x)$:

$$\begin{aligned}
u(y) &= u(x) + \langle \nabla u(x), y-x\rangle + \frac{1}{2}\left\langle y-x, D^2u(x)(y-x)\right\rangle + \|u\|_{C^3(\overline{\Omega})} O(|y-x|^3) \\
\rho(y) &= \rho(x) + \langle \nabla \rho(x), y-x\rangle + O(\varepsilon^2).
\end{aligned}$$

Making a change of variables and using the Taylor expansion for $y = x + \varepsilon z$, we get for that $x \in \Omega_\varepsilon$ it holds

$$\begin{aligned}
\mathcal{L}_\varepsilon u(x) &= \frac{2}{\sigma_\eta \varepsilon^2}\int_\Omega \eta_\varepsilon(|x-y|)(u(y)-u(x))\rho(y)\,\mathrm{d}y \\
&= \frac{2}{\sigma_\eta \varepsilon^2}\int_{B_1(0)} \eta(|z|)\left(u(x+\varepsilon z)-u(x)\right)\rho(x+\varepsilon z)\,\mathrm{d}z \\
&= \frac{2}{\sigma_\eta \varepsilon^2}\int_{B_1(0)} \eta(|z|)\left(\varepsilon\langle\nabla u(x), z\rangle + \frac{\varepsilon^2}{2}\langle z, D^2u(x)z\rangle + \|u\|_{C^3(\overline{\Omega})} O(\varepsilon^3)\right)\big(\rho(x) + \varepsilon\langle\nabla\rho(x), z\rangle + O(\varepsilon^2)\big)\,\mathrm{d}z \\
&= \frac{2}{\sigma_\eta \varepsilon^2}\Bigg[\varepsilon\rho(x)\left\langle \nabla u(x), \int_{B_1(0)}\eta(|z|)z\,\mathrm{d}z\right\rangle + \frac{\varepsilon^2}{2}\rho(x)\sum_{i,j=1}^d \partial_{ij}^2 u(x)\int_{B_1(0)}\eta(|z|)z_iz_j\,\mathrm{d}z \\
&\qquad + \varepsilon^2\int_{B_1(0)}\eta(|z|)\langle\nabla u(x), z\rangle\langle\nabla\rho(x), z\rangle\,\mathrm{d}z + \|u\|_{C^3(\overline{\Omega})} O(\varepsilon^3)\Bigg].
\end{aligned}$$

Now we will show that the first summand is zero and strongly simplify the second and third one.

First summand: We observe that

$$\int_{B_1(0)}\eta(|z|)z\,\mathrm{d}z = -\int_{B_1(0)}\eta(|z|)(-z)\,\mathrm{d}z = -\int_{B_1(0)}\eta(|z|)z\,\mathrm{d}z$$

using the change of variables $-z \mapsto z$. Hence the whole integral is zero.

Second summand: Similarly, we also get that

$$\int_{B_1(0)}\eta(|z|)z_iz_j\,\mathrm{d}z = 0, \quad i \neq j.$$

To see this is suffices to consider the case $d = 2$ where we can make the change of variables $(z_1, -z_2) \mapsto (z_1, z_2)$ to get

$$\int_{B_1(0)}\eta(|z|)z_1z_2\,\mathrm{d}z = -\int_{B_1(0)}\eta(|z|)z_1(-z_2)\,\mathrm{d}z = \int_{B_1(0)}\eta(|z|)z_1z_2\,\mathrm{d}z.$$

Hence, we obtain that

$$\int_{B_1(0)}\eta(|z|)z_iz_j\,\mathrm{d}z = \delta_{ij}\int_{B_1(0)}\eta(|z|)z_i^2\,\mathrm{d}z = \delta_{ij}\int_{\mathbb{R}^d}\eta(|z|)\,|z_1|^2\,\mathrm{d}z = \delta_{ij}\sigma_\eta.$$

Hence, we get

$$\sum_{i,j=1}^d \partial_{ij}^2 u(x)\int_{B_1(0)}\eta(|z|)z_iz_j\,\mathrm{d}z = \sigma_\eta\sum_{i,j=1}^d \partial_{ij}^2 u(x)\delta_{ij} = \sigma_\eta\operatorname{Tr}(D^2u(x)) = \sigma_\eta\Delta u(x)$$

where $\Delta u(x) = \operatorname{div}(\nabla u(x))$.

Third summand: We compute

$$\begin{aligned}
\int_{B_1(0)} \eta(|z|)\langle \nabla u(x), z\rangle\langle\nabla\rho(x), z\rangle \,\mathrm{d}z &= \sum_{i,j=1}^d \partial_i u(x)\partial_j\rho(x)\int_{B_1(0)}\eta(|z|)z_i z_j\,\mathrm{d}z \\
&= \sum_{i=1}^d \partial_i u(x)\partial_i\rho(x)\int_{B_1(0)}\eta(|z|)\,|z_i|^2\,\mathrm{d}z \\
&= \sigma_\eta\langle\nabla u(x),\nabla\rho(x)\rangle
\end{aligned}$$

Final conclusion: Putting things together we arrive at

$$\begin{aligned}
\mathcal{L}_\varepsilon u(x) &= \frac{2}{\sigma_\eta\varepsilon^2}\left[0 + \frac{\varepsilon^2}{2}\sigma_\eta\rho(x)\Delta u(x) + \sigma_\eta\varepsilon^2\langle\nabla u(x),\nabla\rho(x)\rangle + \|u\|_{C^3(\overline{\Omega})}O(\varepsilon^3)\right] \\
&= \rho(x)\Delta u(x) + 2\langle\nabla u(x),\nabla\rho(x)\rangle + \|u\|_{C^3(\overline{\Omega})}O(\varepsilon) \\
&= \frac{1}{\rho(x)}\operatorname{div}\left(\rho(x)^2\nabla u(x)\right) + \|u\|_{C^3(\overline{\Omega})}O(\varepsilon) \\
&= \Delta_\rho u(x) + \|u\|_{C^3(\overline{\Omega})}O(\varepsilon).
\end{aligned}$$

Since $x \in \Omega_\varepsilon$ was arbitrary and the $O(\varepsilon)$ term is independent of x. □

As a corollary of Lemmas 3.6 and 3.7 we obtain pointwise consistency for the graph Laplacian.

Corollary 3.3 (Pointwise consistency)**.** *There are constants $C_1, C_2 > 0$ such that for any $u \in C^3(\overline{\Omega})$ and $0 < \lambda \leq \varepsilon^{-1}$ the event that*

$$\max_{x\in V_n\cap\Omega_\varepsilon}|L_{n,\varepsilon}u(x) - \Delta_\rho u(x)| \leq C_1\|u\|_{C^3(\overline{\Omega})}(\lambda+\varepsilon)$$

has probability at least

$$1 - 2\exp\left(-C_2 n\varepsilon^{d+2}\lambda^2 + \log n\right).$$

Remark 3.4. The choice for λ which leads the best consistency error is in Corollary 3.3 $\lambda = \varepsilon$ and requires the scaling $\varepsilon \gg \left(\frac{\log n}{n}\right)^{\frac{1}{d+4}}$ for the probability to be close to one. In general, one has pointwise consistency (without rate) if $\varepsilon \gg \left(\frac{\log n}{n}\right)^{\frac{1}{d+2}}$.

Exercise 3.6. Find conditions on u and ρ such that we have

$$\max_{x\in\Omega_\varepsilon}|\mathcal{L}_\varepsilon u(x) - \Delta_\rho u(x)| \leq C\varepsilon^2$$

with a suitable constant C that depends on η, d, and the regularity of u and ρ. Also derive the corresponding analogue of Corollary 3.3 in this case and determine the condition of ε to have an overall consistency error of order ε^2.

3.4.3 Convergence rate

We are now ready to prove the main theorem of this section. We first sketch the idea of the proof. Using Corollary 3.3 we have $L_{n,\varepsilon}(u - u_{n,\varepsilon}) = O(\lambda + \varepsilon)$ and hence we cannot directly use the maximum principle for the graph Laplacian from Theorem 3.1 or Lemma 3.1. Hence, we shall replace u by a function $\tilde{u}$ which satisfies $L_{n,\varepsilon}(\tilde{u} - u_{n,\varepsilon}) > 0$ and is uniformly close to u. This allows us to apply the maximum principle to $\tilde{u} - u_{n,\varepsilon}$ and then use the closeness of $\tilde{u}$ and u to bound $u - u_{n,\varepsilon}$.

Proof of Theorem 3.3. For constructing the perturbation we let $\phi \in C^3(\overline{\Omega})$ solve the PDE

$$\begin{cases} -\Delta_\rho \phi = 1 & \text{in } \Omega, \\ \phi = 0 & \text{on } \partial\Omega. \end{cases}$$

It is clear that ϕ attains its minimum on the boundary and hence $\phi \geq 0$. The perturbation is defined as $\tilde{u} := u - K\phi$ where $K > 0$ is to be determined. Note that $\tilde{u} = g$ on $\partial\Omega$. Applying Corollary 3.3 twice and using a union bound, the event that

$$\max_{x \in V_n \cap \Omega_\varepsilon} |L_{n,\varepsilon}\phi(x) + 1| \leq \tilde{C}_1(\lambda + \varepsilon), \tag{3.18}$$

$$\max_{x \in V_n \cap \Omega_\varepsilon} |L_{n,\varepsilon}u(x)| \leq C_1 \|u\|_{C^3(\overline{\Omega})} (\lambda + \varepsilon) \tag{3.19}$$

holds has probability at least $1 - 4\exp\left(-C_2 n\varepsilon^{d+2}\lambda^2 + \log n\right)$ where $\tilde{C}_1 := C_1 \|\phi\|_{C^3(\overline{\Omega})}$, and for the rest of the proof we restrict to this event. If $\tilde{C}_1(\lambda+\varepsilon) \geq \frac{1}{2}$, there is basically nothing to prove since we have

$$|u_{n,\varepsilon}(x) - u(x)| \leq 2\|g\|_\infty \leq 4\tilde{C}_1 \|g\|_\infty (\lambda + \varepsilon) \leq C\left(\|u\|_{C^3(\overline{\Omega})} + 1\right)(\lambda + \varepsilon)$$

for a suitable constant $C > 0$, not depending on u of $u_{n,\varepsilon}$.

Hence, we now assume $\tilde{C}_1(\lambda + \varepsilon) \leq \frac{1}{2}$. Then (3.18) implies that $L_{n,\varepsilon}\phi(x) \leq -\frac{1}{2}$ for all $x \in V_n \cap \Omega_\varepsilon$. Using this together with (3.19) the function $w := \tilde{u} - u_{n,\varepsilon}$ satisfies

$$L_{n,\varepsilon}w = L_{n,\varepsilon}u - KL_{n,\varepsilon}\phi - L_{n,\varepsilon}u_{n,\varepsilon} \geq -C_1 \|u\|_{C^3(\overline{\Omega})} (\lambda + \varepsilon) + \frac{K}{2} \quad \text{in } V_n \cap \Omega_\varepsilon.$$

Setting $K := 2C_1(\|u\|_{C^3(\overline{\Omega})} + 1)(\lambda+\varepsilon)$ we get $L_{n,\varepsilon}w > 0$ and hence Lemma 3.1 implies that $\max_{V_n} w = \max_{V_n \cap \partial_\varepsilon\Omega} w$. Since $u_{n,\varepsilon} = g$ on $X_n \cap \partial_\varepsilon\Omega$, both u and g are Lipschitz, and $\phi \geq 0$ we obtain

$$w = u - K\phi - u_{n,\varepsilon} \leq u - g \leq C\|u\|_{C^3(\overline{\Omega})} \varepsilon \quad \text{in } V_n \cap \partial_\varepsilon\Omega$$

for a suitable constant C, not depending on u. This implies that indeed $w \leq C\|u\|_{C^3(\overline{\Omega})} \varepsilon$ on V_n or equivalently

$$u - u_{n,\varepsilon} \leq K\phi + C\|u\|_{C^3(\overline{\Omega})} \varepsilon \quad \text{in } V_n$$

which proves

$$\max_{x \in V_n} u - u_{n,\varepsilon} \leq C\left(\|u\|_{C^3(\overline{\Omega})} + 1\right)(\lambda + \varepsilon)$$

upon increasing the constant $C > 0$. For the converse direction we apply the same argument to $-w$. $\square$

Here we discuss the pros and cons of the approach:
Pros:

- Elementary proofs;
- Explicit convergence rates;
- Rates in the strong supremum norm;
- Ideas extend to other (nonlinear) graph operators with maximum principles.

Cons:

- High regularity of limiting problem is needed, at least $u \in C^3(\overline{\Omega})$;
- Just works for strong solutions;
- Length scale restrictions $\varepsilon \gg \left(\frac{\log n}{n}\right)^{\frac{1}{d+2}}$ (for convergence) $\varepsilon \gg \left(\frac{\log n}{n}\right)^{\frac{1}{d+4}}$ (for $O(\varepsilon)$ rate) are not sharp due to consistency approach;
- Few operators have a maximum principle.

4 The variational approach

Instead of working with the strong form of the Laplace equation Equation (3.16), which requires at least C^2 regularity of u, one can directly work with the energy associated to the PDE :

$$E(u) := \frac{1}{2}\int_\Omega |\nabla u|^2 \rho^2 \tag{4.1}$$

which requires only C^1 (or actually H^1) regularity to make sense. We can show that any minimizer of Equation (4.1) is a solution to Equation (3.16).

Exercise 4.1. Assume that $\rho \in C^1(\bar{\Omega})$ and that $u \in C^2(\bar{\Omega})$ is a minimizer of Equation (4.1) with $u = g$ on $\partial\Omega$. Show that for all $v \in C_c^\infty(\Omega)$,

$$\int_\Omega \nabla u \cdot \nabla v \rho^2 = 0.$$

Using Green's formula, show that u is a solution to Equation (3.16).

Hence, it makes sense to study the convergence of the graph Dirichlet energies

$$E_{n,\varepsilon}(u) := \frac{1}{\sigma_\eta n^2 \varepsilon^2} \sum_{i,j=1}^n \eta_\varepsilon\left(|x_i - x_j|\right)|u(x_i) - u(x_j)|^2$$

toward E_u. To this purpose, we define the continuous, non-local energy

$$E_\varepsilon(u) := \frac{1}{\sigma_\eta \varepsilon^2}\int_{\Omega\times\Omega} \eta_\varepsilon(|x-y|)|u(x)-u(y)|^2 \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y.$$

4.1 Consistency of the variational setting

In order to study the asymptotic behavior of $E_{n,\varepsilon}$, we will need a new concentration inequality on random variables of the form $U_n = \frac{1}{n(n-1)} \sum_{i\neq j} f(x_i, x_j)$. This is a special instance of the so-called U-statistics, and the good news is that there exist concentration inequalities for this type of random variable:

Theorem 4.1 (Bernstein inequality for U-statistics)**.** *Let $X_1, \dots, X_n$ be i.i.d. random variables and let $f : \mathbb{R}^2 \to \mathbb{R}$ be bounded and symmetric. Let $\mu := \mathbb{E}[f(X_i, X_j)]$, $\sigma^2 := V[f(X_i, X_j)]$ and $b := \|f\|_\infty$. Define*

$$U_n = \frac{1}{n(n-1)} \sum_{i\neq j} f(X_i, X_j).$$

Then for every $t > 0$, we have

$$\mathbb{P}(U_n - \mu \geq t) \leq \exp\left(-\frac{nt^2}{6(\sigma^2 + \frac{bt}{3})}\right).$$

Proof. Let $k \in \mathbb{N}$ such that $n-1 \leq 2k \leq n$ and define

$$V(x_1, x_2, \dots, x_n) = \frac{1}{k}\left(f(x_1, x_2) + f(x_3, x_4) + \dots + f(x_{2k-1}, x_{2k})\right).$$

Then we can write

$$U_n = \frac{1}{n!} \sum_{\tau \in S(n)} V(X_{\tau_1}, X_{\tau_2}, \dots, X_{\tau_n}),$$

where $S(n)$ is the group of permutations of $\{1, \dots, n\}$. Let

$$Y_\tau = V(X_{\tau_1}, X_{\tau_2}, \dots, X_{\tau_n}) - \mu.$$

We use the Chernoff bounding trick to obtain

$$\mathbb{P}(U_n - \mu > t) \leq e^{-st} M_{U_n-\mu}(s) = e^{-st}\mathbb{E}\Big[e^{\frac{s}{n!}\sum_{\tau\in S(n)} Y_\tau}\Big] \leq e^{-st}\frac{1}{n!}\sum_{\tau\in S(n)} \mathbb{E}[e^{sY_\tau}] = e^{-st}\frac{1}{n!}\sum_{\tau\in S(n)} M_{Y_\tau}(s),$$

where the last inequality follows from the convexity of the exponential. Since Y_τ is a sum of k i.i.d. random variables with zero mean, absolute bound b, and σ^2 variance, we can apply Berstein's Lemma Lemma 3.3 to get

$$M_{Y_\tau}(s) \leq \exp\!\left(\frac{k\sigma^2}{b^2}\left(e^{\frac{sb}{k}} - 1 - \frac{sb}{k}\right)\right).$$

Therefore, we obtain

$$\mathbb{P}(U_n - \mu > t) \leq \exp\!\left(-\frac{k\sigma^2}{b^2}\left(\frac{bt}{\sigma^2} - \frac{sb}{k} - \left(e^{\frac{sb}{k}} - 1 - \frac{sb}{k}\right)\right)\right).$$

and we conclude the proof by optimizing over s in the same way as in Theorem 3.5. □

Exercise 4.2. Show that

$$U_n = \frac{1}{n!} \sum_{\tau \in S(n)} V(X_{\tau_1}, X_{\tau_2}, \dots, X_{\tau_n}).$$

Using this property, we can show the consistency of the variational formulation:

Theorem 4.2. *There exists $C_1, C_2 > 0$ such that for any $u \in C^2(\overline{\Omega})$, any $0 < \varepsilon, \lambda \leq 1$, we have:*

$$\mathbb{P}\left(|E_{n,\varepsilon}(u) - E(u)| \leq C_1 \|u\|_{C^2}^2 \left(\frac{1}{n} + \lambda + \varepsilon \right) \right) \geq 1 - 2\exp(-C_2 n \varepsilon^d \lambda^2) \tag{4.2}$$

The proof of this theorem is a simple consequence of the following two lemmas.

Lemma 4.1 (Discrete to non-local consistency)**.** *There exists $C_\eta, C_{\eta,\rho} > 0$ such that for any $0 < \lambda \leq 1$ and any Lipschitz function $u : \Omega \to \mathbb{R}$,*

$$\mathbb{P}\left(|E_{n,\varepsilon}(u) - E_\varepsilon(u)| \leq C_\eta \operatorname{Lip}(u)^2 \left(\frac{1}{n} + \lambda \right) \right) \geq 1 - 2\exp(-C_{\eta,\rho} n \varepsilon^d \lambda^2). \tag{4.3}$$

Proof. Let $f(x,y) := \eta_\varepsilon(|x-y|) \left(\frac{u(x)-u(y)}{\varepsilon} \right)^2$. We can define the U-statistics

$$U_n := \frac{1}{n(n-1)} \sum_{i \neq j} f(x_i, x_j)$$

such that we have $E_{n,\varepsilon}(u) = \frac{n-1}{\sigma_\eta n} U_n$. One can readily see that

$$\mu := \mathbb{E}[f(x_i, x_j)] = \frac{1}{\varepsilon^2} \int_{\Omega \times \Omega} \eta_\varepsilon(|x-y|) |u(x) - u(y)|^2 \rho(x)\rho(y) \, \mathrm{d}x \, \mathrm{d}y = \sigma_\eta E_\varepsilon(u).$$

On the other hand,

$$b := \|f\|_\infty \leq C_\eta \varepsilon^{-d} 1_{(0,\varepsilon)}(|x-y|) \left| \frac{u(x) - u(y)}{\varepsilon} \right|^2 \leq C_\eta \varepsilon^{-d} \operatorname{Lip}(u)^2.$$

Finally,

$$\begin{aligned}
\sigma^2 := V[f(x_i, x_j)] &\leq \mathbb{E}[f(x_i, x_j)^2] \\
&\leq \int_{\Omega \times \Omega} \eta_\varepsilon(|x-y|)^2 \left(\frac{u(x) - u(y)}{\varepsilon} \right)^4 \rho(x)\rho(y) \, \mathrm{d}x \, \mathrm{d}y \\
&\leq \frac{C_{\eta,\rho} \operatorname{Lip}(u)^4}{\varepsilon^{2d}} \int_\Omega \int_{B_\varepsilon(x)} \mathrm{d}y \, \mathrm{d}x \\
&\leq \frac{C_{\eta,\rho} \operatorname{Lip}(u)^4}{\varepsilon^d}.
\end{aligned}$$

Applying Theorem 4.1, we get that

$$\mathbb{P}(|U_n - \sigma_\eta E_\varepsilon(u)| \geq t) \leq 2\exp\left(-\frac{nt^2}{6\left(\frac{C_{\eta,\rho} \operatorname{Lip}(u)^4}{\varepsilon^d} + \frac{C_\eta \operatorname{Lip}(u)^2 t}{\varepsilon^d} \right)} \right).$$

Taking $t = \sigma_\eta \operatorname{Lip}(u)^2\lambda$ for $0 < \lambda \leq 1$, we get

$$\mathbb{P}(|\sigma_\eta^{-1}U_n - E_\varepsilon(u)| \geq \operatorname{Lip}(u)^2\lambda) \leq 2\exp\left(-C_{\eta,\rho}n\varepsilon^d\lambda^2\right).$$

We compute:

$$\begin{aligned}
\left|E_{n,\varepsilon(u)} - E_\varepsilon(u)\right| &= \left|\frac{n-1}{n\sigma_\eta}U_n - E_\varepsilon(u)\right| \\
&= \left|\frac{n-1}{n}\left(\sigma_\eta^{-1}U_n - E_\varepsilon(u)\right) - \frac{1}{n}E_\varepsilon(u)\right| \\
&\leq \frac{n-1}{n}\left|\sigma_\eta^{-1}U_n - E_\varepsilon(u)\right| + \frac{1}{n}\left|E_\varepsilon(u)\right| \\
&\leq \left|\sigma_\eta^{-1}U_n - E_\varepsilon(u)\right| + \frac{C_\eta}{n}\operatorname{Lip}(u)^2
\end{aligned}$$

and hence

$$\mathbb{P}\left(|E_{n,\varepsilon}(u) - E_\varepsilon(u)| \leq C_\eta \operatorname{Lip}(u)^2\left(\frac{1}{n} + \lambda\right)\right) \geq 1 - 2\exp(-C_{\eta,\rho}n\varepsilon^d\lambda^2)$$

$\square$

Lemma 4.2 (Non-local to local consistency)**.** *There exists $C > 0$ such that for all $u \in C^2(\bar{\Omega})$ and all $0 < \varepsilon \leq 1$,*

$$|E_\varepsilon(u) - E(u)| \leq C\|u\|_{C^2}^2\varepsilon. \tag{4.4}$$

Proof. In what follows, we will denote by $O(x)$ any function bounded by 1. Let $x, y \in \Omega$ be such that $|x-y| \leq \varepsilon$. By Taylor expanding u around x, we have that

$$u(y) = u(x) + \nabla u(x)\cdot(y-x) + \|u\|_{C^2}\varepsilon^2 O(x,y).$$

By taking the norm squared, it leads to

$$\begin{aligned}
|u(y)-u(x)|^2 &= |\nabla u(x)\cdot(y-x)|^2 + 2\underbrace{\nabla u(x)\cdot(y-x)}_{\leq \|u\|_{C^2}\varepsilon}\|u\|_{C^2}\varepsilon^2 O(x,y) + \|u\|_{C^2}^2\varepsilon^4 O(x,y)^2 \\
&= |\nabla u(x)\cdot(y-x)|^2 + \|u\|_{C^2}^2\varepsilon^3 O(x,y).
\end{aligned}$$

Hence, we can write

$$\begin{aligned}
E_\varepsilon(u_\varepsilon) =& \frac{1}{\sigma_\eta\varepsilon^2}\int_{\Omega\times\Omega}\eta_\varepsilon(|x-y|)|\nabla u(x)\cdot(y-x)|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
&+ \underbrace{\frac{\varepsilon^4\|u\|_{C^2}^2}{\sigma_\eta\varepsilon^2}\int_{\Omega\times\Omega}\eta_\varepsilon(|x-y|)O(x,y)\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y}_{=\|u\|_{C^2}^2\varepsilon O(1)}
\end{aligned}$$

Since ρ is C^1, we have that $\rho(y) = \rho(x) + \varepsilon O(x,y)$ (where here the big O depends on the

C^1 norm of ρ). Hence:

$$\begin{aligned}
&\frac{1}{\sigma_\eta \varepsilon^2}\int_{\Omega\times\Omega}\eta_\varepsilon(|x-y|)|\nabla u(x)\cdot(y-x)|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
&\qquad = \underbrace{\frac{1}{\sigma_\eta \varepsilon^2}\int_\Omega\int_{\Omega\cap B_\varepsilon(x)}\eta_\varepsilon(|x-y|)|\nabla u(x)\cdot(y-x)|^2\,\mathrm{d}y\rho(x)^2\,\mathrm{d}x}_{(\star)} \\
&\qquad\quad + \underbrace{\frac{\varepsilon}{\sigma_\eta \varepsilon^2}\int_\Omega\int_{\Omega\cap B_\varepsilon(x)}\eta_\varepsilon(|x-y|)|\nabla u(x)\cdot(y-x)|^2 O(x,y)\,\mathrm{d}y\rho(x)\,\mathrm{d}x}_{=\|u\|^2_{C^2}\varepsilon O(1)}\,.
\end{aligned}$$

One must now estimate the first part $(\star)$ of the right-hand side. We will decompose the integral over Ω as a sum of integrals over Ω^ε and $\partial^\varepsilon\Omega$ respectively, and use the fact that for a bounded function f and a smooth Ω,

$$\int_{\partial^\varepsilon\Omega} f = \varepsilon O(1). \tag{4.5}$$

Indeed:

$$\begin{aligned}
(\star) =& \frac{1}{\sigma_\eta \varepsilon^2}\int_{\Omega^\varepsilon}\int_{\Omega\cap B_\varepsilon(x)}\eta_\varepsilon(|x-y|)|\nabla u(x)\cdot(y-x)|^2\,\mathrm{d}y\rho(x)^2\,\mathrm{d}x \\
&+ \frac{1}{\sigma_\eta \varepsilon^2}\int_{\partial^\varepsilon\Omega}\int_{\Omega\cap B_\varepsilon(x)}\eta_\varepsilon(|x-y|)|\nabla u(x)\cdot(y-x)|^2\,\mathrm{d}y\rho(x)^2\,\mathrm{d}x
\end{aligned}$$

For $x\in\Omega^\varepsilon$, $B_\varepsilon(x)\subset\Omega$ and

$$\begin{aligned}
&\frac{1}{\sigma_\eta \varepsilon^2}\int_{\Omega\cap B_\varepsilon(x)}\eta_\varepsilon(|x-y|)|\nabla u(x)\cdot(y-x)|^2\,\mathrm{d}y \\
&\qquad = \frac{1}{\sigma_\eta}\int_{B_\varepsilon(x)}\eta\left(\frac{|x-y|}{\varepsilon}\right)\left|\nabla u(x)\cdot\frac{(y-x)}{\varepsilon}\right|^2\varepsilon^{-d}\,\mathrm{d}y \\
&\qquad = \frac{1}{\sigma_\eta}\int_{B_1(0)}\eta(|z|)\left|\nabla u(x)\cdot z\right|^2\,\mathrm{d}z \qquad \text{by putting } z=\frac{y-x}{\varepsilon} \\
&\qquad \overset{(\star\star)}{=} |\nabla u(x)|^2.
\end{aligned}$$

Therefore, we have (using Equation (4.5)):

$$\begin{aligned}
\frac{1}{\sigma_\eta \varepsilon^2}\int_{\Omega^\varepsilon}\int_{\Omega\cap B_\varepsilon(x)}\eta_\varepsilon(|x-y|)|\nabla u(x)\cdot(y-x)|^2\,\mathrm{d}y\rho(x)^2\,\mathrm{d}x &= \int_{\Omega^\varepsilon}|\nabla u(x)|^2\rho(x)^2\,\mathrm{d}x \\
&= E(u)+\|u\|^2_{C^2}\varepsilon O(1).
\end{aligned}$$

We are almost there ! Actually, we can apply the same arguments as before to show that if $x\in\partial^\varepsilon\Omega$,

$$\frac{1}{\sigma_\eta \varepsilon^2}\int_{\Omega\cap B_\varepsilon(x)}\eta_\varepsilon(|x-y|)|\nabla u(x)\cdot(y-x)|^2\,\mathrm{d}y = \|u\|^2_{C^2}O(1)$$

which - using Equation (4.5) again - implies that

$$\frac{1}{\sigma_\eta \varepsilon^2}\int_{\partial^\varepsilon\Omega}\int_{\Omega\cap B_\varepsilon(x)}\eta_\varepsilon(|x-y|)|\nabla u(x)\cdot(y-x)|^2\,\mathrm{d}y\rho(x)^2\,\mathrm{d}x = \|u\|^2_{C^2}\varepsilon O(1)$$

Finally, putting everything together, we have

$$E_\varepsilon(u) = E(u) + \|u\|_{C^2}^2 \varepsilon O(1),$$

hence the result. □

Exercise 4.3. By a change of variables, show the equality ($\star\star$).

Hence, by remarking that $\mathrm{Lip}(u) \leq \|u\|_{C^2}$ and plugging the two lemmas together, we arrive at Equation (4.2):

$$\mathbb{P}\left(|E_{n,\varepsilon}(u) - E(u)| \leq C_1 \|u\|_{C^2}^2 \left(\frac{1}{n} + \lambda + \varepsilon\right)\right) \geq 1 - 2\exp(-C_2 n \varepsilon^d \lambda^2)$$

Remark 4.1. Contrary to the consistency results we showed before, we have a strange additional term in $\frac{1}{n}$ that appears and prevents us to take λ of ε as small as we would like. This is due to the fact that for f symmetric and $x_1, \ldots, x_n$ independent random variables, the random variable $\hat{\mu} := \frac{1}{n^2}\sum_{i,j} f(x_i, x_j)$ is a *biased* estimator of $\mu = \mathbb{E}f(x_1, x_2)$, meaning that $\mathbb{E}(\hat{\mu}) \neq \mu$. Indeed, counting the non-independent diagonal terms makes the estimator biased: if $\mu_d := \mathbb{E}f(x_1, x_1)$, one can show that

$$E(\hat{\mu}) = \left(1 - \frac{1}{n}\right)\mu + \frac{1}{n}\mu_d.$$

The U-statistic U_n is the correct, unbiased estimator of $\hat{\mu}$. In our case, if we had define

$$E_{n,\varepsilon}(u) := \frac{1}{\sigma_\eta n(n-1)\varepsilon^2} \sum_{i,j=1}^{n} \eta_\varepsilon\left(|x_i - x_j|\right) |u(x_i) - u(x_j)|^2,$$

then Theorem 4.2 would give (for $\lambda = \varepsilon$) that

$$\mathbb{P}\left(|E_{n,\varepsilon}(u) - E(u)| \leq C_1 \|u\|_{C^2}^2 \varepsilon\right) \geq 1 - 2\exp(-C_2 n \varepsilon^{d+2})$$

Let us take some time to unwrap the meaning of Equation (4.2). First of all, we remark that we can not get a convergence faster to $\frac{1}{n}$; hence, no need to take α or ϵ to 0 faster than this rate. Moreover, this rate can not be attained with positive probability, since we would need to take $\lambda_n = \varepsilon_n = \frac{1}{n}$, and the probability would become $1 - 2\exp\left(-\frac{C_2}{n^{d+1}}\right) \to [n \to \infty] - \infty$. To ensure convergence, one can for instance take $\lambda_n = \varepsilon_n = \left(\frac{1}{n}\right)^{\frac{1}{d+3}}$ and apply the Borel-Cantelli lemma to get that

$$E_{n,\varepsilon_n}(u) \xrightarrow[n\to\infty]{} E(u)$$

almost surely.

However, this is not what we wanted in the first place; indeed, we are interested in the convergence of the solution of the graph Laplacian toward the one of the weighted Laplacian. In variational terms, we want to show that the *minimizers* of the graph Dirichlet energy converges to the *minimizers* of the weighted Dirichlet energy, i.e. (informally):

$$\underset{u=g \text{ on } \partial^\varepsilon\Omega}{\arg\min}\ E_{n,\varepsilon}(u) \xrightarrow[n\to\infty,\varepsilon\to 0]{} \underset{u=g \text{ on } \partial\Omega}{\arg\min}\ E(u)$$

This kind convergence of minimizers is ubiquitous in the field of *Calculus of Variations*, and is ensured if we can show that the sequence of functionals $E_{n,\varepsilon}$ converges to E in the sense of Γ-*convergence*.

4.2 Convergence of the minima

Definition 4.1 (Γ-convergence)**.** A sequence of functionals $J_n : X \to [0, \infty]$ defined on a metric space X is said to Γ-converge to $J : X \to [0, \infty]$ if

- The **liminf inequality** holds: For all sequences $(u_n)_{n\in\mathbb{N}}$ converging to some u in X it holds
$$J(u) \leq \liminf_{n\to\infty} J_n(u_n).$$
- The **limsup inequality** holds: For all $u \in X$ there exists a sequence $(u_n)_{n\in\mathbb{N}}$ converging to u such that
$$J(u) \geq \limsup_{n\to\infty} J_n(u_n).$$

Proposition 4.1. *Assume that the functionals $J_n : X \to [0, \infty]$ Γ-converge to $J : X \to [0, \infty]$ and let $u_n \in \arg\min J_n$. If u_n converges to u then $u \in \arg\min J$.*

Proof. To prove the result let us take an arbitrary $v \in X$. Thanks to the limsup inequality there exists a sequence $(v_n)_{n\in\mathbb{N}}$ converging to v such that $\limsup_{n\to\infty} J_n(v_n) \leq J(v)$. Using also the liminf inequality and the minimality of u_n it follows:
$$J(u) \leq \liminf_{n\to\infty} J_n(u_n) \leq \limsup_{n\to\infty} J_n(v_n) \leq J(v).$$
Since v was arbitrary, this proves $u \in \arg\min J$. □

Exercise 4.4. This exercise discusses some useful properties and examples for Γ-convergence.

- Let $J_n : X \to [0, \infty]$ Γ-converge to $J : X \to [0, \infty]$ and $F : X \to [0, \infty]$ be continuous. Then $J_n + F$ Γ-converges to $J + F$.
- Compute the Γ-limit of the functions $J_n : \mathbb{R} \to \mathbb{R}$ defined via
$$J_n(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0 \end{cases} \qquad \forall n \in \mathbb{N}.$$
- Prove that $J_n : \mathbb{R}^d \to \mathbb{R}$ defined by $J_n(x) = \left(\sum_{i=1}^d |x_i|^n\right)^{\frac{1}{n}}$ Γ-converges to $F(x) := \max_{i=1}^d |x_i|$.
- Compute the Γ-limit of the functions $J_n : \mathbb{R} \to \mathbb{R}$ defined via $J_n(x) = \sin(2\pi n x)$.

* Prove that any Γ-limit is lower semicontinuous.

Let us now turn back to the Laplace learning problem. In our case, we will only show the nonlocal to local Γ-convergence. It is also possible to show the discrete to nonlocal one, but it requires tools from optimal transport that are out of the scope of this lecture (see [GS16]).

Let us now apply Γ-convergence to our problem. From now on, we assume that $\Omega \subset \mathbb{R}^d$ is a bounded open set with C^1 boundary and $g : \Omega \to \mathbb{R}$ is Lipschitz. We define the nonlocal, boundary-constrained Dirichlet energies $F_\varepsilon : L^2(\Omega) \to [0, \infty]$:

$$F_\varepsilon(u) := \begin{cases} E_\varepsilon(u) & \text{if } u = g \text{ on } \partial^\varepsilon\Omega, \\ \infty & \text{otherwise} \end{cases}$$

where $u \in L^2(\Omega)$. The limiting Dirichlet energy is defined as

$$F(u) := \begin{cases} E(u) & \text{if } u \in H^1(\Omega) \text{ and } u = g \text{ on } \partial\Omega, \\ \infty & \text{otherwise,} \end{cases}$$

We will show the following theorem:

Theorem 4.3. *Let $\Omega \subset \mathbb{R}^d$ be an open and bounded set with C^1 boundary. Assume that $g \in \mathrm{Lip}(\Omega)$ and $\eta : [0, \infty) \to [0, \infty)$ is such that $\eta \leq C1_{[0,1)}$ for some C. Assume that $\rho \in C^1(\bar{\Omega})$. Then*

$$F_\varepsilon \xrightarrow[\varepsilon \to 0]{\Gamma} F.$$

Remark 4.2. By $F_\varepsilon \xrightarrow[\varepsilon \to 0]{\Gamma} F$, we mean that for every sequence $\varepsilon_n \to 0$, we have that $F_{\varepsilon_n} \xrightarrow[n \to \infty]{\Gamma} F$. In what follows, we assume to have chosen such a sequence ε_n.

Remark 4.3. We can easily relax the assumption on ρ to get $\rho \in \mathrm{Lip}(\Omega)$, and then further relax it to $\rho \in C(\bar{\Omega})$ by approximating it by Lipschitz functions. See [GS16] for more information.

In order to show this theorem, we will need some classical results about Sobolev spaces. The first one is called the Rellich theorem and states that the inclusion operator from H^1 to L^2 is compact:

Theorem 4.4 (Rellich)**.** *Let $\Omega \subset \mathbb{R}^d$ be an open and bounded set with Lipschitz boundary. Let $u_n \in H^1(\Omega)$ be bounded. Then, there exists $u \in H^1(\Omega)$ such that up to a subsequence, $u_n \xrightarrow[n \to \infty]{L^2} u$.*

We won't prove it here, but the idea is to use the density of smooth functions in $H^1(\Omega)$ then use the Arzéla-Ascoli theorem. A full proof can be found in [Bre]. We will also need the following technical lemma:

Lemma 4.3. *Let $u_n \in H^1(\Omega)$ and $u \in H^1(\Omega)$ be such that $u_n \xrightarrow[n \to \infty]{L^2} u$. Then*

$$\int_\Omega |\nabla u|^2 \rho^2 \leq \liminf_{n \to \infty} \int_\Omega |\nabla u_n|^2 \rho^2.$$

Exercise 4.5. We will show the previous lemma as an exercise.

1. Let $\phi \in C_c^\infty(\Omega, \mathbb{R}^d)$. Using Green's formula, show that

$$\int_\Omega (\nabla u_n \cdot \phi)\rho^2 \xrightarrow[n \to \infty]{} \int_\Omega (\nabla u \cdot \phi)\rho^2$$

2. Show that for $u \in H^1(\Omega)$,

$$\int_\Omega |\nabla u|^2 \rho^2 = \sup_{\substack{v \in H^1(\Omega) \\ \int_\Omega |\nabla v|^2 \rho^2 = 1}} \left(\int_\Omega (\nabla u \cdot \nabla v) \rho^2 \right)^2$$

3. Deduce that

$$\int_\Omega |\nabla v|^2 \rho^2 = \sup_{\substack{\phi \in C_c^\infty(\Omega, \mathbb{R}^d) \\ \int_\Omega |\phi|^2 \rho^2 = 1}} \left(\int_\Omega (\nabla v \cdot \phi) \rho^2 \right)^2$$

4. Using the previous observations, show the result.

One must now show the $\Gamma - \liminf$ and $\Gamma - \limsup$ properties. We will start with the easiest one, which in this case is the $\liminf$. For this, we need a technical lemma, which is an adaptation of Lemma 4.2.

Lemma 4.4. *Let $u, u_\varepsilon \in C^2(\bar{\Omega})$ for all $\varepsilon > 0$ be such that $\sup_\varepsilon \|u_\varepsilon\|_{C^2} < \infty$ and $u_\varepsilon \xrightarrow[\varepsilon \to 0]{H^1} u$. Then*

$$E_\varepsilon(u_\varepsilon) \xrightarrow[\varepsilon \to 0]{} E(u).$$

Exercise 4.6. Show Lemma 4.4.

Lemma 4.5. *Let $(u_\varepsilon)_\varepsilon \in L^2(\Omega)$ be a sequence that converges to $u \in L^2$. Then*

$$\liminf_{\varepsilon \to 0} F_\varepsilon(u_\varepsilon) \geq F(u).$$

The idea of the proof is to smooth the functions u_ε in order to be able to apply Lemma 4.4.

Proof. In the case where $\liminf_{\varepsilon \to 0} F_\varepsilon(u_\varepsilon) = \infty$, there is nothing to show. Assume that $\liminf_{\varepsilon \to 0} F_\varepsilon(u_\varepsilon) < \infty$. Then, up to a subsequence, we have that

$$\lim_{\varepsilon \to 0} F_\varepsilon(u_\varepsilon) = \liminf_{\varepsilon \to 0} F_\varepsilon(u_\varepsilon) < \infty.$$

Let $\delta' > \delta > 0$. Assume that we extend every u_ε by (a Lipschitz extension of) g in $\mathbb{R}^d$, and define $u_{\varepsilon,\delta} := J_\delta * u_\varepsilon$ and $u_{0,\delta} := J_\delta * u$ where J is a positive mollifier supported on the unit ball and $J_\delta(x) := \delta^{-d} J(x/\delta)$. Then, we have $\nabla u_{\varepsilon,\delta} = \nabla J_\delta * u_\varepsilon$ and for $x \in \Omega$,

$$|\nabla u_{\varepsilon,\delta}(x)| \leq \int_\Omega |\nabla J_\delta(x - z)||u_\varepsilon(z)| \,\mathrm{d}z \leq \|\nabla J_\delta\|_{L^2} \|u_\varepsilon\|_{L^2}$$

hence $\|\nabla u_{\varepsilon,\delta}\|_\infty \leq \|\nabla J_\delta\|_{L^2} \|u_\varepsilon\|_{L^2}$. Similarly, we can show that $\|D^2 u_{\varepsilon,\delta}\|_\infty \leq \|D^2 J_\delta\|_{L^2} \|u_\varepsilon\|_{L^2}$. It follows that for a fixed δ,

$$\sup_\varepsilon \|u_{\varepsilon,\delta}\|_{C^2} < \infty. \tag{4.6}$$

Using the same estimate as before, we have that

$$\int_\Omega |\nabla u_{\varepsilon,\delta} - \nabla u_{0,\delta}|^2 \leq \int_\Omega \|\nabla J_\delta\|_{L^2}^2 \|u_\varepsilon - u\|_{L^2}^2 = |\Omega| \|\nabla J_\delta\|_{L^2}^2 \|u_\varepsilon - u\|_{L^2}^2.$$

Since by assumption, $\|u_\varepsilon - u\|_{L^2} \to 0$, we have that

$$\|u_{\varepsilon,\delta} - u_{0,\delta}\|_{H^1} \xrightarrow[\varepsilon\to 0]{} 0 \tag{4.7}$$

We can write:

$$\begin{aligned}
E_\varepsilon(u_\varepsilon) &= \frac{1}{\sigma_\eta \varepsilon^2}\int_{\Omega\times\Omega} \eta_\varepsilon(|x-y|)|u_\varepsilon(x)-u_\varepsilon(y)|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y\\
&= \frac{1}{\sigma_\eta \varepsilon^2}\int_{\Omega\times\Omega}\int_{\mathbb{R}^d} J_\delta(z)\eta_\varepsilon(|x-y|)|u_\varepsilon(x)-u_\varepsilon(y)|^2\rho(x)\rho(y)\,\mathrm{d}z\,\mathrm{d}x\,\mathrm{d}y\\
&= \frac{1}{\sigma_\eta \varepsilon^2}\int_{\Omega\times\Omega}\int_{\mathbb{R}^d} J_\delta(z)\eta_\varepsilon(|x-y|)|u_\varepsilon(x)-u_\varepsilon(y)|^2(\rho(x)\rho(y)-\rho(x+z)\rho(y+z))\,\mathrm{d}z\,\mathrm{d}x\,\mathrm{d}y\\
&+ \frac{1}{\sigma_\eta \varepsilon^2}\int_{\Omega\times\Omega}\int_{\mathbb{R}^d} J_\delta(z)\eta_\varepsilon(|x-y|)|u_\varepsilon(x)-u_\varepsilon(y)|^2\rho(x+z)\rho(y+z)\,\mathrm{d}z\,\mathrm{d}x\,\mathrm{d}y\\
&=: a_{\varepsilon,\delta} + b_{\varepsilon,\delta}
\end{aligned}$$

We can estimate $a_{\varepsilon,\delta}$ by using the fact that

$$\begin{aligned}
|\rho(x)\rho(y)-\rho(x+z)\rho(y+z)| &\le |\rho(x)\rho(y)-\rho(x+z)\rho(y)| + |\rho(x+z)\rho(y)-\rho(x+z)\rho(y+z)|\\
&\le 2\|\rho\|_\infty \operatorname{Lip}(\rho)|z| = C|z|.
\end{aligned}$$

This leads to

$$\begin{aligned}
|a_{\varepsilon,\delta}| &\le \frac{1}{\sigma_\eta \varepsilon^2}\int_{\Omega\times\Omega}\int_{B_\delta(0)} J_\delta(z)\eta_\varepsilon(|x-y|)|u_\varepsilon(x)-u_\varepsilon(y)|^2|\rho(x)\rho(y)-\rho(x+z)\rho(y+z)|\,\mathrm{d}z\,\mathrm{d}x\,\mathrm{d}y\\
&\le \frac{C\delta}{\sigma_\eta \varepsilon^2}\int_{\Omega\times\Omega} \eta_\varepsilon(|x-y|)|u_\varepsilon(x)-u_\varepsilon(y)|^2\,\mathrm{d}x\,\mathrm{d}y\\
&\le \frac{C\delta}{c_\rho^2\sigma_\eta \varepsilon^2}\int_{\Omega\times\Omega} \eta_\varepsilon(|x-y|)|u_\varepsilon(x)-u_\varepsilon(y)|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \qquad (\rho \text{ bounded from below})\\
&\le \frac{C\delta}{c_\rho^2} E_\varepsilon(u_\varepsilon) \le C\delta
\end{aligned}$$

since $E_\varepsilon(u_\varepsilon)$ is bounded. One must now estimate $b_{\varepsilon,\delta}$. Using the change of variables $\hat{y} = y+z$ and $\hat{x} = x+z$ and the fact that $\Omega^{\delta'} - z \subset \Omega$ for $|z| \le \delta$, we have

$$b_{\varepsilon,\delta} \ge \frac{1}{\sigma_\eta \varepsilon^2}\int_{\Omega^{\delta'}\times\Omega^{\delta'}}\int_{\mathbb{R}^d} J_\delta(z)\eta_\varepsilon(|\hat{x}-\hat{y}|)|u_\varepsilon(\hat{x}-z)-u_\varepsilon(\hat{y}-z)|^2\rho(\hat{x})\rho(\hat{y})\,\mathrm{d}z\,\mathrm{d}\hat{x}\,\mathrm{d}\hat{y}$$

Using Jensen inequality on the probability measure $J_\delta(z)\,\mathrm{d}z$, we get that

$$\begin{aligned}
&\int_{\mathbb{R}^d} J_\delta(z)|u_\varepsilon(\hat{x}-z)-u_\varepsilon(\hat{y}-z)|^2\,\mathrm{d}z\\
&\qquad\ge \left|\int_{\mathbb{R}^d} J_\delta(z)\eta_\varepsilon(|\hat{x}-\hat{y}|)(u_\varepsilon(\hat{x}-z)-u_\varepsilon(\hat{y}-z))\,\mathrm{d}z\right|^2\\
&\qquad\ge |J_\delta * u_\varepsilon(\hat{x}) - J_\delta * u_\varepsilon(\hat{y})|^2\\
&\qquad= |u_{\varepsilon,\delta}(\hat{x}) - u_{\varepsilon,\delta}(\hat{y})|^2
\end{aligned}$$

which leads to

$$b_{\varepsilon,\delta} \geq \frac{1}{\sigma_\eta \varepsilon^2} \int_{\Omega^{\delta'} \times \Omega^{\delta'}} \eta_\varepsilon(|\hat{x} - \hat{y}|)|u_{\varepsilon,\delta}(\hat{x}) - u_{\varepsilon,\delta}(\hat{y})|^2 \rho(\hat{x})\rho(\hat{y}) \,\mathrm{d}\hat{x} \,\mathrm{d}\hat{y}.$$

Hence, recalling that we have Equation (4.6) and Equation (4.7), we can use Lemma 4.4 on $\Omega^{\delta'}$ to get that

$$\frac{1}{\sigma_\eta \varepsilon^2} \int_{\Omega^{\delta'} \times \Omega^{\delta'}} \eta_\varepsilon(|\hat{x} - \hat{y}|)|u_{\varepsilon,\delta}(\hat{x}) - u_{\varepsilon,\delta}(\hat{y})|^2 \rho(\hat{x})\rho(\hat{y}) \,\mathrm{d}x \,\mathrm{d}y \xrightarrow[\varepsilon \to 0]{} \int_{\Omega^{\delta'}} |\nabla u_{0,\delta}|^2 \rho^2 \,\mathrm{d}x.$$

This leads to :

$$\begin{aligned}
\liminf_\varepsilon E_\varepsilon(u_\varepsilon) &\geq \liminf_\varepsilon a_{\varepsilon,\delta} + \liminf_\varepsilon \frac{1}{\sigma_\eta \varepsilon^2} \int_{\Omega^{\delta'} \times \Omega^{\delta'}} \eta_\varepsilon(|\hat{x} - \hat{y}|)|u_{\varepsilon,\delta}(\hat{x}) - u_{\varepsilon,\delta}(\hat{y})|^2 \rho(\hat{x})\rho(\hat{y}) \,\mathrm{d}x \,\mathrm{d}y \\
&\geq \liminf_\varepsilon a_{\varepsilon,\delta} + \int_{\Omega^{\delta'}} |\nabla u_{0,\delta}|^2 \rho^2 \,\mathrm{d}x
\end{aligned}$$

Since we know that $\liminf_\varepsilon a_{\varepsilon,\delta} \xrightarrow[\delta \to 0]{} 0$, we are left to show that

$$\liminf_\delta \int_{\Omega^{\delta'}} |\nabla u_{0,\delta}|^2 \rho^2 \,\mathrm{d}x \geq \int_{\Omega^{\delta'}} |\nabla u|^2 \rho^2 \,\mathrm{d}x$$

However, we do not know a priori that $u \in H^1(\Omega^{\delta'})$! It is, however, actually the case. Since, for all $\delta > 0$,

$$\int_{\Omega^{\delta'}} |\nabla u_{0,\delta}|^2 \rho^2 \,\mathrm{d}x \leq \liminf_\varepsilon E_\varepsilon(u_\varepsilon)$$

we have that $(u_{0,\delta})_\delta$ is bounded in $H^1(\Omega^{\delta'})$. Hence, using Rellich's Theorem 4.4, there exists $v \in H^1(\Omega^{\delta'})$ such that up to a subsequence, $u_{0,\delta} \xrightarrow[\delta \to 0]{L^2} v$. However, we know that by construction $u_{0,\delta} \xrightarrow[\delta \to 0]{L^2} u$ which means that $v = u \in H^1(\Omega^{\delta'})$. Finally, using Lemma 4.3, we have

$$\liminf \int_{\Omega^{\delta'}} |\nabla u_{0,\delta}|^2 \rho^2 \,\mathrm{d}x \geq \int_{\Omega^{\delta'}} |\nabla u|^2 \rho^2 \,\mathrm{d}x.$$

This leads to

$$\liminf_\varepsilon E_\varepsilon(u_\varepsilon) \geq \int_{\Omega^{\delta'}} |\nabla u|^2 \rho^2 \,\mathrm{d}x$$

and by taking an increasing sequence of $\Omega^{\delta'}$, we get that

$$\liminf_\varepsilon E_\varepsilon(u_\varepsilon) \geq \int_\Omega |\nabla u|^2 \rho^2 \,\mathrm{d}x = E(u)$$

and $u \in H^1(\Omega)$.

Now, we must still show that $u = g$ on $\partial\Omega$. Knowing that $u \in H^1(\Omega)$, we now have that $u_{0,\delta} \xrightarrow[\delta \to 0]{H^1} u$. Hence, for all $\varepsilon, \delta > 0$, we have

$$\|u_{\varepsilon,\delta} - u\|_{H^1} \leq \|u_{\varepsilon,\delta} - u_{0,\delta}\|_{H^1} + \|u_{0,\delta} - u\|_{H^1} \leq C\|u_\varepsilon - u\|_{H^1} + \|u_{0,\delta} - u\|_{H^1}$$

where in the second inequality, we used the Young inequality for convolution to get that

$$\|u_{\varepsilon,\delta} - u_{0,\delta}\|_{L^2} \leq \|u_\varepsilon - u\|_{L^2} \qquad \text{and} \qquad \|\nabla u_{\varepsilon,\delta} - \nabla u_{0,\delta}\|_{L^2} \leq \|\nabla u_\varepsilon - \nabla u\|_{L^2}$$

Since $u_\varepsilon = g$ near the boundary, for $\delta^\varepsilon < \varepsilon$ small enough, we have

$$\int_{\partial\Omega} |u_{\varepsilon,\delta^\varepsilon} - g|^2 \leq \varepsilon$$

Hence $u_{\varepsilon,\delta^\varepsilon} \xrightarrow[\varepsilon\to 0]{H^1} u$ and by continuity of the trace operator in H^1, we have $u = g$ and finally,

$$F(u) = E(u) \leq \liminf_{\varepsilon} E_\varepsilon(u_\varepsilon) \leq \liminf_{\varepsilon} F_\varepsilon(u_\varepsilon).$$

□

Let us now turn to the $\Gamma - \limsup$. Often, we construct the recovery sequence of the $\Gamma - \limsup$ by taking the constant sequence $u_\varepsilon = u$, the limiting function. However, in our case, the boundary conditions on the thinckened boundary $\partial^\varepsilon\Omega$ prevents us from doing that since, in general, $u \neq g$ on $\partial^\varepsilon\Omega$. Hence, we will need to interpolate between u and g near the boundary, and show that we can control the distance between u and g in $\partial^\varepsilon\Omega$. This control will be provided by the Hardy inequality, which we will assume:

Theorem 4.5 (Hardy Inequality)**.** *Let $\Omega \subset \mathbb{R}^n$ be an open set with non-empty boundary. There exists a constant $C > 0$ such that for all $u \in H_0^1(\Omega)$,*

$$\int_\Omega \frac{|u(x)|^2}{\operatorname{dist}(x,\partial\Omega)^2}\, \mathrm{d}x \leq C \int_\Omega |\nabla u|^2. \tag{4.8}$$

Corollary 4.1. *Let $\Omega \subset \mathbb{R}^n$ be an open set with non-empty boundary and $u \in H_0^1(\Omega)$. Then*

$$\frac{1}{\varepsilon^2} \int_{\partial^\varepsilon\Omega} |u|^2 \xrightarrow[\varepsilon\to 0]{} 0.$$

Proof. Using Hardy, we know that $\frac{|u|^2}{d(\cdot,\partial\Omega)} \in L^1(\Omega)$. Moreover, for all $x \in \Omega$,

$$\frac{1}{\varepsilon^2} 1_{\partial^\varepsilon\Omega}(x)|u(x)|^2 \leq \frac{|u(x)|^2}{\operatorname{dist}(x,\partial\Omega)}.$$

Using the dominated convergence theorem, we get the result. □

We can now show the limsup:

Lemma 4.6. *Let $u \in L^2(\Omega)$. Then, there exist a sequence $u_\varepsilon \in L^2(\Omega)$ such that $u_\varepsilon \xrightarrow[\varepsilon\to 0]{L^2} u$ and*

$$F(u) \geq \limsup_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon).$$

The idea is to take the constant sequence $u_\varepsilon \equiv u$. However, we need to be careful about the boundary conditions on the thick boundary $\partial^\varepsilon\Omega$ in order for F_ε to be finite.

Proof. First, if u is such that $F(u) = \infty$, then there is nothing to show. Hence we can consider the case where $F(u) < \infty$. In this case, we know that $u \in H^1(\Omega)$ and $u = g$ on $\partial\Omega$. For $t > 0$, let

$$\phi(t) = \begin{cases} 1 \text{ if } t \in [0,1] \\ 2 - t \text{ if } t \in (1,2) \\ 0 \text{ otherwise} \end{cases}$$

and define $\xi_\varepsilon(x) := \phi\left(\frac{d(x,\partial\Omega)}{2\varepsilon}\right)$. The recovery sequence will then be

$$u_\varepsilon = (1-\xi_\varepsilon)u + \xi_\varepsilon g,$$

which satisfies $u_\varepsilon = g$ on $\partial^\varepsilon\Omega$ and $u_\varepsilon = u$ on $\Omega^{2\varepsilon}$. Using the dominated convergence theorem, one can easily show that $u_\varepsilon \xrightarrow[\varepsilon\to 0]{L^2} u$.

Let us compute

$$F_\varepsilon(u_\varepsilon) = \frac{1}{\sigma_\eta \varepsilon^2} \int_{\Omega\times\Omega} \eta_\varepsilon(|x-y|)|u_\varepsilon(x) - u_\varepsilon(y)|^2 \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y.$$

We can split the integrals into integrals over $\partial^{2\varepsilon}\Omega$ and $\Omega^{2\varepsilon}$ in the following way:

$$\int_\Omega\int_\Omega = \int_{\partial^{2\varepsilon}\Omega}\int_\Omega + \int_{\Omega^{2\varepsilon}}\int_{\partial^{2\varepsilon}\Omega} + \int_{\Omega^{2\varepsilon}}\int_{\Omega^{2\varepsilon}}$$

The first term is then

$$\begin{aligned}
A_\varepsilon :=& \frac{1}{\sigma_\eta\varepsilon^2}\int_{\partial^{2\varepsilon}\Omega}\int_\Omega \eta_\varepsilon(|x-y|)|u_\varepsilon(x) - u_\varepsilon(y)|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
=& \frac{1}{\sigma_\eta\varepsilon^2}\int_{\partial^{2\varepsilon}\Omega}\int_\Omega \eta_\varepsilon(|x-y|)|\xi_\varepsilon(x)(g(x)-u(x)) - \xi_\varepsilon(y)(g(y)-u(y)) + u(x) - u(y)|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
\leq& \frac{2}{\sigma_\eta\varepsilon^2}\int_{\partial^{2\varepsilon}\Omega}\int_\Omega \eta_\varepsilon(|x-y|)|\xi_\varepsilon(x)(g(x)-u(x)) - \xi_\varepsilon(y)(g(y)-u(y))|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \quad \text{(A.1)} \\
&+ \frac{2}{\sigma_\eta\varepsilon^2}\int_{\partial^{2\varepsilon}\Omega}\int_\Omega \eta_\varepsilon(|x-y|)|u(x)-u(y)|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \quad \text{(A.2)}
\end{aligned}$$

where we used that $(a+b)^2 \leq 2(a^2+b^2)$. Set $v := g - u \in H_0^1(\Omega)$ and let study $(A.1)$ first. We have

$$\begin{aligned}
(A.1) :=& \frac{2}{\sigma_\eta\varepsilon^2}\int_{\partial^{2\varepsilon}\Omega}\int_\Omega \eta_\varepsilon(|x-y|)|\xi_\varepsilon(x)v(x) - \xi_\varepsilon(y)v(y)|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
=& \frac{2}{\sigma_\eta\varepsilon^2}\int_{\partial^{2\varepsilon}\Omega}\int_\Omega \eta_\varepsilon(|x-y|)|\xi_\varepsilon(x)v(x) - \xi_\varepsilon(x)v(y) + \xi_\varepsilon(x)v(y) - \xi_\varepsilon(y)v(y)|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
\leq& \frac{4}{\sigma_\eta\varepsilon^2}\int_{\partial^{2\varepsilon}\Omega}\int_\Omega \eta_\varepsilon(|x-y|)|\xi_\varepsilon(x)|^2|v(x)-v(y)|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
&+ \frac{4}{\sigma_\eta\varepsilon^2}\int_{\partial^{2\varepsilon}\Omega}\int_\Omega \eta_\varepsilon(|x-y|)|\xi_\varepsilon(x)-\xi_\varepsilon(y)|^2|v(y)|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
\leq& \frac{4}{\sigma_\eta\varepsilon^2}\int_{\partial^{2\varepsilon}\Omega}\int_\Omega \eta_\varepsilon(|x-y|)|v(x)-v(y)|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
&+ \frac{C}{\sigma_\eta\varepsilon^{2+d}}\int_{\partial^{2\varepsilon}\Omega}\int_{\Omega\cap B_\varepsilon(y)} Lip(\xi_\varepsilon)^2|x-y|^2|v(y)|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y
\end{aligned}$$

The second term of the sum goes to 0 when $\varepsilon \to 0$ since using that $Lip(\xi_\varepsilon) = 1/2\varepsilon$, we have

$$\begin{aligned}
&\frac{C}{\sigma_\eta \varepsilon^{2+d}} \int_{\partial^{2\varepsilon}\Omega} \int_{\Omega \cap B_\varepsilon(y)} Lip(\xi_\varepsilon)^2 |x-y|^2 |v(y)|^2 \rho(x)\rho(y) \,\mathrm{d}x \,\mathrm{d}y \\
&\leq \frac{C}{\sigma_\eta \varepsilon^{2+d}} \int_{\partial^{2\varepsilon}\Omega} |B_\varepsilon(y)| |v(y)|^2 \rho(y) \,\mathrm{d}x \,\mathrm{d}y \\
&\leq \frac{C}{\sigma_\eta \varepsilon^{2+d}} \int_{\partial^{2\varepsilon}\Omega} |B_\varepsilon(y)| |v(y)|^2 \rho(y) \,\mathrm{d}y \\
&\leq \frac{C}{\sigma_\eta \varepsilon^{2}} \int_{\partial^{2\varepsilon}\Omega} |v(y)|^2 \rho(y) \,\mathrm{d}y
\end{aligned}$$

Using Corollary 4.1, we get that this last term goes to 0 with ε. With the same arguments, we get that

$$\frac{4}{\sigma_\eta \varepsilon^2} \int_{\partial^{2\varepsilon}\Omega} \int_\Omega \eta_\varepsilon(|x-y|) |v(x) - v(y)|^2 \rho(x)\rho(y) \,\mathrm{d}x \,\mathrm{d}y \xrightarrow[\varepsilon \to 0]{} 0.$$

Let us consider the $(A.2)$ term. We have

$$\begin{aligned}
(A.2) =& \frac{2}{\sigma_\eta \varepsilon^2} \int_{\partial^{2\varepsilon}\Omega} \int_\Omega \eta_\varepsilon(|x-y|) |u(x) - u(y)|^2 \rho(x)\rho(y) \,\mathrm{d}x \,\mathrm{d}y \\
\leq& \frac{2}{\sigma_\eta \varepsilon^2} \int_{\partial^{2\varepsilon}\Omega} \int_\Omega \eta_\varepsilon(|x-y|) |u(x) - g(x)|^2 \rho(x)\rho(y) \,\mathrm{d}x \,\mathrm{d}y \\
&+ \frac{2}{\sigma_\eta \varepsilon^2} \int_{\partial^{2\varepsilon}\Omega} \int_\Omega \eta_\varepsilon(|x-y|) |g(y) - u(y)|^2 \rho(x)\rho(y) \,\mathrm{d}x \,\mathrm{d}y \\
&+ \frac{2}{\sigma_\eta \varepsilon^2} \int_{\partial^{2\varepsilon}\Omega} \int_\Omega \eta_\varepsilon(|x-y|) |g(x) - g(y)|^2 \rho(x)\rho(y) \,\mathrm{d}x \,\mathrm{d}y
\end{aligned}$$

The first two terms of the sum goes to 0 using once again **[coro:hardy]**. For the third term, we use that g is Lipschitz:

$$\begin{aligned}
&\frac{2}{\sigma_\eta \varepsilon^2} \int_{\partial^{2\varepsilon}\Omega} \int_\Omega \eta_\varepsilon(|x-y|) |g(x) - g(y)|^2 \rho(x)\rho(y) \,\mathrm{d}x \,\mathrm{d}y \\
&\qquad \leq \frac{C}{\sigma_\eta \varepsilon^{2+d}} \int_{\partial^{2\varepsilon}\Omega} \int_{\Omega \cap B_\varepsilon(y)} Lip(g)^2 \varepsilon^2 \rho(x)\rho(y) \,\mathrm{d}x \,\mathrm{d}y \\
&\qquad \leq \frac{C}{\sigma_\eta \varepsilon^{d}} \int_{\partial^{2\varepsilon}\Omega} |B_\varepsilon(y)| \,\mathrm{d}y \\
&\qquad \leq C \int_{\partial^{2\varepsilon}\Omega} \mathrm{d}y \xrightarrow[\varepsilon \to 0]{} 0.
\end{aligned}$$

Consequently,

$$\limsup_{\varepsilon \to 0} A_\varepsilon = 0.$$

Using the same arguments, we can show that the same holds for the third term

$$C_\varepsilon := \frac{1}{\sigma_\eta \varepsilon^2} \int_{\Omega^{2\varepsilon}} \int_{\partial^{2\varepsilon}\Omega} \eta_\varepsilon(|x-y|) |u_\varepsilon(x) - u_\varepsilon(y)|^2 \rho(x)\rho(y) \,\mathrm{d}x \,\mathrm{d}y.$$

Let us now focus on the remaining term

$$B_\varepsilon := \frac{1}{\sigma_\eta \varepsilon^2} \int_{\Omega^{2\varepsilon}} \int_{\Omega^{2\varepsilon}} \eta_\varepsilon(|x-y|)|u_\varepsilon(x) - u_\varepsilon(y)|^2 \rho(x)\rho(y) \,\mathrm{d}x \,\mathrm{d}y$$

For this, we will use the density of smooth functions in H^1 and the consistency that we proved in Lemma 4.2. First, we observe that

$$B_\varepsilon = \frac{1}{\sigma_\eta \varepsilon^2} \int_{\Omega^{2\varepsilon}} \int_{\Omega^{2\varepsilon}} \eta_\varepsilon(|x-y|)|u(x) - u(y)|^2 \rho(x)\rho(y) \,\mathrm{d}x \,\mathrm{d}y \le E_\varepsilon(u).$$

We can easily show that both E_ε and E are continuous funtionals on $H^1(\Omega)$. Hence, for every $u \in H^1(\Omega)$ and every $\delta > 0$, there exists $\delta' > 0$ such that $\|u - v\|_{H^1} < \delta'$ implies that $|E_\varepsilon(u) - E_\varepsilon(v)| < \delta$ and $|E(u) - E(v)| < \delta$. By density, we can find a $v \in C^\infty(\bar{\Omega})$ such that $\|u - v\|_{H^1} < \delta'$. By Lemma 4.2, we have

$$\limsup_{\varepsilon\to 0} B_\varepsilon \le \limsup_{\varepsilon\to 0} E_\varepsilon(u) \le \limsup_{\varepsilon\to 0} E_\varepsilon(v) + \delta \le E(v) + \delta \le E(u) + 2\delta.$$

Since this holds for all $\delta > 0$, we have that

$$\limsup_{\varepsilon\to 0} B_\varepsilon \le E(u).$$

Finally time to put everything together ! What we have obtained is that for $u \in H^1(\Omega)$ and u_ε as previously defined, we have

$$\limsup_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon) \le \limsup_{\varepsilon\to 0} A_\varepsilon + B_\varepsilon + C_\varepsilon \le \limsup_{\varepsilon\to 0} B_\varepsilon \le E(u) = F(u)$$

□

Here are the pros and cons of the approach we just considered:

Pros:

- Much less regularity of limiting solution is needed, just weak solutions in $H^1(\Omega)$ suffice;
- Easily adaptable to more complicated (nonlinear) graph equations, as long as they are variational;

Cons:

- Convergence in weaker L^2-type norms;
- Complete discrete to local Γ-convergence needs TL^p spaces, which are more difficult objects requiring optimal transport;
- Γ-convergence cannot be used to prove rates and is restricted to variational problems.

4.2.1 Other approaches to prove the continuum limit

Viscosity solutions An alternative approach that is closest to the maximum principle approach uses the notion of viscosity solutions of the PDE

$$-\Delta_\rho u = 0 \quad \text{in } \Omega \tag{4.9}$$

which allows to apply the consistency-based argument above to smooth test functions. We refer to [Cal18; Cal19] for uses of this technique in the context of semi-supervised learning.

Definition 4.2. We say that $u \in USC(\Omega)$ is a viscosity subsolution of (4.9) if for every $x_0 \in \Omega$ and for every $\phi \in C^\infty(\mathbb{R}^d)$ such that $u - \phi$ has its global maximum at x_0 it holds $-\Delta_\rho \phi(x_0) \leq 0$. Similarly, $u \in LSC(\Omega)$ is a viscosity supersolution of (4.9) if $-u$ is a viscosity subsolution of (4.9). Finally, We say that $u \in C(\Omega)$ is a viscosity solution of (4.9) if it is a viscosity sub- and supersolution.

Using the notion of viscosity solutions it is pretty straightforward to prove that limits of solutions to the Laplace learning problem are solution to the Laplace equation (4.9). The boundary conditions can be taken into account as well. The hard part, though, is to prove that solutions of the graph problem converge to some limit in the first place.

Proposition 4.2. *Let ε_n satisfy*

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

for some $\sigma > 0$ and $u_n := u_{n,\varepsilon_n}$ be as in Theorem 3.3. Assume that there exists a function $u \in C(\Omega)$ such that almost surely $\max_{x \in V_n} |u_n(x) - u(x)| \to 0$ as $n \to \infty$. Then u is a viscosity solution of (4.9).

Proof. We just prove that u is a viscosity subsolution. The supersolution part works in the same way. Letting $x_0 \in \Omega$ and $\phi \in C^\infty(\mathbb{R}^d)$ be such that $\phi(x_0) = u(x_0)$ and $\phi \geq u$ in Ω, we want to prove that $-\Delta_\rho \phi(x_0) \leq 0$. Using Exercise 3.4 and Lemma 3.7 as well as the Borel–Cantelli lemma we get that

$$\lim_{n\to\infty} \max_{x \in V_n \cap \Omega_{\varepsilon_n}} |L_{n,\varepsilon_n}\phi(x) - \Delta_\rho \phi(x)| = 0 \tag{4.10}$$

holds almost surely and uniformly in ϕ.

By the assumption that u_n converges uniformly with respect to V_n to u, there exists a sequence of points $(x_n)_{n\in\mathbb{N}} \subset V_n$ with $\lim_{n\to\infty} x_n = x_0$ such that $u_n - \phi$ has its global maximum over V_n at x_n (just like $u - \phi$ has its maximum at x_0). This means that $u_n(x_n) - u_n(x) \geq \phi(x_n) - \phi(x)$ for all $x \in V_n$ and as a consequence

$$L_{n,\varepsilon_n} u_n(x_n) \leq L_{n,\varepsilon_n}\phi(x_n). \tag{4.11}$$

Since $x \in \Omega$ there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have $x_n \notin \Gamma_n = V_n \cap \partial_{\varepsilon_n}\Omega$ and we get

$$0 = \lim_{n\to\infty} L_{n,\varepsilon_n} u_n(x_n) \leq \lim_{n\to\infty} L_{n,\varepsilon_n}\phi(x_n) = \Delta_\rho \phi(x_0)$$

where we used (4.10) and (4.11) as well as the continuity of $x \mapsto \Delta_\rho \phi(x)$. This shows $-\Delta_\rho \phi(x_0) \leq 0$ and hence u is a viscosity subsolution. $\square$

Remark 4.4. For certain PDEs (in particular, for (4.9)) that admit a so-called strong uniqueness property, the assumption that the approximating sequence $(u_n)_{n\in\mathbb{N}}$ has a uniform limit can be dropped. The strong uniqueness demands that if $u \in USC(\Omega)$ is a subsolution and $v \in LSC(\Omega)$ is a supersolution with $u \leq v$ on $\partial\Omega$, then $u \leq v$ in Ω. In this case one can define the functions

$$\bar{u}(x) := \limsup_{\substack{V_n \cap \Omega_{\varepsilon_n} \ni y \to x \\ n\to\infty}} u_n(y), \qquad \underline{u}(x) := \liminf_{\substack{V_n \cap \Omega_{\varepsilon_n} \ni y \to x \\ n\to\infty}} u_n(y)$$

which satisfy $\underline{u} \leq \bar{u}$ by definition. We claim that $\bar{u}$ is a subsolution and $\underline{v}$ is a supersolution. If this was true, the strong uniqueness property would imply $\bar{u} \leq \underline{u}$ and therefore $\bar{u} = \underline{u}$ and the limit exists. Replacing u by $\bar{u}$ in the previous proof, one can indeed show this, and analogously the supersolution property of $\underline{u}$. For details we refer to [BS91].

Next we prove that the Laplace equation (4.9) admits a maximum principle even for viscosity solutions. For strong solutions this is obvious by differentiation.

Proposition 4.3. *Let $u \in USC(\Omega)$ be a subsolution of $-\Delta_\rho u \leq 0$ and $v \in C^\infty(\mathbb{R}^d)$ satisfy $-\Delta_\rho v > 0$ in Ω. Then it holds*

$$\max_{\overline{\Omega}}(u - v) = \max_{\partial\Omega}(u - v).$$

Proof. Since $u - v$ is upper semicontinuous, it attains its maximum at some $x_0 \in \overline{\Omega}$. If $x_0 \in \Omega$, then the fact that u is a subsolution and that $v \in C^\infty(\mathbb{R}^d)$ would imply $-\Delta_\rho v(x_0) \leq 0$ which is a contradiction. Hence, it has to hold $x_0 \in \partial\Omega$. □

With more effort one can prove the same statement, just assuming that v is a viscosity supersolution and one can also relax the strictness. Next we also discuss pros and cons of the viscosity solution approach:

Pros:

- Less regularity of limiting solution is needed, continuity is enough;
- Easily adaptable to more complicated (nonlinear) graph equations, as long as they are *monotone* and *consistent*;
- Extends maximum principle idea beyond strong solutions.

Cons:

- Existence of a uniform limit requires some compactness to invoke Arzelà–Ascoli;
- Visosity theory not as elementary;
- Getting rates is much harder (e.g., DoV technique).

Qualitative variational techniques Hence, one can use the notion of Γ-convergence to prove that the respective minimizers converge to each other.

Quantitative variational techniques Finally, one can also to some extend quantify these variational techniques to obtain convergence rates.

Using the strong convexity of the Dirichlet energy one can prove

$$C\left\|v-u\right\|^2_{L^2(\Omega)} \leq E(v)-E(u) \qquad \forall v \in L^2(\Omega),\ v=u \text{ on } \partial\Omega.$$

This estimate will be used for $v=\Lambda_{\varepsilon_n}E_n u_n$, where $E_n : \ell^2(V_n) \to L^2(\Omega)$ is a suitable piecewise constant extension operator, and $\Lambda_{\varepsilon_n} : L^2(\Omega) \to H^1(\Omega)$ is a suitably constructed convolution operator with the property that $\left\|\Lambda_{\varepsilon_n}v-v\right\|_{L^2(\Omega)} \to 0$ as $n\to\infty$. Indeed, with these construction one can prove that $E(\Lambda_{\varepsilon_n}E_n u_n)-E_n(u_n)\to 0$ and using also that $E_n(u_n) \leq E_n(u|_{V_n})$ and $E(u)-E_n(u|_{V_n}) \to 0$ as $n\to\infty$, one can prove rates of convergence.

Pros:

- Can be combined with other techniques to prove rates for more singular problems [Bun+24a].
- Works with less regularity than the maximum principle approach.
- Extends to equations without maximum principle or with non-uniqueness.

Cons:

- Gives worse rates in weaker norms, in the case of the Laplace equation one gets $\frac{1}{n}\sum_{x\in V_n}\left|u_n(x)-u(x)\right|^2 \leq C\varepsilon_n$ if $\varepsilon_n \gg \left(\frac{\log n}{n}\right)^{\frac{1}{2d+2}}$ which is an even stronger assumption on the scaling [Cal20].
- Technically challenging.

4.3 Other models for semi-supervised learning

It turns out that the graph Laplace equation (3.15) only has a well-posed continuum limit if the labeled data set Γ_n is sufficiently large and approximates a $d-1$-dimensional subset of $\overline{\Omega}$, e.g., $\partial\Omega$. Consequently, if one works with finite labeled data even in the continuum limit, i.e., $\Gamma_n = \Gamma$ for all $n\in\mathbb{N}$, one has to resort to different methods.

p-Laplacian and Lipschitz learning One way of maintaining a continuum limit is to replace the graph Laplacian (3.14) in (3.15) by the graph p-Laplacian defined as

$$L_{n,\varepsilon}^{(p)}u(x) = \frac{1}{n\varepsilon^p}\sum_{y\in V_n}\eta_\varepsilon(|x-y|)\,(u(y)-u(x))\,|u(y)-u(x)|^{p-2}\,. \tag{4.12}$$

If $p > d$ is larger than the dimension of the underlying space, one can use Γ-convergence techniques to prove [ST19] that the continuum limit is

$$\begin{cases} \operatorname{div}\left(\rho(x)^2\,|\nabla u(x)|^{p-2}\,\nabla u(x)\right) = 0, & x \in \Omega\setminus\Gamma,\\ u(x) = g(x), & x\in\Gamma. \end{cases}$$

In realistic situations, however, $d \in \mathbb{N}$ is very large and potentially unknown. Therefore a reasonable model to consider is Lipschitz learning which is derived by sending $p \to \infty$ in the above. The graph infinity Laplacian is defined as

$$L_{n,\varepsilon}^{(\infty)}u(x) = \frac{1}{\varepsilon^2}\left(\max_{y\in V_n}\eta_\varepsilon(|x-y|)\,(u(y)-u(x)) + \min_{y\in V_n}\eta_\varepsilon(|x-y|)\,(u(y)-u(x))\right) \tag{4.13}$$

and the continuum limit is

$$\begin{cases} \Delta_\infty u(x) = 0, & x\in\Omega,\\ u(x) = g(x), & x\in\Gamma, \end{cases}$$

in the viscosity sense, where $\Delta_\infty u = \langle \nabla u, D^2 u \nabla u\rangle$ for a smooth function u is the infinity Laplacian. This continuum limit was proved in [Cal19] (see also [RB23]), and the following rates of convergence were shown in [BCR23; BCR24]:

$$\max_{x\in V_n}|u_{n,\varepsilon}(x)-u(x)| \leq C\left(\frac{\delta_n}{\varepsilon}\right)^{\frac{1}{4}}$$

for bandwidths satisfying $\delta_n \lesssim \varepsilon \lesssim \delta_n^{\frac{5}{9}}$, where $\delta_n = \left(\frac{\log n}{n}\right)^{\frac{1}{d}}$.

Note that the pointwise constraint $u = g$ on Γ is meaningful since $W^{1,p}$-functions have continuous representatives for $p > d$. A major disadvantage of these two approaches is that the equations become more and more independent of the data distribution ρ as p grows.

> Lipschitz learning is asymptotically well-posed even for sparse graphs and arbitrary label sets.

Poisson learning An alternative approach to graph-based semi-supervised learning is through a graph Poisson equation of the form

$$-L_{n,\varepsilon}u(x) = n\sum_{y\in\Gamma_n}(g(y)-\overline{g})\,\delta_{y,x}, \qquad x\in V_n, \tag{4.14}$$

where the labels enter through a source term, $\overline{g} = \frac{1}{|\Gamma_n|}\sum_{y\in\Gamma_n} g(y)$ is the label mean, and $\delta_{y,x}$ the Kronecker delta symbol. The equation is complemented with a constraint on the mean value for uniqueness and the final labeling decision is achieved by thresholding.

The continuum limit of Poisson learning is a Poisson equation with measure data which has distributional solutions in $W^{1,p}$ for $p < \frac{d}{d-1}$ and no weak or even classical solutions:

$$-\operatorname{div}\left(\rho^2 \nabla u\right) = \sum_{y\in\Gamma} (g(y) - \overline{g})\, \delta_y \quad \text{in } \Omega. \tag{4.15}$$

The equation is complemented with homogeneous Neumann boundary conditions. Proving this continuum limit is very difficult because of singularities of the limiting solutions around the labels. Recently this was achieved in [Bun+24b] and it was shown that with high probability it holds

$$\frac{1}{n} \sum_{x\in V_n} |u_{n,\varepsilon}(x) - u(x)| \leq C\varepsilon^{\frac{1}{d+2}}, \tag{4.16}$$

where u_n and u are the solutions of (4.14) and (4.15), respectively. The condition on the graph bandwidth for this to hold is that

$$\varepsilon \gg \left(\frac{\log n}{n}\right)^{\frac{1}{3d}}$$

which is a much stricter condition that for the graph Laplace equation.

Poisson learning is asymptotically well-posed for dense graphs and arbitrary label sets.

Exercise 4.7. Prove that any minimizer of

$$\min_{u\in\ell^2(V_n)} \left\{ \frac{1}{4\sigma_\eta n^2\varepsilon^2} \sum_{x,y\in V_n} \eta_\varepsilon(|x-y|)(u(y)-u(x))^2 - \sum_{y\in\Gamma_n} u(y)\,(g(y)-\overline{g}) \right\}$$

solves the Poisson learning problem (4.14). Notably, such variational interpretation does not hold for the continuum limit (4.15), see [Bun+24a].

5 Solutions to the exercises

Solution 5.1 (of Exercise 3.2)**.** Since Z follows a standard normal distribution,

$$\mathbb{P}(Z \geq t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2} dx \leq \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t} e^{-x^2/2} dx = \frac{1}{t\sqrt{2\pi}} \int_t^\infty \left(-e^{-x^2/2}\right)' dx = \frac{1}{t\sqrt{2\pi}} e^{-t^2/2}$$

Solution 5.2 (of Exercise 3.4)**.** In order to show this exercise, we will need some intermediary results. The first one is just a consequence of the order 3 Taylor expansion of a function.

Proposition 5.1. *Let $u \in C^3(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$. There exists $\varepsilon_x : \mathbb{R}^d \to \mathbb{R}$ such that $\|\varepsilon_x\|_\infty \leq 1$ and for all $y \in \mathbb{R}^d$,*

$$u(y) = u(x) + \sum_{i=1}^d \partial_i u(x)(y^i - x^i) + \sum_{i,j=1}^d \partial_{ij}^2 u(x)(y^i - x^i)(y^j - x^j) + \|u\|_{C^3}|y-x|^3\varepsilon_x(y)$$

The next one is convenient to treat sums of variables:

Proposition 5.2. *Let $X_1, \ldots, X_n$ be real-valued random variables defined on the same propability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then for all $t \in \mathbb{R}$,*

$$\mathbb{P}(\left|\sum_{i=1}^n X_i\right| \geq t) \leq \sum_{i=1}^n \mathbb{P}(|X_i| \geq t/d)$$

The proof is immediate by considering the events

$$\left\{\left|\sum_{i=1}^n X_i\right| \geq t\right\} \subset \left\{\sum_{i=1}^n |X_i| \geq t\right\} \subset \bigcup_{i=1}^n \{|X_i| \geq t/d\}.$$

We can now show the desired result. Let $u \in C^3(\overline{\Omega})$ and $x \in \Omega$. Assume that $\varepsilon \leq 1$. Then using the Taylor expansion, we have:

$$\begin{aligned}
L_{n,\varepsilon}u(x) =& \frac{2}{\sigma_\eta \varepsilon^2 n}\sum_{p=1}^n \eta_\varepsilon(|x_p - x|)(u(x_p) - u(x)) \\
=& \frac{2}{\sigma_\eta \varepsilon^2 n}\sum_{i=1}^d \sum_{p=1}^n \eta_\varepsilon(|x_p - x|)\partial_i u(x)(x_p^i - x^i) \\
&+ \frac{2}{\sigma_\eta \varepsilon^2 n}\sum_{i,j=1}^d \sum_{p=1}^n \eta_\varepsilon(|x_p - x|)\partial_{ij}^2 u(x)(x_p^i - x^i)(x_p^j - x^j) \\
&+ \frac{2\|u\|_{C^3}}{\sigma_\eta \varepsilon^2 n}\sum_{p=1}^n \eta_\varepsilon(|x_p - x|)|x_p - x|^3\varepsilon_x(x_p).
\end{aligned}$$

We will use the random variables

$$\begin{aligned}
X_p^i :=& \frac{2}{\sigma_\eta \varepsilon^2}\eta_\varepsilon(|x_p - x|)(x_p^i - x^i) \\
Y_p^{ij} :=& \frac{2}{\sigma_\eta \varepsilon^2}\eta_\varepsilon(|x_p - x|)(x_p^i - x^i)(x_p^j - x^j) \\
Z_p :=& \frac{2}{\sigma_\eta \varepsilon^2}\eta_\varepsilon(|x_p - x|)|x_p - x|^3\varepsilon_x(x_p)
\end{aligned}$$

The idea will be to apply Berstein's inequality on each of these random variables. Let us first consider the X_p^is. Fix $1 \leq i \leq d$ and observe that

$$\mathbb{E}(X_p^i) = \frac{2}{\sigma_\eta \varepsilon^2}\int_\Omega \eta_\varepsilon(|y - x|)(y^i - x^i)\rho(y)\,\mathrm{d}y$$

We can then estimate the variance from above:

$$
\begin{aligned}
V(X_p^i) \le \mathbb{E}((X_p^i)^2) &= \frac{4}{\sigma_\eta^2 \varepsilon^4} \int_\Omega \eta_\varepsilon(|y-x|)^2 (y^i - x^i)^2 \rho(y) \, \mathrm{d}y \\
&\le \frac{4C_\rho}{\sigma_\eta^2 \varepsilon^4} \int_\Omega \eta_\varepsilon(|y-x|)^2 |y-x|^2 \, \mathrm{d}y \\
&\le \frac{4C_\rho}{\sigma_\eta^2 \varepsilon^2} \int_{\Omega \cap B_\varepsilon(x)} \eta_\varepsilon(|y-x|)^2 \, \mathrm{d}y \\
&\le \frac{4C_\rho}{\sigma_\eta^2 \varepsilon^2} \frac{C_\eta}{\varepsilon^d} \le \frac{C_{\rho,\eta}}{\varepsilon^{d+2}}
\end{aligned}
$$

Next we need to estimate the variation to the mean:

$$
\begin{aligned}
|X_p^i - \mathbb{E}(X_p^i)| &\le |X_p^i| + |\mathbb{E}(X_p^i)| \\
&\le \frac{2}{\sigma_\eta \varepsilon^2} \underbrace{\eta_\varepsilon(|x_p - x|)|x_p - x|}_{\le \frac{C_\eta}{\varepsilon^d}\varepsilon} + \frac{2C_\rho}{\sigma_\eta \varepsilon^2} \int_{\Omega \cap B_\varepsilon(x)} \eta_\varepsilon(|y-x|)|y-x| \, \mathrm{d}y \\
&\le \frac{C_{\eta,\rho}}{\varepsilon^{d+1}} + \frac{C_{\eta,\rho}}{\varepsilon} \le \frac{C_{\eta,\rho}}{\varepsilon^{d+1}}
\end{aligned}
$$

where the last inequality comes from the fact that $\varepsilon \le 1$. Applying Bernstein's inequality, we get that for all $0 < t \le \varepsilon^{-1}$ and all $1 \le i \le d$,

$$
\mathbb{P}\left(\left| \frac{1}{n} \sum_{p=1}^n X_p^i - \mathbb{E}(X_1^i) \right| \ge t \right) \le 2 \exp\left(- \frac{nt^2}{2\left(\frac{C_{\eta,\rho}}{\varepsilon^{d+2}} + \frac{tC_{\eta,\rho}}{3\varepsilon^{d+1}} \right)} \right) \le 2\exp(-C_{\eta,\rho} n \varepsilon^{d+2} t^2).
$$

In a similar fashion, we can show that for all $0 < t \le \varepsilon^{-1}$ and all $1 \le i, j \le d$,

$$
\mathbb{P}\left(\left| \frac{1}{n} \sum_{p=1}^n Y_p^{ij} - \mathbb{E}(Y_1^{ij}) \right| \ge t \right) \le 2\exp(-C_{\eta,\rho} n \varepsilon^{d+2} t^2)
$$

and

$$
\mathbb{P}\left(\left| \frac{1}{n} \sum_{p=1}^n Z_p - \mathbb{E}(Z_1) \right| \ge t \right) \le 2\exp(-C_{\eta,\rho} n \varepsilon^{d+2} t^2).
$$

Now, define

$$
\begin{aligned}
X &:= \frac{2}{\sigma_\eta \varepsilon^2 n} \sum_{i=1}^d \sum_{p=1}^n \eta_\varepsilon(|x_p - x|) \partial_i u(x) (x_p^i - x^i) = \sum_{i=1}^d \partial_i u(x) \frac{1}{n} \sum_{p=1}^n X_p^i \\
Y &:= \frac{2}{\sigma_\eta \varepsilon^2 n} \sum_{i,j=1}^d \sum_{p=1}^n \eta_\varepsilon(|x_p - x|) \partial_{ij}^2 u(x) (x_p^i - x^i)(x_p^j - x^j) = \sum_{i,j=1}^d \partial_{ij}^2 u(x) \frac{1}{n} \sum_{p=1}^n Y_p^{ij} \\
Z &:= \frac{2\|u\|_{C^3}}{\sigma_\eta \varepsilon^2 n} \sum_{p=1}^n \eta_\varepsilon(|x_p - x|)|x_p - x|^3 \varepsilon_x(x_p) = \frac{\|u\|_{C^3}}{n} \sum_{p=1}^n Z_p.
\end{aligned}
$$

Then, using the previous proposition, we have that

$$\begin{aligned}\mathbb{P}\left(\forall u \in C^3, |L_{n,\varepsilon}u(x) - \mathcal{L}_\varepsilon u(x)| \geq t\right) &= \mathbb{P}(\forall u \in C^3, |X - \mathbb{E}(X) + Y - \mathbb{E}(Y) + Z - \mathbb{E}(Z)| \geq t) \\ &\leq \mathbb{P}(\forall u \in C^3, |X - \mathbb{E}(X)| \geq t/3) + \mathbb{P}(|Y - \mathbb{E}(Y)| \geq t/3) + \mathbb{P}(|Z - \mathbb{E}(Z)| \geq t/3)\end{aligned}$$

We need to bound every probability appearing in the right hand side:

$$\begin{aligned}\mathbb{P}(\forall u \in C^3, |X - \mathbb{E}(X)| \geq t/3) &= \mathbb{P}\left(\forall u \in C^3, \sum_{i=1}^d \partial_i u(x)\left(\frac{1}{n}\sum_{p=1}^n X_p^i - \mathbb{E}(X_1^i)\right)\right) \\ &\leq \sum_{i=1}^d \mathbb{P}\left(\forall u \in C^3, |\partial_i u(x)|\left|\frac{1}{n}\sum_{p=1}^n X_p^i - \mathbb{E}(X_1^i)\right| \geq \frac{t}{3d}\right) \\ &\leq \sum_{i=1}^d \mathbb{P}\left(\forall u \in C^3, \|u\|_{C^3}\left|\frac{1}{n}\sum_{p=1}^n X_p^i - \mathbb{E}(X_1^i)\right| \geq \frac{t}{3d}\right) \\ &\leq d\mathbb{P}\left(\forall u \in C^3, \left|\frac{1}{n}\sum_{p=1}^n X_p^i - \mathbb{E}(X_1^i)\right| \geq \frac{t}{3d\|u\|_{C^3}}\right)\end{aligned}$$

By putting $t = t\|u\|_{C^3}d$, we get

$$\begin{aligned}\mathbb{P}(\forall u \in C^3, |X - \mathbb{E}(X)| \geq C_d\|u\|_{C^3}t) &\leq C_d\mathbb{P}\left(\forall u \in C^3, \left|\frac{1}{n}\sum_{p=1}^n X_p^i - \mathbb{E}(X_1^i)\right| \geq t\right) \\ &= C_d\mathbb{P}\left(\left|\frac{1}{n}\sum_{p=1}^n X_p^i - \mathbb{E}(X_1^i)\right| \geq t\right) \\ &\leq C_d \exp(-C_{\eta,\rho} n\varepsilon^{d+2}t^2)\end{aligned}$$

and the same arguments same holds for Y and Z, hence

$$\mathbb{P}\left(\forall u \in C^3, |L_{n,\varepsilon}u(x) - \mathcal{L}_\varepsilon u(x)| \geq C_d\|u\|_{C^3}t\right) \leq C_d \exp(-C_{\eta,\rho} n\varepsilon^{d+2}t^2)$$

We can then finish the proof the same way as in Lemma 3.6 to conclude that

$$\mathbb{P}\left(\forall u \in C^3, \max_{x \in V_n} |L_{n,\varepsilon}u(x) - \mathcal{L}_\varepsilon u(x)| \geq C_d\|u\|_{C^3}t\right) \geq 1 - C_d \exp(-C_{\eta,\rho} n\varepsilon^{d+2}t^2 + \log n).$$

Solution 5.3 (of Exercise 4.2)**.** Let begin by the end. we have:

$$\begin{aligned}
\frac{1}{n!}\sum_{\tau\in S(n)} V(x_{\tau_1},\dots,x_{\tau_n}) &= \frac{1}{kn!}\sum_{\tau\in S(n)}\sum_{p=1}^{k} f(x_{\tau_{2p-1}},x_{\tau_{2p}}) \\
&= \frac{1}{kn!}\sum_{p=1}^{k}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{\tau\in S(n)\\ \tau_{2p-1}=i\\ \tau_{2p}=j}} f(x_{\tau_{2p-1}},x_{\tau_{2p}}) \\
&= \frac{1}{kn!}\sum_{p=1}^{k}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{\tau\in S(n)\\ \tau_{2p-1}=i\\ \tau_{2p}=j}} f(x_i,x_j) \\
&= \frac{1}{kn!}\sum_{p=1}^{k}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(n-2)!f(x_i,x_j) \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} f(x_i,x_j) \\
&= U_n
\end{aligned}$$

Solution 5.4 (of Exercise 4.4)**.**

- The first question follows from the definition.
- Let

$$J_n(x) = \begin{cases} 0, & x<0,\\ 1, & x\geq 0\end{cases} \qquad \forall n\in\mathbb{N}.$$

We can check that for $x\in\mathbb{R}\setminus\{0\}$, there is no problem for either the lim inf or lim sup. Now the problem is to check what happens at 0. If we take a sequence $x_n\to 0^-$ in the lim inf, we see that the limiting functional J must verify $J(0)\leq\liminf_n J_n(x_n)=0$. We can show that

$$J(x) = \begin{cases} 0, & x\leq 0,\\ 1, & x> 0\end{cases} \qquad \forall n\in\mathbb{N}.$$

is the correct one.

- Let $J_n(x) = \left(\sum_{i=1}^d |x_i|^n\right)^{\frac{1}{n}} = \|x\|_n$ and $J(x) := \max_{i=1}^d |x_i| = \|x\|_\infty$. Let show the lim sup: for $x\in\mathbb{R}^d$, let $x_n := x$ for all $n\geq 0$. Then

$$\limsup J_n(x_n) = \limsup J_n(x) = \lim \|x\|_n = \|x\|_\infty = J(u).$$

Now let show the lim inf. Let $x \in \mathbb{R}^d$ and let $x_n \in \mathbb{R}^d$ be such that $x_n \to x$. Using the triangle inequalities, we have:

$$|\|x_n\|_n - \|x\|_\infty| \leq |\|x_n\|_n - \|x\|_n| + |\|x\|_n - \|x\|_\infty| \leq \underbrace{\|x_n - x\|_n}_{\leq d^{1/n}\|x_n - x\|_\infty} + \underbrace{|\|x\|_n - \|x\|_\infty|}_{\xrightarrow[n\to\infty]{} 0}$$

which shows the result.

- Let $J_n(x) = \sin(2\pi n x)$. Using the power of intuition, we decide that $J(x) := -1$ (the idea is that sequence of minimizers of J_n should approach minimizers of J, and the minimizers of J_n becomes denser and denser on the real line). Let us show it. For the lim inf, this is trivial. For the lim sup, let $x \in \mathbb{R}$, and let

$$x_n := n^{-1}\left(\lfloor nx \rfloor - \frac{1}{4}\right)$$

We have that $x - \frac{1}{4n} \leq x_n \leq x + \frac{3}{4n}$ hence $x_n \to x$. Moreover,

$$\limsup J_n(x_n) = \sin\left(2\pi\lfloor nx \rfloor - \frac{\pi}{2}\right) = -1 \leq J(x).$$

* Prove that any Γ-limit is lower semicontinuous.

Solution 5.5 (of Exercise 4.5)**.** We follow the steps :

1. Using that $\phi \in C_c^\infty$ and $\rho \in C^1$, we can write that

$$\int_\Omega (\nabla u_n \cdot \phi)\rho^2 = \int_{\partial\Omega} u_n \rho^2 \phi \cdot n - \int_\Omega u_n \operatorname{div}(\rho^2\phi) = -\int_\Omega u_n \operatorname{div}(\rho^2\phi)$$

since $\phi = 0$ on $\partial\Omega$. Using the L^2 convergence of u_n and the Green formula again, we get

$$-\int_\Omega u_n \operatorname{div}(\rho^2\phi) \xrightarrow[n\to\infty]{} -\int_\Omega u \operatorname{div}(\rho^2\phi) = \int_\Omega (\nabla u \cdot \phi)\rho^2$$

2. Let $v \in H^1(\Omega)$ be such that $\int_\Omega |\nabla v|^2\rho^2 = 1$. Using Cauchy-Schwarz, we have that

$$\left(\int_\Omega (\nabla u \cdot \nabla v)\right)^2 \rho^2 \leq \left(\int_\Omega |\nabla u|^2\rho^2\right)\left(\int_\Omega |\nabla v|^2\rho^2\right) = \int_\Omega |\nabla u|^2\rho^2$$

By taking the supremum, we have that

$$\int_\Omega |\nabla u|^2\rho^2 \geq \sup_{\substack{v\in H^1(\Omega) \\ \int_\Omega |\nabla v|^2\rho^2 = 1}} \left(\int_\Omega (\nabla u \cdot \nabla v)\rho^2\right)^2$$

and we can check that the supremum is attained by taking $v = u/\sqrt{\int_\Omega |\nabla u|^2\rho^2}$.

3. This follows directly for the density of $C_c^\infty(\Omega, \mathbb{R}^d)$ into $L^2(\Omega, \mathbb{R}^d)$.
4. Let $\phi \in C_c^\infty(\Omega, \mathbb{R}^d)$. For all $n \in \mathbb{N}$, we have

$$\int_\Omega |\nabla u_n|^2 \rho^2 \geq \left(\int_\Omega (\nabla u_n \cdot \phi) \rho^2 \right)^2$$

By taking the liminf, this leads to

$$\liminf_n \int_\Omega |\nabla u_n|^2 \rho^2 \geq \liminf_n \left(\int_\Omega (\nabla u_n \cdot \phi) \rho^2 \right)^2 = \left(\int_\Omega (\nabla u \cdot \phi) \rho^2 \right)^2$$

and taking the supremum over ϕ leads to the result.

References

[BS91] Guy Barles and Panagiotis E Souganidis. "Convergence of approximation schemes for fully nonlinear second order equations". In: *Asymptotic analysis* 4.3 (1991), pp. 271–283.

[Pag99] Lawrence Page. *The PageRank citation ranking: Bringing order to the web.* Tech. rep. Technical Report, 1999.

[Amg03] Saïd Amghibech. "Eigenvalues of the discrete p-Laplacian for graphs". In: *Ars Combinatoria* 67 (2003), pp. 283–302.

[BH09a] Thomas Bühler and Matthias Hein. "Spectral clustering based on the graph p-Laplacian". In: *Proceedings of the 26th annual international conference on machine learning.* 2009, pp. 81–88.

[BH09b] Thomas Bühler and Matthias Hein. *Supplementary material for "Spectral clustering based on the graph p-Laplacian".* 2009.

[GS16] Nicolás García Trillos and Dejan Slepčev. "Continuum limit of total variation on point clouds". In: *Archive for rational mechanics and analysis* 220 (2016), pp. 193–241.

[Cal18] Jeff Calder. "The game theoretic p-Laplacian and semi-supervised learning with few labels". In: *Nonlinearity* 32.1 (2018), p. 301.

[Cal19] Jeff Calder. "Consistency of Lipschitz learning with infinite unlabeled data and finite labeled data". In: *SIAM Journal on Mathematics of Data Science* 1.4 (2019), pp. 780–812.

[ST19] Dejan Slepcev and Matthew Thorpe. "Analysis of p-Laplacian regularization in semisupervised learning". In: *SIAM Journal on Mathematical Analysis* 51.3 (2019), pp. 2085–2120.

[Cal20] Jeff Calder. *The Calculus of Variations.* Lecture Notes. 2020.

[BCR23] Leon Bungert, Jeff Calder, and Tim Roith. "Uniform convergence rates for Lipschitz learning on graphs". In: *IMA Journal of Numerical Analysis* 43.4 (2023), pp. 2445–2495.

[RB23] Tim Roith and Leon Bungert. "Continuum limit of Lipschitz learning on graphs". In: *Foundations of Computational Mathematics* 23.2 (2023), pp. 393–431.

[Bun+24a] Leon Bungert, Jeff Calder, Max Mihailescu, Kodjo Houssou, and Amber Yuan. *Convergence rates for Poisson learning to a Poisson equation with measure data*. 2024. arXiv: `2407.06783 [math.AP]`.

[Bun+24b] Leon Bungert, Jeff Calder, Max Mihailescu, Kodjo Houssou, and Amber Yuan. *Convergence rates for Poisson learning to a Poisson equation with measure data*. 2024. arXiv: `2407.06783 [math.AP]`.

[BCR24] Leon Bungert, Jeff Calder, and Tim Roith. "Ratio convergence rates for Euclidean first-passage percolation: applications to the graph infinity Laplacian". In: *The Annals of Applied Probability* 34.4 (2024), pp. 3870–3910.

[Bre] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. New York, NY, USA: Springer. ISBN: 978-0-387-70914-7.