The Calculus of Variations

Jeff Calder

University of Minnesota
School of Mathematics
jwcalder@umn.edu

February 6, 2024

Contents

Chapter 1

Introduction

The calculus of variations is a field of mathematics concerned with minimizing (or maximizing) functionals (that is, real-valued functions whose inputs are functions). The calculus of variations has a wide range of applications in physics, engineering, applied and pure mathematics, and is intimately connected to partial differential equations (PDEs).

For example, a classical problem in the calculus of variations is finding the shortest path between two points. The notion of length need not be Euclidean, or the path may be constrained to lie on a surface, in which case the shortest path is called a *geodesic*. In physics, Hamilton's principle states that trajectories of a physical system are critical points of the action functional. Critical points may be minimums, maximums, or saddle points of the action functional. In computer vision, the problem of segmenting an image into meaningful regions is often cast as a problem of minimizing a functional over all possible segmentations—a natural problem in the calculus of variations. Likewise, in image processing, the problem of restoring a degraded or noisy images has been very successfully formulated as a problem in the calculus of variations.

PDEs arise as the necessary conditions for minimizers of functionals. Recall in multi-variable calculus that if a function $u : \mathbb{R}^d \to \mathbb{R}$ has a minimum at $x \in \mathbb{R}^d$ then $\nabla u(x) = 0$. The necessary condition $\nabla u(x) = 0$ can be used to solve for candidate minimizers x. In the calculus of variations, if a function $u : \mathbb{R}^d \to \mathbb{R}$ is a minimizer of a functional $I(u)$ then the necessary condition $\nabla I(u) = 0$ turns out to be a PDE called the *Euler-Lagrange equation*. Studying the Euler-Lagrange equation allows us to explicitly compute minimizers and to study their properties. For this reason, there is a rich interplay between the calculus of variations and the theory of PDEs.

These notes aim to give a brief overview of the calculus of variations at the beginning graduate level. The first four chapters are concerned with smooth solutions of the Euler-Lagrange equations, and finding explicit solutions of classical problems, like the Brachistochrone problem, and exploring applications to image processing and computer vision. These chapters are written at an advanced undergraduate level, and leave out some of the rigorous detail that is expected in a graduate course. These first four chapters require some familiarity with ordinary differential equations (ODEs), and multi-variable calculus. The appendix contains some mathematical preliminaries and notational conventions that are important to review.

The level of rigour increases dramatically in Chapter 4, where we begin a study of the direct method in the Calculus of Variations for proving existence of minimizers. This chapter requires some basic background in functional analysis and, in particular, Sobolev spaces. We provide a very brief overview in Section 4.2, and refer the reader to [28] for more details.

Finally, in Chapter 5 we give an overview of applications of the calculus of variations to prove discrete to continuum results in graph-based semi-supervised learning. The ideas in this chapter are related to Γ-convergence, which is a notion of convergence for functionals that ensures minimizers converge to minimizers.

These notes are just a basic introduction. We refer the reader to [28, Chapter 8] and [22] for a more thorough exposition of the theory behind the Calculus of Variations.

1.1 Examples

We begin with some examples.

Example 1.1 (Shortest path)**.** Let A and B be two points in the plane. What is the shortest path between A and B? The answer depends on how we measure length! Suppose the length of a short line segment near (x, y) is the usual Euclidean length multiplied by a positive scale factor $g(x, y)$. For example, the length of a path could correspond to the length of time it would take a robot to navigate the path, and certain regions in space may be easier or harder to navigate, yielding larger or smaller values of g. Robotic navigation is thus a special case of finding the shortest path between two points.

Suppose A lies to the left of B and the path is a graph $u(x)$ over the x axis. See Figure 1.1. Then the "length" of the path between x and $x + \Delta x$ is approximately

$$L = g(x, u(x))\sqrt{1 + u'(x)^2}\Delta x.$$

If we let $A = (0, 0)$ and $B = (a, b)$ where $a > 0$, then the length of a path $(x, u(x))$ connecting A to B is

$$I(u) = \int_0^a g(x, u(x))\sqrt{1 + u'(x)^2}\, dx.$$

The problem of finding the shortest path from A to B is equivalent to finding the function u that minimizes the functional $I(u)$ subject to $u(0) = 0$ and $u(a) = b$. $\triangle$

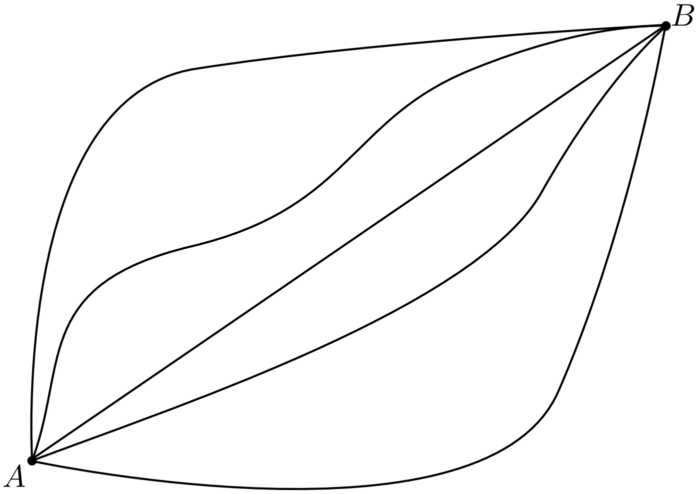


Figure 1.1: In our version of the shortest path problem, all paths must be graphs of functions $u = u(x)$.

Example 1.2 (Brachistochrone problem)**.** In 1696 Johann Bernoulli posed the following problem. Let A and B be two points in the plane with A lying above B. Suppose we connect A and B with a thin wire and allow a bead to slide from A to B under the influence of gravity. Assuming the bead slides without friction, what is the shape of the wire that minimizes the travel time of the bead? Perhaps counter-intuitively, it turns out that the optimal shape is not a straight line! The problem is commonly referred to as the *brachistochrone problem*—the word brachistochrone derives from ancient Greek meaning "shortest time".

Let g denote the acceleration due to gravity. Suppose that $A = (0, 0)$ and $B = (a, b)$ where $a > 0$ and $b < 0$. Let $u(x)$ for $0 \leq x \leq a$ describe the shape of the wire, so $u(0) = 0$ and $u(a) = b$. Let $v(x)$ denote the speed of the bead when it is at position x. When the bead is at position $(x, u(x))$ along the wire, the potential energy stored in the bead is $\text{PE} = mgu(x)$ (relative to height zero), and the kinetic energy is $\text{KE} = \frac{1}{2}mv(x)^2$, where m is the mass of the bead. By conservation of energy

$$\frac{1}{2}mv(x)^2 + mgu(x) = 0,$$

since the bead starts with zero total energy at point A. Therefore

$$v(x) = \sqrt{-2gu(x)}.$$

Between x and $x + \Delta x$, the bead slides a distance of approximately $\sqrt{1 + u'(x)^2}\Delta x$ with a speed of $v(x) = \sqrt{-2gu(x)}$. Hence it takes approximately

$$t = \frac{\sqrt{1 + u'(x)^2}}{\sqrt{-2gu(x)}}\Delta x$$

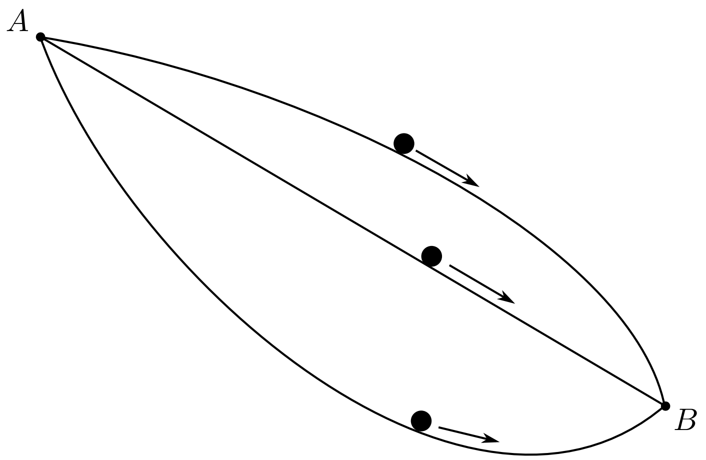


Figure 1.2: Depiction of possible paths for the brachistochrone problem.

time for the bead to move from position x to $x+\Delta x$. Therefore the total time taken for the bead to slide down the wire is given by

$$I(u) = \frac{1}{\sqrt{2g}} \int_0^a \sqrt{\frac{1+u'(x)^2}{-u(x)}}\, dx.$$

The problem of determining the optimal shape of the wire is therefore equivalent to finding the function $u(x)$ that minimizes $I(u)$ subject to $u(0) = 0$ and $u(a) = b$. △

Example 1.3 (Minimal surfaces)**.** Suppose we bend a piece of wire into a loop of any shape we wish, and then dip the wire loop into a solution of soapy water. A soap bubble will form across the loop, and we may naturally wonder what shape the bubble will take. Physics tells us that soap bubble formed will be the one with least surface area, at least locally, compared to all other surfaces that span the wire loop. Such a surface is called a *minimal surface*.

To formulate this mathematically, suppose the loop of wire is the graph of a function $g : \partial U \to \mathbb{R}$, where $U \subset \mathbb{R}^2$ is open and bounded. We also assume that all possible surfaces spanning the wire can be expressed as graphs of functions $u : \overline{U} \to \mathbb{R}$. To ensure the surface connects to the wire we ask that $u = g$ on ∂U. The surface area of a candidate soap film surface u is given by

$$I(u) = \int_U \sqrt{1 + |\nabla u|^2}\, dx.$$

Thus, the minimal surface problem is equivalent to finding a function u that minimizes I subject to $u = g$ on ∂U. △

Example 1.4 (Image restoration)**.** A grayscale image is a function $u : [0,1]^2 \to [0,1]$. For $x \in \mathbb{R}^2$, $u(x)$ represents the brightness of the pixel at location x. In real-world applications, images are often corrupted in the acquisition process or thereafter, and we observe a noisy version of the image. The task of image restoration is to recover the true noise-free image from a noisy observation.

Let $f(x)$ be the observed noisy image. A widely used and very successful approach to image restoration is the so-called total variation (TV) restoration, which minimizes the functional

$$I(u) = \int_U \frac{1}{2}(u-f)^2 + \lambda|\nabla u|\,dx,$$

where $\lambda > 0$ is a parameter and $U = (0,1)^2$. The restored image is the function u that minimizes I (we do not impose boundary conditions on the minimizer). The first term $\frac{1}{2}(u-f)^2$ is a called a fidelity term, and encourages the restored image to be close to the observed noisy image f. The second term $|\nabla u|$ measures the amount of noise in the image and minimizing this term encourages the removal of noise in the restored image. The name TV restoration comes from the fact that $\int_U |\nabla u|\,dx$ is called the total variation of u. Total variation image restoration was pioneered by Rudin, Osher, and Fatemi [49]. △

Example 1.5 (Image segmentation)**.** A common task in computer vision is the segmentation of an image into meaningful regions. Let $f : [0,1]^2 \to [0,1]$ be a grayscale image we wish to segment. We represent possible segmentations of the image by the level sets of functions $u : [0,1]^2 \to \mathbb{R}$. Each function u divides the domain $[0,1]^2$ into two regions defined by

$$R^+(u) = \{x \in [0,1]^2 \,:\, u(x) > 0\} \quad \text{and} \quad R^-(u) = \{x \in [0,1]^2 \,:\, u(x) \leq 0\}.$$

The boundary between the two regions is the level set $\{x \in [0,1]^2 \,:\, u(x) = 0\}$.

At a very basic level, we might assume our image is composed of two regions with different intensity levels $f = a$ and $f = b$, corrupted by noise. Thus, we might propose to segment the image by minimizing the functional

$$I(u,a,b) = \int_{R^+(u)} (f(x)-a)^2\,dx + \int_{R^-(u)} (f(x)-b)^2\,dx,$$

over all possible segmentations u and real numbers a and b. However, this turns out not to work very well since it does not incorporate the geometry of the region in any way. Intuitively, a semantically meaningful object in an image is usually concentrated in some region of the image, and might have a rather smooth boundary. The minimizers of I could be very pathological and oscillate rapidly trying to capture every pixel near a in one region and those near b in another region. For example, if f only takes the values 0 and 1, then minimizing I will try to put all the pixels in the image where f is 0 into one region, and all those where f is 1 into the other region, and will choose $a = 0$ and $b = 1$. This is true regardless of whether the region where f is zero is a nice circle in the center of the image, or if we randomly choose each pixel to be 0 or 1. In the later case, the segmentation u will oscillate wildly and does not give a meaningful result.

A common approach to fixing this issue is to include a penalty on the length of the boundary between the two regions. Let us denote the length of the boundary between $R^+(u)$ and $R^-(u)$ (i.e., the zero level set of u) by $L(u)$. Thus, we seek instead to minimize the functional

$$I(u,a,b) = \int_{R^+(u)} (f(x)-a)^2\,dx + \int_{R^-(u)} (f(x)-b)^2\,dx + \lambda L(u),$$

where $\lambda > 0$ is a parameter. Segmentation of an image is therefore reduced to finding a function $u(x)$ and real numbers a and b minimizing $I(u,a,b)$, which is a problem in the calculus of variations. This widely used approach was proposed by Chan and Vese in 2001 and is called Active Contours Without Edges [20].

The dependence of I on u is somewhat obscured in the form above. Let us write the functional in another way. Recall the Heaviside function H is defined as

$$H(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0. \end{cases}$$

Then the region $R^+(u)$ is precisely the region where $H(u(x)) = 1$, and the region $R^-(u)$ is precisely where $H(u(x)) = 0$. Therefore

$$\int_{R^+(u)} (f(x)-a)^2\,dx = \int_U H(u(x))(f(x)-a)^2\,dx,$$

where $U = (0,1)^2$. Likewise

$$\int_{R^-(u)} (f(x)-b)^2\,dx = \int_U (1 - H(u(x)))\,(f(x)-b)^2\,dx.$$

We also have the identity (see Section A.10.2)

$$L(u) = \int_U |\nabla H(u(x))|\,dx = \int_U \delta(u(x))|\nabla u(x)|\,dx.$$

Therefore we have

$$I(u,a,b) = \int_U H(u)(f-a)^2 + (1-H(u))\,(f-b)^2 + \lambda\delta(u)|\nabla u|\,dx.$$

△

Chapter 2

The Euler-Lagrange equation

We aim to study general functionals of the form

$$I(u) = \int_U L(x, u(x), \nabla u(x))\, dx, \tag{2.1}$$

where $U \subset \mathbb{R}^d$ is open and bounded, and L is a function

$$L : U \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}.$$

The function L is called the *Lagrangian*. We will write $L = L(x, z, p)$ where $x \in U$, $z \in \mathbb{R}$ and $p \in \mathbb{R}^d$. Thus, z represents the variable where we substitute $u(x)$, and p is the variable where we substitute $\nabla u(x)$. Writing this out completely we have

$$L = L(x_1, x_2, \dots, x_n, z, p_1, p_2, \dots, p_n).$$

The partial derivatives of L will be denoted $L_z(x, z, p)$,

$$\nabla_x L(x, z, p) = (L_{x_1}(x, z, p), \dots, L_{x_n}(x, z, p)),$$

and

$$\nabla_p L(x, z, p) = (L_{p_1}(x, z, p), \dots, L_{p_n}(x, z, p)).$$

Each example from Section 1.1 involved a functional of the general form of (2.1). For the shortest path problem $d = 1$ and

$$L(x, z, p) = g(x_1, z)\sqrt{1 + p_1^2}.$$

For the brachistochrone problem $d = 1$ and

$$L(x, z, p) = \sqrt{\frac{1 + p_1^2}{-z}}.$$

For the minimal surface problem $n = 2$ and

$$L(x, z, p) = \sqrt{1 + |p|^2}.$$

For the image restoration problem $n = 2$ and

$$L(x, z, p) = \frac{1}{2}(z - f(x))^2 + \lambda|p|.$$

Finally, for the image segmentation problem

$$L(x, z, p) = H(z)(f(x) - a)^2 + (1 - H(z))\,(f(x) - b)^2 + \lambda\delta(z)|p|.$$

We will always assume that L is smooth, and the boundary condition $g : \partial U \to \mathbb{R}$ is smooth. We now give necessary conditions for minimizers of I.

Theorem 2.1 (Euler-Lagrange equation)**.** *Suppose that $u \in C^2(\overline{U})$ satisfies*

$$I(u) \leq I(v) \tag{2.2}$$

for all $v \in C^2(\overline{U})$ with $v = u$ on ∂U. Then

$$L_z(x, u, \nabla u) - \operatorname{div}(\nabla_p L(x, u, \nabla u)) = 0 \quad \textit{in } U. \tag{2.3}$$

Proof. Let $\varphi \in C_c^\infty(U)$ and set $v = u + t\varphi$ for a real number t. Since $\varphi = 0$ on ∂U we have $u = v$ on ∂U. Thus, by assumption

$$I(u) \leq I(v) = I(u + t\varphi) \quad \text{for all } t \in \mathbb{R}.$$

This means that $h(t) := I(u + t\varphi)$ has a global minimum at $t = 0$, i.e., $h(0) \leq h(t)$ for all t. It follows that $h'(t) = 0$, which is equivalent to

$$\left.\frac{d}{dt}\right|_{t=0} I(u + t\varphi) = 0. \tag{2.4}$$

We now compute the derivative in (2.4). Notice that

$$I(u + t\varphi) = \int_U L(x, u(x) + t\varphi(x), \nabla u(x) + t\nabla\varphi(x))\,dx.$$

For notational simplicity, let us suppress the x arguments from $u(x)$ and $\varphi(x)$. By the chain rule

$$\frac{d}{dt}L(x, u+t\varphi, \nabla u+t\nabla\varphi) = L_z(x, u+t\varphi, \nabla u+t\nabla\varphi)\varphi+\nabla_p L(x, u+t\varphi, \nabla u+t\nabla\varphi)\cdot\nabla\varphi.$$

Therefore

$$\left.\frac{d}{dt}\right|_{t=0} L(x, u + t\varphi, \nabla u + t\nabla\varphi) = L_z(x, u, \nabla u)\varphi + \nabla_p L(x, u, \nabla u) \cdot \nabla\varphi,$$

and we have

$$\begin{aligned}\frac{d}{dt}\Big|_{t=0} I(u+t\varphi) &= \frac{d}{dt}\Big|_{t=0}\int_U L(x,u(x)+t\varphi(x),\nabla u(x)+t\nabla\varphi(x))\,dx\\ &= \int_U \frac{d}{dt}\Big|_{t=0} L(x,u(x)+t\varphi(x),\nabla u(x)+t\nabla\varphi(x))\,dx\\ &= \int_U L_z(x,u,\nabla u)\varphi + \nabla_p L(x,u,\nabla u)\cdot\nabla\varphi\,dx\\ &= \int_U L_z(x,u,\nabla u)\varphi\,dx + \int_U \nabla_p L(x,u,\nabla u)\cdot\nabla\varphi\,dx. \qquad (2.5)\end{aligned}$$

Since $\varphi = 0$ on ∂U we can use the Divergence Theorem (Theorem A.32) to compute

$$\int_U \nabla_p L(x,u,\nabla u)\cdot\nabla\varphi\,dx = -\int_U \mathrm{div}\,(\nabla_p L(x,u,\nabla u))\,\varphi\,dx.$$

Combining this with (2.4) and (2.5) we have

$$0 = \frac{d}{dt}\Big|_{t=0} I(u+t\varphi) = \int_U \big(L_z(x,u,\nabla u) - \mathrm{div}\,(\nabla_p L(x,u,\nabla u))\big)\varphi\,dx.$$

It follows from the vanishing lemma (Lemma A.43 in the appendix) that

$$L_z(x,u,\nabla u) - \mathrm{div}\,(\nabla_p L(x,u,\nabla u)) = 0$$

everywhere in U, which completes the proof. □

Remark 2.2. Theorem 2.1 says that minimizers of the functional I satisfy the PDE (2.3). The PDE (2.3) is called the *Euler-Lagrange* equation for I.

Definition 2.3. A solution u of the Euler-Lagrange equation (2.3) is called a *critical point* of I.

Remark 2.4. In dimension $d = 1$ we write $x = x_1$ and $p = p_1$. Then the Euler-Lagrange equation is

$$L_z(x,u(x),u'(x)) - \frac{d}{dx}L_p(x,u(x),u'(x)) = 0.$$

Remark 2.5. In the proof of Theorem 2.1 we showed that

$$\int_U L_z(x,u,\nabla u)\varphi + \nabla_p L(x,u,\nabla u)\cdot\nabla\varphi\,dx = 0$$

for all $\varphi \in C_c^\infty(U)$. A function $u \in C^1(U)$ satisfying the above for all $\varphi \in C_c^\infty(U)$ is called a *weak* solution of the Euler-Lagrange equation (2.3). Thus, weak solutions of PDEs arise naturally in the calculus of variations.

Example 2.1. Consider the problem of minimizing the Dirichlet energy

$$I(u) = \int_U \frac{1}{2}|\nabla u|^2 - uf\,dx, \tag{2.6}$$

over all u satisfying $u = g$ on ∂U. Here, $f : U \to \mathbb{R}$ and $g : \partial U \to \mathbb{R}$ are given functions, and

$$L(x, z, p) = \frac{1}{2}|p|^2 - zf(x).$$

Therefore

$$L_z(x, z, p) = -f(x) \quad \text{and} \quad \nabla_p L(x, z, p) = p,$$

and the Euler-Lagrange equation is

$$-f(x) - \operatorname{div}(\nabla u) = 0 \ \text{ in } U.$$

This is Poisson's equation

$$-\Delta u = f \ \text{ in } U$$

subject to the boundary condition $u = g$ on ∂U. △

Exercise 2.6. Derive the Euler-Lagrange equation for the problem of minimizing

$$I(u) = \int_U \frac{1}{p}|\nabla u|^p - uf\,dx$$

subject to $u = g$ on ∂U, where $p \geq 1$. △

Example 2.2. The Euler-Lagrange equation in dimension $d = 1$ can be simplified when L has no x-dependence, so $L = L(z, p)$. In this case the Euler-Lagrange equation reads

$$L_z(u(x), u'(x)) = \frac{d}{dx} L_p(u(x), u'(x)).$$

Using the Euler-Lagrange equation and the chain rule we compute

$$\begin{aligned}
\frac{d}{dx} L(u(x), u'(x)) &= L_z(u(x), u'(x))u'(x) + L_p(u(x), u'(x))u''(x) \\
&= u'(x)\frac{d}{dx} L_p(u(x), u'(x)) + L_p(u(x), u'(x))u''(x) \\
&= \frac{d}{dx}\left(u'(x)L_p(u(x), u'(x))\right).
\end{aligned}$$

Therefore

$$\frac{d}{dx}\left(L(u(x), u'(x)) - u'(x)L_p(u(x), u'(x))\right) = 0.$$

It follows that

$$L(u(x), u'(x)) - u'(x)L_p(u(x), u'(x)) = C \tag{2.7}$$

for some constant C. This form of the Euler-Lagrange equation is often easier to solve when L has no x-dependence. △

In some of the examples presented in Section 1.1, such as the image segmentation and restoration problems, we did not impose any boundary condition on the minimizer u. For such problems, Theorem 2.1 still applies, but the Euler-Lagrange equation (2.3) is not uniquely solvable without a boundary condition. Hence, we need some additional information about minimizers in order for the Euler-Lagrange equation to be useful for these problems.

Theorem 2.7. *Suppose that $u \in C^2(\overline{U})$ satisfies*

$$I(u) \leq I(v) \tag{2.8}$$

for all $v \in C^2(\overline{U})$. Then u satisfies the Euler-Lagrange equation (2.3) with boundary condition

$$\nabla_p L(x, u, \nabla u) \cdot \nu = 0 \quad \text{on} \ \ \partial U. \tag{2.9}$$

Proof. By Theorem 2.1, u satisfies the Euler-Lagrange equation (2.3). We just need to show that u also satisfies the boundary condition (2.9).

Let $\varphi \in C^\infty(\overline{U})$ be a smooth function that is not necessarily zero on ∂U. Then by hypothesis $I(u) \leq I(u + t\varphi)$ for all t. Therefore, as in the proof of Theorem 2.1 we have

$$0 = \frac{d}{dt}\bigg|_{t=0} I(u + t\varphi) = \int_U L_z(x, u, \nabla u)\varphi + \nabla_p L(x, u, \nabla u) \cdot \nabla\varphi \, dx. \tag{2.10}$$

Applying the Divergence Theorem (Theorem A.32) we find that

$$0 = \int_{\partial U} \varphi \, \nabla_p L(x, u, \nabla u) \cdot \nu \, dS + \int_U \left(L_z(x, u, \nabla u) - \operatorname{div}\left(\nabla_p L(x, u, \nabla u)\right)\right) \varphi \, dx.$$

Since u satisfies the Euler-Lagrange equation (2.3), the second term above vanishes and we have

$$\int_{\partial U} \varphi \, \nabla_p L(x, u, \nabla u) \cdot \nu \, dS = 0$$

for all test functions $\varphi \in C^\infty(\overline{U})$. By a slightly different version of the vanishing lemma (Lemma A.43 in the appendix) we have that

$$\nabla_p L(x, u, \nabla u) \cdot \nu = 0 \quad \text{on} \ \ \partial U.$$

This completes the proof. □

2.1 The gradient interpretation

We can interpret the Euler-Lagrange equation (2.3) as the gradient of I. That is, in a certain sense (2.3) is equivalent to $\nabla I(u) = 0$.

To see why, let us consider a function $u : \mathbb{R}^d \to \mathbb{R}$. The gradient of u is defined as the vector of partial derivatives

$$\nabla u(x) = (u_{x_1}(x), u_{x_2}(x), \ldots, u_{x_n}(x)).$$

By the chain rule we have

$$\frac{d}{dt}\Big|_{t=0} u(x+tv) = \nabla u(x)\cdot v \tag{2.11}$$

for any vector $v \in \mathbb{R}^d$. It is possible to take (2.11) as the *definition* of the gradient of u. By this, we mean that $w = \nabla u(x)$ is the unique vector satisfying

$$\frac{d}{dt}\Big|_{t=0} u(x+tv) = w\cdot v$$

for all $v \in \mathbb{R}^d$.

In the case of functionals $I(u)$, we showed in the proof of Theorem 2.1 that

$$\frac{d}{dt}\Big|_{t=0} I(u+t\varphi) = \left(L_z(x,u,\nabla u) - \operatorname{div}\left(\nabla_p L(x,u,\nabla u)\right), \varphi\right)_{L^2(U)}$$

for all $\varphi \in C_c^\infty(U)$. Here, the L^2-inner product plays the role of the dot product from the finite dimensional case. Thus, it makes sense to define the *gradient*, also called the *functional gradient*, to be

$$\nabla I(u) := L_z(x,u,\nabla u) - \operatorname{div}\left(\nabla_p L(x,u,\nabla u)\right), \tag{2.12}$$

so that we have the identity

$$\frac{d}{dt}\Big|_{t=0} I(u+t\varphi) = (\nabla I(u), \varphi)_{L^2(U)}. \tag{2.13}$$

The reader should compare this with the ordinary chain rule (2.11). Notice the definition of the gradient ∇I depends on the choice of the L^2-inner product. Using other inner products will result in different notions of gradient.

2.2 The Lagrange multiplier

There are many problems in the calculus of variations that involve constraints on the feasible minimizers. A classic example is the *isoperimetric problem*, which corresponds to finding the shape of a simple closed curve that maximizes the enclosed area given the curve has a fixed length ℓ. Here we are maximizing the area enclosed by the curve subject to the constraint that the length of the curve is equal to ℓ.

Let J be a functional defined by

$$J(u) = \int_U H(x, u(x), \nabla u(x))\, dx.$$

We consider the problem of minimizing $I(u)$ subject to the constraint $J(u) = 0$. The Lagrange multiplier method gives necessary conditions that must be satisfied by any minimizer.

Theorem 2.8 (Lagrange multiplier)**.** *Suppose that $u \in C^2(\overline{U})$ satisfies $J(u) = 0$ and*

$$I(u) \leq I(v) \tag{2.14}$$

for all $v \in C^2(\overline{U})$ with $v = u$ on ∂U and $J(v) = 0$. Then there exists a real number λ such that

$$\nabla I(u) + \lambda \nabla J(u) = 0 \quad \text{in } U. \tag{2.15}$$

Here, ∇I and ∇J are the functional gradients of I and J, respectively, defined by (2.12).

Remark 2.9. The number λ in Theorem 2.8 is called a *Lagrange multiplier*.

Proof. We will give a short sketch of the proof. Let $\varphi \in C_c^\infty(U)$. Then as in Theorem 2.1

$$\frac{d}{dt}\Big|_{t=0} I(u + t\varphi) = (\nabla I(u), \varphi)_{L^2(U)} \quad \text{and} \quad \frac{d}{dt}\Big|_{t=0} J(u + t\varphi) = (\nabla J(u), \varphi)_{L^2(U)}.$$

Suppose that $(\nabla J(u), \varphi)_{L^2(U)} = 0$. Then, up to a small approximation error

$$0 = J(u) = J(u + t\varphi)$$

for small t. Since $\varphi = 0$ on ∂U, we also have $u = u + t\varphi$ on ∂U. Thus by hypothesis

$$I(u) \leq I(u + t\varphi) \quad \text{for all small } t.$$

Therefore, $t \mapsto I(u + t\varphi)$ has a minimum at $t = 0$ and so

$$(\nabla I(u), \varphi)_{L^2(U)} = \frac{d}{dt}\Big|_{t=0} I(u + t\varphi) = 0.$$

Hence, we have shown that for all $\varphi \in C_c^\infty(U)$

$$(\nabla J(u), \varphi)_{L^2(U)} = 0 \implies (\nabla I(u), \varphi)_{L^2(U)} = 0.$$

This says that $\nabla I(u)$ is orthogonal to everything that is orthogonal to $\nabla J(u)$. Intuitively this must imply that $\nabla I(u)$ and $\nabla J(u)$ are co-linear; we give the proof below.

We now have three cases.

1. If $\nabla J(u) = 0$ then $(\nabla I(u), \varphi)_{L^2(U)} = 0$ for all $\varphi \in C^\infty(U)$, and by the vanishing lemma $\nabla I(u) = 0$. Here we can choose any real number for λ.

2. If $\nabla I(u) = 0$ then we can take $\lambda = 0$ to complete the proof.

3. Now we can assume $\nabla I(u) \neq 0$ and $\nabla J(u) \neq 0$. Define

$$\lambda = -\frac{(\nabla I(u), \nabla J(u))_{L^2(U)}}{\|\nabla J(u)\|_{L^2(U)}} \quad \text{and} \quad v = \nabla I(u) + \lambda \nabla J(u).$$

By the definition of λ we can check that

$$(\nabla J(u), v)_{L^2(U)} = 0.$$

Therefore $(\nabla I(u), v)_{L^2(U)} = 0$ and we have

$$\begin{aligned} 0 &= (\nabla I(u), v)_{L^2(U)} \\ &= (v - \lambda \nabla J(u), v)_{L^2(U)} \\ &= (v, v)_{L^2(U)} - \lambda (\nabla J(u), v)_{L^2(U)} \\ &= \|v\|^2_{L^2(U)}. \end{aligned}$$

Therefore $v = 0$ and so $\nabla I(u) + \lambda \nabla J(u) = 0$. This completes the proof. □

Remark 2.10. Notice that (2.15) is equivalent to the necessary conditions minimizers for the augmented functional

$$K(u) = I(u) + \lambda J(u).$$

2.3 Gradient descent

To numerically compute solutions of the Euler-Lagrange equation $\nabla I(u) = 0$, we often use *gradient descent*, which corresponds to solving the PDE

$$\tag{2.16} \begin{cases} u_t(x,t) = -\nabla I(u(x,t)), & \text{in } U \times (0,\infty) \\ u(x,0) = u_0(x), & \text{on } U \times \{t = 0\}. \end{cases}$$

We also impose the boundary condition $u = g$ or $\nabla_p L(x, u, \nabla u) \cdot \nu = 0$ on $\partial U \times (0,\infty)$, depending on which condition was imposed in the variational problem. Gradient descent evolves u in the direction that decreases I most rapidly, starting at some initial guess u_0. If we reach a stationary point where $u_t = 0$ then we have found a solution of the Euler-Lagrange equation $\nabla I(u) = 0$. If solutions of the Euler-Lagrange equation are not unique, we may find different solutions depending on the choice of u_0.

To see that gradient descent actually decreases I, let $u(x,t)$ solve (2.16) and compute

$$\begin{aligned}\frac{d}{dt}I(u) &= \int_U \frac{d}{dt}L(x,u(x,t),\nabla u(x,t))\,dx\\ &= \int_U L_z(x,u,\nabla u)u_t + \nabla_p L(x,u,\nabla u)\nabla u_t\,dx\\ &= \int_U \left(L_z(x,u,\nabla u) - \operatorname{div}\left(\nabla_p L(x,u,\nabla u)\right)\right)u_t\,dx\\ &= (\nabla I(u), u_t)_{L^2(U)}\\ &= (\nabla I(u), -\nabla I(u))_{L^2(U)}\\ &= -\|\nabla I(u)\|_{L^2(U)}\\ &\leq 0.\end{aligned}$$

We used integration by parts in the third line, mimicking the proof of Theorem 2.1. Notice that either boundary condition $u = g$ or $\nabla_p L(x,u,\nabla u)\cdot\nu = 0$ on ∂U ensures the boundary terms vanish in the integration by parts.

Notice that by writing out the Euler-Lagrange equation we can write the gradient descent PDE (2.16) as

$$\left\{\begin{aligned} u_t + L_z(x,u,\nabla u) - \operatorname{div}\left(\nabla_p L(x,u,\nabla u)\right) = 0, &\quad \text{in } U\times(0,\infty)\\ u = u_0, &\quad \text{on } U\times\{t=0\}.\end{aligned}\right. \tag{2.17}$$

Example 2.3. Gradient descent on the Dirichlet energy (2.6) is the heat equation

$$\left\{\begin{aligned} u_t - \Delta u = f, &\quad \text{in } U\times(0,\infty)\\ u = u_0, &\quad \text{on } U\times\{t=0\}.\end{aligned}\right. \tag{2.18}$$

Thus, solving the heat equation is the fastest way to decrease the Dirichlet energy. △

We now prove that gradient descent converges to the unique solution of the Euler-Lagrange equation when I is strongly convex. For simplicity, we consider functionals of the form

$$I(u) = \int_U \Psi(x,u) + \Phi(x,\nabla u)\,dx. \tag{2.19}$$

This includes the regularized variational problems used for image denoising in Example 1.4 and Section 3.6, for example. The reader may want to review definitions and properties of convex functions in Section A.8 before reading the following theorem.

Theorem 2.11. *Assume I is given by (2.19), where $z\mapsto\Psi(x,z)$ is θ-strongly convex for $\theta>0$ and $p\mapsto\Phi(x,p)$ is convex. Let $u^*\in C^2(\overline{U})$ be any function satisfying*

$$\nabla I(u^*) = 0 \quad \textit{in} \ \ U. \tag{2.20}$$

Let $u(x,t) \in C^2(\overline{U} \times [0,\infty))$ *be a solution of the gradient descent equation* (2.16) *and assume that* $u = u^*$ *on* ∂U *for all* $t > 0$. *Then*

$$\|u - u^*\|^2_{L^2(U)} \le Ce^{-2\theta t} \quad \text{for all } t \ge 0, \tag{2.21}$$

where $C = \|u_0 - u^*\|^2_{L^2(U)}$.

Proof. Define the energy

$$e(t) = \frac{1}{2}\|u - u^*\|^2_{L^2(U)} = \frac{1}{2}\int_U (u(x,t) - u^*(x))^2\, dx. \tag{2.22}$$

We differentiate $e(t)$, use the equations solved by u and u^* and the divergence theorem to find that

$$\begin{aligned}
e'(t) &= \int_U (u - u^*)u_t\, dx \\
&= -\int_U (\nabla I(u) - \nabla I(u^*))(u - u^*)\, dx \\
&= -\int_U (\Psi_z(x,u) - \Psi_z(x,u^*))(u - u^*)\, dx \\
&\qquad\qquad - \int_U \operatorname{div}(\nabla_p\Phi(x,\nabla u) - \nabla_p\Phi(x,\nabla u^*))(u - u^*)\, dx \\
&= -\int_U (\Psi_z(x,u) - \Psi_z(x,u^*))(u - u^*)\, dx \\
&\qquad\qquad + \int_U (\nabla_p\Phi(x,\nabla u) - \nabla_p\Phi(x,\nabla u^*)) \cdot (\nabla u - \nabla u^*)\, dx.
\end{aligned}$$

Since $z \mapsto \Psi(x,z)$ is θ-strongly convex, Theorem A.38 yields

$$(\Psi_z(x,u) - \Psi_z(x,u^*))(u - u^*) \ge \theta(u - u^*)^2.$$

Likewise, since $p \mapsto \Phi(x,p)$ is convex, Theorem A.38 implies that

$$(\nabla_p\Phi(x,\nabla u) - \nabla_p\Phi(x,\nabla u^*)) \cdot (\nabla u - \nabla u^*) \ge 0.$$

Therefore

$$e'(t) \le -\theta\int_U (u - u^*)^2\, dx = -2\theta e(t). \tag{2.23}$$

Noting that

$$\frac{d}{dt}(e(t)e^{-2\theta t}) = e'(t)e^{-2\theta t} - 2\theta e(t)e^{-2\theta t} \le 0,$$

we have $e(t) \le e(0)e^{-2\theta t}$, which completes the proof. □

Remark 2.12. Notice that the proof of Theorem 2.11 shows that solutions of $\nabla I(u) = 0$, subject to Dirichlet boundary condition $u = g$ on ∂U, are unique.

Remark 2.13. From an optimization or numerical analysis point of view, the exponential convergence rate (2.21) is a *linear convergence rate*. Indeed, if we consider the decay in the energy $e(t)$ defined in (2.22) over a unit Δt in time, then Eq. (2.23) yields

$$e(t+\Delta t) - e(t) = \int_t^{t+\Delta t} e'(s)\,ds \le -2\theta \int_t^{t+\Delta t} e(s)\,ds \le -2\theta\Delta t e(t+\Delta t),$$

since e is decreasing. Therefore

$$\frac{e(t+\Delta t)}{e(t)} \le \frac{1}{1+2\theta\Delta t} =: \mu.$$

Hence, gradient descent converges with a convergence rate of $\mu \approx 1 - 2\theta\Delta t < 1$. It is important to point out that the convergence rate depends on the time step Δt with which we discretize the equation. This will be discussed further in Section 2.4.

2.4 Accelerated gradient descent

Theorem 2.11 shows that gradient descent converges at a linear rate in the strongly convex setting. This is what one expects from first order optimization methods that only use gradient information. To obtain higher order convergence rates, one usually needs to consider higher order methods, such as Newton's method. However, for large scale problems, computing and storing the Hessian matrix is intractable.

It is possible, however, to *accelerate* the convergence of gradient descent without using second order methods. In some situations, this can yield an order of magnitude better linear convergence rate, while adding no significant additional computational burden. We describe some of these methods, in the PDE setting, here.

Heuristically, one often hears gradient descent explained as a ball rolling down the energy landscape and landing in a valley corresponding to a local or global minimum. However, this is clearly a false analogy, since a rolling ball has momentum, and its velocity cannot change instantaneously, whereas gradient descent $u_t = -\nabla I(u)$ has no notion of momentum and the velocity depends only on the current position. The idea of *accelerated gradient descent* is based on introducing additional momentum terms into the descent equations, to make the equations appear more similar to the equations of motion for a heavy ball with friction rolling down the energy landscape. In optimization, the ideas date back to Polyak's heavy ball method [48] and Nesterov's accelerated gradient descent [45]. Accelerated gradient descent combined with stochastic gradient descent is used to train modern machine learning algorithms, such as deep neural networks [39].

Accelerated gradient descent can be formulated in a Lagrangian setting, with striking connections to the calculus of variations. In Lagrangian mechanics, the equations of motion for the trajectory $\mathbf{x}(t)$ of a system is exactly the Euler-Lagrange equation for an action functional of the form

$$\int_a^b K(\mathbf{x}'(t)) - P(\mathbf{x}(t))\,dt$$

where K is the kinetic energy and P is the potential energy. The total energy $K+P$ is conserved in Lagrangian mechanics, and momentum arises through the kinetic energy K.

Example 2.4 (Mass on a spring)**.** Consider the motion of a mass on a spring with spring constant $k > 0$. Assume one end of the spring is fixed to a wall at position $x = 0$ and the other end attached to a mass at position $x(t) > 0$. The spring is relaxed (at rest) at position $x = a > 0$, and we assume the mass slides horizontally along a frictionless surface. Here, the kinetic energy is $\frac{1}{2}mx'(t)^2$, where m is the mass of the spring. The potential energy is the energy stored in the spring, and is given by $\frac{1}{2}k(x(t) - a)^2$, according to Hooke's law. Hence, the Lagrangian action is

$$\int_a^b \frac{1}{2}mx'(t)^2 - \frac{1}{2}k(x(t) - a)^2\,dt.$$

The corresponding Euler-Lagrange equation is

$$-\frac{d}{dt}(mx'(t)) - k(x(t) - a) = 0,$$

which reduces to the equations of motion $mx''(t) + k(x(t) - a) = 0$. △

Exercise 2.14. Let $f : \mathbb{R} \to \mathbb{R}$ and suppose a bead of mass $m > 0$ is sliding without friction along the graph of f under the force of gravity. Let g denote the acceleration due to gravity, and write the trajectory of the bead as $(x(t), f(x(t)))$. What are the equations of motion satisfied by $x(t)$. [Hint: Follow Example 2.4, using the appropriate expressions for kinetic and potential energy. Note that kinetic energy is not $\frac{1}{2}mx'(t)^2$, since this only measures horizontal movement, and bead is moving vertically as well. The potential energy is $mgf(x(t))$.] △

In our setting, we may take the kinetic energy of $u = u(x,t)$ to be

$$K(u) = \frac{1}{2}\int_U u_t^2\,dx, \tag{2.24}$$

and the potential energy to be the functional $I(u)$ we wish to minimize, defined in (2.1). Then our action functional is

$$\int_a^b \int_U \frac{1}{2}u_t^2 - L(x, u, \nabla u)\,dx\,dt. \tag{2.25}$$

The equations of motion are simply the Euler-Lagrange equations for (2.25), treating the problem in $\mathbb{R}^{n+1}$ with (x, t) as our variables, which are

$$-\frac{\partial}{\partial t} u_t - (L_z - \operatorname{div}_x(\nabla_p L(x, u, \nabla u))) = 0.$$

This reduces to $u_{tt} = -\nabla I(u)$, which is a nonlinear wave equation in our setting. Unfortunately this wave equation will conserve energy, and is thus useless for optimization.

In order to ensure we descend on the energy $I(u)$, we need to introduce some dissipative agent into the problem, such as friction. In the Lagrangian setting, this can be obtained by adding time-dependent functions into the action functional, in the form

$$\int_a^b \int_U e^{at} \left(\frac{1}{2} u_t^2 - L(x, u, \nabla u) \right) dx\, dt, \tag{2.26}$$

where $a > 0$. The Euler-Lagrange equations are now

$$-\frac{\partial}{\partial t}(e^{at} u_t) - e^{at}(L_z - \operatorname{div}_x(\nabla_p L(x, u, \nabla u))) = 0,$$

which reduce to

$$u_{tt} + a u_t = -\nabla I(u). \tag{2.27}$$

Notice that in the descent equation (2.27), the gradient $\nabla I(u)$ acts as a forcing term that causes acceleration, whereas in gradient descent $u_t = -\nabla I(u)$, the gradient is the velocity. The additional term au_t is a damping term, with the physical interpretation of friction for a rolling (or sliding) ball, with $a > 0$ the coefficient of friction. The damping term dissipates energy, allowing the system to converge to a critical point of I. We note it is possible to make other modifications to the action functional (2.26) to obtain a more general class of descent equations (see, e.g., [17]).

The accelerated descent equation (2.27) does not necessarily descend on $I(u)$. However, since we derived the equation from a Lagrangian variational formulation, we do have monotonicity of total energy.

Lemma 2.15 (Energy monotonicity [17, Lemma 1])**.** *Assume $a \geq 0$ and let u satisfy (2.27). Suppose either $u(x, t) = g(x)$ or $\nabla_p L(x, u, \nabla u) \cdot \mathbf{n} = 0$ for $x \in \partial U$ and $t > 0$. Then for all $t > 0$ we have*

$$\frac{d}{dt}\left(K(u) + I(u)\right) = -2aK(u). \tag{2.28}$$

Proof.

$$\begin{aligned}\frac{d}{dt}I(u) &= \int_U \frac{d}{dt}L(x,u,\nabla u)\,dx \\ &= \int_U L_z(x,u,\nabla u)u_t + \nabla_p L(x,u,\nabla u)\nabla u_t\,dx \\ &= \int_U L_z(x,u,\nabla u)u_t - \mathrm{div}_x(\nabla_p L(x,u,\nabla u))u_t\,dx \\ &\qquad\qquad + \int_{\partial U} u_t \nabla_p L(x,u,\nabla u)\cdot\nu\,dS \\ &= \int_U \nabla I(u)u_t\,dx + \int_{\partial U} u_t\nabla_p L(x,u,\nabla u)\cdot\nu\,dS.\end{aligned} \tag{2.29}$$

Due to the boundary condition on ∂U, either $u_t = 0$ or $\nabla_p L(x,u,\nabla u)\cdot\mathbf{n} = 0$ on ∂U. Therefore

$$\frac{d}{dt}I(u) = \int_U \nabla I(u)u_t\,dx. \tag{2.30}$$

Similarly, we have

$$\begin{aligned}\frac{d}{dt}K(u) &= \int_U u_t u_{tt}\,dx \\ &= \int_U u_t(-au_t - \nabla I(u))\,dx \\ &= -2aK(u) - \int_U \nabla I(u)u_t\,dx,\end{aligned} \tag{2.31}$$

where we used the equations of motion (2.27) in the second step above. Adding (2.30) and (2.31) completes the proof. □

It is possible to prove an exponential convergence rate for the accelerated descent equations (2.27) to the solution u^* of $\nabla I(u^*) = 0$ in the strongly convex setting, as we did in Theorem 2.11 for gradient descent. The proof uses energy methods, as in Theorem 2.11, but is more involved, so we refer the reader to [17, Theorem 1].

There is a natural question the reader may have at this point. If both gradient descent and accelerated gradient descent have similar exponential convergence rates, why would acceleration converge faster than gradient descent? The answer is that the accelerated convergence rate is invisible at the level of the continuum descent equations (2.16) and (2.27). The acceleration is realized only when the equations are discretized in space-time, which is necessary for computations. Recall from Remark 2.13 that the convergence rate for a discretization with time step Δt is

$$\mu \approx 1 - c\Delta t,$$

where $c > 0$ is the exponential rate. Hence, the convergence rate μ depends crucially on how large we can take the time step Δt; larger time steps yield much faster convergence rates. The time step is restricted by stability considerations in the discrete schemes, also called Courant-Friedrichs-Lewy (CFL) conditions.

The gradient descent equation (2.16) is first order in time and second order in space, so the ratio $\frac{\Delta t}{\Delta x^2}$ appears in the discretization of the descent equation, where Δx is the spatial resolution. Stability considerations require bounding this ratio and so $\Delta t \leq C\Delta x^2$ turns out to be the stability condition for gradient descent equations in this setting. This is a very restrictive stability condition, and such equations are called *numerically stiff*. This yields a convergence rate of the form

$$\text{Gradient descent rate } \mu \approx 1 - C\Delta x^2 \tag{2.32}$$

for gradient descent.

On the other hand, the accelerated gradient descent equation (2.27) is second order in space *and* time. Hence, the ratio $\frac{\Delta t^2}{\Delta x^2}$ appears in the discrete scheme, and stability conditions amount to bounding $\Delta t \leq C\Delta x$. Another way to see this is that wave equations have bounded speed of propagation c, and the CFL condition amounts to ensuring that the speed of propagation of the discrete scheme is at least as fast as c, that is $\frac{\Delta x}{\Delta t} \geq c$. Thus, accelerated gradient descent has a convergence rate of the form

$$\text{Accelerated gradient descent rate } \mu \approx 1 - C\Delta x. \tag{2.33}$$

Thus, as a function of the grid resolution Δx, the accelerated equation (2.27) has a convergence rate that is a whole order of magnitude better than the gradient descent equation (2.16). At a basic level, gradient descent is a diffusion equation, which is numerically stiff, while accelerated descent is a wave equation, which does not exhibit the same stiffness. The discussion above is only heuristic; we refer the reader to [17] for more details.

We present below the results of a simple numerical simulation to show the difference between gradient descent and accelerated gradient descent. We consider solving the Dirichlet problem

$$\begin{cases} \Delta u = 0 & \text{in } U \\ u = g & \text{on } \partial U, \end{cases} \tag{2.34}$$

in which case $\nabla I(u) = -\Delta u$ in both (2.27) and (2.16). We choose $a = 2\pi$[1], $g(x_1, x_2) = \sin(2\pi x_1^2) + \sin(2\pi x_2^2)$, and use initial condition $u(x, 0) = g(x)$. We discretize using the standard 5-point stencil for Δu in $n = 2$ dimensions, and use forward differences for u_t and centered differences for u_{tt}. We choose the largest stable time step for each scheme and run the corresponding descent equations until the discretization of (2.34)

	Accelerated descent		**Gradient Descent**	
Mesh	Time (s)	Iterations	Time (s)	Iterations
64^2	0.012	399	0.148	8404
128^2	0.05	869	2.4	3872
256^2	0.38	1898	40	174569
512^2	4.8	4114	1032	774606
1024^2	41	8813	23391	3399275

Table 2.1: Comparison of PDE acceleration (2.27) and gradient descent (2.16) for solving the Dirichlet problem (2.34). Runtimes are for C code, and copied from [17].

is satisfied up to an $O(\Delta x^2)$ error. The iteration counts and computation times for C code are shown in Table 2.1 for different grid resolutions.

We note that many other methods are faster for linear equations, such as multigrid or Fast Fourier Transform. The point here is to contrast the difference between gradient descent and accelerated gradient descent. The accelerated descent method works equally well for non-linear problems where multigrid and FFT methods do not work, or in the case of multigrid take substantial work to apply.

Remark 2.16. We have not taken into account discretization errors in our analysis above. In practice, the solution is computed on a grid of spacing Δx in the space dimensions and Δt in time. The discretization errors accumulate over time, and need to be controlled to ensure convergence of the numerical scheme.

For gradient descent, the PDE is first order in space and second order in time, and the discretization errors on $U \times [0, T]$ are bounded by $CT(\Delta t + \Delta x^2)$, which is proved in standard numerical analysis books. Thus, by (2.21) the error between the discrete solution and the minimizer u^* is bounded by

$$\|u_{\Delta t,\Delta x} - u^*\|^2_{L^2(U)} \leq C(e^{-2\theta t} + t^2(\Delta t^2 + \Delta x^4)) \quad \text{for all } t \geq 0,$$

where $u_{\Delta t,\Delta x}$ denotes any piecewise constant extension of the numerical solution to U. Furthermore, since $\Delta t \leq C\Delta x^2$ is necessary for stability and convergence of the gradient descent equation, we have the bound

$$\|u_{\Delta t,\Delta x} - u^*\|^2_{L^2(U)} \leq C(e^{-2\theta t} + t^2 \Delta x^4).$$

Choosing $T = -\log(\Delta x^4)/2\theta$ as our stopping time, we have

$$\|u_{\Delta t,\Delta x}(\cdot, T) - u^*(\cdot, T)\|_{L^2(U)} \leq C\theta^{-1}\log(\Delta x^{-1})\Delta x^2.$$

Hence, our numerical solution computes the true solution u^* up to $O(\Delta x^2)$ accuracy (excluding log factors) in $k = \frac{T}{\Delta t} = \frac{C}{\Delta x^2}$ iterations. The analysis for accelerated

[1] The choice $a = 2\pi$ is optimal. See Exercise 2.18 for an explanation in a simple setting.

gradient descent is identical, except here $\Delta t \leq C\Delta x$ is sufficient for stability and convergence of the scheme, and so only $k = \frac{C}{\Delta x}$ iterations are needed for convergence.

Exercise 2.17. Consider the linear ordinary differential equation (ODE) version of gradient descent

$$\mathbf{x}'(t) = \mathbf{b} - \mathcal{A}\,\mathbf{x}(t), \quad t > 0, \tag{2.35}$$

where $\mathbf{x} : [0, \infty) \to \mathbb{R}^d$, $\mathbf{b} \in \mathbb{R}^d$, and the matrix $\mathcal{A} \in \mathbb{R}^{n\times n}$ is symmetric and positive definite. Prove that for any initial condition $\mathbf{x}(0) = \mathbf{x}_0$, the solution of (3.1) converges exponentially fast (as in Theorem 2.11) to the solution of $\mathcal{A}\,\mathbf{x} = \mathbf{b}$. How does the convergence rate depend on $\mathcal{A}$? [Hint: Since $\mathcal{A}$ is symmetric and positive definite, there is an orthonormal basis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ for $\mathbb{R}^d$ consisting of eigenvectors of $\mathcal{A}$ with corresponding eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, counted with multiplicity. The solution of (2.35) can be expressed in the form

$$x(t) = \sum_{i=1}^{d} d_i(t)\mathbf{v}_i$$

for corresponding functions $d_1(t), \ldots, d_n(t)$. Find the $d_i(t)$ and complete the problem from here.] △

Exercise 2.18. Consider the linear ordinary differential equation (ODE) version of acceleration

$$\mathbf{x}''(t) + a\,\mathbf{x}'(t) = \mathbf{b} - \mathcal{A}\,\mathbf{x}(t), \quad t > 0, \tag{2.36}$$

where $\mathbf{x} : [0, \infty) \to \mathbb{R}^d$, $\mathbf{b} \in \mathbb{R}^d$, and the matrix $\mathcal{A} \in \mathbb{R}^{n\times n}$ is symmetric and positive definite. Prove that for any initial condition $\mathbf{x}(0) = \mathbf{x}_0$, the solution of (3.1) converges exponentially fast (as in Theorem 2.11) to the solution of $\mathcal{A}\,\mathbf{x} = \mathbf{b}$. How does the convergence rate depend on $\mathcal{A}$ and a? What is the optimal choice of the damping coefficient $a > 0$ to obtain the fastest convergence rate, and how does the convergence rate compare with the rate for gradient descent obtained in Exercise 2.17? [Hint: Use the same approach as in Exercise 2.17.] △

Notes

We note the presentation here is an adaptation of the variational interpretation of acceleration due to Wibisono, Wilson, and Jordan [63] to the PDE setting. The framework is called *PDE acceleration*; for more details see [7, 17, 56, 66]. We also mention the dynamical systems perspective is investigated in [3].

Chapter 3

Applications

We now continue the parade of examples by computing and solving the Euler-Lagrange equations for the examples from Section 1.1.

3.1 Shortest path

Recall for the shortest path problem we wish to minimize

$$I(u) = \int_0^a g(x, u(x))\sqrt{1 + u'(x)^2}\, dx,$$

subject to $u(0) = 0$ and $u(a) = b$. Here $d = 1$ and

$$L(x, z, p) = g(x, z)\sqrt{1 + p^2}.$$

Therefore $L_z(x, z, p) = g_z(x, z)\sqrt{1 + p^2}$ and $L_p(x, z, p) = g(x, z)(1 + p^2)^{-\frac{1}{2}}p$. The Euler-Lagrange equation is

$$g_z(x, u(x))\sqrt{1 + u'(x)^2} - \frac{d}{dx}\left(g(x, u(x))(1 + u'(x)^2)^{-\frac{1}{2}}u'(x)\right) = 0.$$

This is in general difficult to solve. In the special case that $g(x, z) = 1$, $g_z = 0$ and this reduces to

$$\frac{d}{dx}\left(\frac{u'(x)}{\sqrt{1 + u'(x)^2}}\right) = 0.$$

Computing the derivative yields

$$\frac{\sqrt{1 + u'(x)^2}u''(x) - u'(x)(1 + u'(x)^2)^{-\frac{1}{2}}u'(x)u''(x)}{1 + u'(x)^2} = 0.$$

Multiplying both sides by $\sqrt{1 + u'(x)^2}$ we obtain

$$(1 + u'(x)^2)u''(x) - u'(x)^2u''(x) = 0.$$

This reduces to $u''(x) = 0$, hence the solution is a straight line! This verifies our intuition that the shortest path between two points is a straight line.

3.2 The brachistochrone problem

Recall for the brachistochrone problem we wish to minimize

$$I(u) = \frac{1}{\sqrt{2g}} \int_0^a \sqrt{\frac{1+u'(x)^2}{-u(x)}}\, dx,$$

subject to $u(0) = 0$ and $u(a) = b$. Here, $d = 1$ and

$$L(x, z, p) = \sqrt{\frac{1+p^2}{-z}}.$$

Therefore

$$L_p(x, z, p) = \frac{p}{\sqrt{-z(1+p^2)}}.$$

Notice in this case that L has no x-dependence. Hence we can use the alternative form of the Euler-Lagrange equation (2.7), which yields

$$\sqrt{\frac{1+u'(x)^2}{-u(x)}} - \frac{u'(x)^2}{\sqrt{-u(x)(1+u'(x)^2)}} = C$$

for a constant C. Making some algebraic simplifications leads to

$$u(x) + u(x)u'(x)^2 = C, \tag{3.1}$$

where the constant C is different than the one on the previous line. The constant C should be chosen to ensure the boundary conditions hold.

Before solving (3.1), let us note that we can say quite a bit about the solution u from the ODE it solves. First, since $u(a) = b < 0$, the left hand side must be negative somewhere, hence $C < 0$. Solving for $u'(x)^2$ we have

$$u'(x)^2 = \frac{C - u(x)}{u(x)}.$$

This tells us two things. First, since the left hand side is positive, so is the right hand side. Hence $C - u(x) \leq 0$ and so $u(x) \geq C$. If u attains its maximum at an interior point $0 < x < a$ then $u'(x) = 0$ and $u(x) = C$, hence u is constant. This is impossible, so the maximum of u is attained at $x = 0$ and we have

$$C \leq u(x) < 0 \quad \text{for } 0 < x \leq a.$$

This means in particular that we must select $C \leq b$.

Second, as $x \to 0^+$ we have $u(x) \to 0^-$ and hence

$$\lim_{x\to 0^+} u'(x)^2 = \lim_{x\to 0^+} \frac{C - u(x)}{u(x)} = \lim_{t\to 0^-} \frac{C-t}{t} = \infty.$$

Since $x = 0$ is a local maximum of u, we have $u'(x) < 0$ for $x > 0$ small. Therefore

$$\lim_{x \to 0^+} u'(x) = -\infty.$$

This says that the optimal curve starts out vertical.

Third, we can differentiate (3.1) to find that

$$u'(x) + u'(x)^3 + 2u(x)u'(x)u''(x) = 0.$$

Therefore

$$1 + u'(x)^2 = -2u(x)u''(x). \tag{3.2}$$

Since the left hand side is positive, so is the right hand side, and we deduce $u''(x) > 0$. Therefore $u'(x)$ is strictly increasing, and u is a convex function of x. In fact, we can say slightly more. Solving for $u''(x)$ and using (3.1) we have

$$u''(x) = -\frac{1 + u'(x)^2}{2u(x)} = -\frac{(1 + u'(x)^2)^2}{2C} \geq -\frac{1}{2C} > 0.$$

This means that u is uniformly convex, and for large enough x we will have $u'(x) > 0$, so the optimal curve could eventually be *increasing*.

In fact, since u' is strictly increasing, there exists (by the intermediate value theorem) a unique point $x^* > 0$ such that $u'(x^*) = 0$. We claim that u is symmetric about this point, that is

$$u(x^* + x) = u(x^* - x).$$

To see this, write $w(x) = u(x^* + x)$ and $v(x) = u(x^* - x)$. Then

$$w'(x) = u'(x^* + x) \quad \text{and} \quad v'(x) = -u'(x^* - x).$$

Using (3.1) we have

$$w(x) + w(x)w'(x) = u(x^* + x) + u(x^* + x)u'(x^* + x)^2 = C$$

and

$$v(x) + v(x)v'(x) = u(x^* - x) + u(x^* - x)u'(x^* - x)^2 = C.$$

Differentiating the eqautions above, we find that both w and v satisfy the second order ODE (3.2). We also note that $w(0) = u(x^*) = v(0)$, and $w'(0) = u'(x_*) = 0$ and $v'(0) = -u'(x_*) = 0$, so $w'(0) = v'(0)$. Since solutions of the initial value problem for (3.2) are unique, provided their values stay away from zero, we find that $w(x) = v(x)$, which establishes the claim. The point of this discussion is that without explicitly computing the solution, we can say quite a bit both quantitatively and qualitatively about the solutions.

We now solve (3.1). Note we can write

$$u(x) = \frac{C}{1 + u'(x)^2}.$$

Since u' is strictly increasing, the angle θ between the tangent vector $(1, u'(x))$ and the vertical $(0, -1)$ is strictly increasing. Therefore, we can parametrize the curve in terms of this angle θ. Let us write $C(\theta) = (x(\theta), y(\theta))$. Then we have $y(\theta) = u(x)$ and

$$\sin^2 \theta = \frac{1}{1 + u'(x)^2}.$$

Therefore

$$y(\theta) = C \sin^2 \theta = -\frac{C}{2}(\cos(2\theta) - 1).$$

Since $y(\theta) = u(x)$

$$\frac{dy}{dx} = u'(x) = -\frac{\cos\theta}{\sin\theta}.$$

By the chain rule

$$x'(\theta) = \frac{dx}{d\theta} = \frac{dx}{dy}\frac{dy}{d\theta} = \left(-\frac{\sin\theta}{\cos\theta}\right)(2C \sin\theta\cos\theta) = -2C\sin^2\theta.$$

Therefore

$$x(\theta) = C\int_0^\theta \cos(2\theta) - 1\, dt = -\frac{C}{2}\left(2\theta - \sin(2\theta)\right).$$

This gives an explicit solution for the brachistochrone problem, where θ is just the parameter of the curve.

There is a nice geometrical interpretation of the brachistochrone curve. Notice that

$$\begin{bmatrix} x(\theta) \\ y(\theta) \end{bmatrix} = -\frac{C}{2}\begin{bmatrix} 2\theta \\ -1 \end{bmatrix} - \frac{C}{2}\begin{bmatrix} -\sin(2\theta) \\ \cos(2\theta) \end{bmatrix}.$$

The first term parametrizes the line $y = C/2$, while the second term traverses the circle of radius $r = -C/2$ in the counterclockwise direction. Thus, the curve is traced by a point on the rim of a circular wheel as the wheel rolls along the x-axis. Such a curve is called a *cycloid*.

Notice that the minimum occurs when $\theta = \frac{\pi}{2}$, and $y = C$ and $x = -\frac{C\pi}{2}$. Hence the minima of all brachistochrone curves lie on the line

$$x + \frac{\pi}{2}y = 0.$$

It follow that if $a + \frac{\pi}{2}b > 0$, then the optimal path starts traveling steeply downhill, reaches a low point, and then climbs uphill before arriving at the final point (a, b). If $a + \frac{\pi}{2}b \leq 0$ then the bead is always moving downhill. See Figure 3.1 for an illustration of the family of brachistochrone curves.

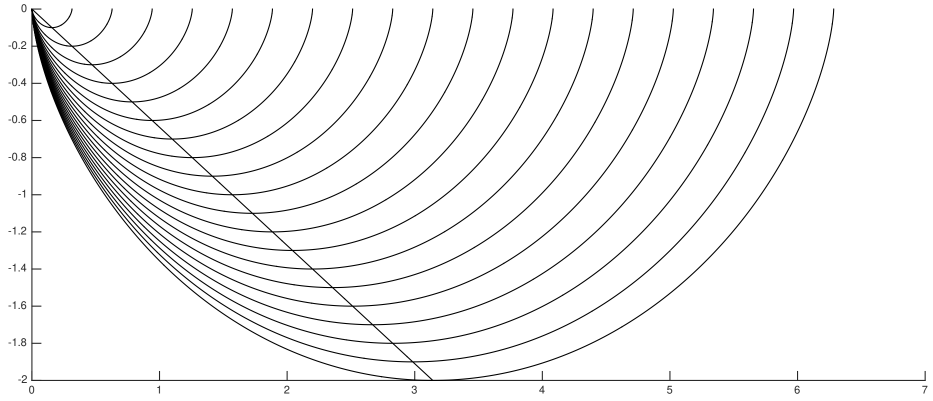


Figure 3.1: Family of brachistochrone curves. The straight line is the line $x + \frac{\pi}{2}y = 0$ passing through the minima of all brachistochrone curves.

Now, suppose instead of releasing the bead from the top of the curve, we release the bead from some position $(x_0, u(x_0))$ further down (but before the minimum) on the brachistochrone curve. How long does it take the bead to reach the lowest point on the curve? It turns out this time is the same regardless of where you place the bead on the curve! To see why, we recall that conservation of energy gives us

$$\frac{1}{2}mv(x)^2 + mgu(x) = mgu(x_0),$$

where $v(x)$ is the velocity of the bead. Therefore

$$v(x) = \sqrt{2g(u(x_0) - u(x))},$$

and the time to descend to the lowest point is

$$T = \frac{1}{\sqrt{2g}} \int_{x_0}^{-\frac{C\pi}{2}} \sqrt{\frac{1 + u'(x)^2}{u(x_0) - u(x)}}\, dx.$$

Recall that

$$1 + u'(x)^2 = \frac{1}{\sin^2\theta}, \;\; u(x) = y(\theta) = C\sin^2\theta, \quad \text{and } dx = -2C\sin^2\theta d\theta,$$

where $u(x_0) = y(\theta_0) = C\sin^2\theta_0$. Making the change of variables $x \to \theta$ yields

$$T = \sqrt{\frac{-2C}{g}} \int_{\theta_0}^{\frac{\pi}{2}} \frac{\sin\theta}{\sqrt{\sin^2\theta - \sin^2\theta_0}}\, d\theta = \sqrt{\frac{-2C}{g}} \int_{\theta_0}^{\frac{\pi}{2}} \frac{\sin\theta}{\sqrt{\cos^2\theta_0 - \cos^2\theta}}\, d\theta.$$

Make the change of variables $t = -\cos\theta/\cos\theta_0$. Then $\cos\theta_0 dt = \sin\theta d\theta$ and

$$T = \sqrt{\frac{-2C}{g}} \int_{-1}^{0} \frac{1}{\sqrt{1-t^2}}\,dt.$$

We can integrate this directly to obtain

$$T = \sqrt{\frac{-2C}{g}}\,(\arcsin(0) - \arcsin(-1)) = \pi\sqrt{\frac{-C}{2g}}.$$

Notice this is independent of the initial position x_0 at which the bead is released! A curve with the property that the time taken by an object sliding down the curve to its lowest point is independent of the starting position is called a *tautochrone*, or *isochrone* curve. So it turns out that the tautochrone curve is the same as the brachistochrone curve. The words tautochrone and isochrone are ancient Greek for same-time and equal-time, respectively.

3.3 Minimal surfaces

Recall for the minimal surface problem we wish to minimize

$$I(u) = \int_U \sqrt{1 + |\nabla u|^2}\,dx$$

subject to $u = g$ on ∂U. Here $n \geq 2$ and

$$L(x, z, p) = \sqrt{1 + |p|^2} = \sqrt{1 + p_1^1 + p_2^2 + \cdots + p_n^2}.$$

Even though minimal surfaces are defined in dimension $n = 2$, it can still be mathematically interesting to consider the general case of arbitrary dimension $n \geq 2$.

From the form of L we see that $L_z(x, z, p) = 0$ and

$$\nabla_p L(x, z, p) = \frac{p}{\sqrt{1 + p^2}}.$$

Therefore, the Euler-Lagrange equation for the minimal surface problem is

$$\text{div}\left(\frac{\nabla u}{\sqrt{1 + |\nabla u|^2}}\right) = 0 \quad \text{in } U \tag{3.3}$$

subject to $u = g$ on ∂U. This is called the *minimal surface equation.* Using the chain rule, we can rewrite the PDE as

$$\nabla\left(\frac{1}{\sqrt{1 + |\nabla u|^2}}\right) \cdot \nabla u + \frac{1}{\sqrt{1 + |\nabla u|^2}}\text{div}(\nabla u) = 0.$$

Notice that

$$\frac{\partial}{\partial x_j}\left(\frac{1}{\sqrt{1+|\nabla u|^2}}\right) = -\frac{1}{2}(1+|\nabla u|^2)^{-\frac{3}{2}}\sum_{i=1}^d 2u_{x_i}u_{x_ix_j}.$$

Therefore the PDE in expanded form is

$$-\frac{1}{(1+|\nabla u|^2)^{\frac{3}{2}}}\sum_{i,j=1}^d u_{x_ix_j}u_{x_i}u_{x_j} + \frac{\Delta u}{\sqrt{1+|\nabla u|^2}} = 0.$$

Multiplying both sides by $(1+|\nabla u|^2)^{\frac{3}{2}}$ we have

$$-\sum_{i,j=1}^d u_{x_ix_j}u_{x_i}u_{x_j} + (1+|\nabla u|^2)\Delta u = 0. \tag{3.4}$$

which is also equivalent to

$$-\nabla u\cdot\nabla^2 u\nabla u + (1+|\nabla u|^2)\Delta u = 0. \tag{3.5}$$

Exercise 3.1. Show that the plane

$$u(x) = a\cdot x + b$$

solves the minimal surface equation on $U = \mathbb{R}^d$, where $a\in\mathbb{R}^d$ and $b\in\mathbb{R}$. △

Exercise 3.2. Show that for $n = 2$ the Scherk surface

$$u(x) = \log\left(\frac{\cos(x_1)}{\cos(x_2)}\right)$$

solves the minimal surface equation on the box $U = (-\frac{\pi}{2},\frac{\pi}{2})^2$. △

Notice that if we specialize to the case of $n = 2$ then we have

$$-u_{x_1x_1}u_{x_1}^2 - 2u_{x_1x_2}u_{x_1}u_{x_2} - u_{x_2x_2}u_{x_2}^2 + (1+u_{x_1}^2+u_{x_2}^2)(u_{x_1x_1}+u_{x_2x_2}) = 0,$$

which reduces to

$$(1+u_{x_2}^2)u_{x_1x_1} - 2u_{x_1x_2}u_{x_1}u_{x_2} + (1+u_{x_1}^2)u_{x_2x_2} = 0. \tag{3.6}$$

It is generally difficult to find solutions of the minimal surface equation (3.4). It is possible to prove that a solution always exists and is unique, but this is outside the scope of this course.

3.4 Minimal surface of revolution

Suppose we place rings of equal radius $r > 0$ at locations $x = -L$ and $x = L$ on the x-axis. What is the resulting minimal surface formed between the rings? In other words, if we dip the rings into a soapy water solution, what is the shape of the resulting soap bubble?

Here, we may assume the surface is a surface of revolution. Namely, the surface is formed by taking a function $u : [-L, L] \to \mathbb{R}$ with $u(-L) = r = u(L)$ and rotating it around the x-axis. The surface area of a surface of revolution is

$$I(u) = 2\pi \int_{-L}^{L} u(x)\sqrt{1 + u'(x)^2}\, dx.$$

Since $L(x, z, p) = z\sqrt{1 + p^2}$ does not have an x-dependence, the Euler-Lagrange equation can be computed via (2.7) and we obtain

$$u(x)\sqrt{1 + u'(x)^2} - \frac{u'(x)^2 u(x)}{\sqrt{1 + u'(x)^2}} = \frac{1}{c}$$

for a constant $c \neq 0$. Multiplying both sides by $\sqrt{1 + u'(x)^2}$ and simplifying we have

$$cu(x) = \sqrt{1 + u'(x)^2}. \tag{3.7}$$

Before solving this, we make the important observation that at a minimum of u, $u'(x) = 0$ and hence $u(x) = \frac{1}{c}$ at the minimum. Since we are using a surface of revolution, we require $u(x) > 0$, hence we must take $c > 0$.

We now square both sides of (3.7) and rearrange to get

$$c^2 u(x)^2 - u'(x)^2 = 1. \tag{3.8}$$

Since $u'(x)^2 \geq 0$, we deduce that $c^2 u(x)^2 \geq 1$ for all. Since $u(L) = r$ we require

$$c \geq \frac{1}{r}. \tag{3.9}$$

To solve (3.8), we use a clever trick: We differentiate both sides to find

$$2c^2 u(x)u'(x) - 2u'(x)u''(x) = 0$$

or

$$u''(x) = c^2 u(x).$$

Notice we have converted a nonlinear ODE into a linear ODE! Since $c^2 > 0$ the general solution is

$$u(x) = \frac{A}{2}e^{cx} + \frac{B}{2}e^{-cx}.$$

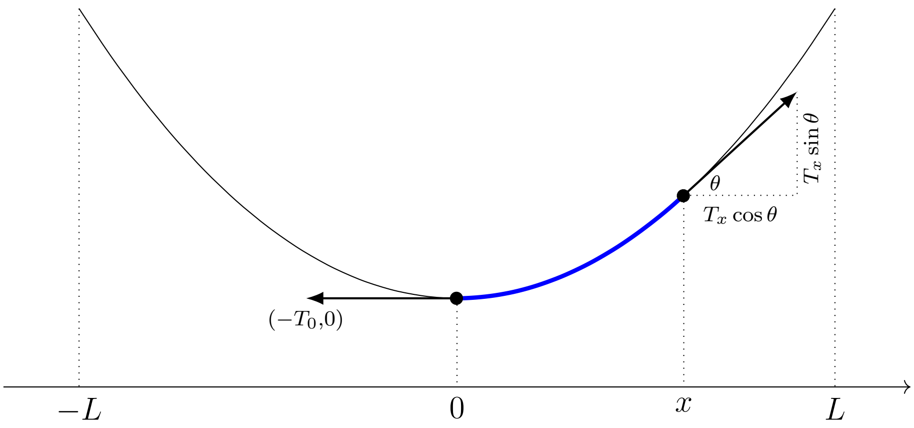


Figure 3.2: Depicition of some quantities from the derivation of the shape of a hanging chain. The shape of the curve is deduced by balancing the force of gravity and tension on the blue segment above.

It is possible to show, as we did for the brachistochrone problem, that u must be an even function. It follows that $A = B$ and

$$u(x) = A\frac{e^{cx} + e^{-cx}}{2} = A\cosh(cx).$$

The minimum of u occurs at $x = 0$, so

$$A = u(0) = \frac{1}{c}.$$

Therefore the solution is

$$u(x) = \frac{1}{c}\cosh(cx). \tag{3.10}$$

This curve is called a *catenary* curve, and the surface of revolution obtained by rotating it about the x-axis is called a *catenoid.*

Remark 3.3 (The shape of a hanging chain)**.** We pause for a moment to point out that the catenary curve is also the shape that an idealized hanging chain or cable assumes. Due to this property, an inverted catenary, or *catenary arch*, has been used in architectural designs since ancient times.

To see this, we assume the chain has length ℓ, with $\ell > 2L$, and is supported at $x = \pm L$ at a height of r, thus the shape of the chain $w(x)$ is a positive even function satisfying $w(\pm L) = r$. Consider the segment of the chain lying above the interval $[0, x]$ for $x > 0$. Let $s(x)$ denote the arclength of this segment, given by

$$s(x) = \int_0^x \sqrt{1 + w'(t)^2}\, dt.$$

If the chain has density ρ, then the force of gravity acting on this segment of the chain is $F = -\rho s(x) g$ in the vertical direction, where g is acceleration due to gravity. We also have tension forces within the rope acting on the segment at both ends of the interval. Let T_x denote the magnitude of the tension vector in the rope at position x, noting that the tension vector is always directed tangential to the curve. Since $w'(0) = 0$, as w is even, the tension vector at 0 is $(-T_0, 0)$. At the right point x of the interval, the tension vector is $T_x = (T_x \cos\theta, T_x \sin\theta)$, where θ is the angle between the tangent vector to w and the x-axis. Since the chain is at rest, the sum of the forces acting on the segment $[0, x]$ of rope must vanish, which yields

$$T_0 = T_x \cos\theta \quad \text{and} \quad \rho s(x) g = T_x \sin\theta.$$

Therefore

$$w'(x) = \tan\theta = \frac{T_x \sin\theta}{T_x \cos\theta} = \frac{\rho g}{T_0} s(x).$$

Setting $c = \rho g / T_0$ and differentiating above we obtain $w''(x) = c\sqrt{1 + w'(x)^2}$, and so

$$w''(x)^2 - c^2 w'(x)^2 = c^2.$$

Noting the similarity to (3.8), we can use the methods above to obtain the solution

$$w'(x) = \frac{cA}{2} e^{cx} + \frac{cB}{2} e^{-cx},$$

for some constants A and B. Since $w(x)$ is even, $w'(x)$ is odd, and so $B = -A$ and we have $w'(x) = cA \sinh(cx)$. Integrating again yields

$$w(x) = A\cosh(cx) + C$$

for an integration constant C. The difference now with the minimal surface problem is that A, C are the free parameters, while c is fixed.

Note that for fixed A, we can choose $C > 0$ so that $w(L) = r$, yielding

$$w(x) = A(\cosh(cx) - \cosh(cL)) + r.$$

We set the parameter A by asking that the curve has length ℓ, that is

$$\ell = 2\int_0^L \sqrt{1 + w'(x)^2}\, dx = 2\int_0^L \sqrt{1 + A^2c^2 \sinh^2(cx)}\, dx. \tag{3.11}$$

Unless $Ac = 1$, the integral on the right is not expressible in terms of simple functions. Nevertheless, we can see that the length of w is continuous and strictly increasing in A, and tends to ∞ as $A \to \infty$. Since the length of w for $A = 0$ is $2L$, and $\ell > 2L$, there exists, by the intermediate value theorem, a unique value of A for which (3.11) holds.

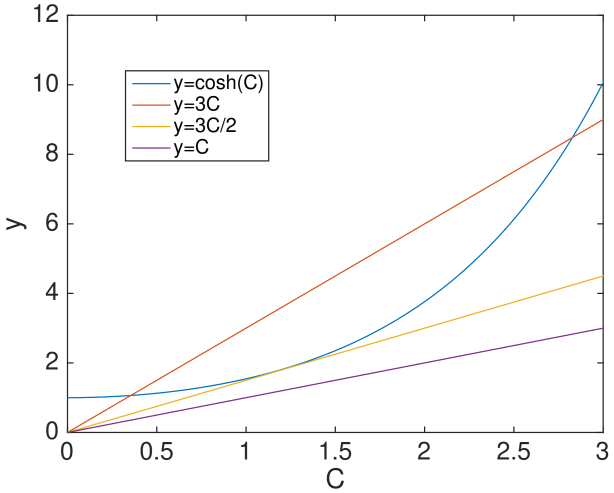


Figure 3.3: When $\theta > \theta_0$ there are two solutions of (3.12). When $\theta = \theta_0$ there is one solution, and when $\theta < \theta_0$ there are no solutions. By numerical computations, $\theta_0 \approx 1.509$.

We now return to the minimal surface problem. We need to see if it is possible to select $c > 0$ so that

$$u(-L) = r = u(L).$$

Since u is even, we only need to check that $u(L) = r$. This is equivalent choosing $c > 0$ so that $\cosh(cL) = cr$. Let us set $C = cL$ and $\theta = \frac{r}{L}$. Then we need to choose $C > 0$ such that

$$\cosh(C) = \theta C. \tag{3.12}$$

This equation is not always solvable, and depends on the value of $\theta = \frac{r}{L}$, that is, on the ratio of the radius r of the rings to L, which is half of the separation distance. There is a threshold value θ_0 such that for $\theta > \theta_0$ there are two solutions $C_1 < C_2$ of (3.12). When $\theta = \theta_0$ there is one solution C, and when $\theta < \theta_0$ there are no solutions. See Figure 3.3 for an illustration. To rephrase this, if $\theta < \theta_0$ or $r < L\theta_0$, then the rings are too far apart and there is no minimal surface spanning the two rings. If $r \geq L\theta_0$ then the rings are close enough together and a minimal surface exists. From numerical computations, $\theta_0 \approx 1.509$.

Now, when there are two solutions $C_1 < C_2$, which one gives the smallest surface area? We claim it is C_1. To avoid complicated details, we give here a heuristic argument to justify this claim. Let $c_1 < c_2$ such that $C_1 = c_1 L$ and $C_2 = c_2 L$. So we have two potential solutions

$$u_1(x) = \frac{1}{c_1}\cosh(c_1 x) \quad \text{and} \quad u_2(x) = \frac{1}{c_2}\cosh(c_2 x).$$

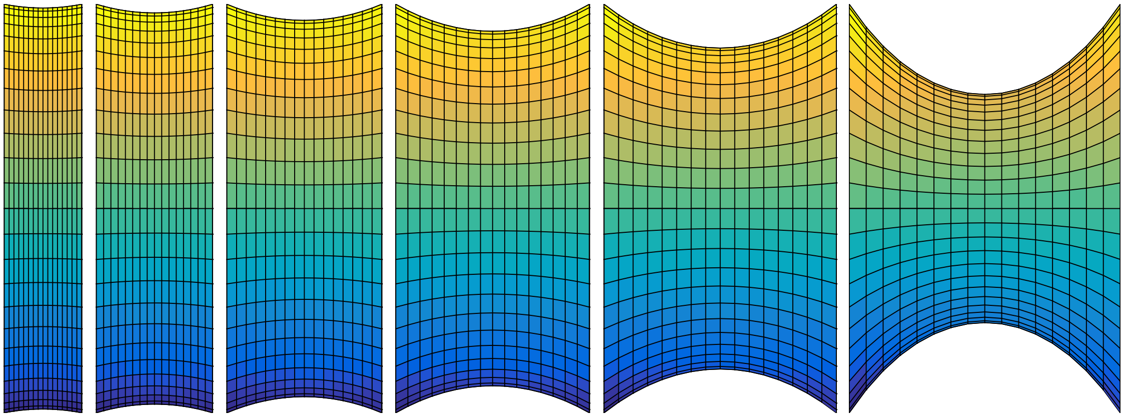


Figure 3.4: Simulation of the minimal surface of revolution for two rings being slowly pulled apart. The rings are located at $x = -L$ and $x = L$ where L ranges from (left) $L = 0.095$ to (right) $L = 0.662$, and both rings have radius $r = 1$. For larger L the soap bubble will collapse.

Since $u_1(0) = \frac{1}{c_1} \geq \frac{1}{c_2} = u_2(0)$, we have $u_1 \geq u_2$. In other words, as we increase c the solution decreases. Now, as we pull the rings apart we expect the solution to decrease (the soap bubble becomes thinner), so the value of c should increase as the rings are pulled apart. As the rings are pulled apart L is increasing, so $\theta = r/L$ is decreasing. From Figure 3.3 we see that C_2 is decreasing as θ decreases. Since $C_2 = c_2 L$ and L is increasing, c_2 must be decreasing as the rings are pulled apart. In other words, u_2 is *increasing* as the rings are pulled apart, so u_2 is a non-physical solution. The minimal surface is therefore given by u_1. Figure 3.4 shows the solutions u_1 as the two rings pull apart, and Figure 3.5 shows non-physical solutions u_2.

We can also explicitly compute the minimal surface area for $c = c_1$. We have

$$I(u) = 2\pi \int_{-L}^{L} u(x)\sqrt{1 + u'(x)^2}\, dx = 4\pi c \int_0^L u(x)^2\, dx, \tag{3.13}$$

where we used the Euler-Lagrange equation (3.7) and the fact that u is even in the last step above. Substituting $u''(x) = c^2 u(x)$ and integrating by parts we have

$$\begin{aligned} I(u) &= \frac{4\pi}{c}\int_0^L u''(x)u(x)\, dx \\ &= \frac{4\pi}{c} u'(x)u(x)\Big|_0^L - \frac{4\pi}{c}\int_0^L u'(x)^2\, dx \\ &= \frac{4\pi}{c} u'(L)u(L) - \frac{4\pi}{c}\int_0^L u'(x)^2\, dx. \end{aligned}$$

Using (3.8) we have $u'(x)^2 = c^2 u(x)^2 - 1$ and so

$$I(u) = \frac{4\pi u(L)}{c}\sqrt{c^2 u(L)^2 - 1} - \frac{4\pi}{c}\int_0^L c^2 u(x)^2 - 1\, dx.$$

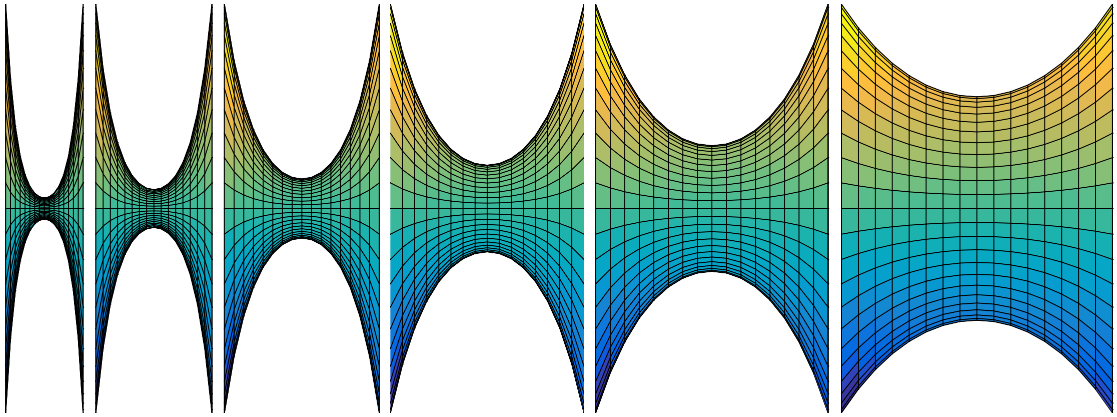


Figure 3.5: Illustration of solutions of the minimal surface equation that do not minimize surface area. The details are identical to Figure 3.4, except that we select c_2 instead of c_1. Notice the soap bubble is growing as the rings are pulled apart, which is the opposite of what we expect to occur.

Since $u(L) = r$ we have

$$I(u) = \frac{4\pi r}{c}\sqrt{c^2r^2 - 1} - 4\pi c \int_0^L u(x)^2 \, dx + \frac{4\pi}{c}\int_0^L \, dx.$$

Recalling (3.13) we have

$$I(u) = \frac{4\pi r}{c}\sqrt{c^2r^2 - 1} - I(u) + \frac{4\pi L}{c}.$$

Solving for $I(u)$ we have

$$I(u) = \frac{2\pi}{c}\left(r\sqrt{c^2r^2 - 1} + L\right). \tag{3.14}$$

Notice that we have at no point used the explicit formula $u(x) = \frac{1}{c}\cosh(cx)$. We have simply used the ODE that u satisfies, some clever integration by parts, and the boundary condition $u(L) = r$. There is an alternative expression for the surface area. Recall c is chosen so that $cr = \cosh(cL)$. Thus

$$c^2r^2 - 1 = \cosh^2(cL) - 1 = \sinh^2(cL),$$

and we have

$$I(u) = \frac{2\pi}{c}\left(r\sinh(cL) + L\right).$$

While it is not possible to analytically solve for c_1 and c_2, we can numerically compute the values to arbitrarily high precision with our favorite root-finding algorithm.

Most root-finding algorithms require one to provide an initial interval in which the solution is to be found. We already showed (see (3.9)) that $c \geq 1/r$. For an upper bound, recall we have the Taylor series

$$\cosh(x) = 1 + \frac{x^2}{2} + \frac{x^4}{4!} + \cdots = \sum_{k=0}^{\infty} \frac{x^{2n}}{(2k)!},$$

and so $\cosh(x) \geq \frac{x^2}{2}$. Recall also that c_1 and c_2 are solutions of

$$\cosh(cL) = cr.$$

Therefore, if $\frac{c^2L^2}{2} > cr$ we know that $\cosh(cL) > cr$. This gives the bounds

$$\frac{1}{r} \leq c_i \leq \frac{2r}{L^2} \qquad (i = 1, 2).$$

Furthermore, the point c^* where the slope of $c \mapsto \cosh(cL)$ equals r lies between c_1 and c_2. Therefore

$$c_1 < c^* < c_2,$$

where $L \sinh(c^* L) = r$, or

$$c^* = \frac{1}{L} \sinh^{-1}\left(\frac{r}{L}\right).$$

Thus, if $\cosh(c^*L) = c^*r$, then there is exactly one solution $c_1 = c_2 = c^*$. If $\cosh(c^*L) < c^*r$ then there are two solutions

$$\frac{1}{r} \leq c_1 < c^* < c_2 \leq \frac{2r}{L^2}. \tag{3.15}$$

Otherwise, if $\cosh(c^*L) > c^*r$ then there are no solutions. Now that we have the bounds (3.15) we can use any root-finding algorithm to determine the values of c_1 and c_2. In the code I showed in class I used a simple bisection search.

3.5 Isoperimetric inequality

Let C be a simple closed curve in the plane $\mathbb{R}^2$ of length ℓ, and let A denote the area enclosed by C. How large can A be, and what shape of curve yields the largest enclosed area? This question, which is called the *isoperimetric problem*, and similar questions have intrigued mathematicians for many thousands of years. The origin of the isoperimetric problem can be traced back to a Greek mathematician Zenodorus sometime in the second century B.C.E.

Let us consider a few examples. If C is a rectangle of width w and height h, then $\ell = 2(w + h)$ and $A = wh$. Since $w = \frac{1}{2}\ell - h$ we have

$$A = \frac{1}{2}\ell h - h^2,$$

where $h < \frac{1}{2}\ell$. The largest area for this rectangle is attained when $\frac{1}{2}\ell - 2h = 0$, or $h = \frac{1}{4}\ell$. That is, the rectangle is a square! The area of the square is

$$A = \frac{\ell^2}{16}.$$

Can we do better? We can try regular polygons with more sides, such as a pentagon, hexagon, etc., and we will find that the area increases with the number of sides. In the limit as the number of sides tends to infinity we get a circle, so perhaps the circle is a good guess for the optimal shape. If C is a circle of radius $r > 0$ then $2\pi r = \ell$, so $r = \frac{\ell}{2\pi}$ and

$$A = \pi r^2 = \frac{\ell^2}{4\pi}.$$

This is clearly better than the square, since $\pi < 4$.

The question again is: Can we do better still? Is there some shape we have not thought of that would give larger area than a circle while having the same perimeter? We might expect the answer is no, but lack of imagination is not a convincing proof.

We will prove shortly that, as expected, the circle gives the largest area for a fixed perimeter. Thus for any simple closed curve C we have the *isoperimetric inequality*

$$4\pi A \leq \ell^2, \tag{3.16}$$

where equality holds only when C is a circle of radius $r = \frac{\ell}{2\pi}$.

We give here a short proof of the isoperimetric inequality (3.16) using the Lagrange multiplier method in the calculus of variations. Let the curve C be parametrized by $(x(t), y(t))$ for $t \in [0,1]$. For notational simplicity we write $x = x_1$ and $y = x_2$ in this section. Since C is a simple closed curve, we may also assume without loss of generality that

$$x(0) = y(0) = x(1) = y(1) = 0. \tag{3.17}$$

Let U denote the interior of C. Then the area enclosed by the curve C is

$$A(x,y) = \int_U dx = \int_U \operatorname{div}(F)\, dx,$$

where F is the vector field $F(x,y) = \frac{1}{2}(x,y)$. By the divergence theorem

$$A(x,y) = \int_{\partial U} F \cdot \nu \, dS,$$

where ν is the outward normal to ∂U. Since $\partial U = C$ we have

$$\nu(t) = \frac{(y'(t), -x'(t))}{\sqrt{x'(t)^2 + y'(t)^2}},$$

and

$$dS = \sqrt{x'(t)^2 + y'(t)^2}\,dt,$$

provided we take the curve to have positive orientation. Therefore

$$A(x,y) = \int_0^1 \frac{1}{2}(x(t), y(t)) \cdot (y'(t), -x'(t))\,dt = \frac{1}{2}\int_0^1 x(t)y'(t) - x'(t)y(t)\,dt.$$

The length of C is given by

$$\ell(x,y) = \int_0^1 \sqrt{x'(t)^2 + y'(t)^2}\,dt.$$

Thus, we wish to find functions $x, y : [0,1] \to \mathbb{R}$ that maximize $A(x,y)$ subject to $\ell(x,y) = \ell$ and the boundary conditions (3.17).

The area and length functionals depend on two functions $x(t)$ and $y(t)$, which is a situation we have not encountered yet. Similar to the case of partial differentiation of functions on $\mathbb{R}^d$, we can freeze one input, say $y(t)$, and take the functional gradient with respect to $x(t)$. So we treat A and ℓ as functions of $x(t)$ only, and hence $z = x(t)$ and $p = x'(t)$. The gradient $\nabla_x A$, or Euler-Lagrange equation, is then given by

$$\nabla_x A(x,y) = \frac{1}{2}y'(t) - \frac{d}{dt}\left(-\frac{1}{2}y(t)\right) = y'(t)$$

while $\nabla_x \ell(x,y)$ is given by

$$\nabla_x \ell(x,y) = -\frac{d}{dt}\left(\frac{x'(t)}{\sqrt{x'(t)^2 + y'(t)^2}}\right).$$

Similarly

$$\nabla_y A(x,y) = -\frac{1}{2}x'(t) - \frac{d}{dt}\left(\frac{1}{2}x(t)\right) = -x'(t),$$

and

$$\nabla_y \ell(x,y) = -\frac{d}{dt}\left(\frac{y'(t)}{\sqrt{x'(t)^2 + y'(t)^2}}\right).$$

Then the gradients of A and ℓ are defined as

$$\nabla A(x,y) = \begin{bmatrix}\nabla_x A(x,y)\\ \nabla_y A(x,y)\end{bmatrix} \quad \text{and} \quad \nabla \ell(x,y) = \begin{bmatrix}\nabla_x \ell(x,y)\\ \nabla_y \ell(x,y)\end{bmatrix}.$$

Following (2.15), the necessary conditions for our constrained optimization problem are

$$\nabla A(x,y) + \lambda \nabla \ell(x,y) = 0,$$

where λ is a Lagrange multiplier. This is a set of two equations

$$y'(t) - \frac{d}{dt}\left(\frac{\lambda x'(t)}{\sqrt{x'(t)^2 + y'(t)^2}}\right) = 0,$$

and

$$-x'(t) - \frac{d}{dt}\left(\frac{\lambda y'(t)}{\sqrt{x'(t)^2 + y'(t)^2}}\right) = 0.$$

Integrating both sides we get

$$y(t) - \frac{\lambda x'(t)}{\sqrt{x'(t)^2 + y'(t)^2}} = a \quad \text{and} \quad x(t) + \frac{\lambda y'(t)}{\sqrt{x'(t)^2 + y'(t)^2}} = b,$$

for constants a and b. Therefore

$$\begin{aligned}(x(t) - a)^2 + (y(t) - b)^2 &= \left(-\frac{\lambda y'(t)}{\sqrt{x'(t)^2 + y'(t)^2}}\right)^2 + \left(\frac{\lambda x'(t)}{\sqrt{x'(t)^2 + y'(t)^2}}\right)^2 \\ &= \frac{\lambda^2 y'(t)^2}{x'(t)^2 + y'(t)^2} + \frac{\lambda^2 x'(t)^2}{x'(t)^2 + y'(t)^2} \\ &= \lambda^2 \frac{x'(t)^2 + y'(t)^2}{x'(t)^2 + y'(t)^2} \\ &= \lambda^2.\end{aligned}$$

This means that the curve $C(t) = (x(t), y(t))$ is a circle of radius λ centered at (a, b). Hence, as we expected, the circle is shape with largest area given a fixed perimeter. Since the perimeter is $\ell(x, y) = \ell$ we have $\lambda = \frac{\ell}{2\pi}$. Since $x(0) = y(0) = 0$, a and b must be chosen to satisfy

$$a^2 + b^2 = \lambda^2 = \frac{\ell^2}{4\pi^2}.$$

That is, the circle must pass through the origin, due to our boundary conditions (3.17).

3.6 Image restoration

Recall the total variation (TV) image restoration problem is based on minimizing

$$I(u) = \int_U \frac{1}{2}(f - u)^2 + \lambda|\nabla u|\, dx \tag{3.18}$$

over all $u : U \to \mathbb{R}$, where $U = (0, 1)^2$. The function f is the original noisy image, and the minimizer u is the denoised image. Here, the Lagrangian

$$L(x, z, p) = \frac{1}{2}(f(x) - z)^2 + \lambda|p|$$

is not differentiable at $p = 0$. This causes some minor issues with numerical simulations, so it is common to take an approximation of the TV functional that is differentiable. One popular choice is

$$I_\varepsilon(u) = \int_U \frac{1}{2}(f-u)^2 + \lambda\sqrt{|\nabla u|^2 + \varepsilon^2}\,dx, \tag{3.19}$$

where $\varepsilon > 0$ is a small parameter. When $\varepsilon = 0$ we get the TV functional (3.18). In this case the Lagrangian is

$$L_\varepsilon(x, z, p) = \frac{1}{2}(f(x) - z)^2 + \lambda\sqrt{|p|^2 + \varepsilon^2},$$

which is differentiable in both z and p. It is possible to prove that minimizers of I_ε converge to minimizers of I as $\varepsilon \to 0$, but the proof is very technical and outside the scope of this course.

So the idea is to fix some small value of $\varepsilon > 0$ and minimize I_ε. To compute the Euler-Lagrange equation note that

$$L_{\varepsilon,z}(x, z, p) = z - f(x) \quad \text{and} \quad \nabla_p L_\varepsilon(x, z, p) = \frac{\lambda p}{\sqrt{|p|^2 + \varepsilon^2}}.$$

Therefore the Euler-Lagrange equation is

$$u - \lambda \operatorname{div}\left(\frac{\nabla u}{\sqrt{|\nabla u|^2 + \varepsilon^2}}\right) = f \quad \text{in } U \tag{3.20}$$

with homogeneous Neumann boundary conditions $\frac{\partial u}{\partial \nu} = 0$ on ∂U. It is almost always impossible to find a solution of (3.20) analytically, so we are left to use numerical approximations.

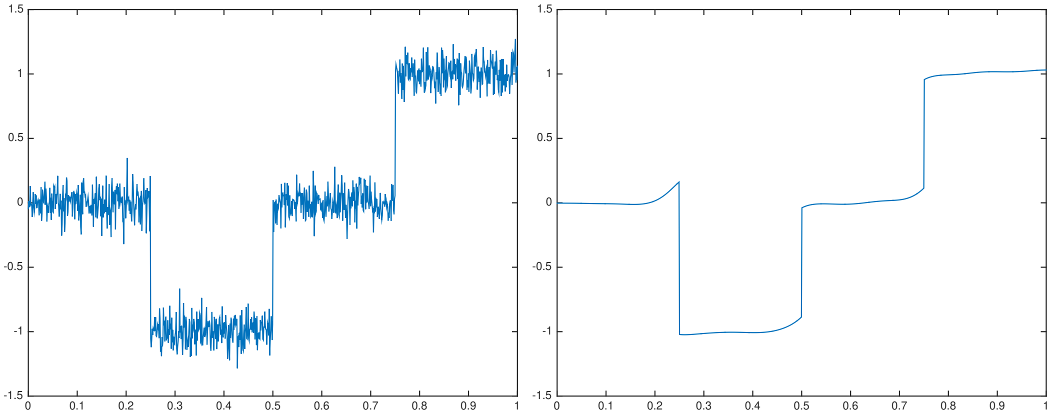
3.6.1 Gradient descent and acceleration

A standard numerical method for computing solutions of (3.20) is gradient descent, as described in Section 2.1. The gradient descent partial differential equation is

$$u_t + u - \lambda \operatorname{div}\left(\frac{\nabla u}{\sqrt{|\nabla u|^2 + \varepsilon^2}}\right) = f \quad \text{for } x \in U,\ t > 0,$$

with initial condition $u(x, 0) = f(x)$ and boundary conditions $\frac{\partial u}{\partial \nu} = 0$ on ∂U. This is a nonlinear heat equation where the thermal conductivity

$$\kappa = \frac{1}{\sqrt{|\nabla u|^2 + \varepsilon^2}}$$



(a) Noisy signal (b) TV denoising

Figure 3.6: Example of denoising a noisy signal with the total variations restoration algorithm. Notice the noise is removed while the edges are preserved.

depends on ∇u. Discretized on a grid with spacing Δt and Δx, the stability condition for the scheme is

$$\Delta t \leq \frac{\varepsilon \Delta x^2}{4\lambda}.$$

Hence, the scheme is numerically stiff when Δx is small or when $\varepsilon > 0$ is small.

We can also solve (3.20) with accelerated gradient descent, described in Section 2.4, which leads to the accelerated descent equation

$$u_{tt} + au_t + u - \lambda \operatorname{div}\left(\frac{\nabla u}{\sqrt{|\nabla u|^2 + \varepsilon^2}}\right) = f \quad \text{for } x \in U,\, t > 0,$$

where $a > 0$ is the coefficient of friction (damping). The CFL stability condition for accelerated descent is

$$\Delta t^2 \leq \frac{\varepsilon \Delta x^2}{4\lambda + \varepsilon \Delta x^2}.$$

The scheme is only numerically stiff when $\varepsilon > 0$ is small or $\lambda > 0$ is large. However, for accelerated descent, a nonlinear phenomenon allows us to violate the stability condition and roughly set $\Delta t \approx \Delta x$, independent of $\varepsilon > 0$. We refer the reader to [7] for details.

Figure 3.6 shows a one dimensional example of denoising a signal with the TV restoration algorithm. Notice the algorithm can remove noise while preserving the sharp discontinuities in the signal. This suggests that minimizers of the TV restoration functional can have discontinuities. Figure 3.7 shows an example of TV image restoration of a noisy image of Vincent hall. The top image is the noisy image, the middle image is TV restoration with a small value of λ, and the bottom image is TV restoration with a large value of λ. Notice that as λ is increased, the images becomes

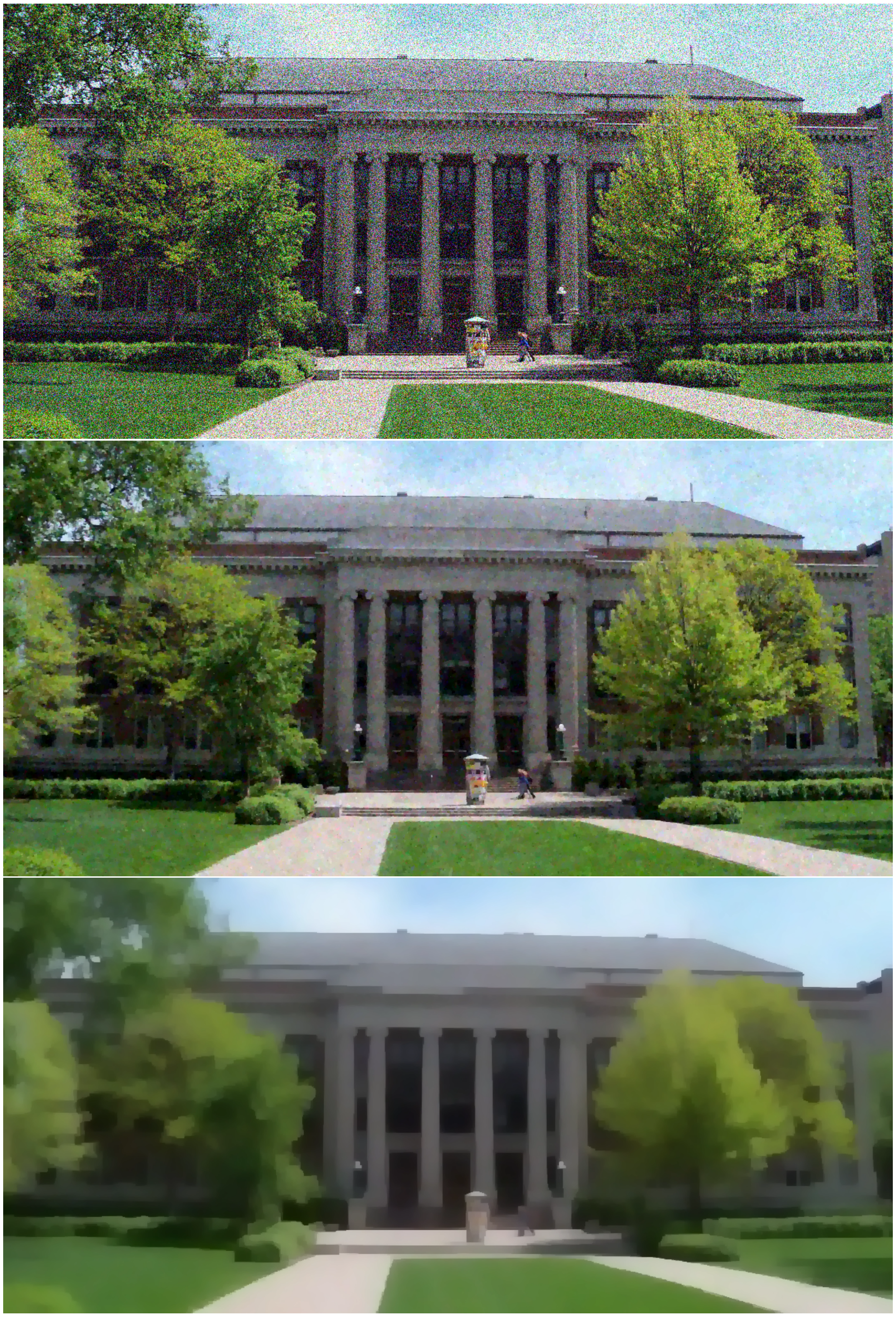


Figure 3.7: Example of TV image restoration: (top) noisy image (middle) TV restoration with small value for λ, and (bottom) TV restoration with large value for λ.

smoother, and many fine details are removed. We also observe that for small λ the algorithm is capable of preserving edges and fine details while removing unwanted noise.

3.6.2 Primal dual approach

Since gradient descent for solving Total Variation regularized problems requires smoothing out the functional with a parameter $\varepsilon > 0$, and the equations are numerically stiff when $\varepsilon > 0$ is small, many works have developed approaches to solving Total Variation problems without smoothing the functional, and are hence less numerically stiff. One such approach belongs to the class of primal dual approaches to Total Variation restoration. The ideas originally appeared in [18], and were later expanded upon to handle general problems in [19], among many other works.

We define the convex dual, or Legendre-Fenchel transform, of a function $\Phi : \mathbb{R}^d \to \mathbb{R}$ by

$$\Phi^*(p) := \sup_{x\in\mathbb{R}^d} \{x \cdot p - \Phi(x)\}. \tag{3.21}$$

If Φ is convex, then by convex duality we have $\Phi^{**} := (\Phi^*)^* = \Phi$, that is we have the representation

$$\Phi(x) := \sup_{p\in\mathbb{R}^d} \{x \cdot p - \Phi^*(p)\}. \tag{3.22}$$

Exercise 3.4. Prove the assertion above for $d = 1$ and Φ smooth and strongly convex. △

We assume $\Phi : \mathbb{R}^d \to \mathbb{R}$ is convex and consider for simplicity the problem

$$\min_u \int_U \Psi(x, u) + \Phi(\nabla u)\, dx, \tag{3.23}$$

where $u : U \to \mathbb{R}$, which includes the case of Total Variation restoration. A primal dual algorithm for solving (3.23) expresses Φ through its convex dual Φ^* giving the initially more looking complicated formation

$$\min_u \max_p \int_U \Psi(x, u) + p \cdot \nabla u - \Phi^*(p)\, dx, \tag{3.24}$$

where $p : U \to \mathbb{R}^d$. Here, u is the primal variable, p is the dual variable, and (3.24) is called a *saddle point formulation*. Notice we are dualizing only in the gradient variable ∇u, in which the original functional is nonsmooth, and leaving the variable u, in which the functional is smooth, unchanged.

Provided $u\, p \cdot \mathbf{n} = 0$ on ∂U, we can integrate by parts to express the problem as

$$\min_u \max_p \int_U \Psi(x, u) - u \operatorname{div}(p) - \Phi^*(p)\, dx. \tag{3.25}$$

In general, primal dual methods work by alternatively taking small steps in the direction of minimizing with respect to u and maximizing with respect to p, with either gradient descent steps, or *proximal updates*, which corresponds to implicit gradient descent. Either (3.24) or (3.25) may be used in the descent equations, and often (3.24) is more convenient for the dual update while (3.25) is more convenient for the primal update.

For some application, things can be simplified significantly. We consider Total Variation image restoration where $\Psi(x,z) = \frac{1}{2\lambda}(z - f(x))^2$ and $\Phi(p) = |p|$. The convex dual is easily computed to be

$$\Phi^*(v) = \sup_{p\in\mathbb{R}^d}\{p\cdot v - |p|\} = \begin{cases} 0, & \text{if } |v| \leq 1 \\ \infty, & \text{if } |v| > 1, \end{cases}$$

and the representation of Φ via convex duality in (3.22) is the familiar formula

$$|p| = \max_{|v|\leq 1} p\cdot v.$$

Thus, the saddle point formulation (3.25) becomes

$$\min_u \max_{|p|\leq 1} \int_U \frac{1}{2\lambda}(u-f)^2 + u\,\mathrm{div}(p)\,dx, \tag{3.26}$$

where we swapped p for $-p$ and assume the boundary condition $p\cdot\mathbf{n} = 0$ on ∂U. Notice we have turned a problem with a nonsmooth dependence on ∇u, into a smooth problem with a constraint $|p| \leq 1$. Thus, primal dual methods allow us to avoid the non-smoothness in the dependence on the gradient.

In fact, in (3.26), we can swap the min and max, due to strong convexity in u, to obtain the equivalent problem

$$\max_{|p|\leq 1} \min_u \int_U \frac{1}{2\lambda}(u-f)^2 + u\,\mathrm{div}(p)\,dx. \tag{3.27}$$

We can now explicitly solve for u, finding that

$$u = f - \lambda\mathrm{div}(p), \tag{3.28}$$

and yielding the dual problem

$$\max_{|p|\leq 1} T(p) := \int_U f\,\mathrm{div}(p) - \frac{\lambda}{2}\mathrm{div}(p)^2\,dx, \tag{3.29}$$

subject to $p\cdot\mathbf{n} = 0$ on ∂U. Since $p : U \to \mathbb{R}^d$ is vector-valued, we do not have the Euler-Lagrange equations for T on hand, and must compute them from scratch. To do this, we compute a variation in the direction of $\varphi \in C^\infty(U;\mathbb{R}^d)$

$$\left.\frac{d}{dt}\right|_{t=0} T(p+t\varphi) = \int_U f\,\mathrm{div}(\varphi) - \lambda\,\mathrm{div}(p)\mathrm{div}(\varphi)\,dx.$$

Taking $\varphi \cdot \mathbf{n} = 0$ on ∂U we can integrate by parts to find that

$$\frac{d}{dt}\Big|_{t=0} T(p + t\varphi) = -\int_U \nabla(f - \lambda\,\mathrm{div}(p)) \cdot \varphi\, dx.$$

Therefore

$$\nabla T(p) = \nabla(\lambda\,\mathrm{div}(p) - f).$$

To handle the constraint $|p| \leq 1$ we can do a type of projected gradient ascent

$$p^{k+1} = \frac{p^k + \tau\nabla(\mathrm{div}(p^k) - f/\lambda)}{1 + \tau|\mathrm{div}(p^k) - f/\lambda)|}, \tag{3.30}$$

with time step $\tau > 0$, starting from an initial guess p^0 with $|p^0| < 1$, and taking appropriate discretizations of div and ∇ on a grid with spacing Δx above. Notice we absorbed λ into the choice of time step, to be consistent with other presentations. This primal dual approach, and its convergence properties, is studied closely in [18]. The iteration above converges when τ is sufficiently small, and the solution of the Total Variation problem is obtained by evaluating (3.28).

3.6.3 The split Bregman method

Another popular method for solving L^1 regularized problems is the *split Bregman method* [31]. Our presentation follows [31] closely.

The split Bregman method aims to handle the L^1 norm by introducing a new function p and solving the constrained problem

$$\min_{u,p} \int_U \Psi(x,u) + |p|\, dx \quad \text{subject to } \nabla u = p, \tag{3.31}$$

where $\Psi(x,z) = \frac{1}{2\lambda}(z - f(x))^2$. Since the exact form of Ψ is not needed, we will proceed in generality. So far nothing has changed. The constraint $\nabla u = p$ could be handled by a penalty method

$$\min_{u,p} \int_U \Psi(x,u) + |p| + \frac{\mu}{2}|p - \nabla u|^2\, dx. \tag{3.32}$$

In penalty methods we have to take $\mu \to \infty$ to ensure the constraint is satisfied, and the problem would become ill-conditioned and difficult to solve in this regime.

The split Bregman iteration essentially allows one to solve something similar to (3.32) and obtain a solution of the constrained problem (3.31) *without* sending $\mu \to \infty$. The steps of the split Bregman method are outlined below. We introduce a new variable b and iterate

$$\begin{cases} (u^{k+1}, p^{k+1}) = \displaystyle\arg\min_{u,p} \int_U \Psi(x,u) + |p| + \frac{\mu}{2}|p - \nabla u - b^k|^2\, dx \\ \qquad\qquad b^{k+1} = b^k + \nabla u^{k+1} - p^{k+1}. \end{cases}$$

The Bregman iteration for solving constrained problems was originally proposed in [9], and the novelty of the split Bregman method is in splitting the terms $|p|$ and Ψ and applying the Bregman iteration only to $|p|$.

We will not prove convergence of the Bregman iteration here, and refer the reader to [47]. However, it is easy to see that *if* the method converges to a triple (u^*, p^*, b^*), then the pair (u^*, p^*) is a solution of the constrained problem (3.31). To see this, we note that since $b^k \to b^*$ we have $\nabla u^* = p^*$, so the constraint is satisfied. The pair (u^*, p^*) is clearly a solution of

$$(u^*, p^*) = \operatorname*{arg\,min}_{u,p} \int_U \Psi(x,u) + |p| + \frac{\mu}{2}|p - \nabla u - b^*|^2\,dx \tag{3.33}$$

Let $(\hat{u}, \hat{p})$ be any pair of functions satisfying $\nabla\hat{u} = \hat{p}$, that is, a feasible pair for the constrained problem (3.31). By definition of (u^*, p^*) we have

$$\int_U \Psi(x,u^*) + |p^*| + \frac{\mu}{2}|p^* - \nabla u^* - b^*|^2\,dx \le \int_U \Psi(x,\hat{u}) + |\hat{p}| + \frac{\mu}{2}|\hat{p} - \nabla\hat{u} - b^*|^2\,dx.$$

Due to the constraints $\nabla\hat{u} = \hat{p}$ and $\nabla u^* = p^*$ this simplifies to

$$\int_U \Psi(x,u^*) + |p^*|\,dx \le \int_U \Psi(x,\hat{u}) + |\hat{p}|\,dx.$$

Thus, (u^*, p^*) is a solution of the constrained problem (3.31). This holds independently of the choice of $\mu > 0$. One may think of the spit Bregman method as an iteration scheme to find for any $\mu > 0$, a function b^* so that the penalized problem (3.33) is equivalent to the original constrained problem (3.31).

The efficiency of the split Bregman method relies on the ability to solve the problem

$$\operatorname*{arg\,min}_{u,p} \int_U \Psi(x,u) + |p| + \frac{\mu}{2}|p - \nabla u - b^k|^2\,dx \tag{3.34}$$

efficiently. We focus on the L^2 penalty case where $\Psi(x,z) = \frac{1}{2\lambda}(z - f(x))^2$. Here, the problem (3.34) can be efficiently solved with the iteration scheme

$$\begin{cases} u^{j+1} = \operatorname*{arg\,min}_{u} \displaystyle\int_U \frac{1}{2\lambda}(u-f)^2 + \frac{\mu}{2}|p^j - \nabla u - b^k|^2\,dx \\ p^{j+1} = \operatorname*{arg\,min}_{p} \left\{|p| + \dfrac{\mu}{2}|p - \nabla u^{j+1} - b^k|^2\right\} \end{cases} \tag{3.35}$$

The Euler-Lagrange equation for the first problem is

$$u - \mu\lambda\Delta u = f + \mu\lambda\operatorname{div}(b^k - p^j) \quad \text{in } U. \tag{3.36}$$

This is a linear Poisson-type equation that can be solved very efficiently with either a fast Fourier transform (FFT) or the multigrid method. Alternatively, in [31], the authors report using only a handful of steps of gradient descent and can get by with only approximately solving the problem.

The efficiency of the split Bregman method stems largely from the fact that the second problem in (3.35) is pointwise, and can be solved *explicitly.* Note the objective is strongly convex in p, so the minimum is attained and unique, and the critical point, if it exists, is the minimizer. We can differentiate in p to find that the minimizer p^{j+1} satisfies

$$\frac{p^{j+1}}{|p^{j+1}|} + \mu p^{j+1} = \mu(\nabla u^{j+1} + b^k),$$

provided $p^{j+1} \neq 0$. Taking norms on both sides we have

$$|p^{j+1}| = |\nabla u^{j+1} + b^k| - \frac{1}{\mu}.$$

Clearly p^{j+1} points in the same direction as $\nabla u^{j+1} + b^k$, and so we have

$$p^{j+1} = \frac{\nabla u^{j+1} + b^k}{|\nabla u^{j+1} + b^k|}\left(|\nabla u^{j+1} + b^k| - \frac{1}{\mu}\right)$$

provided $|\nabla u^{j+1} + b^k| > \frac{1}{\mu}$. If $|\nabla u^{j+1} + b^k| \leq \frac{1}{\mu}$, then there are no critical points where the objective is smooth, and the minimizer is $p^{j+1} = 0$, i.e., the only non-smooth point. The formula for p^{j+1} can be expressed compactly as

$$p^{j+1} = \operatorname{shrink}(\nabla u^{j+1} + b^k, 1/\mu), \tag{3.37}$$

where the *shrinkage* operator is defined as

$$\operatorname{shrink}(x, \gamma) := \frac{x}{|x|}\max\{|x| - \gamma, 0\}. \tag{3.38}$$

The shrinkage operator is extremely fast and requires only a few operations per element of p^j, which is the key to the computational efficiency of the split Bregman method.

3.6.4 Edge preservation properties of Total Variation restoration

The simulations presented in Figures 3.6 and 3.7 suggest that minimizers of the TV restoration functional I can be discontinuous. This may present as counter-intuitive, since the derivative of u is very large near a discontinuity and we are in some sense minimizing the derivative. It is important to keep in mind, however, that we are minimizing the *integral* of the derivative, and while the derivative may be large at some points, its integral can still be small.

As an example, let us consider the one dimensional case and ignore the fidelity term, since it does not involve the derivative of u. Hence, we consider the functional

$$J_p(u) = \int_{-1}^{1} |u'(x)|^p \, dx,$$

where $p \geq 1$. The TV functional corresponds to $p = 1$, but it is interesting to consider other values of p to understand why $p = 1$ is preferred in signal processing communities. Suppose we want to minimize J_p subject to $u(-1) = 0$ and $u(1) = 1$. It is not difficult to convince yourself that the minimizer should be an increasing function, so we may write

$$J_p(u) = \int_{-1}^{1} u'(x)^p \, dx,$$

provided we restrict $u'(x) \geq 0$. If $p > 1$ then the Euler-Lagrange equation is

$$\frac{d}{dx}\left(pu'(x)^{p-1}\right) = 0,$$

which expands to

$$p(p-1)u'(x)^{p-2}u''(x) = 0.$$

The straight line $u(x) = \frac{1}{2}x + \frac{1}{2}$ is a solution of the Euler-Lagrange equation, and hence a minimizer since J_p is convex. When $p = 1$ the Euler-Lagrange equation is

$$\frac{d}{dx}(1) = 0$$

which does not even involve u! This means *every* increasing function is a solution of the Euler-Lagrange equation. A Lagrangian $L(x, z, p)$ for which every function solves the Euler-Lagrange equation is called a *null Lagrangian*, and they have many important applications in analysis.

Notice that when $p = 1$ and u is increasing

$$J_1(u) = \int_{-1}^{1} u'(x) \, dx = u(1) - u(-1) = 1 - 0 = 1.$$

Hence, the functional $J_1(u)$ actually only depends on the boundary values $u(1)$ and $u(-1)$, provided u is increasing. This is the reason why the Euler-Lagrange equation is degenerate; every increasing function satisfying the boundary conditions $u(-1) = 0$ and $u(1) = 1$ is a minimizer. Thus, the TV functional does not care *how* the function gets from $u(-1) = 0$ to $u(1) = 1$, provided the function is increasing. So the linear function $u(x) = \frac{1}{2}x + \frac{1}{2}$ is a minimizer, but so is the sequence of functions

$$u_n(x) = \begin{cases} 0, & \text{if } -1 \leq x \leq 0 \\ nx, & \text{if } 0 \leq x \leq \frac{1}{n} \\ 1, & \text{if } \frac{1}{n} \leq x \leq 1. \end{cases}$$

The function u_n has a sharp transition from zero to one with slope n between $x = 0$ and $x = 1/d$. For each n

$$J_1(u_n) = \int_0^{\frac{1}{n}} n\,dx = 1,$$

so each u_n is a minimizer. The pointwise limit of u_n as $n \to \infty$ is the Heaviside function

$$H(x) = \begin{cases} 0, & \text{if } -1 \leq x \leq 0 \\ 1, & \text{if } 0 \leq x \leq 1. \end{cases}$$

So in some sense, the discontinuous function $H(x)$ is also a minimizer. Indeed, we can compute

$$J_1(H) = \int_{-1}^{1} H'(x)\,dx = \int_{-1}^{1} \delta(x)\,dx = 1,$$

where $\delta(x)$ is the Delta function. This explains why minimizers of the TV functional can admit discontinuities.

Notice that if $p > 1$ then

$$J_p(u_n) = \int_0^{\frac{1}{n}} n^p\ dx = n^{p-1},$$

hence $J_p(u_n) \to \infty$ as $n \to \infty$. This means that a discontinuous function cannot minimize J_p for $p > 1$, and the only sensible value for $J_p(H)$ is ∞ when $p > 1$. Thus, if we used a version of TV restoration where $|\nabla u|^p$ appeared with $p > 1$, we would not expect that edges and fine details would be preserved, since discontinuities have infinite cost in this case.

3.7 Image segmentation

Recall in the image segmentation problem we aim to minimize

$$I(u, a, b) = \int_U H(u)(f - a)^2 + (1 - H(u))\,(f - b)^2 + \lambda\delta(u)|\nabla u|\,dx \tag{3.39}$$

over $u : U \to \mathbb{R}$ and real numbers a and b. Let us assume for the moment that a and b are fixed, and I is a function of only u.

The Lagrangian

$$L(x, z, p) = H(z)(f(x) - a)^2 + (1 - H(z))(f(x) - b)^2 + \lambda\delta(z)|p|$$

is not even continuous, due to the presence of the Heaviside function $H(z)$ and the delta function $\delta(z)$. This causes problems numerically, hence in practice we usually replace L by a smooth approximation. For $\varepsilon > 0$ we define the smooth approximate Heaviside function

$$H_\varepsilon(x) = \frac{1}{2}\left(1 + \frac{2}{\pi}\arctan\left(\frac{x}{\varepsilon}\right)\right). \tag{3.40}$$

The approximation to δ is then

$$\delta_\varepsilon(x) := H'_\varepsilon(x) = \frac{1}{\pi}\frac{\varepsilon}{\varepsilon^2 + x^2}. \tag{3.41}$$

We then form the approximation

$$I_\varepsilon(u) = \int_U H_\varepsilon(u)(f-a)^2 + (1 - H_\varepsilon(u))\,(f-b)^2 + \lambda\delta_\varepsilon(u)|\nabla u|\,dx. \tag{3.42}$$

The Lagrangian for I_ε is

$$L_\varepsilon(x,z,p) = H_\varepsilon(z)(f(x)-a)^2 + (1-H_\varepsilon(z))(f(x)-b)^2 + \lambda\delta_\varepsilon(z)|p|.$$

Therefore

$$L_{\varepsilon,z}(x,z,p) = \delta_\varepsilon(z)\left((f(x)-a)^2 - (f(x)-b)^2\right) + \lambda\delta'_\varepsilon(z)|p|$$

and

$$\nabla_p L_\varepsilon(x,z,p) = \frac{\lambda\delta_\varepsilon(z)p}{|p|}.$$

By the chain and product rules

$$\begin{aligned}
\operatorname{div}\left(\nabla_p L(x,u(x),\nabla u(x))\right) &= \lambda \operatorname{div}\left(\frac{\delta_\varepsilon(u(x))\nabla u(x)}{|\nabla u(x)|}\right) \\
&= \lambda\frac{\nabla\delta_\varepsilon(u(x))\cdot\nabla u(x)}{|\nabla u(x)|} + \lambda\operatorname{div}\left(\frac{\nabla u(x)}{|\nabla u(x)|}\right) \\
&= \lambda\frac{\delta'_\varepsilon(u(x))\nabla u(x)\cdot\nabla u(x)}{|\nabla u(x)|} + \lambda\delta_\varepsilon(u(x))\operatorname{div}\left(\frac{\nabla u(x)}{|\nabla u(x)|}\right) \\
&= \lambda\delta'_\varepsilon(u(x))|\nabla u(x)| + \lambda\delta_\varepsilon(u(x))\operatorname{div}\left(\frac{\nabla u(x)}{|\nabla u(x)|}\right)
\end{aligned}$$

Therefore, the Euler-Lagrange equation is

$$\delta_\varepsilon(u)\left[(f-a)^2 - (f-b)^2 - \lambda\operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right)\right] = 0 \quad \text{in } U \tag{3.43}$$

subject to homogeneous Neumann boundary conditions $\frac{\partial u}{\partial \nu} = 0$ on ∂U.

As with the image restoration problem, it is nearly impossible to solve the Euler-Lagrange equation (3.43) analytically. Thus we are left to devise numerical algorithms to find solutions. Here, we are minimizing over u, a, and b, which is a situation we have not encountered before. Note that if u is fixed, then minimizing with respect to a and b is easy. Indeed, differentiating I_ε with respect to a yields

$$0 = -2\int_U H_\varepsilon(u)(f-a)\,dx.$$

Therefore the optimal value for a is

$$a = \frac{\int_U H_\varepsilon(u) f\,dx}{\int_U H_\varepsilon(u)\,dx}, \tag{3.44}$$

which is approximately the average of f in the region where $u > 0$. Similarly, if u is fixed, the optimal choice of b is

$$b = \frac{\int_U (1 - H_\varepsilon(u)) f\,dx}{\int_U 1 - H_\varepsilon(u)\,dx}. \tag{3.45}$$

Since it is easy to minimize over a and b, the idea now is to consider an alternating minimization algorithm, whereby one fixes $a, b \in \mathbb{R}$ and takes a small gradient descent step in the direction of minimizing I_ε with respect to u, and then one freezes u and updates a and b according to (3.44) and (3.45). We repeat this iteratively until the values of a, b, and u remain unchanged with each new iteration.

Gradient descent on I_ε with respect to u is the partial differential equation

$$u_t + \delta_\varepsilon(u)\left[(f-a)^2 - (f-b)^2 - \lambda\,\mathrm{div}\left(\frac{\nabla u}{|\nabla u|}\right)\right] = 0 \quad \text{for } x \in U,\ t > 0 \tag{3.46}$$

subject to homogeneous Neumann boundary conditions $\frac{\partial u}{\partial \nu} = 0$ on ∂U and an initial condition $u(x,0) = u_0(x)$. As with the image restoration problem, we normally replace the non-differentiable norm $|\nabla u|$ by a smooth approximation, so instead we solve the partial differential equation

$$u_t + \delta_\varepsilon(u)\left[(f-a)^2 - (f-b)^2 - \lambda\,\mathrm{div}\left(\frac{\nabla u}{\sqrt{|\nabla u|^2 + \varepsilon^2}}\right)\right] = 0 \quad \text{for } x \in U,\ t > 0. \tag{3.47}$$

At each iteration of solving (3.47) numerically, we update the values of a and b according to (3.44) and (3.45).

Figure 3.8 shows the result of segmenting the cameraman image. The bottom four images in the figure show the evolution of the zero level set of u throughout the gradient descent procedure resulting in the segmentation obtained in the lower right image. Figure 3.9 shows that results of segmenting blurry and noisy versions of the cameraman image, to illustrate that the algorithm is robust to image distortions.

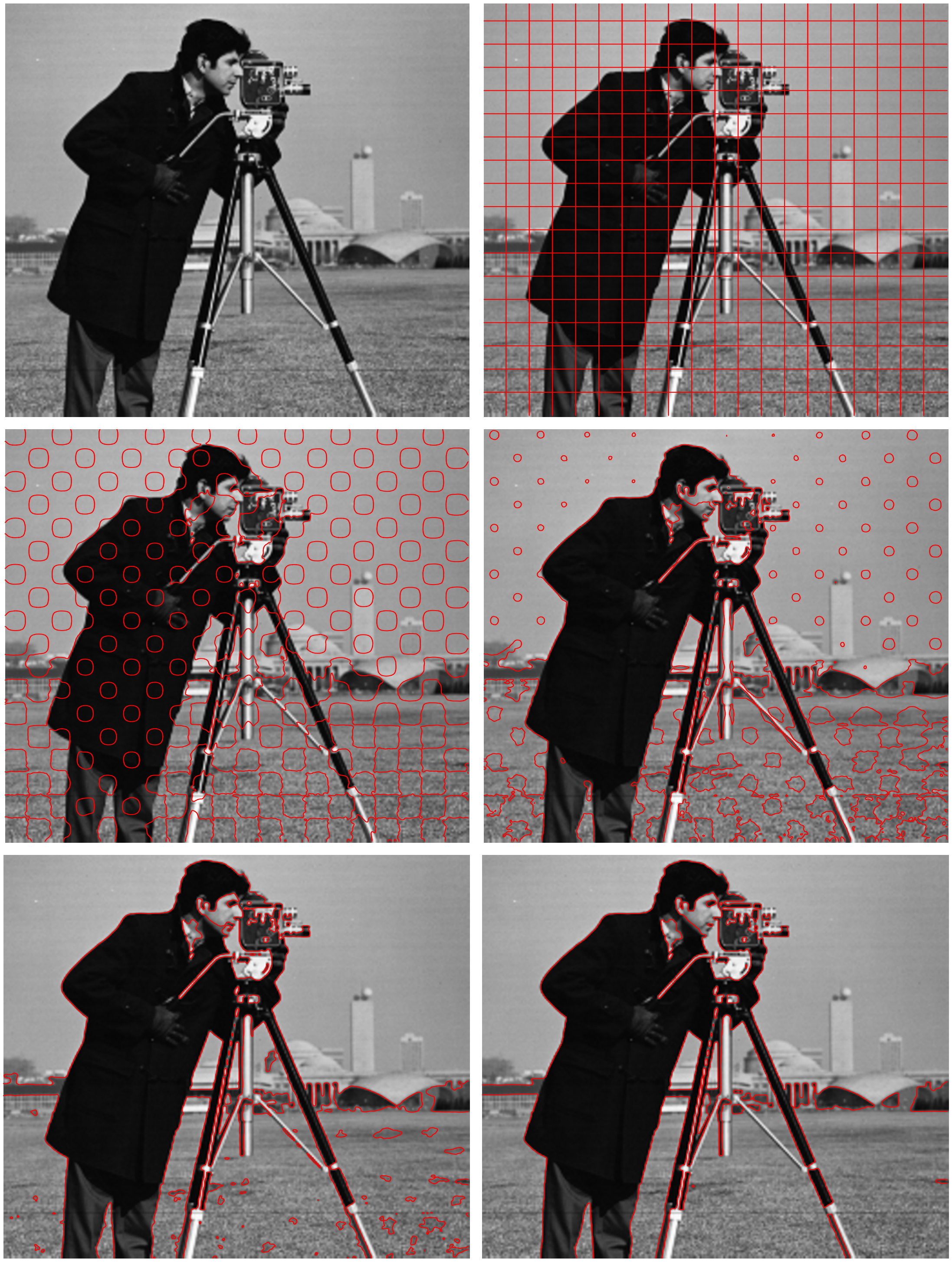


Figure 3.8: Illustration of gradient descent for segmenting the cameraman image. Top left is the original image, and top right is the initialization of the gradient descent algorithm. The lower four images show the evolution of the zero level set of u.

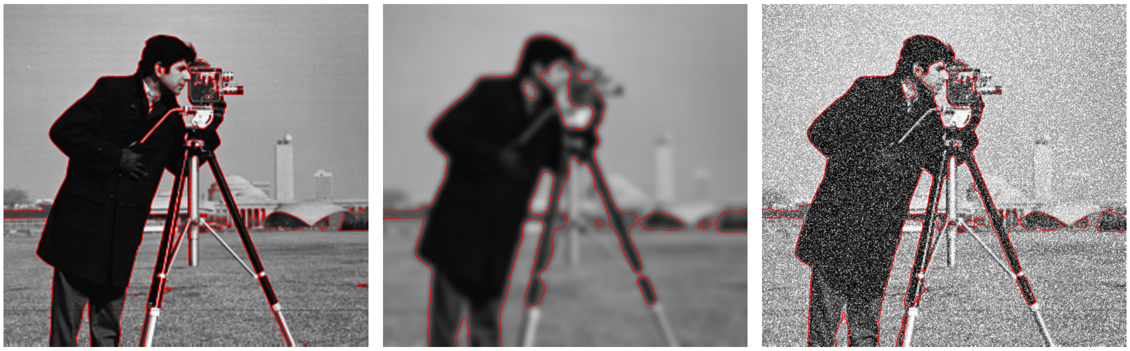


Figure 3.9: Segmentation of clean, blurry, and noisy versions of the cameraman image.

3.7.1 Ginzburg-Landau approximation

Employing gradient descent (3.47) to minimize image segmentation energies can be extremely slow when asking for accurate results (i.e., $\varepsilon > 0$ is small), and/or when imposing a large amount of regularization (i.e., $\lambda \gg 1$). Here, we describe some approximations based on Ginzburg-Landau functional that yield very fast approximate algorithms for image segmentation. We follow [27] closely.

For $\varepsilon > 0$ the Ginzburg-Landau functional is defined as

$$G_\varepsilon(u) = \int_U \varepsilon|\nabla u|^2 + \frac{1}{\varepsilon}W(u)\,dx, \tag{3.48}$$

where W is a double-well potential defined by $W(t) = t^2(1-t)^2$. When $\varepsilon > 0$ is small, the double-well potential term forces u to be either 1 or 0 in most of the domain with a sharp $O(\varepsilon)$ transition between. It can be shown that the Ginzburg-Landau functional G_ε converges (in the Γ-sense that we will discuss later) to the total variation $\int_U |\nabla u|\,dx$. Thus, we can (formally for now) make the approximation

$$\int_U |\nabla H(u)|\,dx \approx \int_U \varepsilon|\nabla u|^2 + \frac{1}{\varepsilon}W(u)\,dx$$

in the segmentation energy (3.39). Since the double-well forces u to be close to 1 or 0 everywhere, we can approximate $H(u)$ by u^2 and $1 - H(u)$ by $(1-u)^2$. This leads to the approximate energy

$$I_\varepsilon(u,a,b) = \int_U \mu\left\{u^2(f-a)^2 + (1-u)^2(f-b)^2\right\} + \varepsilon|\nabla u|^2 + \frac{1}{\varepsilon}W(u)\,dx. \tag{3.49}$$

The energy (3.49) is called a *phase-field* approximation of (3.39). Notice that $\mu = \lambda^{-1}$ from the original energy (3.39).

The phase-field approximation (3.49) can be efficiently solved by splitting into two pieces, the double-well potential and the remaining portions, and alternatively taking small steps of minimizing over each piece. Gradient descent on the double-well potential part of the energy $\frac{1}{\varepsilon}W(u)$ would be $u_t = -\frac{1}{\varepsilon}W'(u)$. When $\varepsilon > 0$ is small, we approximate this step by sending $t \to \infty$, and u converges to stable equillibrium points $W'(u) = 0$ and $W''(u) > 0$, i.e., $u = 0$ or $u = 1$. By examining $W'(u) = 2u(1-u)(1-2u)$, we see that $W'(u) > 0$ for $u > 1/2$ and $W'(u) < 0$ for $u < 1/2$. So any value of u larger than $1/2$ gets sent to 1, while value smaller than $1/2$ is sent to 0. This step is thus *thresholding*.

The other part of the energy is

$$\int_U \mu \left\{ u^2(f-a)^2 + (1-u)^2(f-b)^2 \right\} + \varepsilon |\nabla u|^2 \, dx.$$

Gradient descent on this portion of the energy corresponds to

$$u_t = 2\varepsilon \Delta u - 2\mu \left(u(f-a)^2 - (1-u)(f-b)^2 \right).$$

The scheme runs the gradient descent equation above for a short time, alternating with the thresholding step described above. The steps are outlined below.

Given: Image f, initial guess u^0, and time step $\tau > 0$. For $k = 1, 2, \ldots$

1. Solve the equation

$$\text{(3.50)} \qquad \begin{cases} w_t = \Delta w - \dfrac{\mu}{\sqrt{\pi\tau}} \left(w(f-a)^2 - (1-w)(f-b)^2 \right), & \text{for } 0 < t \le \tau \\ w = u^k, & \text{at } t = 0. \end{cases}$$

2. Set

$$u^{k+1}(x) = \begin{cases} 0, & \text{if } w(x,\tau) \le \frac{1}{2} \\ 1, & \text{if } w(x,\tau) > \frac{1}{2}. \end{cases}$$

The constants a, b are updated after each thresholding step according to the formulas:

$$a = \frac{\int_U uf \, dx}{\int_U u \, dx} \quad \text{and} \quad b = \frac{\int_U (1-u) f \, dx}{\int_U (1-u) \, dx}.$$

The algorithm above is an example of *threshold dynamics*, which was invented in [41] for approximation motion by mean curvature, and has been widely used for other problems since. The scheme is *unconditionally stable*, which means that the scheme is stable and convergent for arbitrarily large time step $\tau > 0$. This breaks the stiffness in the segmentation energy and allows very fast computations of image segmentations.

Chapter 4

The direct method

We now turn to the problem of determining conditions under which the functional

$$I(u) = \int_U L(x, u(x), \nabla u(x))\, dx \tag{4.1}$$

attains its minimum value. That is, we want to understand under what conditions we can construct a function u_* so that $I(u_*) \leq I(u)$ for all u in an appropriate function space. We will always assume L is a smooth function of all variables.

Example 4.1 (Optimization on $\mathbb{R}^d$)**.** Suppose $f : \mathbb{R}^d \to [0, \infty)$ is continuous. Then f attains its minimum value over any closed and bounded set $D \subset \mathbb{R}^d$. Indeed, we can always take a sequence x_m with $f(x_m) \to \inf_D f$ as $m \to \infty$. Then by the Bolzano-Weierstrauss Theorem, there is a subsequence x_{m_j} and a point $x_0 \in D$ such that $x_{m_j} \to x_0$. By continuity of f

$$f(x_0) = \lim_{j\to\infty} f(x_{m_j}) = \inf_D f.$$

Therefore, f attains its minimum value over D at x_0. △

The situation is much different in the *infinite dimensional* setting, as the following examples illustrate.

Example 4.2 (Weierstrauss)**.** Consider minimizing

$$I(u) = \int_{-1}^{1} (xu'(x))^2\, dx$$

subject to $u(-1) = -1$ and $u(1) = 1$. Here, the Lagrangian $L(x, p) = x^2p^2$ is convex in p, though not strongly convex at $x = 0$. Let $u_n(x) = \max\{\min\{nx, 1\}, -1\}$. We easily compute that $I(u) = n^2 \int_{-1/d}^{1/d} x^2\, dx = \frac{2}{3n}$, which shows that $\inf I = 0$ over the space of Lipschitz functions. However, there is no Lipschitz function u for which $u(-1) = -1$, $u(1) = 1$ and $I(u) = 0$, since any Lipschitz functions for which $I(u) = 0$ are constant. This example is originally due to Weierstrauss. △

In example above, the sequence u_n is converging to a step function $u(x) = 1$ for $x > 0$ and $u(x) = -1$ for $x < 0$. One may try to interpret the step function as a solution of the variational problem in some appropriate way, so the lack of existence of minimizers stems from looking in the wrong function space (i.e., Lipschitz functions). The next example shows that the situation can be far worse when L is not convex.

Example 4.3 (Bolza problem)**.** Consider the problem of minimizing

$$I(u) = \int_0^1 u(x)^2 + (u'(x)^2 - 1)^2 \, dx,$$

subject to $u(0) = 0 = u(1)$. We claim there is no solution to this problem. To see this, let $u_k(x)$ be a sawtooth wave of period $2/k$ and amplitude $1/k$. That is, $u_k(x) = x$ for $x \in [0, 1/k]$, $u_k(x) = 2/k - x$ for $x \in [1/k, 2/k]$, $u_k(x) = x - 2/k$ for $x \in [2/k, 3/k]$ and so on. The function u_k satisfies $u_k'(x) = 1$ or $u_k'(x) = -1$ at all but a finite number of points of non-differentiability. Also $0 \le u_k(x) \le 1/k$. Therefore

$$I(u_k) \le \int_0^1 \frac{1}{k^2} \, dx = \frac{1}{k^2}.$$

If the minimum exists, then

$$0 \le \min_u I(u) \le \frac{1}{k^2}$$

for all natural numbers k. Therefore $\min_u I(u)$ would have to be zero. However, there is no function u for which $I(u) = 0$. Such a function would have to satisfy $u(x)^2 = 0$ and $u'(x)^2 = 1$ for all x, which are not compatible conditions. Therefore, there is no minimizer of I, and hence no solution to this problem. △

In contrast with Example 4.2, the minimizing sequence u_k constructed for the Bolza problem converges to a function (the zero function) that cannot be easily interpreted as a feasible candidate for a minimizer of the variational problem. The gradient u_k' is oscillating wildely and is nonconvergent everywhere in any strong sense.

The issue underlying both examples above is that the function space we are optimizing over is *infinite dimensional*, and so the notion of closed and bounded is not relevant. We will need a stronger compactness criterion. Indeed, the functional I in the Bolza example is a smooth function of u and u', and the minimizing sequence u_k constructed in Example 4.3 satisfies $|u_k| \le 1$ and $|u_k'| \le 1$, so we are minimizing I over a closed and bounded set. These examples show that infinite dimensional spaces are much different and we must proceed more carefully. In particular, we require further structure assumptions on L beyond smoothness to ensure minimizers exist.

4.1 Coercivity and lower semicontinuity

It is useful to first study a more abstract problem

$$\inf_{x \in X} f(x), \tag{4.2}$$

where X is a *topological* space and $f : X \to [0,\infty)$. A topological space is the most general space where we can make sense of limits $\lim_{n\to\infty} x_n = x$. Beyond this, it is not so important that the reader knows what a topological space is, so we will not give a definition, other than to say that all spaces considered in these notes are toplogical spaces with some additional structure (e.g., normed linear space, Hilbert space, etc).

The main question is: What conditions should we place on f to ensure its minimum value is attained over X, that is, that a minimizer exists in (4.2)?

It is useful to consider how one may go about proving the existence of a minimizer. Since $f \geq 0$, we have $\inf_{x\in X} f(x) \geq 0$, and so there exists a *minimizing sequence* (x_n) satisfying

$$\lim_{n\to\infty} f(x_n) = \inf_{x\in X} f(x). \tag{4.3}$$

The limit of the sequence (x_n), or any subsequence thereof, is presumably a leading candidate for a minimizer of f. If (x_n) is convergent in X, so there exists $x_0 \in X$ so that $x_n \to x_0$ as $n \to \infty$, then we would like to show that $f(x_0) = \inf_{x\in X} f(x)$. If f is continuous, then this follow directly from (4.3). However, functionals f of interest are often not continuous in the correct topology. It turns out we can get by with *lower semicontinuity*.

Definition 4.1. A function $f : X \to [0,\infty)$ is *lower semicontinuous* on X if for every $x \in X$

$$f(x) \leq \liminf_{y\to x} f(y). \tag{4.4}$$

If (x_n) is a minimizing sequence, $x_n \to x_0$ as $n \to \infty$, and f is lower semicontinuous, then it follows that

$$f(x_0) \leq \liminf_{n\to\infty} f(x_n) = \lim_{n\to\infty} f(x_n) = \inf_{x\in X} f(x).$$

Therefore $f(x_0) = \inf_{x\in X} f(x)$.

Now we turn to the question of how to guarantee that (x_n), or a subsequence thereof, is convergent. This is the question of *compactness*.

Definition 4.2. A subset $A \subset X$ is *sequentially compact* if every sequence $(x_n) \subset A$ has a convergent subsequence converging to a point in A. We call the set A *sequentially precompact* if the subsequence converges in X, but not necessarily in A.

For topological spaces, there is another notion called *compactness*, which has a definition in terms of open sets. In metric spaces, sequentially compact is equivalent to compact, and in that setting we will just use the term *compact*. In general topological spaces the notions are not the same, and for minimizing functionals we require sequential compactness. We recall that when $X = \mathbb{R}^d$, a subset $A \subset \mathbb{R}^d$ is sequentially compact (or just compact) if and only if A is closed and bounded, due to the celebrated Bolzano-Weierstrauss Theorem.

To show that (x_n) has a convergent subsequence, we just need to ensure that the sequence belongs to a sequentially compact subset of X.

Definition 4.3. We say that $f : X \to [0, \infty)$ is *coercive* if there exists $\lambda > \inf_{x\in X} f(x)$ such that the set

$$A = \{x \in X \,:\, f(x) < \lambda\} \tag{4.5}$$

is sequentially precompact.

If f is coercive, then any minimizing sequence (x_n) eventually belongs to a sequentially compact subset of X, for n sufficiently large, and so (x_n) admits a convergent subsequence, which converges to a minimizer of f if f is lower semicontinuous.

The discussion above is a proof of the following theorem.

Theorem 4.4. *If $f : X \to [0, \infty)$ is coercive and lower semicontinuous, then there exists $x_0 \in X$ such that*

$$f(x_0) = \min_{x\in X} f(x). \tag{4.6}$$

Theorem 4.4 is known as the *direct method* in the calculus of variations for proving existence of minimizers. It is important to understand how the minimizer is constructed as the *limit* of a minimizing sequence. In our setup, a minimizing sequence u_k has bounded energy, that is

$$\sup_{k\geq 1} \int_U L(x, u_k, \nabla u_k)\, dx < \infty. \tag{4.7}$$

From this, we must prove estimates on appropriate norms of the sequence u_k to obtain *coercivity*—convergence of a subsequence u_{k_j} in some appropriate sense. Since (4.7) only gives estimates on the integral of functions of u_k and ∇u_k, the best we can hope for is that u_k and ∇u_k are uniformly bounded in $L^p(U)$ for some $p \geq 1$. In the best case we have $p = \infty$, and by the Arzelà-Ascoli Theorem (proof below) there exists a subsequence u_{k_j} converging uniformly to a continuous (in fact Lipschitz) function u. However, since u is constructed though a limit of a sequence of functions, even if we ask that $u_k \in C^1(U)$, the uniform limit u need not be in $C^1(U)$.

Exercise 4.5. Give an example of a sequence of functions $u_k \in C^\infty((0, 1))$ such that u_k is uniformly convergent to some continuous function $u \in C((0, 1))$, but $u \notin C^1((0, 1))$. Hence $C^1((0, 1))$ is not closed under the topology of uniform convergence. △

This means that coercivity cannot hold when $X = C^1(U)$ in Theorem 4.4, and so the classical function spaces $C^k(U)$ are entirely incompatible with the direct method. While we may expect minimizers to belong to $C^k(U)$, proving this directly via the direct method is intractable. Instead, we have to work on a larger space of functions where it is easier to find minimizers, and treat the question of *regularity* of minimizers separately from existence.

We give a brief sketch of the proof of the Arzelà-Ascoli Theorem in a special case.

Lemma 4.6. *Let $u_k : (0,1) \to \mathbb{R}$ be a sequence of smooth functions with*

$$|u_k(x)| + |u_k'(x)| \leq \lambda$$

for all $k \geq 1$ and $x \in (0,1)$. Then there exists a subseuqence u_{k_j} and a function $u \in C((0,1))$ satisfying

$$|u(x) - u(y)| \leq C|x - y|$$

such that $u_{k_j} \to u$ uniformly on $(0,1)$ (that is, $\lim_{j\to\infty} \max_{0\leq x\leq 1} |u_{k_j}(x) - u(x)| = 0$).

Proof. We claim there exists a subsequence u_{k_j} such that $u_{k_j}(x)$ is a convergent sequence for all *rational* coordinates x. To prove this, we enumerate the rationals $r_1, r_2, r_3, \ldots,$ and note that each sequence $(u_k(r_i))_k$ is a bounded seqeunce, and by Bolzano-Weierstrauss has a convergent subsequence. We construct the subsequence k_j inductively as follows. Let ℓ_i^1 such that $u_{\ell_i^1}(r_1)$ is convergent (by Bolzano-Weierstrauss) and set $k_1 = \ell_1^1$. Since $u_{\ell_i^1}(r_2)$ is a bounded sequence, there is a convergent subsequence $(u_{\ell_{i_j}^1}(r_2))_j$, and we write $\ell_q^2 = \ell_{i_q}^1$ and $k_2 = \ell_2^2$. Given we have constructed ℓ_i^j so that $u_{\ell_i^j}(r_j)$ is convergent as $i \to \infty$, and $k_i = \ell_i^i$, we construct k_{i+1} and ℓ_i^{j+1} as follows: Since $u_{\ell_i^j}(r_{j+1})$ is a bounded sequence, there is a convergent subsequence $(u_{\ell_{i_q}^j}(r_{j+1}))_q$. We write $\ell_q^{j+1} = \ell_{i_q}^j$ and $k_{j+1} = \ell_{j+1}^{j+1}$. We have constructed k_i so that $(k_i) \subset (\ell_i^j)_i$ for all j. Since $k_i \geq i$ we have that $u_{k_i}(r_j)$ is convergent for all j, which establishes the claim. The argument above is called a *diagonal* argument, due to Cantor.

We now claim there is a unique continuous function $\overline{u} : (0,1) \to \mathbb{R}$ such that $\overline{u}(x) = u(x)$ for all rational x. To see this, let $x \in (0,1)$ be irrational. Let r_{i_j} be an approximating rational sequence, so that $\lim_{j\to\infty} r_{i_j} = x$. We have

$$|u(r_{i_j}) - u(r_{i_\ell})| = \lim_{j\to\infty} |u_{k_j}(r_{i_j}) - u_{k_j}(r_{i_\ell})| \leq \lambda |r_{i_j} - r_{i_\ell}|.$$

It follows that the sequence $(u(r_{i_j}))_j$ is a Cauchy sequence, and hence convergent. The value of the limit is independent of the choice of approximating rational sequence, since if r_{ℓ_j} is another such sequence with $r_{\ell_j} \to x$ as $j \to \infty$, and argument similar to above shows that

$$|u(r_{i_j}) - u(r_{\ell_j})| \leq \lambda |r_{i_j} - r_{i_\ell}| \longrightarrow 0$$

as $j \to \infty$. Therefore, we define $u(x) = \lim_{j\to\infty} u(r_{i_j})$, where r_{i_j} is any rational sequence convergening to x as $j \to \infty$. To see that u is continuous, let $x, y \in (0,1)$ and let $r_{i_j} \to x$ and $r_{\ell_j} \to y$ as $j \to \infty$. Then we have

$$|u(x) - u(y)| = \lim_{j\to\infty} |u(r_{i_j}) - u(r_{\ell_j})| \le \lambda|x-y|.$$

Therefore u is Lipschitz continuous.

Finally, we show that $u_{k_j} \to u$ uniformly. Let $M \ge 1$ be an integer and $x_j = j/M$. Fix $x \in (0,1)$ choose x_i with $0 \le i \le M$ and $|x - x_i| \le= 1/M$. Then

$$\begin{aligned}
|u_{k_j}(x) - u(x)| &= |u_{k_j}(x) - u_{k_j}(x_i) + u_{k_j}(x_i) - u(x_i) + u(x_i) - u(x)| \\
&\le |u_{k_j}(x) - u_{k_j}(x_i)| + |u_{k_j}(x_i) - u(x_i)| + |u(x_i) - u(x)| \\
&\le \frac{2\lambda}{M} + \max_{0\le i\le M} |u_{k_j}(x_i) - u(x_i)|
\end{aligned}$$

Sending $j \to \infty$ we have that

$$\limsup_{j\to\infty} |u_{k_j}(x) - u(x)| \le \frac{2\lambda}{M}.$$

Sending $M \to \infty$ completes the proof. □

4.2 Brief review of Sobolev spaces

Sobolev spaces are the natural function spaces in which to look for minimizers of many functionals in the form (4.1). Section A.5 has a brief introduction to some function spaces such as $C^k(U)$ and $L^2(U)$. Here, we briefly introduce the L^p spaces before giving a very brief overview of the necessary Sobolev space theory. For more details on Sobolev spaces we refer the reader to [28].

4.2.1 Definitions and basic properties

Let $1 \le p < \infty$ and let $U \subset \mathbb{R}^d$ be open and bounded with a smooth boundary ∂U. For $u : U \to \mathbb{R}$ we define

$$\|u\|_{L^p(U)} = \left(\int_U |u|^p dx\right)^{1/p}, \tag{4.8}$$

and

$$L^p(U) = \{\text{Lebesgue measurable } u : U \to \mathbb{R} \text{ such that } \|u\|_{L^p(U)} < \infty\}. \tag{4.9}$$

The space $L^p(U)$ is a Banach space[1] equipped with the norm $\|u\|_{L^p(U)}$, provided we identify functions in $L^p(U)$ that agree up to sets of measure zero (i.e., almost everywhere). As in any normed linear space, we say that $u_k \to u$ in $L^p(U)$ provided

$$\|u - u_k\|_{L^p(U)} \longrightarrow 0 \quad \text{as } k \to \infty.$$

If $u_k \to u$ in $L^p(U)$, there exists a subsequence u_{k_j} such that $\lim_{j\to\infty} u_{k_j}(x) = u(x)$ for almost every $x \in U$.

When $p = 2$, the norm $\|u\|_{L^2(U)}$ arises from an inner product

$$(u, v)_{L^2(U)} = \int_U uv\, dx, \tag{4.10}$$

in the sense that $\|u\|_{L^2(U)} = \sqrt{(u, v)_{L^2(U)}}$. The space $L^2(U)$ is thus a *Hilbert space.* The borderline cases $p = 1$ and $p = \infty$ deserve special treatment in the Calculus of Variations, so we will not consider them here.

The norms on the L^p spaces (and any Banach space for that matter), satisfy the triangle inequality

$$\|u + v\|_{L^p(U)} \le \|u\|_{L^p(U)} + \|v\|_{L^p(U)}. \tag{4.11}$$

Another important inequality is Hölder's inequality

$$(u, v)_{L^2(U)} \le \|u\|_{L^p(U)}\|v\|_{L^q(U)}, \tag{4.12}$$

which holds for any $1 < p, q < \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$. When $p = q = 2$, Hölder's inequality is called the Cauchy-Schwarz inequality.

To define Sobolev spaces, we need to define the notion of a *weak derivative.* We write $u \in L^1_{\text{loc}}(U)$ when $u \in L^1(V)$ for all $V \subset\subset U$.

Definition 4.7 (Weak derivative)**.** Suppose $u, v \in L^1_{\text{loc}}(U)$ and $i \in \{1, \ldots, n\}$. We say that v is the *x_i-weak partial derivative of* u, written $v = u_{x_i}$, provided

$$\int_U u\varphi_{x_i}\, dx = -\int_U v\varphi\, dx \tag{4.13}$$

for all test functions $\varphi \in C^\infty_c(U)$.

The weak derivative, when it exists, is unique, and when $u \in C^1(U)$ the weak and regular derivatives coincide. When $u \in L^1_{\text{loc}}$ has weak partial derivatives in all coordinates $x_1, \ldots, x_n$ we say that u is *weakly differentiable*, and we will write $\nabla u = (u_{x_i}, \ldots, u_{x_n})$ for the vector of all weak partial derivatives.

For $1 \le p < \infty$ and weakly differentiable $u \in L^1_{\text{loc}}$ we define

$$\|u\|_{W^{1,p}(U)} = \left(\int_U |u|^p + |\nabla u|^p\, dx\right)^{1/p}. \tag{4.14}$$

[1] A Banach space is a normed linear space that is complete, so that all Cauchy sequences are convergent.

Definition 4.8 (Sobolev space)**.** We define the Sobolev space $W^{1,p}(U)$ as

$$W^{1,p}(U) = \{\text{Weakly differentiable } u \in L^p(U) \text{ for which } \|u\|_{W^{1,p}(U)} < \infty\}. \tag{4.15}$$

The Sobolev spaces $W^{1,p}(U)$ are Banach spaces, and in particular, their norms satisfy the triangle inequality (4.11). For $p = 2$ we write $H^1(U) = W^{1,2}(U)$. The space $H^1(U)$ is a Hilbert space with inner product

$$(u, v)_{H^1(U)} = \int_U u \cdot v + \nabla u \cdot \nabla v \, dx. \tag{4.16}$$

The smooth functions $C^\infty(\overline{U})$ are dense in $W^{1,p}(U)$ and $L^p(U)$. That is, for any $u \in W^{1,p}(U)$, there exists a sequence $u_m \in C^\infty(\overline{U})$ such that $u_m \to u$ in $W^{1,p}(U)$ (that is $\lim_{m\to\infty} \|u_m - u\|_{W^{1,p}(U)} = 0$). However, a general function $u \in W^{1,p}(U)$ need not be continuous, much less infinitely differentiable.

Example 4.4. Let $u(x) = |x|^\alpha$ for $\alpha \neq 0$. If $\alpha > -n$ then u is summable, and hence a candidate for weak differentiability. Let us examine the values of α for which u is weakly differentiable and belongs to $W^{1,p}(U)$, where $U = B(0, 1)$. Provided $x \neq 0$, u is classically differentiable and $\nabla u(x) = \alpha|x|^{\alpha-2}x$. Define $v(x) = \alpha|x|^{\alpha-2}x_i$. Let $\varphi \in C_c^\infty(U)$ and compute via the divergence theorem that

$$\begin{aligned}\int_U u\, \varphi_{x_i} \, dx &= \int_{U-B(0,\varepsilon)} u\, \varphi_{x_i} \, dx + \int_{B(0,\varepsilon)} u\, \varphi_{x_i} \, dx \\ &= -\int_{U-B(0,\varepsilon)} u_{x_i} \varphi \, dx - \int_{\partial B(0,\varepsilon)} u \varphi \nu_i \, dS + \int_{B(0,\varepsilon)} u\, \varphi_{x_i} \, dx,\end{aligned}$$

where ν is the outward normal to $\partial B(0, \varepsilon)$. Noting that $|u_{x_i}\varphi| \leq C|x|^{\alpha-1}$ and $|x|^{\alpha-1} \in L^1(U)$ when $\alpha > 1 - n$ we have

$$\lim_{\varepsilon\to 0} \int_{U-B(0,\varepsilon)} u_{x_i} \varphi \, dx = \int_U v\, \varphi \, dx$$

when $\alpha > 1 - n$. Similarly, $|u\varphi\nu_i| \leq C|x|^\alpha = C\varepsilon^\alpha$ for $x \in \partial B(0, \varepsilon)$, and so

$$\left| \int_{\partial B(0,\varepsilon)} u\varphi\nu_i \, dS \right| \leq C\varepsilon^{\alpha+n-1} \to 0$$

as $\varepsilon \to 0$ provided $\alpha > 1 - n$. Similarly, since u is summable we have

$$\int_{B(0,\varepsilon)} u\, \varphi_{x_i} \, dx \to 0$$

as $\varepsilon \to 0$. Therefore

$$\int_U u\, \varphi_{x_i} \, dx = -\int_U v\varphi \, dx$$

for all $\varphi \in C_c^\infty(U)$ provided $\alpha > 1 - n$. This is exactly the definition of weak differentiability, so u is weakly differentiable and $\nabla u = \alpha |x|^{\alpha-2} x_i$ in the weak sense when $\alpha > 1-n$. We have $|u_{x_i}|^p \leq C|x|^{p(\alpha-1)}$, and so $u_{x_i} \in L^p(U)$ provided $\alpha > 1 - \frac{n}{p}$. Therefore, when $\alpha > 1 - \frac{n}{p}$ we have $u \in W^{1,p}(U)$. Notice the u is not continuous and, in fact, unbounded at $x = 0$, when $\alpha < 0$, which is allowed when $p < n$.

In fact, the situation can be much worse. Provided $\alpha > 1 - \frac{n}{p}$ the function

$$u(x) = \sum_{k=1}^{\infty} \frac{1}{2^k} |x - r_k|^\alpha$$

belongs to $W^{1,p}(U)$, where $r_1, r_2, r_3, \ldots$ is any enumeration of the points in $B(0,1)$ with rational coordinates. When $p < n$, so we can choose $\alpha < 0$, the function u is unbounded on every open subset of $B(0,1)$. △

Finally, we need to consider the notion of boundary condition in Sobolev spaces. For continuous functions $u \in C(\overline{U})$, the values of $u(x)$ for $x \in \partial U$ are uniquely defined. This is less clear for Sobolev space functions $u \in W^{1,p}(U)$, which are defined only up to sets of measure zero. Fortunately we have the notion of a *trace*. That is, one can prove that there exists a bounded linear operator

$$T : W^{1,p}(U) \to L^p(\partial U) \tag{4.17}$$

such that $Tu = u|_{\partial U}$ whenever $u \in W^{1,p}(U) \cap C(\overline{U})$. The trace operator allows us to make sense of the values of a Sobolev space function $u \in W^{1,p}(U)$ on the boundary ∂U.

There are a couple of important points to make about trace-zero functions.

Definition 4.9. We denote by $W_0^{1,p}(U)$ the closure of $C_c^\infty(U)$ in $W^{1,p}(U)$, and write $H_0^1(U) = W_0^{1,p}(U)$.

Thus, $u \in W_0^{1,p}(U)$ if and only if there exist functions $u_m \in C_c^\infty(U)$ such that $u_m \to u$ in $W^{1,p}(U)$. The spaces $W_0^{1,p}(U)$ are closed subspaces of $W^{1,p}(U)$. A theorem in the theory of Sobolev spaces states that

$$Tu \equiv 0 \text{ on } \partial U \quad \text{if and only if} \quad u \in W_0^{1,p}(U). \tag{4.18}$$

Thus, we can interpret $W_0^{1,p}(U)$ as the subspace of functions in $W^{1,p}(U)$ for which "$u \equiv 0$ on ∂U." When we consider boundary values $u = g$ on ∂U, we will always consider the boundary values in the trace sense, so we mean that $Tu = g$ with $g \in L^p(\partial U)$.

4.2.2 Weak compactness in L^p spaces

Recalling the discussion in Section 4.1, we need to examine the appropriate notion of compactness in L^p and Sobolev spaces. However, in Banach spaces, compact sets in the norm topology are hard to come by.

Example 4.5. The closed unit ball $B = \{u \in L^2((0,1)) : \|u\|_{L^2((0,1))} \leq 1\}$ is not compact in $L^2((0,1))$. To see this, consider the sequence $u_m(x) = \sqrt{2}\sin(\pi m x)$. We have

$$\|u_m\|^2_{L^2((0,1))} = \int_0^1 2\sin^2(\pi m x)\,dx = \int_0^1 1 - \cos(2\pi m x)\,dx = 1.$$

Therefore $u_m \in B$. However, u_m has no convergent subsequences. Indeed, by the Riemann-Lebesgue Lemma in Fourier analysis

$$\lim_{m\to\infty}(u_m, v)_{L^2((0,1))} = \lim_{m\to\infty}\sqrt{2}\int_0^1 \sin(\pi m x)v(x)\,dx = 0 \tag{4.19}$$

for all $v \in L^2((0,1))$. Essentially, the sequence u_m is oscillating very rapidly, and when measured on average against a test function v, the oscillations average out to zero in the limit. Therefore, if any subsequence of u_{m_j} was convergent in $L^2((0,1))$ to some $u_0 \in L^2((0,1))$, then we would have

$$\|u_0\|_{L^2((0,1))} = \lim_{j\to\infty}\|u_{m_j}\|_{L^2((0,1))} = 1,$$

and

$$(u_0, v)_{L^2((0,1))} = \lim_{j\to\infty}(u_{m_j}, v)_{L^2((0,1))} = 0$$

for all $v \in L^2((0,1))$. Taking $v = u_0$ is a contradiction. △

For any $1 \leq p < \infty$, define p' so that $\frac{1}{p} + \frac{1}{p'} = 1$, taking $p' = \infty$ when $p = 1$. Note that by the Hölder inequality (4.12), the pairing $(u, v)_{L^2(U)}$ is well-defined when $u \in L^p(U)$ and $v \in L^{p'}(U)$. In this setting, u, v may not be $L^2(U)$ functions, so we will write (u, v) in place of $(u, v)_{L^2(U)}$, so that $(u, v) = \int_U u\,v\,dx$.

While we must abandon compactness in the norm topology, Eq. (4.19) suggests that in some weaker topology, the sequence u_m constructed in Example 4.5 should converge to 0.

Definition 4.10 (Weak convergence)**.** Let $1 \leq p < \infty$ and $u_m, u \in L^p(U)$. We say that u_m *converges weakly* to u in $L^p(U)$, written $u_m \rightharpoonup u$ as $m \to \infty$, provided

$$\lim_{m\to\infty}(u_m, v) = (u, v) \quad \text{for all } v \in L^{p'}(U). \tag{4.20}$$

Any weakly convergent sequence is norm-bounded, that is, if $u_m \rightharpoonup u$ in $L^p(U)$ then u_m is bounded in $L^p(U)$ ($\sup_{m\geq 1}\|u_m\|_{L^p(U)} < \infty$), and furthermore

$$\|u\|_{L^p(U)} \leq \liminf_{m\to\infty}\|u_m\|_{L^p(U)}. \tag{4.21}$$

The inequality above is in general strict, as we saw in Example 4.5.

Taking $p = q = 2$, we see that the sequence u_m from Example 4.5 converges weakly to 0 in $L^2((0,1))$. This is a special case of a more general theorem.

Definition 4.11. Let $1 \le p < \infty$. We say a subset $B \subset L^p(U)$ is *weakly sequentially compact* if for every bounded sequence $u_m \in B$, there is a subsequence u_{m_j} and some $u \in B$ such that $u_{m_j} \rightharpoonup u$ as $j \to \infty$.

Theorem 4.12 (Banach-Alaoglu Theorem)**.** *Let $1 < p < \infty$. The unit ball $B = \{u \in L^p(U) : \|u\|_{L^p(U)} \le 1\}$ is weakly sequentially compact.*

Of course, it follows that any ball $\{u \in L^p(U) : \|u\|_{L^p(U)} \le r\}$ is weakly sequentially compact as well.

Example 4.6. Theorem 4.12 does not hold for $p = 1$. We can take, for example, $U = (-1, 1)$ and the sequence of functions $u_m(x) = m/2$ for $|x| \le 1/m$ and $u_m(x) = 0$ for $|x| > 1/m$. Note that $\|u_m\|_{L^1(U)} = 1$, so the sequence is bounded in $L^1(U)$, and that $u_m(x) \to 0$ pointwise for $x \neq 0$ (actually, uniformly away from $x = 0$). So any weak limit in $L^1(U)$, if it exists, must be identically zero. However, for any smooth function $\varphi : U \to \mathbb{R}$ we have

$$\lim_{m\to\infty} (u_m, \varphi)_{L^2(U)} = \frac{m}{2}\int_{-1/m}^{1/m} \varphi(x)\, dx = \varphi(0).$$

Hence, the sequence u_m does not converge weakly to zero. In fact, the sequence u_m converges to a measure, called the Dirac delta function. Note that $\|u_m\|_{L^p(U)} = (m/2)^{1-1/p}$, and so the sequence is unbounded in $L^p(U)$ for $p > 1$.

We note that Theorem 4.12 does hold when $p = \infty$, provided we interpret weak convergence in the correct sense (which is weak-* convergence here). The issue with these borderline cases is that L^1 and L^∞ are not *reflexive* Banach spaces. △

While weak convergence seems to solve some of our problems, it immediately introduces others.

Example 4.7. If $u_m \rightharpoonup u$ in $L^p(U)$, it is in general *not true* that $f(u_m) \rightharpoonup f(u)$, even for smooth and bounded $f : \mathbb{R} \to \mathbb{R}$. Take, for example, the setting of Example 4.5 and $f(t) = t^2$. Then

$$f(u_m) = 2\sin^2(\pi m x) = 1 - 2\cos(2\pi m x).$$

Arguing as in Example 4.5 we have $u_m \rightharpoonup 0$ and $f(u_m) \rightharpoonup 1 \neq f(0)$. △

Example 4.8. If $u_m \rightharpoonup u$ in $L^p(U)$ and $w_m \rightharpoonup w$ in $L^{p'}(U)$, it is in general *not true* that

$$(u_m, w_m) \longrightarrow (u, w) \quad \text{as } m \to \infty.$$

Indeed, we can take $u_m = w_m$ from Example 4.7. △

The examples above show that weak convergence does not behave well with function composition, even in the setting of the bilinear form given by the L^2 inner-product. The consequence is that our functional I defined in (4.1) is in general *not continuous* in the weak topology. So by weakening the topology we have gained a crucial compactness condition, but have lost continuity.

The issues above show how it is, in general, difficult to pass to limits in the weak topology when one is dealing with nonlinear problems. This must be handled carefully in our analysis of minimizers of Calculus of Variations problems. We record below a useful lemma for passing to limits with weakly convergent sequences.

Lemma 4.13. *Let $1 \leq p < \infty$. If $u_m \rightharpoonup u$ in $L^p(U)$ and $w_m \to w$ in $L^{p'}(U)$, then*

$$(u_m, w_m) \longrightarrow (u, w) \quad \text{as } m \to \infty. \tag{4.22}$$

Notice the difference with Example 4.8 is that w_m converges strongly (and not just weakly) to w in $L^{p'}(U)$.

Proof. We write

$$\begin{aligned} |(u_m, w_m) - (u, w)| &= |(u_m, w) - (u, w) + (u_m, w_m) - (u_m, w)| \\ &\leq |(u_m, w) - (u, w)| + |(u_m, w_m) - (u_m, w)| \\ &= |(u_m - u, w)| + |(u_m, w_m - w)| \\ &\leq |(u_m - u, w)| + \|u_m\|_{L^p(U)} \|w_m - w\|_{L^{p'}(U)}, \end{aligned}$$

where the last line follows from Hölder's inequality. Since $u_m \rightharpoonup w$, the first term vanishes as $m \to \infty$ by definition. Since u_m is bounded and $w_m \to w$ in $L^{p'}(U)$, the second term vanishes as $m \to \infty$ as well, which completes the proof. □

4.2.3 Compactness in Sobolev spaces

To address calculus of variations problems in $\mathbb{R}^d$, we need to extend the notion of weak convergence to Sobolev spaces.

Definition 4.14 (Weak convergence in $W^{1,p}$)**.** Let $1 \leq p < \infty$ and $u_m, u \in W^{1,p}(U)$. We say that u_m *converges weakly* to u in $W^{1,p}(U)$, written $u_m \rightharpoonup u$ as $m \to \infty$, provided

$$u_m \rightharpoonup u \quad \text{and} \quad (u_m)_{x_i} \rightharpoonup u_{x_i} \quad \text{in } L^p(U) \tag{4.23}$$

for all $i \in \{1, \ldots, n\}$.

Note we are using $u_m \rightharpoonup u$ for weak convergence in L^p and $W^{1,p}$ spaces. It will be clear from the context which is intended.

As a warning, subsets of $W^{1,p}(U)$ that are closed in the strong topology are not in general closed in the weak topology. By Mazur's Theorem [50], subsets of $W^{1,p}(U)$

that are strongly closed and *convex*, are also weakly closed. For example, the closed linear subspace $W_0^{1,p}(U)$ is weakly closed in $W^{1,p}(U)$. That is, if $u_k \rightharpoonup u$ in $W^{1,p}(U)$ and $u_k \in W_0^{1,p}(U)$ for all k, then $u \in W_0^{1,p}(U)$.

It follows from the Banach-Alaoglu Theorem that the unit ball in $W^{1,p}(U)$ is weakly compact. However, we gain an additional, and very important, strong compactness in $L^p(U)$ due to bounding the weak gradient in $L^p(U)$. This is essentially what makes it possible to pass to limits with weak convergence, since we end up in the setting where we can apply tools like Lemma 4.13.

Theorem 4.15 (Weak/strong compactness). *Let $1 \leq p \leq \infty$ and let $u_m \in W^{1,p}(U)$ be a bounded sequence in $W^{1,p}(U)$ (i.e., $\sup_{m\geq 1} \|u_m\|_{W^{1,p}(U)} < \infty$). Then there exists a subsequence u_{m_j} and a function $u \in L^p(U)$ such that*

$$u_{m_j} \to u \quad \text{in } L^p(U). \tag{4.24}$$

Furthermore, if $1 < p < \infty$, we can pass to a further subsequence, if necessary, so that $u \in W^{1,p}(U)$ and

$$u_{m_j} \rightharpoonup u \quad \text{in } W^{1,p}(U). \tag{4.25}$$

Notice the convergence $u_{m_j} \to u$ is ordinary (strong) convergence in $L^p(U)$ and not weak convergence. This is the crucial part of the compact Sobolev embeddings that make them far stronger statements than the Banach-Alaoglu Theorem. In other words, we are saying that the unit ball in $W^{1,p}(U)$ is a *compact* subset of $L^p(U)$. We say that $W^{1,p}(U)$ is compactly embedded in $L^p(U)$, which is written $W^{1,p}(U) \subset\subset L^p(U)$. The reader should compare Theorem 4.15 with the Arzelá-Ascoli Theorem, which gives a similar result for classes of equicontinuous functions (in fact, the Arzelá-Ascoli Theorem is one of the main tools in the proof of Theorem 4.15).

An immediate application of compactness is the Poincaré inequality

Theorem 4.16 (Trace-zero Poincaré). *Let $1 \leq p \leq \infty$. There exists a constant $C > 0$, depending on p and U, such that*

$$\|u\|_{L^p(U)} \leq C\|\nabla u\|_{L^p(U)} \tag{4.26}$$

for all $u \in W_0^{1,p}(U)$.

Proof. Assume, by way of contradiction, that for each $k \geq 1$ there exists $u_k \in W_0^{1,p}(U)$ such that

$$\|u_k\|_{L^p(U)} > k\|\nabla u_k\|_{L^p(U)}. \tag{4.27}$$

We may assume $\|u_k\|_{L^p(U)} = 1$ so that $\|\nabla u_k\|_{L^p(U)} < 1/k$ and $\nabla u_k \to 0$ in $L^p(U)$ as $k \to \infty$. By Theorem 4.15 there exists a subsequence u_{k_j} and $u \in L^p(U)$ such that $u_{k_j} \to u$ in $L^p(U)$. Since $\|u_k\|_{L^p(U)} = 1$ we have that $\|u\|_{L^p(U)} = 1$.

For any $\varphi \in C_c^\infty(U)$ we have

$$\int_U u\varphi_{x_i}\,dx = \lim_{j\to\infty}\int_U u_{k_j}\varphi_{x_i}\,dx = \lim_{j\to\infty} -\int_U u_{k_j,x_i}\varphi\,dx = 0.$$

Therefore $u \in W^{1,p}(U)$ and $\nabla u = 0$. It follows that $u_{k_j} \to u$ in $W^{1,p}(U)$, and so $u \in W_0^{1,p}(U)$, as $u_{k_j} \in W_0^{1,p}(U)$. Since $\nabla u = 0$ in the weak sense, and $u = 0$ on ∂U in the trace sense, we have $u = 0$ almost everywhere in U (exercise), which contradicts that $\|u\|_{L^p(U)} = 1$. □

For some domains we can say more about the constant C. Let $B_r \subset \mathbb{R}^d$ denote the ball $B(0, r)$ of radius $r > 0$ centered at zero.

Corollary 4.17. *Let $1 \le p \le \infty$. There exists a constant $C > 0$, depending only on p, such that for all $r > 0$ we have*

$$\|u\|_{L^p(B_r)} \le Cr\|\nabla u\|_{L^p(B_r)} \tag{4.28}$$

for all $u \in W_0^{1,p}(B_r)$.

Proof. By Theorem 4.16, there exists $C > 0$ such that

$$\|w\|_{L^p(B_1)} \le C\|\nabla w\|_{L^p(B_1)}$$

for all $w \in W_0^{1,p}(B_1)$.

Let $u \in W_0^{1,p}(B_r)$, and set $w(y) = u(ry)$. Then $w \in W_0^{1,p}(B_1)$. By a change of variables $y = x/r$ we have

$$\int_{B_r} |u|^p\,dx = r^d\int_{B_1} |w|^p\,dx,$$

and

$$\int_{B_r} |\nabla u|^p\,dx = r^{d-p}\int_{B_1} |\nabla w|^p\,dx.$$

Therefore

$$\|u\|_{L^p(B_r)}^p = r^d\|w\|_{L^p(B_1)}^p \le C^p r^d\|\nabla w\|_{L^p(U)}^p = C^p r^p\|\nabla u\|_{L^p(B_r)}^p,$$

which completes the proof. □

Poincaré-type inequalities are true in many other situations.

Exercise 4.18. Let $1 \le p \le \infty$. Show that there exists a constant $C > 0$, depending on p and U, such that

$$\|u\|_{L^p(U)} \le C(\|Tu\|_{L^p(\partial U)} + \|\nabla u\|_{L^p(U)}) \tag{4.29}$$

for all $u \in W^{1,p}(U)$, where $T : W^{1,p}(U) \to L^p(\partial U)$ is the trace operator. [Hint: Use a very similar argument to the proof of Theorem 4.16.] △

Exercise 4.19. Let $1 \leq p \leq \infty$. Show that there exists a constant $C > 0$, depending on p and U, such that

$$\|u - (u)_U\|_{L^p(U)} \leq C\|\nabla u\|_{L^p(U)} \tag{4.30}$$

for all $u \in W^{1,p}(U)$, where $(u)_U := \fint_U u\, dx$. [Hint: Use a very similar argument to the proof of Theorem 4.16.] △

Exercise 4.20. Let $1 \leq p \leq \infty$. Show that there exists a constant $C > 0$, depending on p, such that

$$\|u - (u)_{B_r}\|_{L^p(B_r)} \leq Cr\|\nabla u\|_{L^p(B_r)} \tag{4.31}$$

for all $u \in W^{1,p}(B_r)$. △

4.3 Lower semicontinuity

We now examine conditions on L for which (4.1) is lower semicontinuous in the weak topology.

Definition 4.21. We say that a functional I, defined by (4.1), is *(sequentially) weakly lower semicontinuous* on $W^{1,q}(U)$, provided

$$I(u) \leq \liminf_{k\to\infty} I[u_k]$$

whenever

$$u_k \rightharpoonup u \text{ weakly in } W^{1,q}(U).$$

Theorem 4.22 (Weak lower semicontinuity). *Let $1 < q < \infty$. Assume L is bounded below and that $p \mapsto L(x, z, p)$ is convex for each $z \in \mathbb{R}$ and $x \in U$. Then I is weakly lower semicontinuous on $W^{1,q}(U)$.*

We split up the proof of Theorem 4.22 into two parts. The first part, Lemma 4.23, contains the main ideas of the proof under the stronger condition of uniform convergence, while the proof of Theorem 4.22, given afterwards, shows how to reduce to this case using standard tools in analysis.

Lemma 4.23. *Let $1 < q < \infty$. Assume L is bounded below and convex in p. Let $u_k \in W^{1,q}(U)$ and $u \in W^{1,p}(U)$ such that $u_k \rightharpoonup u$ weakly in $W^{1,q}(U)$, $u_k \to u$ uniformly on U, and $|\nabla u|$ is uniformly bounded on U. Then*

$$I(u) \leq \liminf_{k\to\infty} I(u_k). \tag{4.32}$$

Proof. Since L is convex in p, it follows from Theorem A.38 (iii) that

$$L(x,u_k,\nabla u_k) \geq L(x,u_k,\nabla u) + \nabla_p L(x,u_k,\nabla u)\cdot(\nabla u_k - \nabla u).$$

Integrating over U we have

$$I(u_k) \geq \int_U L(x,u_k,\nabla u)\,dx + \int_U \nabla_p L(x,u_k,\nabla u)\cdot(\nabla u_k - \nabla u)\,dx. \tag{4.33}$$

Let $M > 0$ such that $|u|, |\nabla u| \leq M$ a.e. on U. Since L is smooth

$$|L(x,u_k,\nabla u) - L(x,u,\nabla u)| \leq \sup_{\substack{x\in U\\ |z|\leq M\\ |p|\leq M}} |L_z(x,z,p)||u_k - u|,$$

a.e. in U. Since $u_k \to u$ uniformly, we deduce that $L(x,u_k,\nabla u) \to L(x,u,\nabla u)$ uniformly on U and so

$$\int_U L(x,u_k,\nabla u)\,dx \longrightarrow I(u)$$

as $k\to\infty$. Similarly we can show that $\nabla_p L(x,u_k,\nabla u) \to \nabla_p L(x,u,\nabla u)$ uniformly on U, and hence in $L^p(U)$ for any $1 \leq p \leq \infty$. Since $\nabla u_k \rightharpoonup \nabla u$ in $W^{1,q}(U)$, we can use Lemma 4.13 to deduce that

$$\int_U \nabla_p L(x,u_k,\nabla u)\cdot(\nabla u_k - \nabla u)\,dx \longrightarrow 0$$

as $k\to\infty$. Inserting these observations into (4.33) completes the proof. □

Remark 4.24. It is important to understand how convexity of $p \mapsto L(x,z,p)$ is used to deal with weak convergence in the proof of Lemma 4.23, by allowing the gradient ∇u_k to appear *linearly*. The uniform convergence $u_k \to u$ is much stronger, and hence we do not need any convexity assumption on $z \mapsto L(x,z,p)$.

We now give the proof of Theorem 4.22.

Proof of Theorem 4.22. Let $u_k \in W^{1,q}(U)$ be any sequence with $u_k \rightharpoonup w$ weakly in $W^{1,q}(U)$. Then we have

$$\sup_k \|w_k\|_{W^{1,q}(U)} < \infty. \tag{4.34}$$

It follows from the Sobolev compact embedding (Theorem 4.15) that, upon passing to a subsequence if necessary, $u_k \to u$ strongly in $L^q(U)$. Passing to another subsequence we have

$$u_k \to u \quad a.e. \text{ in } U.$$

Let $m > 0$. By Egoroff's Theorem (Theorem A.29) there exists $E_m \subset U$ such that $|U - E_m| \leq 1/m$ and $u_k \to u$ uniformly on E_m. We define

$$F_m := \{x \in U \,:\, |u(x)| + |\nabla u(x)| \leq m\},$$

and

$$G_m = F_m \cap (E_1 \cup E_2 \cup \cdots \cup E_m).$$

Then

$$u_k \to u \quad \text{uniformly on } G_m \tag{4.35}$$

for each m. Furthermore, the sequence of sets G_m is monotone, i.e.,

$$G_1 \subset G_2 \subset G_3 \subset \cdots \subset G_m \subset G_{m+1} \subset \cdots,$$

and

$$|U - G_m| \leq |U - F_m| + |U - E_m| \leq |U - F_m| + \frac{1}{m}.$$

By the Dominated Convergence Theorem (Theorem A.28) we have

$$\lim_{m\to\infty} |U - F_m| = \lim_{m\to\infty} \int_U 1 - \chi_{F_m}(x)\, dx = 0,$$

where $\chi_{F_m}(x) = 1$ when $x \in F_m$ and zero otherwise. Therefore $|U - G_m| \to 0$ as $m \to \infty$.

Since L is bounded below, there exists $\beta \in \mathbb{R}$ such that $L(x, z, p) + \beta \geq 0$ for all $x \in U$, $z \in \mathbb{R}$, and $p \in \mathbb{R}^d$. We have

$$I(u_k) + \beta|U| = \int_U L(x, u_k, \nabla u_k) + \beta\, dx \geq \int_{G_m} L(x, u_k, \nabla u_k) + \beta\, dx. \tag{4.36}$$

Since $u_k \to u$ uniformly on G_m, and $|\nabla u| \leq m$ on G_m, we can apply Lemma 4.23 to deduce that

$$\liminf_{k\to\infty} \int_{G_m} L(x, u_k, \nabla u_k)\, dx \geq \int_{G_m} L(x, u, \nabla u)\, dx.$$

Inserting this into (4.36) we have

$$\liminf_{k\to\infty} I(u_k) + \beta|U| \geq \int_{G_m} L(x, u, \nabla u) + \beta\, dx.$$

By the Monotone Convergence Theorem (Theorem A.27) we have

$$\begin{aligned} \lim_{m\to\infty} \int_{G_m} L(x, u, \nabla u) + \beta\, dx &= \lim_{m\to\infty} \int_U \chi_{G_m}(x)(L(x, u, \nabla u) + \beta)\, dx \\ &= \int_U L(x, u, \nabla u) + \beta\, dx. \end{aligned}$$

Therefore

$$\liminf_{k\to\infty} I(u_k) + \beta|U| \geq \int_U L(x, u, \nabla u) + \beta\, dx = I(u) + \beta|U|,$$

which completes the proof. □

4.4 Existence and uniqueness of minimizers

Armed with lower semicontinuity, we can now turn to the question of existence of a minimizer. We need a further assumption on L. We will assume that

$$\text{(4.37)}\qquad \begin{cases}\text{there exists } \alpha > 0, \beta \geq 0 \text{ such that}\\ \qquad L(x,z,p) \geq \alpha|p|^q - \beta\\ \text{for all } x \in U, z \in \mathbb{R}, p \in \mathbb{R}^d.\end{cases}$$

We call (4.37) a *coercivity condition*, since it plays the role of coercivity from Section 4.1. We also need to allow for the imposition of constraints on the function u. Thus, we write

$$\text{(4.38)}\qquad \mathcal{A} = \{w \in W^{1,q}(U) \,:\, w = g \text{ on } \partial U \text{ in the trace sense.}\},$$

for given $g \in L^q(\partial U)$.

Theorem 4.25. *Let $1 < q < \infty$. Assume that L is bounded below, satisfies the coercivity condition (4.37), and is convex in p. Suppose also that the admissible set $\mathcal{A}$ is nonempty. Then there exists $u \in \mathcal{A}$ such that*

$$\text{(4.39)}\qquad I(u) = \min_{w\in\mathcal{A}} I(w).$$

Proof. We may assume $\inf_{w\in\mathcal{A}} I(w) < \infty$. Select a minimizing sequence $u_k \in \mathcal{A}$, so that $I(u_k) \to \inf_{w\in\mathcal{A}}$ as $k \to \infty$. We can pass to a subsequence, if necessary, so that $I(u_k) < \infty$ for all k, so that $\sup_{k\geq 1} I(u_k) < \infty$. By the coercivity condition (4.37)

$$I(u_k) \geq \int_U \alpha|\nabla u_k|^q - \beta\, dx,$$

and so

$$\sup_{k\geq 1} \int_U |\nabla u_k|^q\, dx < \infty.$$

Fix any function $w \in \mathcal{A}$. Then $w = u_k$ on ∂U in the trace sense, and so $w - u_k \in W^{1,q}_0(U)$. By the Poincaré inequality (Theorem 4.16) we have

$$\|w - u_k\|_{L^q(U)} \leq C\|\nabla w - \nabla u_k\|_{L^q(U)},$$

and so

$$\begin{aligned}\|u_k\|_{L^q(U)} &\leq \|w - u_k\|_{L^q(U)} + \|w\|_{L^q}\\ &\leq C\|\nabla w - \nabla u_k\|_{L^q(U)} + \|w\|_{L^q}\\ &\leq C\|\nabla w\|_{L^q}(U) + C\|\nabla u_k\|_{L^q(U)} + \|w\|_{L^q}.\end{aligned}$$

Therefore $\sup_{k\geq 1}\|u_k\|_{W^{1,q}(U)} < \infty$, and so there exists a subsequence u_{k_j} and a function $u \in W^{1,q}(U)$ such that $u_{k_j} \rightharpoonup u$ weakly in $W^{1,q}(U)$ as $j \to \infty$. By Theorem 4.22 we have

$$I(u) \leq \liminf_{j\to\infty} I(u_{k_j}) = \inf_{w\in\mathcal{A}} I(w).$$

To complete the proof, we need to show that $u \in \mathcal{A}$, that is, that $u = g$ on ∂U in the trace sense. Since $W_0^{1,q}(U)$ is weakly closed in $W^{1,q}(U)$ and $w - u_k \in W_0^{1,q}(U)$ for all k, we have $w - u \in W_0^{1,q}(U)$. Therefore $w = u = g$ on ∂U in the trace sense, which completes the proof. □

We finally turn to the problem of uniqueness of minimizers. There can, in general, be many minimizers, and to ensure uniqueness we need further assumptions on L. For simplicity, we present the results for L of the form

$$L(x,z,p) = \Psi(x,z) + \Phi(x,p). \tag{4.40}$$

Theorem 4.26. *Suppose L is given by (4.40), and assume $z \mapsto \Psi(x,z)$ is θ_1-strongly convex and $p \mapsto \Phi(x,p)$ is θ_2-strongly convex for all $x \in U$, where $\theta_1, \theta_2 \geq 0$. If $u \in \mathcal{A}$ is a minimizer of I, then*

$$\frac{1}{2}\int_U \theta_1(u-w)^2 + \theta_2|\nabla u - \nabla w|^2\, dx \leq I(w) - I(u) \tag{4.41}$$

for all $w \in \mathcal{A}$.

Remark 4.27. If $\theta_1 > 0$ or $\theta_2 > 0$, then (4.41) shows that minimizers of I are unique. However, the estimate (4.41) is stronger than uniqueness. It shows that any function $w \in \mathcal{A}$ whose energy $I(w)$ is close to minimal is in fact close to the unique minimizer of I in the $H^1(U)$ norm.

Proof. Let $u \in \mathcal{A}$ be a minimizer of I and let $w \in \mathcal{A}$. By Exercise A.36 we have

$$\Psi\left(x, \lambda u + (1-\lambda)w)\right) + \frac{\theta_1}{2}\lambda(1-\lambda)(u-w)^2 \leq \lambda\Psi(x,u) + (1-\lambda)\Psi(x,w),$$

and

$$\Phi\left(x, \lambda\nabla u + (1-\lambda)\nabla w)\right) + \frac{\theta_2}{2}\lambda(1-\lambda)|\nabla u - \nabla w|^2 \leq \lambda\Phi(x,\nabla u) + (1-\lambda)\Phi(x,\nabla w),$$

for any $\lambda \in (0,1)$. Adding the two equations above and integrating both sides over U yields

$$I\left(\lambda u + (1-\lambda)w)\right) + \frac{1}{2}\lambda(1-\lambda)\int_U \theta_1(u-w)^2 + \theta_2|\nabla u - \nabla w|^2\, dx \leq \lambda I(u) + (1-\lambda)I(w).$$

Since $I\left(\lambda u + (1-\lambda)w)\right) \geq I(u)$ we find that

$$\frac{1}{2}\lambda(1-\lambda)\int_U \theta_1(u-w)^2 + \theta_2|\nabla u - \nabla w|^2\, dx \leq (1-\lambda)(I(w) - I(u)).$$

Divide by $1-\lambda$ on both sides and send $\lambda \to 1$ to complete the proof. □

Exercise 4.28. Show that minimizers of the q-Dirichlet energy

$$I(u) = \int_U |\nabla u|^q \, dx$$

are unique for $1 < q < \infty$. [Hint: Note that while $L(x,z,p) = |p|^q$ is not strongly convex in p, it is strictly convex. Use a proof similar to Theorem 4.26.] △

4.5 The Euler-Lagrange equation

We now examine the sense in which a minimizer u of $I(u)$ satisfies the boundary-value problem for the Euler-Lagrange equation

$$\left\{\begin{aligned} L_z(x,u,\nabla u) - \operatorname{div}(\nabla_p L(x,u,\nabla u) &= 0 \quad \text{in } U\\ u &= g \quad \text{on } \partial U\end{aligned}\right. \tag{4.42}$$

Since our candidate solutions live in $W^{1,q}(U)$, we need a weaker notion of solution that does not require differentiating u twice. For motivation, recall when we derived the Euler-Lagrange equation in Theorem 2.1 we arrive at Eq. (2.5), which reads

$$\int_U L_z(x,u,\nabla u)\varphi + \nabla_p L(x,u,\nabla u)\cdot \nabla\varphi \, dx = 0, \tag{4.43}$$

where $\varphi \in C_c^\infty(U)$ is any test function. We obtained the Euler-Lagrange equation by integrating by parts and using a vanishing lemma. If u is not twice differentiable, we cannot complete this last step, and are left with (4.43).

Definition 4.29. We say $u \in \mathcal{A}$ is a *weak solution* of the boundary-value problem (4.42) provided

$$\int_U L_z(x,u,\nabla u)v + \nabla_p L(x,u,\nabla u)\cdot \nabla v \, dx = 0 \tag{4.44}$$

for all $v \in W_0^{1,q}(U)$.

Remark 4.30. Notice the only difference between (4.43) and (4.44) is the choice of function space for the test function v. Since $C_c^\infty(U)$ is dense in $W_0^{1,p}(U)$ for any p, it is worth examining momentarily why we chose $p = q$ in Definition 4.29. The typical example of a functional on $W^{1,q}(U)$ is

$$I(u) = \frac{1}{q}\int_U |u|^q + |\nabla u|^q \, dx,$$

in which case $L(x,z,p) = \frac{1}{q}(|z|^q + |p|^q)$. For any $u \in W^{1,q}(U)$ we have

$$|\nabla_p L(x,u,\nabla u)| = |\nabla u|^{q-1},$$

and so $\nabla_p L(x,u,\nabla u) \in L^{q'}(U)$, where $q' = \frac{q}{q-1}$, $\frac{1}{q} + \frac{1}{q'} = 1$. Hence, we need $\nabla v \in L^q(U)$ to ensure the integral in (4.44) exists (by Hölder's inequality). Similarly $|L_z(x,u,\nabla u)| = |z|^{q-1}$, and so $L_z(x,u,\nabla u) \in L^{q'}(U)$, which also necessitates $v \in L^q(U)$.

Following the discussion in Remark 4.30, we must place some assumptions on L so that the weak form (4.44) of the Euler-Lagrange equation is well-defined. We assume that

$$\left\{\begin{aligned} |L(x,z,p)| &\le C(|p|^q + |z|^q + 1), \\ |L_z(x,z,p)| &\le C(|p|^{q-1} + |z|^{q-1} + 1), \text{ and} \\ |\nabla_p L(x,z,p)| &\le C(|p|^{q-1} + |z|^{q-1} + 1), \end{aligned}\right. \tag{4.45}$$

hold for some constant $C > 0$ and all $p \in \mathbb{R}^d$, $z \in \mathbb{R}$, and $x \in U$.

Theorem 4.31. *Assume L satisfies (4.45) and $u \in \mathcal{A}$ satisfies*

$$I(u) = \min_{w \in \mathcal{A}} I(w).$$

Then u is a weak solution of the Euler-Lagrange equation (4.42).

Proof. Let $v \in W_0^{1,q}(U)$ and for $t \neq 0$ define

$$w_t(x) = \frac{1}{t}\left[L(x,u(x)+tv(x),\nabla u(x) + t\nabla v(x)) - L(x,u(x),\nabla u(x))\right].$$

Then we have

$$\begin{aligned} w_t(x) &= \frac{1}{t}\int_0^t \frac{d}{ds} L(x,u(x)+sv(x),\nabla u(x)+s\nabla v(x))\,ds \\ &= \frac{1}{t}\int_0^t L_z(x,u+sv,\nabla u+s\nabla v)v + \nabla_p L(x,u+sv,\nabla u+s\nabla v)\cdot\nabla v\,ds. \end{aligned}$$

Applying Young's inequality $ab \le \frac{a^q}{q} + \frac{b^{q'}}{q'}$ (see (A.2)) we deduce

$$\begin{aligned} |w_t(x)| \le \frac{C}{t}\int_0^t |L_z(x,u+sv,&\nabla u+s\nabla v)|^{q'} + |v|^q + \\ &|\nabla_p L(x,u+sv,\nabla u+s\nabla v)|^{q'} + |\nabla v|^q\,ds. \end{aligned}$$

Since $s \mapsto s^{q'}$ is convex, we have

$$(a+b)^{q'} = 2^{q'}(\tfrac{1}{2}a + \tfrac{1}{2}b)^{q'} \le 2^{q'-1}(a^{q'} + b^{q'})$$

for all $a,b > 0$. Invoking (4.45) we have

$$\begin{aligned} |w_t(x)| &\le \frac{C}{t}\int_0^t |v|^q + |\nabla v|^q + |u|^q + |\nabla u|^q + 1\,ds \\ &\le C(|v|^q + |\nabla v|^q + |u|^q + |\nabla u|^q + 1) \in L^1(U) \end{aligned}$$

for every nonzero $t \in [-1,1]$. Thus, we can apply the Dominated Convergence Theorem (Theorem A.28) to find that

$$\begin{aligned} 0 = \frac{d}{dt}\Big|_{t=0} I(u+tv) &= \lim_{t\to 0} \int_U w_t(x)\,dx \\ &= \int_U \lim_{t\to 0} w_t(x)\,dx \\ &= \int_U L_z(x,u,\nabla u) + \nabla_p L(x,u,\nabla u)\cdot \nabla v\,dx, \end{aligned}$$

since u is a minimizer of I. This completes the proof. □

So far, we have only shown that minimizers of I satisfy the Euler-Lagrange equation (4.42). However, the Euler-Lagrange equation may have other solutions that are not minimizers (e.g., maximizers or saddle points). Consider, for example, the minimal surface of revolution problem discussed in Section 3.4, where we found two solutions of the Euler-Lagrange equation, but only one solution yielded the least area.

So a natural question concerns sufficient conditions for a weak solution of the Euler-Lagrange equation to be a minimizer. Convexity in some form must play a role, and it turns out we require *joint convexity* in z and p. Namely, we assume that

$$(z,p) \mapsto L(x,z,p) \text{ is convex for all } x \in U. \tag{4.46}$$

If L is jointly convex in (z,p), then it is easy to check that $z \mapsto L(x,z,p)$ and $p \mapsto L(x,z,p)$ are convex. The converse is not true.

Example 4.9. The function $u(x) = x_1x_2$ is not convex, since

$$\sum_{i=1}^{2}\sum_{j=1}^{2} u_{x_ix_j} v_i v_j = 2v_1v_2$$

is clearly not positive for, say, $v_1 = 1$ and $v_2 = -1$. However, $x_1 \mapsto u(x_1,x_2)$ is convex, since $u_{x_1x_1} = 0$, and $x_2 \mapsto u(x_1,x_2)$ is convex, since $u_{x_2x_2} = 0$. Therefore, convexity of $x \mapsto u(x)$ is not equivalent to convexity in each variable x_1 and x_2 independently. △

Theorem 4.32. *Assume the joint convexity condition (4.46) holds. If $u \in \mathcal{A}$ is a weak solution of the Euler-Lagrange equation (4.42) then*

$$I(u) = \min_{w\in\mathcal{A}} I(w).$$

Proof. Let $w \in \mathcal{A}$. Since $(z,p) \mapsto L(x,z,p)$ is convex, we can use Theorem A.38 (iii) to obtain

$$L(x,w,\nabla w) \geq L(x,u,\nabla u) + L_z(x,u,\nabla u)(w-u) + \nabla_p L(x,u,\nabla u)\cdot(\nabla w - \nabla u).$$

We now integrate both sides over U and write $\varphi = w - u$ to deduce

$$I(w) \geq I(u) + \int_U L_z(x, u, \nabla u)(w - u) + \nabla_p L(x, u, \nabla u) \cdot \nabla(w - u)\, dx.$$

Since u is a weak solution of the Euler-Lagrange equation (4.42) and $w - u \in W_0^{1,q}(U)$ we have

$$\int_U L_z(x, u, \nabla u)(w - u) + \nabla_p L(x, u, \nabla u) \cdot \nabla(w - u)\, dx = 0$$

and so $I(w) \geq I(u)$. □

Exercise 4.33. Show that $L(x, z, p) = \frac{1}{2}|p|^2 - zf(x)$ is jointly convex in z and p. △

Exercise 4.34. Show that $L(x, z, p) = zp_1$ is not jointly convex in z and p. △

4.5.1 Regularity

It is natural to ask whether the weak solution we constructed to the Euler-Lagrange equation (4.42) is in fact smooth, or has some additional regularity beyond $W^{1,q}(U)$. In fact, this was one of the famous 23 problems posed by David Hilbert in 1900. Hilbert's 19th problem asked "Are the solutions of regular problems in the calculus of variations necessarily analytic"? The term "regular problems" refers to strongly convex Lagrangians L for which the Euler-Lagrange equation is uniformly elliptic, and it is sufficient to interpret the term "analytic" as "smooth".

The problem was resolved affirmatively by de Giorgi [23] and Nash [44], independently. Moser later gave an alternative proof [43], and the resulting theory is often called the de Giorgi-Nash-Moser theory. We should note that the special case of $n = 2$ is significantly easier, and was established earlier by Morrey [42].

We will not give proofs of regularity here, but will say a few words about the difficulties and the main ideas behind the proofs. We consider the simple calculus of variations problem

$$\min_u \int_U L(\nabla u)\, dx, \tag{4.47}$$

where L is smooth and θ-strongly convex with $\theta > 0$. The corresponding Euler-Lagrange equation is

$$\operatorname{div}(\nabla_p L(\nabla u)) = \sum_{j=1}^d \partial_{x_j}(L_{p_j}(\nabla u)) = 0 \quad \text{in } U. \tag{4.48}$$

The standard way to obtain regularity in PDE theory is to differentiate the equation to get a PDE satisfied by the derivative u_{x_k}. For simplicity we assume u is smooth. Differentiate (4.48) in x_k to obtain

$$\sum_{i,j=1}^d \partial_{x_j}(L_{p_i p_j}(\nabla u) u_{x_k x_i}) = 0 \quad \text{in } U.$$

Setting $a_{ij}(x) = L_{p_i p_j}(\nabla u(x))$ and $w = u_{x_k}$ we have

$$\sum_{i,j=1}^{d} \partial_{x_j}(a_{ij} w_{x_i}) = 0 \quad \text{in } U. \tag{4.49}$$

Since L is strongly convex and smooth we have

$$\theta|\eta|^2 \leq \sum_{i,j=1}^{d} a_{ij}\eta_i\eta_j \leq \Theta|\eta|^2, \tag{4.50}$$

for some $\Theta > 0$. Thus (4.49) is uniformly elliptic. Notice we have converted our problem of regularity of minimizers of (4.47) into an elliptic regularity problem.

However, since $u \in W^{1,q}$ is not smooth, we at best have $a_{ij} \in L^\infty(U)$, and in particular, the coefficients a_{ij} are *not a priori continuous*, as is necessary to apply elliptic regularity theory, such as the Schauder estimates [30]. All we know about the coefficients is the ellipticity condition (4.50). Nonetheless, the remarkable conclusion is that $w \in C^{0,\alpha}$ for some $\alpha > 0$. This was proved separately by de Giorgi [23] and Nash [44]. Since $w = u_{x_k}$ we have $u \in C^{1,\alpha}$ and so $a_{ij} \in C^{0,\alpha}$. Now we are in the world of classical elliptic regularity. A version of the Schauder estimates shows that $u \in C^{2,\alpha}$, and so $a_{ij} \in C^{1,\alpha}$. But another application of a different Schauder estimate gives $u \in C^{3,\alpha}$ and so $a_{ij} \in C^{2,\alpha}$. Iterating this argument is called *bootstrapping*, and leads to the conclusion that $u \in C^\infty$ is smooth.

4.6 Minimal surfaces

Our discussion in the preceding sections does not cover the minimal surface energy

$$I(u) = \int_U \sqrt{1 + |\nabla u|^2}\, dx,$$

since the Lagrangian $L(p) = \sqrt{1 + |p|^2}$ is *coercive* (see (4.37)) with $q = 1$, and we required $q > 1$ for weak compactness. In this section, we show how to problems like this with *a priori* gradient estimates. We will consider the general functional

$$I(u) = \int_U L(\nabla u)\, dx, \tag{4.51}$$

where L is strictly convex, which means

$$L(\lambda p + (1-\lambda)q) < \lambda L(p) + (1-\lambda)L(q) \tag{4.52}$$

for all $p, q \in \mathbb{R}^d$ and $\lambda \in (0,1)$, and consider minimizing L over the constraint set

$$\mathcal{A} = \{u \in C^{0,1}(U) \,:\, u = g \text{ on } \partial U\}, \tag{4.53}$$

where $g \in C^{0,1}(\partial U)$.

We recall that $C^{0,1}(U)$ is the space of Lipschitz continuous functions $u : U \to \mathbb{R}$ with norm

$$\|u\|_{C^{0,1}(U)} := \|u\|_{C(U)} + \mathrm{Lip}(u), \tag{4.54}$$

where $\|u\|_{C(U)} = \sup_{x\in U} |u(x)|$ and

$$\mathrm{Lip}(u) := \sup_{\substack{x,y\in U\\ x\neq y}} \frac{|u(x)-u(y)|}{|x-y|}. \tag{4.55}$$

The Lipschitz constant $\mathrm{Lip}(u)$ is the smallest number $C > 0$ for which

$$|u(x) - u(y)| \leq C|x-y|$$

holds for all $x, y \in U$. By Radamacher's Theorem [28], every Lipschitz function is differentiable almost everywhere in U, and $C^{0,1}(U)$ is continuously embedded into $W^{1,q}(U)$ for every $q \geq 1$, which means

$$\|u\|_{W^{1,q}(U)} \leq C\|u\|_{C^{0,1}(U)}. \tag{4.56}$$

Our strategy to overcome the lack of coercivity will be to modify the constraint set to one of the form

$$\mathcal{A}_m = \{u \in C^{0,1}(U) \,:\, \mathrm{Lip}(u) \leq m \text{ and } u = g \text{ on } \partial U\}. \tag{4.57}$$

By (4.56), minimizing sequences are bounded in $W^{1,q}(U)$, which was all that coercivity was used for. Then we will show that minimizers over $\mathcal{A}_m$ satisfy $\|u\|_{C^{0,1}(U)} < m$ by proving *a priori* gradient bounds on u, which will show that minimizers of $\mathcal{A}_m$ are the same as over $\mathcal{A}$.

Theorem 4.35. *Assume L is convex and bounded below. Then there exists $u \in \mathcal{A}_m$ such that $I(u) = \min_{w\in\mathcal{A}_m} I(w)$.*

Proof. Let $u_k \in \mathcal{A}_m$ be a minimizing sequence. By the continuous embedding (4.56), the sequence u_k is bounded in $H^1(U) = W^{1,2}(U)$. Thus, we can apply the argument from Theorem 4.25, and the Sobolev compact embeddings 4.15, to show there is a subsequence u_{k_j} and a function $u \in H^1(U)$ such that $u = g$ on ∂U in the trace sense, $u_{k_j} \rightharpoonup u$ weakly in $H^1(U)$, $u_{k_j} \to u$ in $L^2(U)$, and

$$I(u) \leq \liminf_{j\to\infty} I\left(k_j\right) = \inf_{w\in I} I(w).$$

By passing to a further subsequence, if necessary, $u_{k_j} \to u$ a.e. in U. It follows that

$$\frac{|u(x)-u(y)|}{|x-y|} = \lim_{j\to\infty} \frac{|u_{k_j}(x) - u_{k_j}(y)|}{|x-y|},$$

for almost every $x, y \in U$, and so (exercise) u has a continuous version (e.g., can be redefined on a set of measure zero) with $u \in C^{0,1}(U)$ and $\mathrm{Lip}(u) \leq m$. Therefore $u \in \mathcal{A}_m$, which completes the proof. □

We now turn to proving *a priori* estimates on minimizers. For this, we need the comparison principle and notions of sub and supersolutions. We recall $C^{0,1}_c(U)$ is the subset of $C^{0,1}(U)$ consisting of functions with compact support in U, that is, each $u \in C^{0,1}_c(U)$ is identically zero in some neighborhood of ∂U.

Definition 4.36. We say that $u \in C^{0,1}(U)$ is a *subsolution* if

$$I(u-\varphi) \geq I(u) \quad \text{for all } \varphi \in C^{0,1}_c(U),\ \varphi \geq 0. \tag{4.58}$$

Similarly, we say that $v \in C^{0,1}(U)$ is a *supersolution* if

$$I(v+\varphi) \geq I(v) \quad \text{for all } \varphi \in C^{0,1}_c(U),\ \varphi \geq 0. \tag{4.59}$$

Note that any minimizer of I over $\mathcal{A}$ is both a sub- and supersolution.

We have the following comparison principle.

Theorem 4.37. *Assume L is strictly convex. Let u and v be sub- and supersolutions, respectively, and assume $u \leq v$ on ∂U. Then $u \leq v$ in U.*

Proof. Assume by way of contradiction that $u(x) > v(x)$ for some $x \in U$. Define

$$K = \{x \in U \,:\, u(x) > v(x)\},$$

and consider the functions

$$\overline{u}(x) = \begin{cases} u(x), & \text{if } x \in U \setminus K \\ v(x), & \text{if } x \in K, \end{cases}$$

and

$$\overline{v}(x) = \begin{cases} v(x), & \text{if } x \in U \setminus K \\ u(x), & \text{if } x \in K, \end{cases}$$

Since $u \leq v$ on ∂U, K is an open set compactly contained in U.

Note that $\overline{u} \leq u$ and define $\varphi := u - \overline{u}$. Then $\overline{u} = u - \varphi$ and $\varphi \geq 0$ is supported on K, so $\varphi \in C^{0,1}_c(U)$. By the definition of subsolution we have $I(\overline{u}) \geq I(u)$. By a similar argument we have $I(\overline{v}) \geq I(v)$. Therefore

$$\int_U L(\nabla u)\,dx \leq \int_U L(\nabla \overline{u})\,dx = \int_{U\setminus K} L(\nabla u)\,dx + \int_K L(\nabla v)\,dx,$$

and so

$$\int_K L(\nabla u)\,dx \leq \int_K L(\nabla v)\,dx.$$

The opposite inequality is obtained by a similar argument and so we have

$$\int_K L(\nabla v)\,dx = \int_K L(\nabla u)\,dx. \tag{4.60}$$

Define now

$$w(x) = \begin{cases} u(x), & \text{if } x \in U \setminus K \\ \frac{1}{2}(u(x) + v(x)), & \text{if } x \in K. \end{cases}$$

As above, the subsolution property yields $I(w) \geq I(u)$ and so we can use the argument above and (4.60) to obtain

$$\begin{aligned} \int_K \frac{1}{2}(L(\nabla u) + L(\nabla v))\, dx &= \frac{1}{2}\int_K L(\nabla u)\, dx + \frac{1}{2}\int_U L(\nabla v))\, dx \\ &= \int_K L(\nabla u)\, dx \\ &\leq I(w) = \int_K L\left(\frac{\nabla u + \nabla v}{2}\right)\, dx. \end{aligned}$$

Therefore

$$\int_K L\left(\frac{\nabla u + \nabla v}{2}\right) - \frac{1}{2}(L(\nabla u) + L(\nabla v))\, dx \geq 0.$$

By convexity of L we conclude that

$$L\left(\frac{\nabla u + \nabla v}{2}\right) = \frac{1}{2}(L(\nabla u) + L(\nabla v))$$

a.e. in K. Since L is strictly convex, we have

$$L\left(\frac{p+q}{2}\right) \leq \frac{1}{2}(L(p) + L(q)).$$

with equality if and only if $p = q$. Therefore $\nabla u = \nabla v$ a.e. in K, and so $u - v$ is constant in K. Since $u = v$ on ∂K, we find that $u \equiv v$ in K, which is a contradiction to the definition of K. □

Remark 4.38. Notice that if $u, v \in \mathcal{A}_m$ in the proof of Theorem 4.37, then $\overline{u}, \overline{v}, w \in \mathcal{A}_m$. Hence, the comparison principle holds when $u, v \in \mathcal{A}_m$, and the notion of sub- and supersolution is modified so that we require $u + \varphi \in \mathcal{A}_m$ and $u - \varphi \in \mathcal{A}_m$.

As a corollary, we have the maximum principle.

Corollary 4.39. *Let u and v be sub- and supersolutions, respectively. Then we have*

$$\sup_U (u - v) \leq \sup_{\partial U} (u - v). \tag{4.61}$$

In particular, if u is both a sub- and supersolution, then

$$\sup_U |u| = \sup_{\partial U} |u|. \tag{4.62}$$

Proof. Since $I(u+\lambda) = I(u)$ for any constant λ, adding constants does not change the property of being a sub- or supersolution. Therefore $w := v + \sup_{\partial U}(u-v)$ is a supersolution satisfying $u \le w$ on ∂U. By Theorem 4.37 we have $u \le w$ in U, which proves (4.61).

Note that for any $\varphi \in C_c^{0,1}(U)$ we have by convexity of L that

$$\int_U L(0 \pm D\varphi)\,dx \ge \int_U L(0) \pm \nabla_p L(0)\cdot\nabla\varphi\,dx = \int_U L(0),$$

where we used integration by parts in the last equality. Therefore, any constant function is both a sub- and supersolution. Using (4.61) with $v=0$ we have $\sup_U u \le \sup_{\partial U} u$. If u is also a supersolution, then the opposite inequality is also obtained, yielding (4.62). □

Lemma 4.40. *Let $u \in \mathcal{A}$ be a minimizer of I over $\mathcal{A}$. Then*

$$\sup_{x,y\in U}\frac{|u(x)-u(y)|}{|x-y|} = \sup_{x\in U, y\in\partial U}\frac{|u(x)-u(y)|}{|x-y|}. \tag{4.63}$$

Proof. Let $x_1, x_2 \in U$ with $x_1 \neq x_2$. Define $\tau = x_2 - x_1$,

$$u_\tau(x) = u(x+\tau), \quad \text{and} \quad U_\tau = \{x\in\mathbb{R}^d : x+\tau\in U\}.$$

Both u and u_τ are sub- and supersolutions in $U\cap U_\tau$. By Corollary 4.39 there exists $z\in\partial(U\cap U_\tau)$ such that

$$u(x_1)-u(x_2) = u(x_1)-u_\tau(x_1) \le u(z)-u_\tau(z) = u(z)-u(z+\tau).$$

Note that $z, z+\tau\in\overline{U}$, and either $z\in\partial U$ or $z+\tau\in\partial U$. Therefore

$$\frac{u(x_1)-u(x_2)}{|x_2-x_1|} \le \sup_{x\in U, y\in\partial U}\frac{|u(x)-u(y)|}{|x-y|}.$$

The opposite inequality is obtained similarly. Since $x_1, x_2\in U$ are arbitrary, the proof is completed. □

Lemma 4.40 reduces the problem of *a priori* estimates to estimates near the boundary (or, rather, relative to boundary points). To complete the *a priori* estimates, we need some further assumptions on ∂U and g. The simplest assumption is the *bounded slope condition*, which is stated as follows:

(BSC) A function $u\in\mathcal{A}$ satisfies the *bounded slope condition* (BSC) if there exists $m>0$ such that for every $x_0\in\partial U$ there exist affine functions w, v with $|Dv|\le m$ and $|Dw|\le m$ such that

(a) $v(x_0) = w(x_0) = g(x_0)$, and

(b) $v(x) \leq g(x)$, $w(x) \geq g(x)$ for every $x \in \partial U$.

The bounded slope condition is rather restrictive, and requires U to be convex. Nonetheless, we can obtain the following *a priori* estimates.

Theorem 4.41. *Let L be strictly convex and assume g satisfies the bounded slope condition* **(BSC)** *for a constant $m > 0$. Then any minimizer $u \in \mathcal{A}$ of I over $\mathcal{A}$ satisfies Lip$(u) \leq m$.*

Proof. Let $x_0 \in \partial U$. As in the proof of Corollary 4.39, any function with constant gradient is both a super and subsolution. Therefore, the affine functions v and w in the definition of the bounded slope condition are sub- and supersolutions, respectively, and $v \leq u \leq w$ on ∂U. By the comparison principle (Theorem 4.37) we have $v \leq u \leq w$ in U, and so

$$u(x) - u(x_0) \leq w(x) - w(x_0) \leq m|x - x_0|$$

and

$$u(x) - u(x_0) \geq v(x) - v(x_0) \geq -m|x - x_0|.$$

Therefore

$$\frac{|u(x) - u(x_0)|}{|x - x_0|} \leq m,$$

for all $x \in U$, and so

$$\sup_{x \in U, y \in \partial U} \frac{|u(x) - u(y)|}{|x - y|} \leq m.$$

Combining this with Lemma 4.40 completes the proof. □

Finally, we prove existence and uniqueness of minimizers over $\mathcal{A}$.

Theorem 4.42. *Let L be strictly convex and assume g satisfies the bounded slope condition* **(BSC)** *for a constant $m > 0$. Then there is a unique $u \in \mathcal{A}$ such that $I(u) = \min_{w \in \mathcal{A}} I(w)$ and furthermore Lip$(u) \leq m$.*

Proof. By Theorem 4.35 there exists $u \in \mathcal{A}_{m+1}$ such that

$$I(u) = \min_{w \in \mathcal{A}_{m+1}} I(w).$$

By Remark 4.38, Theorems 4.37 and 4.41 apply to minimizers in $\mathcal{A}_{m+1}$ with similar arguments, and so we find that $\mathrm{Lip}(u) \leq m$. Let $w \in \mathcal{A}$. Then for small enough $t > 0$ we have that

$$(1 - t)u + tw \in \mathcal{A}_{m+1}$$

and so using convexity of L we have

$$I(u) \leq I((1 - t)u + tw) = \int_U L((1 - t)\nabla u + t\nabla w)\, dx \leq (1 - t)I(u) + tI(w).$$

Therefore $I(u) \leq I(w)$, which shows that u minimizes I over $\mathcal{A}$.

If $w \in \mathcal{A}$ is any other minimizer then the maximum principle (Corollary 4.39) gives that

$$\max_{U} |u - v| \leq \max_{\partial U} |u - v| = 0$$

since $u = v = g$ on ∂U. Therefore the minimizer u is unique. □

Chapter 5

Discrete to continuum in graph-based learning

Machine learning algorithms learn relationships between data and labels, which can be used to automate everyday tasks that were once difficult for computers to perform, such as image annotation or facial recognition. To describe the setup mathematically, we denote by $\mathcal{X}$ the space in which our data lives, which is often $\mathbb{R}^d$, and by $\mathcal{Y}$ the label space, which may be $\mathbb{R}^k$. Machine learning algorithms are roughly split into three categories: (i) fully supervised, (ii) semi-supervised, and (iii) unsupervised. *Fully supervised* algorithms use labeled training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ to *learn* a function $f : \mathcal{X} \to \mathcal{Y}$ that appropriately generalizes the rule $x_i \mapsto y_i$. The learned function $f(x)$ *generalizes well* if it gives the correct label for datapoints $x \in \mathcal{X}$ that it was not trained on (e.g., the testing or validation set). For a computer vision problem, each datapoint x_i would be a digital image, and the corresponding label y_i could, for example, indicate the class the image belongs to, what type of objects are in the image, or encode a caption for the image. Convolutional neural networks trained to classify images [37] is a modern example of fully supervised learning, though convolutional nets are used in semi-supervised and unsupervised settings as well. *Unsupervised* learning algorithms, such as clustering [36] or autoencoders [24], make use of only the data $x_1, \ldots, x_n$ and do not assume access to any label information.

An important and active field of current research is *semi-supervised learning*, which refers to algorithms that learn from both labeled and unlabeled data. Here, we are given some labeled data $(x_1, y_1), \ldots, (x_m, y_m)$ and some unlabeled data $x_{m+1}, \ldots, x_n$, where usually $m \ll n$, and the task is to use properties of the unlabeled data to obtain enhanced learning outcomes compared to using labeled data alone. In many applications, such as medical image analysis or speech recognition, unlabeled data is abundant, while labeled data is costly to obtain. Thus, it is important to have efficient and reliable semi-supervised learning algorithms that are able to properly utilize the ever-increasing amounts of unlabeled data that is available for learning tasks.

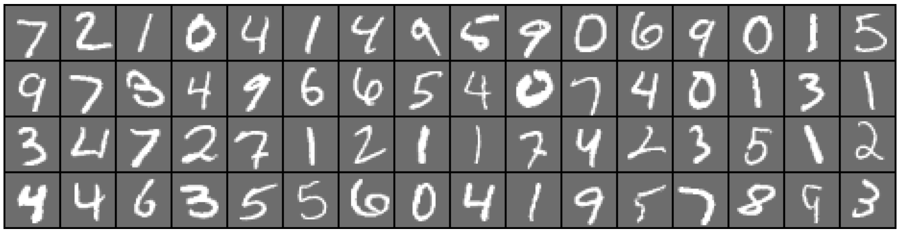


Figure 5.1: Example of some of the handwritten digits from the MNIST dataset [40].

In both the unsupervised and semi-supervised settings, data is often modeled as a graph that encodes similarities between nearest neighbors. To describe the setup mathematically, let $\mathcal{G} = (\mathcal{X}, \mathcal{W})$ denote a graph with vertices (datapoints) $\mathcal{X}$ and nonnegative edge weights $\mathcal{W} = (w_{xy})_{x,y\in\mathcal{X}}$, which encode the *similarity* of pairs of datapoints. The edge weight w_{xy} is chosen to be small or zero when x and y are dissimilar data points, and chosen to be large (e.g., $w_{xy} = 1$) when the data points are similar. In these notes, we always assume the graph is symmetric, so $w_{xy} = w_{yx}$. Nonsymmetric (also called directed) graphs are also very important, modeling connections between webpages and link prediction, but there is less theory and understanding in the directed case. In a multi-label classification problem with k classes, the label space is usually chosen as $\mathcal{Y} = \{e_1, e_2, \dots, e_k\}$, where $e_i \in \mathbb{R}^k$ is the i^{th} standard basis vector in $\mathbb{R}^k$ and represents the i^{th} class. We are given labels $g : \Gamma \to \mathcal{Y}$ on a subset $\Gamma \subset \mathcal{X}$ of the graph, and the task of graph-based semi-supervised learning is to extend the labels to the rest of the vertices $\mathcal{X} \setminus \Gamma$. In an unsupervised clustering problem, one is tasked with grouping the vertices in some sensible way so that the edge weights w_{xy} are large when x and y are in the same cluster, and small otherwise.

In semi-supervised learning, the guiding principle is the *semi-supervised smoothness assumption* [21], which stipulates that similar vertices in dense regions of the graph should have similar labels. In practice, this is often enforced by solving the optimization problem of the form

$$\left\{\begin{aligned}&\text{Minimize } \mathcal{E}(u) \text{ over } u : \mathcal{X} \to \mathbb{R}^k\\ &\text{subject to } u = g \text{ on } \Gamma,\end{aligned}\right. \tag{5.1}$$

where $\mathcal{E}$ is a functional that measures the smoothness of a potential labeling function $u : \mathcal{X} \to \mathbb{R}^k$, and $u(x) = (u_1(x), \dots, u_k(x))$. The learned label $\ell(x)$ for $x \in \mathcal{X} \setminus \Gamma$ is computed by solving (5.1) and setting

$$\ell(x) = \operatorname*{arg\,max}_{i\in\{1,\dots,k\}} u_i(x). \tag{5.2}$$

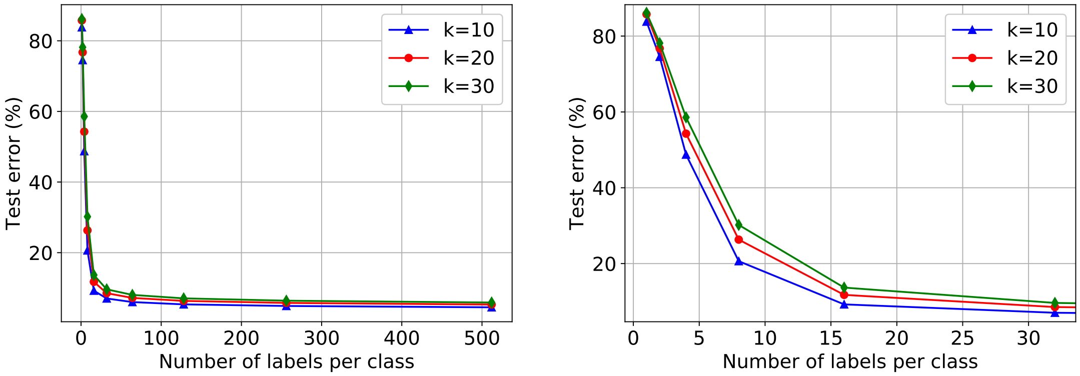


Figure 5.2: Error plots for MNIST experiment showing testing error versus number of labels, averaged over 100 trials.

One of the most widely used smoothness functionals [73] is the graph Dirichlet energy

$$\mathcal{E}_2(u) := \frac{1}{2}\sum_{x,y\in\mathcal{X}} w_{xy}|u(y)-u(x)|^2. \tag{5.3}$$

When $\mathcal{E} = \mathcal{E}_2$ in (5.1), the approach is called *Laplacian regularization*. Laplacian regularization was originally proposed in [73], and has been widely used in semi-supervised learning [1, 68, 69, 71] and manifold ranking [32, 33, 62, 64, 65, 72], among many other problems.

Let us give an example application to image classification. We use the MNIST dataset, which consists of 70,000 grayscale 28x28 pixel images of handwritten digits 0–9 [40]. See Figure 5.1 for some examples of MNIST digits. The graph is constructed as a symmetrized k-nearest neighbor graph using Euclidean distance between images to define the weights. We used 10 different labeling rates, labeling $m = 1, 2, 4, 8, 16, 32, 64, 128,$ $256, 512$ images *per class*, which correspond to labeling rates of 0.014% up to 7.3%. The test error plots are shown in Figure 5.2 for $k = 10, 20, 30$ nearest neighbors. The code for this example is part of the Python GraphLearning package available here: `https://github.com/jwcalder/GraphLearning`.

There are many variants on Laplacian regularization. Belkin and Niyogi [5, 6] proposed the problem

$$\min_{u:\mathcal{X}\to\mathbb{R}^k}\left\{\sum_{x\in\Gamma}(u(x)-g(x))^2+\lambda\mathcal{E}_2(u)\right\},$$

while Zhou et al. [67–69] proposed

$$\min_{u:\mathcal{X}\to\mathbb{R}^k}\left\{\sum_{x\in\mathcal{X}}(u(x)-g(x))^2+\lambda\sum_{x,y\in\mathcal{X}} w_{xy}\left(\frac{u(x)}{\sqrt{d(x)}}-\frac{u(y)}{\sqrt{d(y)}}\right)\right\},$$

where $d(x) = \sum_{y\in\mathcal{X}} w_{xy}$ is the *degree* of vertex x and we take $g(x) = 0$ if $x \not\in \Gamma$. Both of the models above allow for uncertainty in the labels by allowing for $u(x) \neq g(x)$ at a labeled point $x \in \Gamma$, which can be desirable if the labels are noisy or adversarial.

Recently, p-Laplace regularization has been employed, which uses $\mathcal{E} = \mathcal{E}_p$, where

$$\mathcal{E}_p(u) = \frac{1}{2}\sum_{x,y\in\mathcal{X}} w_{xy}|u(y) - u(x)|^p.$$

The p-Laplace problem was introduced originally in [70] for semi-supervised learning. It was proposed more recently in [26] with large p as a superior regularizer in problems with very few labels where Laplacian regularization gives poor results. The well-posedness of the p-Laplacian models with very few labels was proved rigorously in [55] for the variational graph p-Laplacian, and in [11] for the game-theoretic graph p-Laplacian. In addition, in [11] discrete regularity properties of the graph p-harmonic functions are proved, and the $p = \infty$ case, called Lipschitz learning [38], is studied in [13]. Efficient algorithms for solving these problems are developed in recent work [29]. We also note there are alternative re-weighting approaches that have been successful for problems with low label rates [15, 53].

Graph-based techniques area also widely used in unsupervised learning algorithms, such as clustering. The binary clustering problem is to split $\mathcal{X}$ into two groups in such a way that similar data points belong to the same group. Let $A \subset \mathcal{X}$ be one group and $A^c := \mathcal{X} \setminus A$ be the other. One approach to clustering is to minimize the *graph cut* energy

$$\mathrm{MC}(A) = \mathrm{Cut}(A, A^c), \tag{5.4}$$

where

$$\mathrm{Cut}(A, B) = \sum_{x\in A, y\in B} w_{xy}. \tag{5.5}$$

The graph cut energy $\mathrm{MC}(A)$ is the sum of the weights of edges that have to be cut to partition the graph into A and A^c. Minimizing the graph cut energy is known as the *min-cut* problem and the solution can be obtained efficiently. However, the min-cut problem often yields poor clusterings, since the energy is often minimized with $A = \{x\}$ for some $x \in \mathcal{X}$, which is usually undesirable.

This can be addressed by normalizing the graph cut energy by the size of the clusters. One approach is the *normalized cut* [52] energy

$$\mathrm{NC}(A) = \frac{\mathrm{Cut}(A, A^c)}{V(A)V(A^c)}, \tag{5.6}$$

where $V(A) = \mathrm{Cut}(A, \mathcal{X})$ measures the size of the cluster A. There are many other normalizations one can use, such as $\min\{V(A), V(A^c)\}$ [2]. Minimizing the normalized cut energy $\mathrm{NC}(A)$ turns out to be an NP-hard problem.

To address hard computational problems, we often seek ways to approximate the hard problem with problems that are easier to solve. To this end, let

$$\mathbb{1}_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{otherwise,} \end{cases}$$

and note that

$$\mathrm{Cut}(A, A^c) = \frac{1}{2} \sum_{x,y \in \mathcal{X}} w_{xy}(\mathbb{1}_A(x) - \mathbb{1}_A(y))^2 = \mathcal{E}_2(\mathbb{1}_A).$$

Likewise, we have that

$$V(A) = \sum_{x,y \in \mathcal{X}} w_{xy} \mathbb{1}_A(x) = \sum_{x \in \mathcal{X}} d(x) \mathbb{1}_A(x).$$

Therefore,

$$V(A)V(A^c) = \sum_{x,y \in \mathcal{X}} d(x)d(y)\mathbb{1}_A(x)(1 - \mathbb{1}_A(y)),$$

and the normalized cut problem can be written as

$$\min_{A \subset \mathcal{X}} \frac{\sum_{x,y\in\mathcal{X}} w_{xy}(\mathbb{1}_A(x) - \mathbb{1}_A(y))^2}{\sum_{x,y\in\mathcal{X}} d(x)d(y)\mathbb{1}_A(x)(1 - \mathbb{1}_A(y))}.$$

So far, we have changed nothing besides notation, so the problem is still NP-hard. However, this reformulation allows us to view the problem as optimizing over binary functions, in the form

$$\min_{u:\mathcal{X}\to\{0,1\}} \frac{\sum_{x,y\in\mathcal{X}} w_{xy}(u(x) - u(y))^2}{\sum_{x,y\in\mathcal{X}} d(x)d(y)u(x)(1 - u(y))}, \tag{5.7}$$

since binary functions are exactly the characteristic functions. Later we will cover relaxations of (5.7), which essentially proceed by allowing u to take on all real values $\mathbb{R}$, but this must be done carefully. These relaxations lead to spectral methods for dimension reduction and clustering, such as spectral clustering [46] and Laplacian eigenmaps [4].

We remark that all of the methods above can be viewed as discrete calculus of variations problems on graphs. A natural question to ask is what relationship they have with their continuous counterparts. Such relationships can give us insights about how much labeled data is necessary to get accurate classifiers, and which algorithms and hyperparameters are appropriate in different situations. Answering these question requires a modeling assumption on our data. A standard assumption is that the data points are a sequence of independent and identically distributed random variables and the graph is constructed as a *random geometric graph*. The next few sections are a review of calculus on graphs, and some basic probability, so that we can address the random geometric graph model rigorously.

5.1 Calculus on graphs

We now give a basic review of calculus on graphs, to make the discrete variational problems look more like their continuous counterparts. Let $\mathcal{G} = (\mathcal{X}, \mathcal{W})$ be a symmetric graph, which means $w_{xy} = w_{yx}$. We recall the weights are nonnegative and $w_{xy} = 0$ indicates there is no edge between x and y. We also recall the degree is given by $d(x) = \sum_{y\in\mathcal{X}} w_{xy}$.

Let $\ell^2(\mathcal{X})$ denote the space of functions $u : \mathcal{X} \to \mathbb{R}$ equipped with the inner product

$$(u, v)_{\ell^2(\mathcal{X})} = \sum_{x\in\mathcal{X}} u(x)v(x). \tag{5.8}$$

This induces a norm $\|u\|^2_{\ell^2(\mathcal{X})} = (u, u)_{\ell^2(\mathcal{X})}$. We define a *vector field* on the graph to be an *antisymmetric* function $V : \mathcal{X}^2 \to \mathbb{R}$ (i.e., $V(x, y) = -V(y, x)$). Thus, vector fields are defined along edges of the graph. The *gradient* of a function $u \in \ell^2(\mathcal{X})$, denoted for simplicity as ∇u, is the vector field

$$\nabla u(x, y) = u(y) - u(x). \tag{5.9}$$

We define an inner product between vector fields V and W as

$$(V, W)_{\ell^2(\mathcal{X}^2)} = \frac{1}{2} \sum_{x,y\in\mathcal{X}} w_{xy} V(x, y)W(x, y). \tag{5.10}$$

This again induces a norm $\|V\|^2_{\ell^2(\mathcal{X}^2)} = (V, V)_{\ell^2(\mathcal{X}^2)}$

The *graph divergence* is an operator taking vector fields to functions in $\ell^2(\mathcal{X})$, and is defined as the negative adjoint of the gradient. That is, for a vector field $V : \mathcal{X}^2 \to \mathbb{R}$, the graph divergence $\mathrm{div}V : \mathcal{X} \to \mathbb{R}$ is defined so that the identity

$$(\nabla u, V)_{\ell^2(\mathcal{X}^2)} = -(u, \mathrm{div}V)_{\ell^2(\mathcal{X})} \tag{5.11}$$

holds for all $u \in \ell^2(\mathcal{X})$. Compare this with the continuous Divergence Theorem (Theorem A.32). To find an expression for $\mathrm{div}V(x)$, we compute

$$\begin{aligned}
(\nabla u, V)_{\ell^2(\mathcal{X}^2)} &= \frac{1}{2} \sum_{x,y\in\mathcal{X}} w_{xy}(u(y) - u(x))V(x, y) \\
&= \frac{1}{2} \sum_{x,y\in\mathcal{X}} w_{xy}u(y)V(x, y) - \frac{1}{2} \sum_{x,y\in\mathcal{X}} w_{xy}u(x)V(x, y) \\
&= \frac{1}{2} \sum_{x,y\in\mathcal{X}} w_{xy}u(x)V(y, x) - \frac{1}{2} \sum_{x,y\in\mathcal{X}} w_{xy}u(x)V(x, y) \\
&= -\sum_{x,y\in\mathcal{X}} w_{xy}u(x)V(x, y),
\end{aligned}$$

where we used the anti-symmetry of V in the last line. Therefore we have

$$(\nabla u, V)_{\ell^2(\mathcal{X}^2)} = -\sum_{x\in\mathcal{X}}\left(\sum_{y\in\mathcal{X}} w_{xy}V(x,y)\right)u(x).$$

This suggests we should define the divergence of a vector field V to be

$$\text{div}V(x) = \sum_{y\in\mathcal{X}} w_{xy}V(x,y). \tag{5.12}$$

Using this definition, the integration by parts formula (5.11) holds, and div is the negative adjoint of ∇.

5.1.1 The Graph Laplacian and maximum principle

The graph Laplacian $\mathcal{L}u$ of a function $u \in \ell^2(\mathcal{X})$ is the composition of gradient and divergence. That is

$$\mathcal{L}u(x) = \text{div}(\nabla u)(x) = \sum_{y\in\mathcal{X}} w_{xy}(u(y)-u(x)). \tag{5.13}$$

Using (5.11) we have

$$(-\mathcal{L}u, v)_{\ell^2(\mathcal{X})} = (-\text{div}\nabla u, v)_{\ell^2(\mathcal{X})} = (\nabla u, \nabla v)_{\ell^2(\mathcal{X}^2)}. \tag{5.14}$$

In particular

$$(\mathcal{L}u, v)_{\ell^2(\mathcal{X})} = (u, \mathcal{L}v)_{\ell^2(\mathcal{X})}, \tag{5.15}$$

and so the graph Laplacian $\mathcal{L}$ is self-adjoint as an operator $\mathcal{L} : \ell^2(\mathcal{X}) \to \ell^2(\mathcal{X})$. We also note that (5.14) implies

$$(-\mathcal{L}u, u)_{\ell^2(\mathcal{X})} = (\nabla u, \nabla u)_{\ell^2(\mathcal{X}^2)} = \mathcal{E}_2(u) \geq 0, \tag{5.16}$$

which shows that the graph Dirichlet energy $\mathcal{E}_2(u)$, defined in (5.3), is closely related to the graph Laplacian. Eq. (5.16) also shows that $-\mathcal{L}$ is a positive semi-definite operator.

Now, we consider the Laplacian learning problem

$$\min_{u\in\mathcal{A}} \mathcal{E}_2(u), \tag{5.17}$$

where

$$\mathcal{A} = \{u \in \ell^2(\mathcal{X}) \,:\, u(x) = g(x) \text{ for all } x \in \Gamma\},$$

where $\Gamma \subset \mathcal{X}$ and $g : \Gamma \to \mathbb{R}$ are given.

Lemma 5.1 (Existence)**.** *There exists $u \in \mathcal{A}$ such that $\mathcal{E}_2(u) \leq \mathcal{E}_2(w)$ for all $w \in \mathcal{A}$, and*

$$\min_{x\in\Gamma} g(x) \leq u \leq \max_{x\in\Gamma} g(x). \tag{5.18}$$

Proof. First, we note that $\mathcal{E}_2$ decreases by truncation, that is if $u \in \mathcal{A}$ and

$$w(x) := \min\left\{\max\left\{u(x), a\right\}, b\right\}$$

for any $a < b$, then $\mathcal{E}_2(w) \leq \mathcal{E}_2(u)$. Indeed, this follows from

$$|w(x) - w(y)| \leq |u(x) - u(y)|,$$

which holds for all $x, y \in \mathcal{X}$. By setting $a = \min_{x\in\Gamma} g(x)$ and $b = \max_{x\in\Gamma} g(x)$, we have that $w \in \mathcal{A}$ provided $u \in \mathcal{A}$. Therefore, we may look for a minimizer of $\mathcal{E}_2$ over the constrained set

$$\mathcal{B} := \{u \in \mathcal{A} \,:\, \min_{x\in\Gamma} g(x) \leq u \leq \max_{x\in\Gamma} g(x)\}.$$

Since $\mathcal{E}_2(u) \geq 0$ for all $u \in \ell^2(\mathcal{X})$, we have $\inf_{w\in\mathcal{B}} \mathcal{E}_2(w) \geq 0$. For every integer $k \geq 1$, let $u_k \in \mathcal{B}$ such that $\mathcal{E}_2(u_k) \leq \inf_{w\in\mathcal{A}} \mathcal{E}_2(w) + 1/k$. Then u_k is a minimizing sequence, meaning that $\lim_{k\to\infty} \mathcal{E}_2(u_k) = \inf_{w\in\mathcal{A}} \mathcal{E}_2(w)$. By the Bolzano-Weierstrauss Theorem, there exists a subsequence u_{k_j} and $u \in \mathcal{B}$ such that $\lim_{j\to\infty} u_{k_j}(x) = u(x)$ for all $x \in \mathcal{X}$. Since $\mathcal{E}_2(u)$ is a continuous function of $u(x)$ for all $x \in \mathcal{X}$, we have

$$\mathcal{E}_2(u) = \lim_{j\to\infty} \mathcal{E}_2(u_{k_j}) = \inf_{w\in\mathcal{A}} \mathcal{E}_2(w),$$

which completes the proof. □

We now turn to necessary conditions. To derive the Euler-Lagrange equation that u satisfies, let $v \in \ell^2(\mathcal{X})$ such that $v = 0$ on Γ. Then $u + tv \in \mathcal{A}$ for all t, and we compute, as usual, the variation

$$\begin{aligned}
0 &= \frac{d}{dt}\Big|_{t=0} \mathcal{E}_2(u + tv) \\
&= \frac{d}{dt}\Big|_{t=0} (-\mathcal{L}(u + tv), u + tv)_{\ell^2(\mathcal{X})} \\
&= \frac{d}{dt}\Big|_{t=0} \left((-\mathcal{L}u, u)_{\ell^2(\mathcal{X})} + t(-\mathcal{L}u, v)_{\ell^2(\mathcal{X})} + t(u, -\mathcal{L}v)_{\ell^2(\mathcal{X})} + t^2(-\mathcal{L}v, v)_{\ell^2(\mathcal{X})}\right) \\
&= (-\mathcal{L}u, v)_{\ell^2(\mathcal{X})} + (u, -\mathcal{L}v)_{\ell^2(\mathcal{X})} = 2(-\mathcal{L}u, v)_{\ell^2(\mathcal{X})},
\end{aligned}$$

where we used the fact that $\mathcal{L}$ is self-adjoint (Eq. (5.15)) in the last line. Therefore, any minimizer u of (5.17) must satisfy

$$(\mathcal{L}u, v)_{\ell^2(\mathcal{X})} = 0 \quad \text{for all } v \in \ell^2(\mathcal{X}) \text{ with } v = 0 \text{ on } \Gamma.$$

We now choose

$$v(x) = \begin{cases} \mathcal{L}u(x), & \text{if } x \in \mathcal{X} \setminus \Gamma \\ 0, & \text{if } x \in \Gamma, \end{cases}$$

to find that

$$0 = (\mathcal{L}u, v)_{\ell^2(\mathcal{X})} = \sum_{x \in \mathcal{X} \setminus \Gamma} |\mathcal{L}u(x)|^2.$$

Therefore $\mathcal{L}u(x) = 0$ for all $x \in \mathcal{X} \setminus \Gamma$. Thus, any minimizer $u \in \ell^2(\mathcal{X})$ of (5.17) satisfies the boundary value problem (or Euler-Lagrange equation)

$$\begin{cases} \mathcal{L}u = 0 & \text{in } \mathcal{X} \setminus \Gamma, \\ \quad u = g & \text{on } \Gamma. \end{cases} \tag{5.19}$$

It is interesting to note that $\mathcal{L}u(x) = 0$ can be expressed as

$$0 = \sum_{y \in \mathcal{X}} w_{xy}(u(y) - u(x)) = \sum_{y \in \mathcal{X}} w_{xy}u(y) - d(x)u(x),$$

and so $\mathcal{L}u(x) = 0$ is equivalent to

$$u(x) = \frac{1}{d(x)} \sum_{y \in \mathcal{X}} w_{xy}u(y). \tag{5.20}$$

Eq. (5.20) is a *mean value property* on the graph, and says that any graph harmonic function (i.e., $\mathcal{L}u = 0$) is equal to its average value over graph neighbors.

Uniqueness of solutions of the boundary value problem (5.19) follow from the following maximum principle.

Theorem 5.2 (Maximum Principle). *Let $u \in \ell^2(\mathcal{X})$ such that $\mathcal{L}u(x) \geq 0$ for all $x \in \mathcal{X} \setminus \Gamma$. If the graph $\mathcal{G} = (\mathcal{X}, \mathcal{W})$ is connected to Γ, then*

$$\max_{x \in \mathcal{X}} u(x) = \max_{x \in \Gamma} u(x). \tag{5.21}$$

In the theorem, we say the graph is *connected to* Γ if for every $x \in \mathcal{X} \setminus \Gamma$, there exists $y \in \Gamma$ and a sequence of points $x = x_0, x_1, x_2, \ldots, x_m = y$ such that $x_i \in \mathcal{X}$ and $w_{x_i, x_{i+1}} > 0$ for all i.

Proof. We first note that $\mathcal{L}u(x) \geq 0$ is equivalent to

$$u(x) \leq \frac{1}{d(x)} \sum_{y \in \mathcal{X}} w_{xy}u(y), \tag{5.22}$$

by an argument similar to the derivation of (5.20).

Let $x_0 \in \mathcal{X}$ such that $u(x_0) = \max_{x\in\mathcal{X}} u(x)$. We may assume that $x_0 \in \mathcal{X}\setminus\Gamma$, or else (5.21) is immediate. By (5.22) and the fact that $u(y) \leq u(x_0)$ for all y we have

$$u(x_0) \leq \frac{1}{d(x_0)}\sum_{y\in\mathcal{X}} w_{x_0y}u(y) \leq \frac{1}{d(x_0)}\sum_{y\in\mathcal{X}} w_{x_0y}u(x_0) = u(x_0).$$

Therefore, we must have equality throughout, and so

$$\sum_{y\in\mathcal{X}} w_{x_0y}(u(x_0)-u(y)) = 0.$$

Since $w_{x_0y}(u(x_0)-u(y)) \geq 0$, we conclude that

$$w_{x_0y}(u(x_0)-u(y)) = 0 \quad \text{for all } y\in\mathcal{X}.$$

In particular, we have

$$u(x_0) = u(y) \quad \text{for all } y\in\mathcal{X} \text{ such that } w_{x_0y} > 0. \tag{5.23}$$

Since the graph is connected to Γ, there exists $y\in\Gamma$ and a sequence of points $x_0, x_1, x_2, \dots, x_m = y$ such that $x_i \in \mathcal{X}$ and $w_{x_i,x_{i+1}} > 0$ for all i. By (5.23) we have $u(x_0) = u(x_1) = \max_{x\in\mathcal{X}} u(x)$. But now the argument above can be applied to x_1 to obtain $u(x_0) = u(x_1) = u(x_2) = \max_{x\in\mathcal{X}} u(x)$. Continuing by induction we see that

$$\max_{x\in\mathcal{X}} u(x) = u(x_0) = u(x_m) = u(y) \leq \max_{x\in\Gamma} u(x).$$

This completes the proof. □

Uniqueness of solutions to (5.19) is an immediate consequence of the following corollary.

Corollary 5.3. *Let $u, v \in \ell^2(\mathcal{X})$ such that $\mathcal{L}u(x) = \mathcal{L}v(x) = 0$ for all $x\in\mathcal{X}\setminus\Gamma$. If the graph $\mathcal{G} = (\mathcal{X},\mathcal{W})$ is connected to Γ, then*

$$\max_{x\in\mathcal{X}} |u(x)-v(x)| = \max_{x\in\Gamma} |u(x)-v(x)|. \tag{5.24}$$

In particular, if $u = v$ on Γ, then $u \equiv v$.

Proof. Since $\mathcal{L}(u-v)(x) = 0$ for $x\in\mathcal{X}\setminus\Gamma$, we have by Theorem 5.2 that

$$\max_{x\in\mathcal{X}}(u(x)-v(x)) = \max_{x\in\Gamma}(u(x)-v(x)) \leq \max_{x\in\Gamma}|u(x)-v(x)|$$

and

$$\max_{x\in\mathcal{X}}(v(x)-u(x)) = \max_{x\in\Gamma}(v(x)-u(x)) \leq \max_{x\in\Gamma}|u(x)-v(x)|,$$

from which we immediately conclude (5.24). □

For use later on, we record another version of the maximum principle that does not require connectivity of the graph.

Lemma 5.4 (Maximum principle)**.** *Let $u \in \ell^2(\mathcal{X})$ such that $\mathcal{L}u(x) > 0$ for all $x \in \mathcal{X} \setminus \Gamma$. Then*

$$\max_{x\in\mathcal{X}} u(x) = \max_{x\in\Gamma} u(x). \tag{5.25}$$

Proof. Let $x_0 \in \mathcal{X}$ such that $u(x_0) = \max_{x\in\mathcal{X}} u(x)$. Since $u(x_0) \geq u(y)$ for all $y \in \mathcal{X}$, we have

$$\mathcal{L}u(x_0) = \sum_{y\in\mathcal{X}} w_{xy}(u(y) - u(x_0)) \leq 0.$$

Since $\mathcal{L}u(x) > 0$ for all $x \in \mathcal{X} \setminus \Gamma$, we must have $x_0 \in \Gamma$, which completes the proof. □

5.2 Concentration of measure

As we will be working with random geometric graphs, we will require some basic probabilistic estimates, referred to as concentration of measure, to control the random behavior of the graph. In this section, we review some basic, and very useful, concentration of measure results. It is a good idea to review the Section A.9 for a review of basic probability before reading this section.

Let $X_1, X_2, \ldots, X_n$ be a sequence of n independent and identically distributed real-valued random variables and let $S_n = \frac{1}{n}\sum_{i=1}^n X_i$. In Section A.9.4 we saw how to use Chebyshev's inequality to obtain bounds of the form

$$\mathbb{P}(|S_n - \mu| \geq t) \leq \frac{\sigma^2}{nt^2} \tag{5.26}$$

for any $t > 0$, where $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathrm{Var}\,(X_i)$. Without further assumptions on the random variables X_i, these estimates are essentially tight. However, if the random variables X_i are almost surely bounded (i.e., $\mathbb{P}(|X_i| \leq b) = 1$ for some $b > 0$), which is often the case in practical applications, then we can obtain far sharper exponential bounds.

To see what to expect, we note that the Central Limit Theorem says (roughly) that

$$S_n = \mu + \frac{1}{\sqrt{n}} N(0,\sigma^2) + o\left(\frac{1}{\sqrt{n}}\right) \quad \text{as } n \to \infty$$

where $N(0,\sigma^2)$ represents a normally distributed random variable with mean zero and variance σ^2. Ignoring error terms, this says that $Y_n := \sqrt{n}(S_n - \mu)$ is approximately $N(0,\sigma^2)$, and so we may expect Gaussian-like estimates of the form

$$\mathbb{P}(|Y_n| \geq x) \leq C \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

for $x > 0$. Setting $x = \sqrt{n}t$ we can rewrite this as

$$\mathbb{P}(|S_n - \mu| \geq t) \leq C \exp\left(-\frac{nt^2}{2\sigma^2}\right) \tag{5.27}$$

for any $t > 0$. Bounds of the form (5.26) and (5.27) are called *concentration inequalities*, or *concentration of measure*. In this section we describe the main ideas for proving exponential bounds of the form (5.27), and prove the Hoeffding and Bernstein inequalities, which are some of the most useful concentration inequalities. For more details we refer the reader to [8].

One normally proves exponential bounds like (5.27) with the Chernoff bounding trick. Let $s > 0$ and note that

$$\mathbb{P}(S_n - \mu \geq t) = \mathbb{P}(s(S_n - \mu) \geq st) = \mathbb{P}\left(e^{s(S_n-\mu)} \geq e^{st}\right).$$

The random variable $Y = e^{s(S_n-\mu)}$ is nonnegative, so we can apply Markov's inequality (see Proposition A.39) to obtain

$$\begin{aligned}\mathbb{P}(S_n - \mu \geq t) &\leq e^{-st}\mathbb{E}\left[e^{s(S_n-\mu)}\right]\\ &= e^{-st}\mathbb{E}\left[e^{\frac{s}{n}\sum_{i=1}^n (X_i-\mu)}\right]\\ &= e^{-st}\mathbb{E}\left[\prod_{i=1}^n e^{\frac{s}{n}(X_i-\mu)}\right].\end{aligned}$$

Applying (A.30) yields

$$\mathbb{P}(S_n - \mu \geq t) \leq e^{-st}\prod_{i=1}^n \mathbb{E}\left[e^{\frac{s}{n}(X_i-\mu)}\right] = e^{-st}\mathbb{E}\left[e^{\frac{s}{n}(X_1-\mu)}\right]^n. \tag{5.28}$$

This bound is the main result of the Chernoff bounding trick. The key now is to obtain bounds on the *moment generating function*

$$M_X(\lambda) := \mathbb{E}[e^{\lambda(X-\mu)}],$$

where $X = X_1$.

In the case where the X_i are Bernoulli random variables, we can compute the moment generating function explicitly, and this leads to the Chernoff bounds. Before giving them, we present some preliminary technical propositions regarding the function

$$h(\delta) = (1 + \delta)\log(1 + \delta) - \delta, \tag{5.29}$$

which appears in many concentration inequalities.

Proposition 5.5. *For any $\delta > 0$ we have*

$$\max_{x \geq 0} \{\delta x - (e^x - 1 - x)\} = h(\delta). \tag{5.30}$$

and for any $0 \leq \delta < 1$ we have

$$\max_{x \geq 0} \big\{\delta x - \big(e^{-x} - 1 + x\big)\big\} = h(-\delta). \tag{5.31}$$

Proof. Let $f(x) = \delta x - (e^x - 1 - x)$ and check that

$$f'(x) = \delta - e^x + 1 = 0$$

when $x = \log(1 + \delta)$. Since $f''(x) = -e^x$, the critical point is a maximum and we have

$$f(\log(1 + \delta)) = h(\delta),$$

which completes the proof of (5.30). The proof of (5.31) is similar. □

Proposition 5.6. *For every $\delta \geq -1$ we have*

$$h(\delta) \geq \frac{\delta^2}{2(1 + \frac{1}{3}\delta_+)}, \tag{5.32}$$

where $\delta_+ := \max\{\delta, 0\}$.

Proof. We first note that $h'(0) = h(0) = 0$ and $h''(\delta) = 1/(1 + \delta)$. We define

$$f(\delta) = \frac{\delta^2}{2(1 + \frac{1}{3}\delta)},$$

and check that $f'(0) = f(0) = 0$ and

$$f''(\delta) = \frac{1}{(1 + \frac{1}{3}\delta)^3} \leq \frac{1}{1 + \delta} = h''(\delta)$$

for all $\delta > 0$. Therefore, for $\delta > 0$ we have

$$h(\delta) = \int_0^\delta \int_0^t h''(s)\, ds\, dt \geq \int_0^\delta \int_0^t f''(s)\, ds\, dt = f(\delta).$$

For $\delta < 0$ we have that $h''(s) \geq 1$ for $\delta \leq s \leq 0$ and so

$$h(\delta) = \int_\delta^0 \int_t^0 h''(s)\, ds\, dt \geq \int_\delta^0 \int_t^0 ds\, dt = \frac{1}{2}\delta^2,$$

which completes the proof. □

We now state and prove the Chernoff bounds.

Theorem 5.7 (Chernoff bounds)**.** *Let $X_1, X_2 \dots, X_n$ be a sequence of* i.i.d. *Bernoulli random variables with parameter $p \in [0,1]$ (i.e., $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = 1-p$). Then for any $\delta > 0$ we have*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq (1+\delta)np\right) \leq \exp\left(-\frac{np\,\delta^2}{2(1+\frac{1}{3}\delta)}\right), \tag{5.33}$$

and for any $0 \leq \delta < 1$ we have

$$\mathbb{P}\left(\sum_{i=1}^n X_i \leq (1-\delta)np\right) \leq \exp\left(-\frac{1}{2}np\,\delta^2\right), \tag{5.34}$$

Proof. We can explicitly compute the moment generating function

$$\begin{aligned} M_X(\lambda) &= e^{-\lambda p}\mathbb{E}[e^{\lambda X}] = e^{-\lambda p}(pe^{\lambda} + (1-p)e^0) \\ &= e^{-\lambda p}(1 + p(e^{\lambda} - 1)) \\ &\leq \exp\left(-\lambda p + p(e^{\lambda}-1)\right), \end{aligned}$$

where we used the inequality $1 + x \leq e^x$ in the last line. Writing $S_n = \frac{1}{n}\sum_{i=1}^n X_i$ and using the Chernoff bounding method (5.28) we have

$$\mathbb{P}(S_n - p \geq t) \leq \exp\left(-st - sp + np(e^{\frac{s}{n}} - 1)\right),$$

since $\mu = \mathbb{E}[X_i] = p$. Set $t = \delta p$ and rearrange to find that

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^n X_i \geq (1+\delta)np\right) &\leq \exp\left(-s\delta p - sp + np(e^{\frac{s}{n}} - 1)\right) \\ &= \exp\left(-np\left(\delta\tfrac{s}{n} - \left(e^{\frac{s}{n}} - 1 - \tfrac{s}{n}\right)\right)\right), \end{aligned}$$

for any $s > 0$. We use (5.30) from Proposition 5.5 to optimize over $s > 0$ yielding

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq (1+\delta)np\right) \leq \exp\left(-np\,h(\delta)\right).$$

The proof of (5.33) is completed by applying Proposition 5.6.

To prove (5.34) we again use the Chernoff bounding method (5.28) to obtain

$$\mathbb{P}(S_n - p \leq -t) \leq e^{-st}\mathbb{E}\left[e^{-\frac{s}{n}(X_1 - p)}\right]^n \leq \exp\left(-st + sp + np(e^{-\frac{s}{n}} - 1)\right).$$

Set $t = \delta p$ to obtain

$$\mathbb{P}\left(\sum_{i=1}^n X_i \leq (1-\delta)np\right) \leq \exp\left(-s\delta p + sp + np(e^{-\frac{s}{n}} - 1)\right).$$

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \leq (1-\delta)np\right) \leq \exp\left(-np\left(\delta\tfrac{s}{n} - \left(e^{-\frac{s}{n}} - 1 + \tfrac{s}{n}\right)\right)\right).$$

We use (5.31) to optimize over $s > 0$ yielding

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \leq (1-\delta)np\right) \leq \exp\left(-np\, h(-\delta)\right).$$

Invoking again Proposition 5.6 completes the proof. □

Example 5.1. As an application of the Chernoff bounds, we prove that the computational complexity of quicksort is $O(n\log(n))$ with high probability. Quicksort is a randomized algorithm for sorting a list of numbers $x_1, x_2, \ldots, x_n \in \mathbb{R}$ that is one of (if not the) fastest sorting algorithms. Let $X = \{x_1, \ldots, x_n\}$. To sort X, the quicksort algorithm picks a random point $p \in X$, called the *pivot*, and splits X into $X_+(p) := X \cap (-\infty, p]$ and $X_-(p) := \cap(p, \infty]$, and recursively calls quicksort on the sets $X_+(p)$ and $X_-(p)$.

To analyze quicksort, we note that the process generates a tree, where the set X is the parent of children $X_+(p)$ and $X_-(p)$. The tree has exactly n leaves and n paths from the root X to the leaves, one for each point $x_1, \ldots, x_n$. Let Z denote the depth of the tree. Since each layer of the tree contributes n operations, the complexity of quicksort is $O(Zn)$. We call a pivot a *good* pivot if both $X_+(p)$ and $X_-(p)$ contain at least one third of the points in X, and otherwise we call p a *bad* pivot. The probability of a bad pivot is $p = 2/3$, and every good pivot reduces the size of the dataset by the factor $2/3$. Thus, along any path from the root of the tree to a leaf, there can be no more than

$$k = \frac{\log n}{\log(\frac{3}{2})} \leq 3\log n$$

good pivots, before the size of the remaining set is less than one (and $\log(3/2) \geq 1/3$). Let P denote any path from the root to a leaf, and let $|P|$ denote its length. Let Y denote the number of bad pivots along the path, and let $C > 0$ be a constant to be determined. If $|P| \geq (C+3)\log n$, then since there are at most $3\log n$ good pivots we have $Y \geq C\log n$. By the Chernoff bounds with $\delta = 1/3$ we have

$$\mathbb{P}\left(Y \geq \frac{4}{3}mp\right) \leq \exp\left(-\frac{1}{20}mp\right),$$

where m is the largest integer smaller than $(C+3)\log n$ and $p = 2/3$. Let us for simplicity assume $n \geq e$ so $\log n \geq 1$. Then we have

$$(C+2)\log n \leq m \leq (C+3)\log n.$$

Since $4p/3 = 8/9$ we have

$$\mathbb{P}\left(Y \geq \tfrac{8}{9}(C+3)\log n\right) \leq \exp\left(-\frac{1}{30}(C+2)\log n\right).$$

We now choose $C+2=60$, or $C=58$. Then we check that $\frac{8}{9}(C+3)\leq C$ and so

$$\mathbb{P}(|P|\geq 61\log n)\leq \mathbb{P}(Y\geq 58\log n)\leq \frac{1}{n^2}.$$

Since here are n paths from leaves to the root, we union bound over all paths to find that

$$\mathbb{P}(Z\geq 61\log n)\leq \frac{1}{n}.$$

Therefore, with probability at least $1-\frac{1}{n}$, quicksort takes at most $O(n\log n)$ operations to complete. △

In general, we cannot compute the moment generating function explicitly, and are left to derive upper bounds. The first bound is due to Hoeffding.

Lemma 5.8 (Hoeffding Lemma)**.** *Let X be a real-valued random variable for which $|X-\mu|\leq b$ almost surely for some $b>0$, where $\mu=\mathbb{E}[X]$. Then we have*

$$M_X(\lambda)\leq e^{\frac{\lambda^2 b^2}{2}}. \tag{5.35}$$

Proof. Since $x\mapsto e^{sx}$ is a convex function, we have

$$e^{\lambda x}\leq e^{-\lambda b}+\frac{x+b}{b}\sinh(\lambda b)$$

provided $|x|\leq b$ (the right hand side is the secant line from $(-b,e^{-\lambda b})$ to $(b,e^{\lambda b})$; recall $\sinh(t)=(e^t-e^{-t})/2$ and $\cosh(t)=(e^t+e^{-t})/2$). Therefore we have

$$\begin{aligned}M_X(\lambda)=\mathbb{E}\left[e^{\lambda(X-\mu)}\right]&\leq \mathbb{E}\left[e^{-\lambda b}+\frac{X-\mu+b}{b}\sinh(\lambda b)\right]\\&=e^{-\lambda b}+\frac{\mathbb{E}[X]-\mu+b}{b}\sinh(\lambda b)\\&=e^{-\lambda b}+\sinh(\lambda b)=\cosh(\lambda b).\end{aligned}$$

The proof is completed by the elementary inequality $\cosh(x)\leq e^{\frac{x^2}{2}}$ (compare Taylor series). □

Combining the Hoeffding Lemma with (5.28) yields

$$\mathbb{P}(S_n-\mu\geq t)\leq e^{-st}\mathbb{E}\left[e^{\frac{s}{n}(X_1-\mu)}\right]^n=\exp\left(-st+\frac{s^2b^2}{2n}\right),$$

provided $|X_i-\mu|\leq b$ almost surely. Optimizing over $s>0$ we find that $s=nt/b^2$, which yields the following result.

Theorem 5.9 (Hoeffding inequality)**.** *Let $X_1, X_2 \ldots, X_n$ be a sequence of* i.i.d. *real-valued random variables with finite expectation $\mu = \mathbb{E}[X_i]$, and write $S_n = \frac{1}{n}\sum_{i=1}^n X_i$. Assume there exists $b > 0$ such that $|X - \mu| \leq b$ almost surely. Then for any $t > 0$ we have*

$$\mathbb{P}(S_n - \mu \geq t) \leq \exp\left(-\frac{nt^2}{2b^2}\right). \tag{5.36}$$

Remark 5.10. Of course, the opposite inequality

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

holds by a similar argument. Thus, by the union bound we have

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

The Hoeffding inequality is tight if $\sigma^2 \approx b^2$, so that the right hand side looks like the Gaussian distribution in (5.27), up to constants. For example, if X_i are uniformly distributed on $[-b, b]$ then

$$\sigma^2 = \frac{1}{2b}\int_{-b}^{b} x^2\,dx = \frac{b^2}{3}.$$

However, if $\sigma^2 \ll b^2$, then one would expect to see σ^2 in place of b^2 as in (5.27), and the presence of b^2 leads to a suboptimal bound.

Example 5.2. Let

$$Y = \max\left\{1 - \frac{|X|}{\varepsilon}, 0\right\}.$$

where X is uniformly distributed on $[-1, 1]$, as above, and $\varepsilon \ll b$. Then $|Y| \leq 1$, so $b = 1$, but we compute

$$\sigma^2 \leq \frac{1}{2}\int_{-\varepsilon}^{\varepsilon} dx = \varepsilon.$$

Hence, $\sigma^2 \ll 1$ when ε is small, and we expect to get sharper concentration bounds than are provided by the Hoeffding inequality. This example is similar to what we will see later in consistency of graph Laplacians. △

The Bernstein inequality gives the sharper bounds that we desire, and follows from Bernstein's Lemma.

Lemma 5.11 (Bernstein Lemma)**.** *Let X be a real-valued random variable with finite mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = Var(X)$, and assume that $|X - \mu| \leq b$ almost surely for some $b > 0$. Then we have*

$$M_X(\lambda) \leq \exp\left(\frac{\sigma^2}{b^2}(e^{\lambda b} - 1 - \lambda b)\right). \tag{5.37}$$

Proof. Note that for $|x| \leq b$ we have

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

Therefore

$$M_X(\lambda) \leq \mathbb{E}\left[1 + \lambda(X-\mu) + \frac{(X-\mu)^2}{b^2}(e^{\lambda b} - 1 - \lambda b)\right] = 1 + \frac{\sigma^2}{b^2}(e^{\lambda b} - 1 - \lambda b).$$

The proof is completed by using the inequality $1 + x \leq e^x$. □

We now prove the Bernstein inequality.

Theorem 5.12 (Bernstein Inequality)**.** *Let $X_1, X_2 \ldots, X_n$ be a sequence of* i.i.d. *real-valued random variables with finite expectation $\mu = \mathbb{E}[X_i]$ and variance $\sigma^2 = Var(X_i)$, and write $S_n = \frac{1}{n}\sum_{i=1}^n X_i$. Assume there exists $b > 0$ such that $|X - \mu| \leq b$ almost surely. Then for any $t > 0$ we have*

$$\mathbb{P}(S_n - \mu \geq t) \leq \exp\left(-\frac{nt^2}{2(\sigma^2 + \frac{1}{3}bt)}\right). \tag{5.38}$$

Proof. Combining the Bernstein Lemma with the Chernoff bound (5.28) yields

$$\mathbb{P}(S_n - \mu \geq t) \leq e^{-st}\mathbb{E}\left[e^{\frac{s}{n}(X_1-\mu)}\right]^n \leq \exp\left(-\frac{n\sigma^2}{b^2}\left(\tfrac{bt}{\sigma^2}\cdot\tfrac{sb}{n} - \left(e^{\frac{sb}{n}} - 1 - \tfrac{sb}{n}\right)\right)\right).$$

We use (5.30) to optimize over $s > 0$ and obtain

$$\mathbb{P}(S_n - \mu \geq t) \leq \exp\left(-\frac{n\sigma^2}{b^2}h\left(\frac{bt}{\sigma^2}\right)\right). \tag{5.39}$$

The proof is completed by invoking Proposition 5.6. □

Remark 5.13. The Bernstein inequality has two distinct parameter regimes. If $bt \leq \sigma^2$, then Bernstein's inequality yields

$$\mathbb{P}(S_n - \mu \geq t) \leq \exp\left(-\frac{3nt^2}{8\sigma^2}\right),$$

which is, up to constants, the correct Gaussian tails. If $bt \geq \sigma^2$ then we have

$$\mathbb{P}(S_n - \mu \geq t) \leq \exp\left(-\frac{3nt}{8b}\right),$$

which should be compared with the Hoeffding inequality (5.36).

Remark 5.14. As with the Hoeffding inequality, we can obtain two-sided estimates of the form

$$\mathbb{P}(|S_n - \mu| \geq t) \leq 2\exp\left(-\frac{nt^2}{2(\sigma^2 + \frac{1}{3}bt)}\right).$$

Later in the notes, we will encounter double sums of the form

$$U_n = \frac{1}{n(n-1)}\sum_{i\neq j} f(X_i, X_j), \tag{5.40}$$

where $X_1, \dots, X_n$ is a sequence of *i.i.d.* random variables. The random variable U_n is called a U-statistics of order 2. We record here the Bernstein inequality for U-statistics. We assume f is symmetric, so $f(x,y) = f(y,x)$.

Theorem 5.15 (Bernstein for U-statistics)**.** *Let $X_1, \dots, X_n$ is a sequence of* i.i.d. *random variables and let $h : \mathbb{R}^2 \to \mathbb{R}$ be symmetric (i.e., $f(x,y) = f(y,x)$). Let $\mu = \mathbb{E}[f(X_i, X_j)]$, $\sigma^2 = Var(f(X_i, X_j)) = \mathbb{E}[(f(X_i,X_j) - \mu)^2]$, and $b := \|f\|_\infty$, and define the U-statistic U_n by (5.40). Then for every $t > 0$ we have*

$$\mathbb{P}(U_n - \mu \geq t) \leq \exp\left(-\frac{nt^2}{6(\sigma^2 + \frac{1}{3}bt)}\right). \tag{5.41}$$

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

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

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

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

where the last line follows from Jensen's inequality. Since Y_τ is a sum of k *i.i.d.* random variables with zero mean, absolute bound b, and σ^2 variance, the same application of Bernstein's Lemma as used in Theorem 5.12 yields

$$\mathbb{E}[e^{sY_\tau}] \leq \exp\left(\frac{k\sigma^2}{b^2}\left(e^{\frac{sb}{k}} - 1 - \frac{sb}{k}\right)\right).$$

Therefore, we obtain

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

Using (5.30) to optimize over $s > 0$ yields

$$\mathbb{P}(U_n - \mu \geq t) \leq \exp\left(-\frac{k\sigma^2}{b^2}h\left(\frac{bt}{\sigma^2}\right)\right), \tag{5.42}$$

and the proof is completed by applying Proposition 5.6 and noting that $k \geq n/3$ for all $n \geq 2$. □

We pause to give an application to Monte Carlo numerical integration, which is a randomized numerical method for approximating integrals of the form

$$I(u) = \int_{[0,1]^d} u(x)\,dx. \tag{5.43}$$

The method is based on approximating $I(u)$ by an empirical mean

$$I_n(u) = \frac{1}{n}\sum_{i=1}^n u(X_i), \tag{5.44}$$

where $X_1, X_2, X_3, \ldots, X_n$ is a sequence of *i.i.d.* random variables uniformly distributed on $[0,1]^d$. The usefulness of Monte Carlo integration is illustrated by the following error estimates.

Theorem 5.16 (Monte Carlo error estimate)**.** *Let $u : [0,1]^d \to \mathbb{R}$ be continuous and let $\sigma^2 = Var(u(X_i))$. Then for all $0 < \lambda \leq \sigma\sqrt{n}\|u\|^{-1}_{L^\infty([0,1]^d)}$ we have*

$$|I(u) - I_n(u)| \leq \frac{\lambda\sigma}{\sqrt{n}} \tag{5.45}$$

with probability at least $1 - 2\exp\left(-\frac{1}{4}\lambda^2\right)$.

Remark 5.17. Let $\delta > 0$ and choose $\lambda > 0$ so that $\delta = 2\exp\left(-\frac{1}{4}\lambda^2\right)$; that is $\lambda = \sqrt{4\log\left(\frac{2}{\delta}\right)}$. Then Theorem 5.16 can be rewritten to say that

$$|I(u) - I_n(u)| \leq \sqrt{\frac{4\sigma^2\log\left(\frac{2}{\delta}\right)}{n}} \tag{5.46}$$

holds with probability at least $1-\delta$. This is a more common form to see Monte Carlo error estimates.

The reader should contrast this with the case where X_i form a uniform grid over $[0,1]^d$. In this case, for Lipschitz functions the numerical integration error is $O(\Delta x)$, where Δx is the grid spacing. For n points on a uniform grid in $[0,1]^d$ the grid spacing is $\Delta x \sim n^{-1/d}$, which suffers from the *curse of dimensionality* when d is large. The Monte Carlo error estimate (5.46), on the other hand, is remarkable in that it is independent of dimension d! Thus, Monte Carlo integration overcomes the *curse of dimensionality* by simply replacing a uniform discretization grid by random variables. Monte Carlo based techniques have been used to solve PDEs in high dimensions via sampling random walks or Brownian motions.

Proof of Theorem 5.16. Let $Y_i = u(X_i)$. We apply Bernstein's inequality with $S_n = \frac{1}{n}\sum_{i=1}^n Y_i = I_n(u)$, $\sigma^2 = \mathrm{Var}(Y_i)$ and $b = 2\|u\|_{L^\infty([0,1]^d)}$ to find that

$$|I(u) - I_n(u)| \leq t$$

with probability at least $1 - 2\exp\left(-\frac{nt^2}{2\left(\sigma^2 + \frac{1}{3}bt\right)}\right)$ for any $t > 0$. Set $t = \lambda\sigma/\sqrt{n}$ for $\lambda > 0$ to find that

$$|I(u) - I_n(u)| \leq \frac{\lambda\sigma}{\sqrt{n}}$$

with probability at least $1 - 2\exp\left(-\frac{\sigma^2\lambda^2}{2\left(\sigma^2 + \frac{b\lambda\sigma}{3\sqrt{n}}\right)}\right)$. Restricting $\lambda \leq 3\sigma\sqrt{n}/b$ completes the proof. □

We conclude this section with the Azuma/McDiarmid inequality. This is slightly more advanced and is not used in the rest of these notes, so the reader may skip ahead without any loss. It turns out that the Chernoff bounding method used to prove the Chernoff, Hoeffding, and Bernstein inequalities does not use in any essential way the linearity of the sum defining S_n. Indeed, what matters is that S_n does not depend too much on any particular random variable X_i. Using these ideas leads to far more general (and more useful) concentration inequalities for functions of the form

$$Y_n = f(X_1, X_2, \ldots, X_n) \tag{5.47}$$

that may depend *nonlinearly* on the X_i. To express that Y_n does not depend too much on any of the X_i, we assume that f satisfies the following bounded differences condition: There exists $b > 0$ such that

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, \widetilde{x}_i, \ldots, x_n)| \leq b \tag{5.48}$$

for all $x_1, \ldots, x_n$ and $\widetilde{x}_i$. In this case we have the following concentration inequality.

Theorem 5.18 (Azuma/McDiarmid inequality)**.** *Define* Y_n *by* (5.47), *where* $X_1, \ldots, X_n$ *are* i.i.d. *random variables satisfying* $|X_i| \leq M$ *almost surely, and assume* f *satisfies* (5.48). *Then for any* $t > 0$

$$\mathbb{P}(Y_n - \mathbb{E}[Y_n] \geq t) \leq \exp\left(-\frac{t^2}{2nb^2}\right). \tag{5.49}$$

Proof. The proof uses conditional probability, which we have not developed in these notes, so we give a sketch of the proof. For $2 \leq k \leq n$ we define

$$Z_k = \mathbb{E}[Y_n \,|\, X_1, \ldots, X_k] - \mathbb{E}[Y_n \,|\, X_1, \ldots, X_{k-1}].$$

and set $Z_1 = \mathbb{E}[Y_n \,|\, X_1] - \mathbb{E}[Y_n]$. Since

$$Y_n = \mathbb{E}[Y_n \,|\, X_1, \ldots, X_n],$$

we have the telescoping sum

$$Y_n - \mathbb{E}[Y_n] = \sum_{k=1}^n Z_k.$$

The random variables Z_k record how much the conditional expectation changes when we add information about X_k. While the Z_k are not independent, they form a *martingale difference sequence*, which allows us to essentially treat them as independent and use a similar proof to Hoeffding's inequality. The useful martingale difference property is the identity

$$\mathbb{E}[Z_k \,|\, X_1, \ldots, X_{k-1}] = 0$$

for $k \geq 2$, and $\mathbb{E}[Z_1] = 0$, which follow from the law of iterated expectations.

We now follow the Chernoff bounding method and law of iterated expectations to obtain

$$\begin{aligned}
\mathbb{P}(Y_n - \mathbb{E}[Y_n] \geq t) &= \mathbb{P}(e^{s\sum_{k=1}^n Z_k} \geq e^{st}) \\
&\leq e^{-st}\mathbb{E}[e^{s\sum_{k=1}^n Z_k}] \\
&= e^{-st}\mathbb{E}[\mathbb{E}[e^{s\sum_{k=1}^n Z_k} \,|\, X_1, \ldots, X_{n-1}]] \\
&= e^{-st}\mathbb{E}[e^{s\sum_{k=1}^{n-1} Z_k}\mathbb{E}[e^{sZ_n} \,|\, X_1, \ldots, X_{n-1}]].
\end{aligned}$$

Define

$$U_k = \sup_{|x| \leq M} \mathbb{E}[f(X_1, \ldots, X_{k-1}, x, X_{k+1}, \ldots X_n) - Y_n \,|\, X_1, \ldots, X_{k-1}],$$

and

$$L_k = \inf_{|x| \leq M} \mathbb{E}[f(X_1, \ldots, X_{k-1}, x, X_{k+1}, \ldots X_n) - Y_n \,|\, X_1, \ldots, X_{k-1}].$$

Then $L_k \leq Z_k \leq U_k$. By (5.48) we have $U_k \leq b$ and $L_k \geq -b$, and so $|Z_k| \leq b$. Following a very similar argument as in the proof of Lemma 5.8 we have

$$\begin{aligned}
\mathbb{E}[e^{sZ_k} \,|\, X_1, \ldots, X_{k-1}] &\leq \mathbb{E}\left[e^{-sb} + \frac{Z_k + b}{b}\sinh(sb) \,|\, X_1, \ldots, X_{k-1}\right] \\
&= e^{-st} + \sinh(sb) = \cosh(sb) \leq e^{\frac{s^2b^2}{2}},
\end{aligned}$$

since $\mathbb{E}[Z_k \,|\, X_1, \ldots, X_{k-1}] = 0$. Inserting this above we have

$$\mathbb{P}(Y_n - \mathbb{E}[Y_n] \geq t) \leq e^{-st + \frac{s^2 b^2}{2}} \mathbb{E}[e^{s \sum_{k=1}^{n-1} Z_k}].$$

Continuing by induction we find that

$$\mathbb{P}(Y_n - \mathbb{E}[Y_n] \geq t) \leq \exp\left(-st + \frac{ns^2b^2}{2}\right).$$

Optimizing over $s > 0$ completes the proof. □

5.3 Random geometric graphs and basic setup

To prove discrete to continuum convergence of learning problems on graphs, we must make a modeling assumption for the graph. Let $X_1, X_2, \ldots, X_n$ be a sequence of *i.i.d.* random variables on an open connected domain $U \subset \mathbb{R}^d$ with a probability density $\rho : U \to [0, \infty)$. We assume the boundary ∂U is smooth, $\rho \in C^2(\overline{U})$ and there exists $\alpha > 0$ such that

$$\alpha \leq \rho(x) \leq \alpha^{-1} \quad \text{for all } x \in U. \tag{5.50}$$

The vertices of our graph are

$$\mathcal{X}_n := \{X_1, X_2, \ldots, X_n\}. \tag{5.51}$$

Let $\eta : [0, \infty) \to [0, \infty)$ be smooth and nonincreasing such that $\eta(t) \geq 1$ for $0 \leq t \leq \frac{1}{2}$, and $\eta(t) = 0$ for $t > 1$. For $\varepsilon > 0$ define $\eta_\varepsilon(t) = \frac{1}{\varepsilon^d}\eta\left(\frac{t}{\varepsilon}\right)$ and set $\sigma_\eta = \int_{\mathbb{R}^d} |z_1|^2 \eta(|z|)\, dz$. The weight w_{xy} between x and y is given by

$$w_{xy} = \eta_\varepsilon\left(|x - y|\right). \tag{5.52}$$

The graph $\mathcal{G}_{n,\varepsilon} = (\mathcal{X}_n, \mathcal{W}_{n,\varepsilon})$ where $\mathcal{W}_{n,\varepsilon} = (w_{xy})_{x,y \in \mathcal{X}_n}$ is called a *random geometric graph*. It will be more convenient notationally to write out the weights in terms of η_ε, so we will not use the w_{xy} notation in the remainder of this chapter.

Remark 5.19. It is also common in machine learning to make the *manifold assumption*, where $X_1, X_2, \ldots, X_n$ are a sequence of *i.i.d.* random variables sampled from an m-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^d$, where $m \ll d$. The rest of the construction is identical, in particular, the weights are defined using the ambient Euclidean distance in $\mathbb{R}^d$. The random variables are assumed to have a continuous density with respect to the volume form on the manifold, so $X_i \sim \mu$ where $d\mu = \rho\, dVol_{\mathcal{M}}$, where $\rho \in C(\mathcal{M})$ and $dVol_{\mathcal{M}}$ is the volume form on $\mathcal{M}$. All results in these notes extend to the manifold setting, the only difference being additional technical details. We note there are also other graph constructions of interest, such as k-nearest neighbor graphs, and much of the analysis can be extended to this setting, though the proofs can be much different.

We define the normalized graph Dirichlet energy $\mathcal{E}_{n,\varepsilon} : \ell^2(\mathcal{X}_n) \to \mathbb{R}$ by

$$\mathcal{E}_{n,\varepsilon}(u) = \frac{1}{\sigma_\eta n^2 \varepsilon^2} \sum_{x,y\in\mathcal{X}_n} \eta_\varepsilon(|x-y|)(u(x)-u(y))^2, \tag{5.53}$$

and the corresponding normalized graph Laplacian

$$\mathcal{L}_{n,\varepsilon}u(x) = \frac{2}{\sigma_\eta n \varepsilon^2} \sum_{y\in\mathcal{X}_n} \eta_\varepsilon(|x-y|)(u(y)-u(x)). \tag{5.54}$$

As discussed in Section 5.1.1, the Laplacian learning problem can be posed as a variational problem, by minimizing $\mathcal{E}_{n,\varepsilon}$, or as a boundary value problem by solving $\mathcal{L}_{n,\varepsilon}u = 0$. These are the variational and PDE interpretations of Laplace learning, respectively, and each gives its own set of tools and techniques for proving discrete to continuum converge.

In Section 5.1 we defined an L^2 structure and calculus on graphs. In the random geometric setting, we rescale these definitions accordingly, so they agree in the continuum, as $n \to \infty$ and $\varepsilon \to 0$, with their continuum counterparts. For $u \in \ell^2(\mathcal{X}_n)$ we define the *normalized* graph L^2-norm $\|u\|_{\ell^2(\mathcal{X}_n)}$ by

$$\|u\|^2_{\ell^2(\mathcal{X}_n)} := \frac{1}{n} \sum_{x\in\mathcal{X}_n} u(x)^2. \tag{5.55}$$

The L^2 norm is, as before, induced by the L^2-inner product

$$(u,v)_{\ell^2(\mathcal{X}_n)} := \frac{1}{n} \sum_{x\in\mathcal{X}_n} u(x)v(x). \tag{5.56}$$

We define the gradient vector field $\nabla_{n,\varepsilon}u$ for a function $u \in \ell^2(\mathcal{X}_n)$ by

$$\nabla_{n,\varepsilon}u(x,y) := \frac{1}{\varepsilon}(u(y)-u(x)). \tag{5.57}$$

The L^2 norm of the gradient is then defined as

$$\|\nabla_{n,\varepsilon}u\|^2_{\ell^2(\mathcal{X}_n^2)} = \frac{1}{\sigma_\eta n^2} \sum_{x,y\in\mathcal{X}_n} \eta_\varepsilon(|x-y|)|\nabla_{n,\varepsilon}u(x,y)|^2. \tag{5.58}$$

In this notation, $\mathcal{E}_{n,\varepsilon}(u) = \|\nabla_{n,\varepsilon}u\|^2_{\ell^2(\mathcal{X}_n^2)}$. The L^2 norm of the gradient is of course induced by the corresponding inner product

$$(V,W)_{\ell^2(\mathcal{X}_n^2)} := \frac{1}{\sigma_\eta n^2} \sum_{x,y\in\mathcal{X}_n} \eta_\varepsilon(|x-y|)V(x,y)W(x,y). \tag{5.59}$$

We define the $H^1(\mathcal{X}_n)$ inner product by

$$(u,v)_{H^1(\mathcal{X}_n)} = (u,v)_{\ell^2(\mathcal{X}_n)} + (\nabla_{n,\varepsilon}u, \nabla_{n,\varepsilon}u)_{\ell^2(\mathcal{X}_n^2)}, \tag{5.60}$$

and $H^1(\mathcal{X}_n)$ norm by $\|u\|^2_{H^1(\mathcal{X}_n)} = (u,u)_{H^1(\mathcal{X}_n)}$. Notice that

$$\|u\|^2_{H^1(\mathcal{X}_n)} = \|u\|^2_{\ell^2(\mathcal{X}_n)} + \mathcal{E}_{n,\varepsilon}(u). \tag{5.61}$$

Finally, for a vector field V we define, as before, the divergence of V to be

$$\mathrm{div}_{n,\varepsilon} V(x) = \frac{2}{\sigma_\eta n\varepsilon} \sum_{y\in\mathcal{X}_n} \eta_\varepsilon(|x-y|)V(x,y). \tag{5.62}$$

With these definitions, the graph Laplacian (5.54) is given by $\mathcal{L}_{n,\varepsilon} = \mathrm{div}_{n,\varepsilon}(\nabla_{n,\varepsilon} u)$ for $u \in \ell^2(\mathcal{X}_n)$, and the integration by parts formula

$$(\nabla_{n,\varepsilon} u, V)_{\ell^2(\mathcal{X}_n^2)} = -(u, \mathrm{div}_{n,\varepsilon} V)_{\ell^2(\mathcal{X}_n)} \tag{5.63}$$

holds, due to 5.11, where V is a vector field on $\mathcal{X}_n^2$.

We say the graph $\mathcal{G}_{n,\varepsilon}$ is *connected* if for every $x,y \in \mathcal{X}_n$ there exists $x = x_0, x_1, \ldots, x_m = y$ such that $x_i \in \mathcal{X}_n$ and $\eta_\varepsilon(|x_i - x_{i+1}|) > 0$ for all i. For the reader's convenience and interest, we record a standard result below on graph connectivity. While uniqueness of the graph learning problem requires connectivity of the graph, our discrete to continuum results do not explicitly use connectivity, so the result below is presented for interest only.

Lemma 5.20. *The graph $\mathcal{G}_{n,\varepsilon}$ is connected with probability at least $1 - Cn\exp\left(-cn\varepsilon^d\right)$, for constants $C, c > 0$ depending only on α, d, and U.*

Proof. For simplicity, we give the proof for a box $U = [0,1]^d$; the extension to arbitrary domains U with smooth boundaries is straightforward. We assume $0 < \varepsilon \leq 1$

Let Ω be the event that the graph $\mathcal{G}_{n,\varepsilon}$ is not connected. We partition the box U into $M = h^{-d}$ sub boxes $B_1, B_2, \ldots, B_n$ of side length h, and let n_i denote the number of points from $\mathcal{X}_n$ falling in B_i. If $\varepsilon \geq 4\sqrt{d}h$, then every point in B_i is connected to all points in neighboring boxes B_j. Thus, if we set $h = \varepsilon/4\sqrt{d}$ then the graph is connected if all boxes B_i contain at least one point from $\mathcal{X}_n$, and paths between pairs of points are constructed by hopping between neighboring boxes. Hence, if the graph is not connected, then some box B_i must be empty. Letting A_i denote the event that $n_i = 0$, we have by the union bound

$$\mathbb{P}(\Omega) \leq \mathbb{P}\left(\bigcup_{i=1}^M A_i\right) \leq \sum_{i=1}^M \mathbb{P}(A_i) = \sum_{i=1}^M \mathbb{P}(n_i = 0).$$

Since $X_1, \ldots, X_n$ are *i.i.d.* random variables with density $\rho \geq \alpha > 0$, we have that

$$\mathbb{P}(n_i = 0) = \mathbb{P}\left(\forall j \in \{1,\ldots,n\},\, X_j \notin B_i\right) = \prod_{j=1}^n \mathbb{P}(X_j \notin B_i) = \mathbb{P}(X_1 \notin B_i)^n.$$

By direct computation we have

$$\mathbb{P}(X_1 \not\in B_i) = \int_{U\setminus B_i} \rho\, dx = 1 - \int_{B_i} \rho\, dx \leq 1 - \alpha|B_i| = 1 - \alpha h^d.$$

Since $h = \varepsilon/4\sqrt{d}$, there exists $0 < c \leq 1$, depending only on n and α, such that

$$\mathbb{P}(n_i = 0) \leq (1 - c\varepsilon^d)^n \leq \exp\left(-cn\varepsilon^d\right).$$

Therefore

$$\mathbb{P}(\Omega) \leq M \exp\left(-cn\varepsilon^d\right) = C\varepsilon^{-d} \exp\left(-cn\varepsilon^d\right).$$

If $n\varepsilon^d \geq 1$, then

$$\mathbb{P}(\Omega) \leq Cn \exp\left(-cn\varepsilon^d\right).$$

If $n\varepsilon^d \leq 1$, then

$$Cn \exp\left(-cn\varepsilon^d\right) \geq Cn \exp(-c) \geq 1 \geq \mathbb{P}(\Omega)$$

provided we choose C larger, if necessary. This completes the proof. □

Remark 5.21. By Lemma 5.20, there exists $C > 0$ such that if $n\varepsilon^d \geq C\log(n)$, then the graph is connected with probability at least $1 - \frac{1}{n^2}$. Note we can rewrite this condition as

$$\varepsilon \geq K\left(\frac{\log(n)}{n}\right)^{1/d}, \tag{5.64}$$

where $K = C^{1/d}$. If $n \to \infty$ and $\varepsilon = \varepsilon_n \to 0$ so that (5.64) holds, then the graph is connected for large enough n almost surely. This gives our first length scale restriction on ε. We will see more severe restrictions in the following sections.

5.4 The maximum principle approach

The maximum principle is an incredibly useful tool for passing to limits in many kinds of problems. In the graph setting, the graph Laplacian has a discrete maximum principle, and so it is natural to make use of this for discrete to continuum convergence.

For the maximum principle approach, we interpret Laplace learning as the boundary value problem

$$\begin{cases} \mathcal{L}_{n,\varepsilon} u(x) = 0 & \text{if } x \in \mathcal{X}_n \setminus \partial_\varepsilon U \\ u(x) = g(x) & \text{if } x \in \mathcal{X}_n \cap \partial_\varepsilon U, \end{cases} \tag{5.65}$$

where $g : U \to \mathbb{R}$ is a given Lipschitz function and

$$\partial_\varepsilon U = \{x \in U \,:\, \operatorname{dist}(x, \partial U) \leq \varepsilon\}. \tag{5.66}$$

We also write $U_\varepsilon = U \setminus \partial_\varepsilon U$. Note we have chosen our labeled points to be all points within ε of the boundary. Other choices of boundary conditions can be used, and it is an interesting, and somewhat open, problem to determine the fewest number of labeled points for which discrete to continuum convergence holds.

The continuum version of (5.65) is

$$\begin{cases} \operatorname{div}(\rho^2 \nabla u) = 0 & \text{in } U \\ u = g & \text{on } \partial U. \end{cases} \tag{5.67}$$

The goal of this section is to use the maximum principle to prove convergence, with a rate, of the solution of (5.65) to the solution of (5.67) as $\varepsilon \to 0$ and $n \to \infty$.

We first prove pointwise consistency of graph Laplacians. The proof utilizes the nonlocal operator

$$\mathcal{L}_\varepsilon u(x) = \frac{2}{\sigma_\eta \varepsilon^2} \int_U \eta_\varepsilon(|x-y|)(u(y)-u(x))\rho(y)\,dy. \tag{5.68}$$

Lemma 5.22 (Discrete to nonlocal)**.** *Let $u : U \to \mathbb{R}$ be Lipschitz continuous and $\varepsilon > 0$ with $n\varepsilon^d \geq 1$. Then for any $0 < \lambda \leq \varepsilon^{-1}$*

$$\max_{x \in \mathcal{X}_n} |\mathcal{L}_{n,\varepsilon} u(x) - \mathcal{L}_\varepsilon u(x)| \leq C\,Lip(u)(\lambda + \varepsilon). \tag{5.69}$$

with probability at least $1 - C\exp\left(-cn\varepsilon^{d+2}\lambda^2 + 3\log(n)\right)$.

Proof. Fix $x \in U$ and write

$$Y_i = \eta_\varepsilon(|X_i - x|)(u(X_i) - u(x)),$$

so that

$$\mathcal{L}_{n,\varepsilon} u(x) = \frac{2}{\sigma_\eta \varepsilon^2} \frac{1}{n} \sum_{i=1}^n Y_i.$$

We compute

$$|Y_i| = \eta_\varepsilon(|X_i - x|)|u(X_i) - u(x)| \leq C\mathrm{Lip}(u)\varepsilon^{1-d} =: b,$$

$$\mathbb{E}[Y_i] = \int_U \eta_\varepsilon(|x-y|)(u(y)-u(x))\rho(y)\,dy,$$

and

$$\begin{aligned} \sigma^2 = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2] \leq \mathbb{E}[Y_i^2] &= \int_{B(x,\varepsilon)\cap U} \eta_\varepsilon(|x-y|)^2 (u(y)-u(x))^2 \rho(y)\,dy \\ &\leq C\mathrm{Lip}(u)^2 \varepsilon^2 \int_{B(x,\varepsilon)\cap U} \eta_\varepsilon(|x-y|)^2\,dy \\ &\leq C\mathrm{Lip}(u)^2 \varepsilon^{2-d}. \end{aligned}$$

By Bernstein's inequality (Theorem 5.12)

$$\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \int_U \eta_\varepsilon(|x-y|)(u(y)-u(x))\rho(y)\,dy\right| \leq t$$

holds with probability at least $1 - 2\exp\left(-c\mathrm{Lip}(u)^{-2}n\varepsilon^{d-2}t^2\right)$ for $0 < t < \frac{\sigma_\eta}{2}\mathrm{Lip}(u)\varepsilon$. Set $t = \frac{\sigma_\eta}{2}\mathrm{Lip}(u)\varepsilon^2\lambda$ and multiply both sides by $2/(\sigma_\eta\varepsilon^2)$ to find that

$$|\mathcal{L}_{n,\varepsilon}u(x) - \mathcal{L}_\varepsilon u(x)| \leq \mathrm{Lip}(u)\lambda \tag{5.70}$$

holds with probability at least $1 - 2\exp\left(-cn\varepsilon^{d+2}\lambda^2\right)$ for any $0 < \lambda < \varepsilon^{-1}$.

The rest of the proof is completed by either conditioning on $X_i = x$, or by a covering argument. Since we have not covered conditional probability, we will give the covering argument here. Let $U_h = \mathbb{Z}^d_h \cap U$ and $N(x) = \#B(x, 2\varepsilon) \cap \mathcal{X}_n$. By the Chernoff bounds (Theorem 5.7) we have $\mathbb{P}(N(x) \geq Cn\varepsilon^d) \leq \exp(-cn\varepsilon^d)$ for constants $C, c > 0$. Let Ω_h denote the event that (5.70) holds and $N(x) \leq Cn\varepsilon^d$ for all $x \in U_h$. Since U_h has at most Ch^{-d} points we can use the union bound to obtain

$$\mathbb{P}(\Omega_h) \geq 1 - Ch^{-d}\exp\left(-cn\varepsilon^{d+2}\lambda^2\right).$$

For each X_i, let $x_i \in U_h$ denote the closest point to X_i in U_h, which satisfies $|X_i - x_i| \leq Ch$ for some constant $C > 0$. We assume $Ch \leq \varepsilon$. Since $y \mapsto \eta_\varepsilon(|x-y|)(u(y)-u(x))$ is Lipschitz with constant at most $C\varepsilon^{-n}$, we have

$$|\mathcal{L}_\varepsilon u(x_i) - \mathcal{L}_\varepsilon u(X_i)| \leq \frac{Ch}{\varepsilon^{d+2}}\int_{B(x,\varepsilon)\cap U} dx = C\mathrm{Lip}(u)h\varepsilon^{-2},$$

and

$$|\mathcal{L}_{n,\varepsilon}u(x_i) - \mathcal{L}_{n,\varepsilon}u(X_i)| \leq \frac{Ch}{n\varepsilon^{d+2}}\sum_{y\in\mathcal{X}_n\cap B(x_i,2\varepsilon)} 1 \leq C\mathrm{Lip}(u)h\varepsilon^{-2}.$$

Therefore, if Ω_h holds, we have

$$|\mathcal{L}_{n,\varepsilon}u(X_i) - \mathcal{L}_\varepsilon u(X_i)| \leq \mathrm{Lip}(u)\lambda + C\mathrm{Lip}(u)h\varepsilon^{-2}.$$

for all $1 \leq i \leq n$. Setting $h = \varepsilon^3$ yields

$$\max_{x\in\mathcal{X}_n}|\mathcal{L}_{n,\varepsilon}u(x) - \mathcal{L}_\varepsilon u(x)| \leq C\mathrm{Lip}(u)(\lambda + \varepsilon)$$

with probability at least $1 - C\varepsilon^{-3d}\exp\left(-cn\varepsilon^{d+2}\lambda^2\right)$ for any $0 < \lambda \leq \varepsilon^{-1}$. We complete the proof by using $n\varepsilon^d \geq 1$ to obtain $\varepsilon^{-3d} \leq n^3 = \exp(3\log(n))$. □

Remark 5.23. Lemma 5.22 can be strengthen to read

$$\mathbb{P}\left(\forall u \in C^3(\overline{U}),\ \max_{x\in\mathcal{X}_n}|\mathcal{L}_{n,\varepsilon}u(x) - \mathcal{L}_\varepsilon u(x)| \leq C\|u\|_{C^3(\overline{U})}(\lambda+\varepsilon)\right)$$
$$\geq 1 - C\exp\left(-cn\varepsilon^{d+2}\lambda^2 + 3\log(n)\right)$$

for any $0 < \lambda \leq \varepsilon^{-1}$, provided $u \in C^3(\overline{U})$. The proof is similar to Lemma 5.22, but involves Taylor expanding u before applying concentration of measure so that the application of Bernstein's inequality does not depend on u. The uniformity over $u \in C^3(\overline{U})$ is required when using the viscosity solution machinery to prove discrete to continuum convergence in the non-smooth setting. We refer the reader to [11, Theorem 5] for more details. Also, when one uses conditional probability instead of the covering argument in proving Lemma 5.22, the term $3\log(n)$ can be improved to $\log(n)$. This is inconsequential for the results below.

We now turn to comparing the nonlocal operator $\mathcal{L}_\varepsilon$ to its continuum counterpart

$$\Delta_\rho u := \rho^{-1}\mathrm{div}(\rho^2 \nabla u). \tag{5.71}$$

Lemma 5.24 (Nonlocal to local)**.** *There exists $C > 0$ such that for every $u \in C^3(\overline{U})$ and $x \in U$ with dist$(x, \partial U) \geq \varepsilon$ we have*

$$|\mathcal{L}_\varepsilon u(x) - \Delta_\rho u(x)| \leq C\|u\|_{C^3(U)}\varepsilon \tag{5.72}$$

Proof. We first make a change of variables $z = (y - x)/\varepsilon$ to find that

$$\mathcal{L}_\varepsilon u(x) = \frac{2}{\sigma_\eta \varepsilon^2}\int_{B(0,1)} \eta(|z|)(u(x+z\varepsilon) - u(x))\rho(x+z\varepsilon)\,dz.$$

Let $\beta = \|u\|_{C^3(U)}$. We use the Taylor expansions

$$u(x+z\varepsilon) - u(x) = \nabla u(x)\cdot z\varepsilon + \frac{\varepsilon^2}{2}z^T\nabla^2 u(x)z + O(\beta\varepsilon^3)$$

and

$$\rho(x+z\varepsilon) = \rho(x) + \nabla\rho(x)\cdot z\varepsilon + O(\varepsilon^2),$$

for $|z| \leq 1$, to obtain

$$\begin{aligned}
\mathcal{L}_\varepsilon u(x) &= \frac{2}{\sigma_\eta\varepsilon^2}\int_{B(0,1)}\eta(|z|)\left(\nabla u(x)\cdot z\varepsilon + \frac{\varepsilon^2}{2}z^T\nabla^2 u(x)z + O(\beta\varepsilon^3)\right) \\
&\qquad\qquad\qquad\qquad \times\left(\rho(x) + \nabla\rho(x)\cdot z\varepsilon + O(\varepsilon^2)\right)\,dz \\
&= \frac{2}{\sigma_\eta}\int_{B(0,1)}\eta(|z|)\Bigg(\rho(x)\nabla u(x)\cdot z\varepsilon^{-1} + \frac{1}{2}\rho(x)z^T\nabla^2 u(x)z \\
&\qquad\qquad\qquad\qquad + (\nabla u(x)\cdot z)(\nabla\rho(x)\cdot z)\Bigg)\,dz + O(\beta\varepsilon) \\
&= \frac{2}{\sigma_\eta}\int_{B(0,1)}\eta(|z|)\left(\frac{1}{2}\rho(x)z^T\nabla^2 u(x)z + (\nabla u(x)\cdot z)(\nabla\rho(x)\cdot z)\right)\,dz + O(\beta\varepsilon) \\
&=: A + B + O(\beta\varepsilon),
\end{aligned}$$

where we used the fact that

$$\int_{B(0,1)} \eta(|z|)\nabla u(x)\cdot z\,dz = 0,$$

since $z \mapsto \nabla u(x)\cdot z$ is odd.

To compute A, we write

$$\begin{aligned}
A &= \frac{1}{\sigma_\eta}\rho(x)\sum_{i,j=1}^d u_{x_ix_j}(x)\int_{B(0,1)} \eta(|z|)z_iz_j\,dz\\
&= \frac{1}{\sigma_\eta}\rho(x)\sum_{i=1}^d u_{x_ix_i}(x)\int_{B(0,1)} \eta(|z|)z_i^2\,dz\\
&= \rho(x)\Delta u(x),
\end{aligned}$$

where we used that $z \mapsto z_iz_j$ is odd for $i \neq j$, so the integrals vanish for $i \neq j$. To compute B, we have

$$\begin{aligned}
B &= \frac{2}{\sigma_\eta}\sum_{i,j=1}^d u_{x_i}(x)\rho_{x_j}(x)\int_{B(0,1)} \eta(|z|)z_iz_j\,dz\\
&= \frac{2}{\sigma_\eta}\sum_{i=1}^d u_{x_i}(x)\rho_{x_i}(x)\int_{B(0,1)} \eta(|z|)z_i^2\,dz\\
&= 2\nabla\rho(x)\cdot\nabla u(x).
\end{aligned}$$

Combining this with the main argument above we have

$$\begin{aligned}
\mathcal{L}_\varepsilon u(x) &= \rho(x)\Delta u(x) + 2\nabla\rho(x)\cdot\nabla u(x) + O(\beta\varepsilon)\\
&= \rho(x)^{-1}\left(\rho(x)^2\Delta u(x) + 2\rho(x)\nabla\rho(x)\cdot\nabla u(x)\right) + O(\beta\varepsilon)\\
&= \rho^{-1}\mathrm{div}(\rho^2\nabla u) + O(\beta\varepsilon)\\
&= \Delta_\rho u(x) + O(\beta\varepsilon),
\end{aligned}$$

which completes the proof. □

Remark 5.25. Combining Lemmas 5.22 and 5.24, for any $u \in C^3(\overline{U})$ and $0 < \lambda \leq \varepsilon^{-1}$ we have that

$$\max_{x\in\mathcal{X}_n\setminus\partial_\varepsilon U} |\mathcal{L}_{n,\varepsilon}u(x) - \Delta_\rho u(x)| \leq C(\lambda+\varepsilon) \tag{5.73}$$

hold with probability at least $1 - C\exp\left(-cn\varepsilon^{d+2}\lambda^2 + 3\log(n)\right)$. This is called *pointwise consistency* for graph Laplacians. Notice that for the bound to be non-vacuous, we require $n\varepsilon^{d+2} \gg \log(n)$. To get a linear $O(\varepsilon)$ pointwise consistency rate, we set

$\lambda = \varepsilon$ and require $n\varepsilon^{d+4} \geq C\log(n)$ for a large constant $C > 0$. Written a different way, we require

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

for pointwise consistency of graph Laplacians and

$$\varepsilon \geq K\left(\frac{\log(n)}{n}\right)^{1/(d+4)}$$

for pointwise consistency with a linear $O(\varepsilon)$ rate. We note these are larger length scale restrictions than for connectivity of the graph (compare with (5.64)). In the length scale range

$$\left(\frac{\log(n)}{n}\right)^{1/d} \ll \varepsilon \ll \left(\frac{\log(n)}{n}\right)^{1/(d+2)} \tag{5.74}$$

the graph is connected, and Laplacian learning well-posed, but pointwise consistency does not hold and current PDE techniques cannot say anything about discrete to continuum convergence.

Remark 5.26. If $u \in C^4(U)$, then Lemma 5.24 can be sharpened to read

$$|\mathcal{L}_\varepsilon u(x) - \Delta_\rho u(x)| \leq C\|u\|_{C^4(U)}\varepsilon^2.$$

We leave this as an exercise for the reader. Combining this with Lemma 5.22 we have we have that

$$\max_{x\in\mathcal{X}_n\setminus\partial_\varepsilon U} |\mathcal{L}_{n,\varepsilon}u(x) - \Delta_\rho u(x)| \leq C(\lambda + \varepsilon^2) \tag{5.75}$$

hold with probability at least $1 - C\exp\left(-cn\varepsilon^{d+2}\lambda^2 + 3\log(n)\right)$ for all $0 < \lambda \leq \varepsilon^{-1}$ and all $u \in C^4(U)$. To obtain the second order convergence rate, we must choose $\lambda = \varepsilon^2$, and so we require the stricter length scale restriction

$$\varepsilon \geq C\left(\frac{\log n}{n}\right)^{1/(d+6)}.$$

We now show how to use the maximum principle and pointwise consistency to establish discrete to continuum convergence with a rate.

Theorem 5.27 (Discrete to continuum convergence)**.** *Let $0 < \varepsilon \leq 1$ and $n \geq 1$. Let $u_{n,\varepsilon} \in \ell^2(\mathcal{X}_n)$ be a solution of (5.65) satisfying (5.18) with $\Gamma = \mathcal{X}_n \cap \partial_\varepsilon U$, and let $u \in C^3(\overline{U})$ be the solution of (5.67). Then for any $0 < \lambda \leq 1$*

$$\max_{x\in\mathcal{X}_n} |u_{n,\varepsilon}(x) - u(x)| \leq C(\|u\|_{C^3(U)} + 1)(\lambda + \varepsilon) \tag{5.76}$$

holds with probability at least $1 - C\exp\left(-cn\varepsilon^{d+2}\lambda^2 + 3\log(n)\right)$.

Proof. Let $\varphi \in C^3(\overline{U})$ be the solution of

$$\begin{cases} -\Delta_\rho \varphi = -\rho^{-1}\mathrm{div}(\rho^2 \nabla\varphi) = 1 & \text{in } U \\ \varphi = 0 & \text{on } \partial U. \end{cases} \tag{5.77}$$

Since $\Delta_\rho\varphi(x) \geq 0$ at any point $x \in U$ where φ attains its minimum value, and $\Delta_\rho\varphi(x) = -1$ for all $x \in U$, the minimum value of φ is attained on the boundary ∂U, and so $\varphi \geq 0$. Combining Lemmas 5.22 and 5.24, and recalling $\Delta_\rho u = 0$, we have that

$$\max_{x\in\mathcal{X}_n\setminus\partial_\varepsilon U} |\mathcal{L}_{n,\varepsilon}\varphi(x) + 1| \leq C_1(\lambda+\varepsilon) \tag{5.78}$$

and

$$\max_{x\in\mathcal{X}_n\setminus\partial_\varepsilon U} |\mathcal{L}_{n,\varepsilon}u(x)| \leq C_1\|u\|_{C^3(U)}(\lambda+\varepsilon) \tag{5.79}$$

hold with probability at least $1 - C_2\exp\left(-cn\varepsilon^{d+2}\lambda^2 + 3\log(n)\right)$ for any $0 < \lambda \leq \varepsilon^{-1}$. For the rest of the proof, we assume (5.78) and (5.79) hold for some fixed $0 < \lambda, \varepsilon \leq 1$.

First, if $C_1(\lambda+\varepsilon) \geq \frac{1}{2}$ then we have

$$|u_{n,\varepsilon}(x) - u(x)| \leq |u_{n,\varepsilon}(x)| + |u(x)| \leq 2\|g\|_{L^\infty(U)} \leq 4C_1\|g\|_{L^\infty(U)}(\lambda+\varepsilon),$$

and the proof is complete. Therefore, we may assume that $C_1(\lambda+\varepsilon) \leq \frac{1}{2}$, and so by (5.78) we have $\mathcal{L}_{n,\varepsilon}\varphi(x) \leq -\frac{1}{2}$ for all $x \in \mathcal{X}_n \setminus \partial_\varepsilon U$. Let $K > 0$ and define $w = u - u_{n,\varepsilon} - K\varphi$. Then we have

$$\mathcal{L}_{n,\varepsilon}w(x) = \mathcal{L}_{n,\varepsilon}u(x) - \mathcal{L}_{n,\varepsilon}u_{n,\varepsilon}(x) - K\mathcal{L}_{n,\varepsilon}\varphi(x) \geq \frac{K}{2} - C_1\|u\|_{C^3(U)}(\lambda+\varepsilon)$$

for $x \in \mathcal{X}_n \setminus \partial_\varepsilon U$. Setting $K = 2C_1(\|u\|_{C^3(U)} + 1)(\lambda+\varepsilon)$ we have $\mathcal{L}_{n,\varepsilon}w(x) > 0$ for all $x \in \mathcal{X}_n \setminus \partial_\varepsilon U$. By the maximum principle (Lemma 5.4) we have

$$\max_{x\in\mathcal{X}_n} w(x) = \max_{x\in\mathcal{X}_n\cap\partial_\varepsilon U} w(x).$$

Since $u(x) = u_{n,\varepsilon}(x)$ for $x \in \mathcal{X}_n \cap \partial_\varepsilon U$ and $\varphi \geq 0$, we have $w(x) \leq 0$ for $x \in \mathcal{X}_n \cap \partial_\varepsilon U$ and so $w(x) \leq 0$ for all $x \in \mathcal{X}_n$, which yields

$$u(x) - u_{n,\varepsilon}(x) \leq K\varphi(x)$$

for all $x \in \mathcal{X}_n$. For the other direction, we can define $w = u_{n,\varepsilon} - u - K\varphi$ and make a similar argument to obtain

$$u_{n,\varepsilon}(x) - u(x) \leq K\varphi(x)$$

for all $x \in \mathcal{X}_n$, which completes the proof. □

Notes and references: Pointwise consistency for graph Laplacians was established in [34, 35] for a random geometric graph on a manifold. Pointwise consistency for k-nearest neighbor graphs was established without rates in [57], and with convergence rates in [14]. The maximum principle is a well-established tool for passing to limits with convergence rates, and has been well-used in numerical analysis. The theory presented here requires regularity of the solution u of the continuum PDE. In some problems, especially nonlinear problems, the solution is not sufficiently regular, and the theory of viscosity solutions has been developed for precisely this purpose. The papers [11] and [13] use the maximum principle, pointwise consistency and the viscosity solution machinery (see [12] for more on viscosity solutions) to prove discrete to continuum convergence for p-Laplace and ∞-Laplace semi-supervised learning problems with very few labels. The maximum principle is used in [16, 54] to prove convergence rates for Laplacian regularized semi-supervised learning, and in [59] to prove rates for Laplacian regularized regression on graphs.

5.5 The variational approach

The maximum principle approach described in the previous section is very powerful for proving convergence rates, when applicable. However, there are many problems where the maximum principle cannot be used. First, since pointwise consistency requires a large length scale restriction on ε (see Remark 5.25), the maximum principle is, with current techniques, not useful when ε is in the range (5.74). Second, the PDE maximum principle approach requires the continuum PDE to have a comparison principle and be well-posed. There are many important problems, such as spectral or total variation problems, where the continuum limit problem does not have a unique solution. In these situations, we can often use variational tools to prove discrete to continuum convergence, and we illustrate some of the main ideas in this section for the Laplace learning problem.

Variational techniques prove convergence rates by comparing the optimal values of the energies for the discrete and continuum problems. As with pointwise consistency for graph Laplacians, we pass through a nonlocal problem. We define the nonlocal operator

$$I_{\varepsilon,\delta}(u) = \frac{1}{\sigma_\eta \varepsilon^2}\int_U\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)(u(x)-u(y))^2\rho(x)\rho(y)\,dx\,dy, \tag{5.80}$$

for $u \in L^2(U)$ and $\varepsilon, t > 0$. We write $I_\varepsilon(u) = I_{\varepsilon,0}(u)$. We also define the local energy

$$I(u) = \int_U |\nabla u|^2\rho^2\,dx \tag{5.81}$$

for $u \in H^1(U)$.

5.5.1 Consistency for smooth functions

Our first establish consistency for smooth functions, in similar spirit to pointwise consistency results for graph Laplacians from the previous sections. The proof is split into two lemmas. We recall $\mathcal{E}_{n,\varepsilon}$ is defined in (5.53).

Lemma 5.28 (Discrete to nonlocal)**.** *There exists $C, c > 0$ such that for any $0 < \lambda \leq 1$ and Lipschitz continuous $u : U \to \mathbb{R}$ we have*

$$|\mathcal{E}_{n,\varepsilon}(u) - I_\varepsilon(u)| \leq C Lip(u)^2 \left(\frac{1}{n} + \lambda\right) \tag{5.82}$$

with probability at least $1 - 2\exp\left(-cn\varepsilon^d\lambda^2\right)$.

Proof. Define the U-statistic

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

where

$$f(x,y) = \eta_\varepsilon(|x-y|) \left(\frac{u(x) - u(y)}{\varepsilon}\right)^2,$$

and note that $\mathcal{E}_{n,\varepsilon}(u) = \frac{n-1}{\sigma_\eta n} U_n$. Since u is Lipschitz, $|\eta_\varepsilon| \leq C\varepsilon^{-d}$, and $\eta_\varepsilon(|x-y|) = 0$ for $|x-y| > \varepsilon$ we have

$$b = \sup_{x,y \in U} |f(x,y)| \leq C\mathrm{Lip}(u)^2 \varepsilon^{-d}.$$

We also have

$$\mathbb{E}[f(X_i, X_j)] = \int_U \int_U \eta_\varepsilon(|x-y|) \left(\frac{u(x) - u(y)}{\varepsilon}\right)^2 \rho(x)\rho(y)\, dx\, dy = \sigma_\eta I_\varepsilon(u),$$

and

$$\begin{aligned}
\sigma^2 &:= \mathrm{Var}\left(f(X_i, X_j)\right) \\
&\leq \int_U \int_U \eta_\varepsilon(|x-y|)^2 \left(\frac{u(x) - u(y)}{\varepsilon}\right)^4 \rho(x)\rho(y)\, dxdy \\
&\leq C\mathrm{Lip}(u)^4 \varepsilon^{-2d} \int_U \int_{B(y,\varepsilon)} dx\, dy \\
&\leq C\mathrm{Lip}(u)^4 \varepsilon^{-d}.
\end{aligned}$$

Applying Bernstein's inequality for U-statistics (Theorem 5.15) yields

$$\mathbb{P}(|U_n - \sigma_\eta I_\varepsilon(u)| \geq \mathrm{Lip}(u)^2 \lambda) \leq 2\exp\left(-cn\varepsilon^d\lambda^2\right)$$

for any $0 < \lambda \leq 1$. Since $\mathcal{E}_{n,\varepsilon}(u) = \frac{n-1}{\sigma_\eta n} U_n$ we have

$$\begin{aligned}
|\mathcal{E}_{n,\varepsilon}(u) - I_\varepsilon(u)| &= \frac{1}{\sigma_\eta} \left|(1 - \tfrac{1}{n})U_n - \sigma_\eta I_\varepsilon(u)\right| \\
&= \frac{1}{\sigma_\eta} \left|(1 - \tfrac{1}{n})(U_n - \sigma_\eta I_\varepsilon(u)) - \tfrac{\sigma_\eta}{n} I_\varepsilon(u)\right| \\
&\leq \frac{1}{\sigma_\eta} |U_n - \sigma_\eta I_\varepsilon(u)| + \frac{1}{n} I_\varepsilon(u).
\end{aligned}$$

Since u is Lipschitz we have $I_\varepsilon(u) \leq \mathrm{Lip}(u)^2$. Therefore we can apply the result of Bernstein above to obtain that

$$|\mathcal{E}_{n,\varepsilon}(u) - I_\varepsilon(u)| \leq C\mathrm{Lip}(u)^2 \left(\lambda + \frac{1}{n}\right)$$

holds with probability at least $1 - 2\exp\left(-cn\varepsilon^d\lambda^2\right)$, which completes the proof. □

Lemma 5.29 (Nonlocal to local)**.** *There exist $C > 0$ such that for all $u \in C^2(\overline{U})$ and $0 < \varepsilon \leq 1$*

$$|I_\varepsilon(u) - I(u)| \leq C Lip(u)\|u\|_{C^2(U)}\varepsilon. \tag{5.83}$$

Proof. Let $\beta = \|u\|_{C^2(U)}$. We first use the Taylor expansion

$$u(y) - u(x) = \nabla u(x) \cdot (y - x) + O(\beta\varepsilon^2)$$

for $|x - y| \leq \varepsilon$ to obtain

$$\begin{aligned}
(u(y) - u(x))^2 &= (u(y) - u(x))(\nabla u(x) \cdot (y - x) + O(\beta\varepsilon^2)) \\
&= (u(y) - u(x))\nabla u(x) \cdot (y - x) + O(\beta\mathrm{Lip}(u)\varepsilon^3) \\
&= (\nabla u(x) \cdot (y - x) + O(\beta\varepsilon^2))\nabla u(x) \cdot (y - x) + O(\beta\mathrm{Lip}(u)\varepsilon^3) \\
&= |\nabla u(x) \cdot (y - x)|^2 + O(\beta\mathrm{Lip}(u)\varepsilon^3)
\end{aligned}$$

for $|x - y| \leq \varepsilon$. We combine this with the expansion

$$\rho(y) = \rho(x)(1 + O(\varepsilon))$$

for $|x - y| \leq \varepsilon$ to obtain

$$\begin{aligned}
I_\varepsilon(u) &= \frac{1}{\sigma_\eta\varepsilon^2} \int_U \int_{B(x,\varepsilon)\cap U} \eta_\varepsilon\left(|x - y|\right)(|\nabla u(x) \cdot (y - x)|^2 + O(\beta\mathrm{Lip}(u)\varepsilon^3))\rho(x)^2(1 + O(\varepsilon))\,dy\,dx \\
&= \frac{1}{\sigma_\eta\varepsilon^2} \int_U \int_{B(x,\varepsilon)\cap U} \eta_\varepsilon\left(|x - y|\right)\left(|\nabla u(x) \cdot (y - x)|^2 + O(\beta\mathrm{Lip}(u)\varepsilon^3)\right)\rho(x)^2\,dy\,dx \\
&= \frac{1}{\sigma_\eta\varepsilon^2} \int_U \int_{B(x,\varepsilon)\cap U} \eta_\varepsilon\left(|x - y|\right)|\nabla u(x) \cdot (y - x)|^2\rho(x)^2\,dy\,dx + O(\beta\mathrm{Lip}(u)\varepsilon).
\end{aligned}$$

Let $A \in \mathbb{R}^{d\times d}$ be an orthogonal matrix such that

$$A\nabla u(x) = |\nabla u(x)| e_1$$

where $e_1 = (1, 0, 0, \dots, 0) \in \mathbb{R}^d$, and make the change of variables $z = x + A(y-x)$ in the inner integral. Then since A is orthogonal we have $\eta_\varepsilon(|x-y|) = \eta_\varepsilon(|x-z|)$ and

$$\nabla u(x) \cdot (y-x) = A\nabla u(x) \cdot A(y-x) = |\nabla u(x)| e_1 \cdot (z-x) = |\nabla u(x)|(z_1 - x_1).$$

Therefore

$$I_\varepsilon(u) = \frac{1}{\sigma_\eta \varepsilon^2} \int_U |\nabla u(x)|^2 \rho(x)^2 \int_{B(x,\varepsilon)\cap V} \eta_\varepsilon\left(|x-y|\right)|z_1 - x_1|^2\, dz\, dx + O(\beta \mathrm{Lip}(u)\varepsilon),$$

where $V = x + A(U-x)$. If $\mathrm{dist}(x, \partial U) \geq \varepsilon$ then

$$\int_{B(x,\varepsilon)\cap V} \eta_\varepsilon\left(|x-y|\right)|z_1 - x_1|^2\, dz = \varepsilon^2 \int_{B(0,1)} \eta(|y|) y_1^2\, dy = \sigma_\eta \varepsilon^2,$$

where we used that $B(x,\varepsilon)\cap V = B(x,\varepsilon)$ and the change of variables $y = (z-x)/\varepsilon$. For all $x \in U$, a similar computation yields

$$\int_{B(x,\varepsilon)\cap V} \eta_\varepsilon\left(|x-y|\right)|z_1 - x_1|^2\, dz \leq \sigma_\eta \varepsilon^2.$$

Therefore

$$\begin{aligned}|I_\varepsilon(u) - I(u)| &\leq 2\int_{\partial_\varepsilon U} |\nabla u(x)|^2 \rho(x)^2 dx + C\beta \mathrm{Lip}(u)\varepsilon \\ &\leq 2\mathrm{Lip}(u)^2 \int_{\partial_\varepsilon U} \rho(x)^2 dx + C\beta \mathrm{Lip}(u)\varepsilon \\ &\leq C\mathrm{Lip}(u)^2 \varepsilon + C\beta\mathrm{Lip}(u)\varepsilon,\end{aligned}$$

since ∂U is smooth. This completes the proof. □

Remark 5.30. Combining Lemma 5.28 and 5.29 we have that for any $u \in C^2(\overline{U})$

$$|\mathcal{E}_{n,\varepsilon}(u) - I(u)| \leq C\mathrm{Lip}(u)\left[\mathrm{Lip}(u)\left(\frac{1}{n} + \lambda\right) + \|u\|_{C^2(U)}\varepsilon\right] \tag{5.84}$$

with probability at least $1 - 2\exp\left(-cn\varepsilon^d\lambda^2\right)$ for any $\lambda > 0$. The estimate (5.84) is a pointwise consistency estimate for the graph Dirichlet energy. In Section 5.4 we saw that pointwise consistency of the graph Laplacian was sufficient to prove discrete to continuum convergence, since we made use of powerful maximum principle tools. For variational proofs of convergence, were the maximum principle is not applicable, the pointwise consistency result (5.84) is not enough to ensure minimizers of $\mathcal{E}_{n,\varepsilon}$ converge to minimizers of I as $n \to \infty$ and $\varepsilon \to 0$. The pointwise consistency result (5.84) shows that when restricting a smooth function $u \in C^2(\overline{U})$ to the graph, the graph Dirichlet energy $\mathcal{E}_{n,\varepsilon}(u)$ is close to the continuum energy $I(u)$. To use variational methods to prove convergence we have to go in the other direction as well; that is, we need to take a function $u \in \ell^2(\mathcal{X}_n)$ and extend u to a function on the domain U for which $I(u)$ and $\mathcal{E}_{n,\varepsilon}(u)$ are similar. This direction is much harder, since u has no *a priori* regularity. This issue is the focus of the next few sections.

5.5.2 Discrete to nonlocal via transportation maps

In this section, we give an alternative method for passing from discrete to nonlocal energies that does not use concentration of measure, as in the previous sections, and does not require regularity.

We say ρ is a *probability density* on U if $\rho \in L^1(U)$, $\rho \geq 0$ and $\int_U \rho \, dx = 1$. We say that $U_1, U_2, \dots, U_n \subset U$ is a *partition* of U if $U_i \cap U_j = \varnothing$ for $i \neq j$, $U = \bigcup_{i=1}^n U_i$, and each U_i has nonempty interior.

Lemma 5.31. *Assume the probability density ρ is Lipschitz continuous on $\overline{U}$. Let $\delta > 0$ such that $n\delta^d \geq 1$, and let $X_1, X_2, \dots, X_n$ be a sequence of* i.i.d. *random variables with density ρ. Then exists a probability density $\rho_\delta \in L^\infty(U)$ and a partition $U_1, U_2, \dots, U_n$ of U such that*

$$U_i \subset B(X_i, \delta) \quad \textit{and} \quad \int_{U_i} \rho_\delta \, dx = \frac{1}{n} \tag{5.85}$$

hold for $1 \leq i \leq n$, and

$$\mathbb{P}\left(\|\rho_\delta - \rho\|_{L^\infty(U)} \leq C(\lambda + \delta)\right) \geq 1 - Cn\exp(-cn\delta^d\lambda^2) \tag{5.86}$$

holds for any $0 \leq \lambda \leq 1$.

Proof. There exists a universal constants $K > 0$ such that for each $h > 0$ there is a partition $B_1, B_2, \dots, B_M$ of U for which $M \leq Ch^{-n}$ and $B_i \subset B(x_i, Kh)$ for some $x_i \in B_i$ for all i. Let $\delta > 0$ and set $h = \delta/2K$ so that $B_i \subset B(x_i, \delta/2)$. Let ρ_δ be the histogram density estimator

$$\rho_\delta(x) = \frac{1}{n}\sum_{i=1}^M \frac{n_i}{|B_i|}\mathbb{1}_{B_i}(x),$$

where n_i is the number of points from $X_1, \dots, X_n$ that fall in B_i. We easily check that

$$\int_U \rho_\delta(x) \, dx = \frac{1}{n}\sum_{i=1}^M n_i = 1.$$

To construct the sets $U_1, U_2, \dots, U_n$, we construct a partition $B_{i,1}, B_{i,2}, \dots, B_{i,n_i}$ of each B_i consisting of sets of equal measure. Let $k_{i,1}, \dots, k_{i,n_i}$ denote the indices of the random variables that fall in B_i and set $U_{k_{i,j}} = B_{i,j}$. That is, the partition of B_i is assigned one-to-one with the random samples that fall in B_i. Then for some $1 \leq j \leq M$ and $1 \leq k \leq n_j$ we have

$$\int_{U_i} \rho_\delta(x)\,dx = \frac{n_j}{n|B_j|}\int_{B_{j,k}} dx = \frac{1}{n}.$$

Fix $1 \leq j \leq n$. Since $X_j \in B_i$ for some i, and $B_i \subset B(x_i, \delta/2)$, we have $|X_j - x_i| \leq \delta/2$, and so $U_j \subset B_i \subset B(x_i,\delta) \subset B(X_j, \delta)$. This proves (5.85).

To prove (5.86), note that each n_i has the form

$$n_i = \sum_{k=1}^n \mathbb{1}_{B_i}(X_k),$$

where $\mathbb{1}_{B_i}(X_k)$ is Bernoulli with parameter $p_i = \int_{B_i}\rho(x)\,dx$. By the Chernoff bounds (Theorem 5.7) we have

$$\mathbb{P}\left(|n_i - np_i| \geq \lambda n p_i\right) \leq 2\exp\left(-\frac{3}{8}np_i\lambda^2\right)$$

for $0 < \lambda \leq 1$. We note that

$$p_i = \int_{B_i}\rho(x)\,dx \geq \int_{B(x_i, c\delta)\cap U}\rho(x)\,dx \geq c\delta^d,$$

and

$$p_i = \int_{B_i}\rho(x)\,dx \leq C|B_i|.$$

Union bounding over $i = 1, \dots, M$, for any $0 < \lambda \leq 1$, with probability at least $1 - Cn\exp\left(-cn\delta^d\lambda^2\right)$, we have

$$\left|\frac{n_i}{n} - \int_{B_i}\rho(x)\,dx\right| \leq C\lambda|B_i| \tag{5.87}$$

for all $i = 1, \dots, M$, where we used that $n\delta^d \geq 1$. Let $x \in U$. Then $x \in B_i$ for some i and given that (5.87) holds we have

$$\begin{aligned}
|\rho(x) - \rho_\delta(x)| &= \left|\rho(x) - \frac{n_i}{n|B_i|}\right| \\
&= \left|\rho(x) - \frac{1}{|B_i|}\int_{B_i}\rho(x)\,dx + \frac{1}{|B_i|}\int_{B_i}\rho(x)\,dx - \frac{n_i}{n|B_i|}\right| \\
&\leq \left|\rho(x) - \frac{1}{|B_i|}\int_{B_i}\rho(x)\,dx\right| + \left|\frac{1}{|B_i|}\int_{B_i}\rho(x)\,dx - \frac{n_i}{n|B_i|}\right| \\
&\leq C(\delta + \lambda),
\end{aligned}$$

which completes the proof. □

Let $\delta > 0$ with $n\delta^d \geq 1$ and $0 < \lambda \leq 1$. Let $U_1, U_2, \ldots, U_n$ and ρ_δ be the partition and probability density provided by Lemma 5.31 satisfying (5.86) and (5.85), which exists with probability at least $1 - Cn\exp(-cn\delta^d\lambda^2)$. We define the extension operator $E_\delta : \ell^2(\mathcal{X}_n) \to L^2(U)$ by

$$E_\delta u(x) = \sum_{i=1}^n u(X_i)\mathbb{1}_{U_i}(x). \tag{5.88}$$

Since (5.85) holds we have

$$\frac{1}{n}\sum_{i=1}^n u(X_i) = \sum_{i=1}^n \int_{U_i} u(X_i)\rho_\delta(x)\,dx = \int_U (E_\delta u)\rho_\delta\,dx \tag{5.89}$$

for every $u \in \ell^2(\mathcal{X}_n)$. Hence, the extension operator allow us to relate empirical discrete sums with their continuum integral counterparts uniformly over $L^2(U)$ and $\ell^2(\mathcal{X}_n)$, without using concentration of measure as we did for pointwise consistency of graph Laplacians.

It will often be more be convenient in analysis to define the associated *transportation map* $T_\delta : U \to \mathcal{X}_n$, which is defined by $T_\delta(x) = X_i$ if and only if $x \in U_i$. Then the extension map $E_\delta : \ell^2(\mathcal{X}_n) \to L^2(U)$ can be written as $E_\delta u = u \circ T_\delta$. In this notation, (5.89) becomes

$$\frac{1}{n}\sum_{i=1}^n u(X_i) = \int_U u(T_\delta(x))\rho_\delta(x)\,dx. \tag{5.90}$$

Since $T_\delta(U_i) = X_i$ and $U_i \subset B(X_i, \delta)$, we have $|T_\delta(x) - x| \leq \delta$. This implies that

$$|T_\delta(x) - T_\delta(y)| \leq |x - y| + 2\delta \tag{5.91}$$

for all $x, y \in U$.

To see why T_δ is called a transportation map, note that $U_i = T_\delta^{-1}(\{X_i\})$, and so property (5.85) becomes

$$\int_{T^{-1}(\{X_i\})} \rho_\delta(x)\,dx = \frac{1}{n}.$$

We define the empirical distribution $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(x)$. Then for any $A \subset U$

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \in A}$$

and so

$$\mu_n(A) = \sum_{i=1}^n \mathbb{1}_{X_i \in A} \int_{T^{-1}(\{X_i\})} \rho_\delta(x)\,dx = \int_{T^{-1}(A)} \rho_\delta\,dx.$$

By definition, this means that T_δ pushes forward the probability density ρ_δ to the empirical distribution μ_n, written $T_{\delta\#}\rho_\delta = \mu_n$. In other words, the map T_δ is a transportation map, transporting the distribution ρ_δ to μ_n.

We can also extend this to double integrals. In the transportation map notation we have

$$\frac{1}{n^2}\sum_{i,j=1}^{n} \Phi(X_i, X_j) = \int_U \int_U \Phi(T_\delta(x), T_\delta(y))\rho_\delta(x)\rho_\delta(y)\,dx\,dy. \tag{5.92}$$

We are now ready to prove discrete to nonlocal estimates.

Lemma 5.32 (Discrete to nonlocal)**.** *Let $n \geq 1$ and $\delta > 0$ with $n\delta^d \geq 1$. There exists $c > 0$ such that for $0 < \lambda \leq 1$ with $\lambda + \delta \leq c$ the event that*

$$I_{\varepsilon,\delta}(E_\delta u) \leq (1 + C(\delta + \lambda))\mathcal{E}_{n,\varepsilon}(u) \tag{5.93}$$

holds for all $u \in \ell^2(\mathcal{X}_n)$ has probability at least $1 - Cn\exp(-cn\delta^d\lambda^2)$.

Proof. Let $u_\delta = E_\delta u = u \circ T_\delta$. Applying (5.92) we have

$$\begin{aligned}
\mathcal{E}_{n,\varepsilon}(u) &= \frac{1}{\sigma_\eta n^2 \varepsilon^2} \sum_{x,y\in\mathcal{X}_n} \eta_\varepsilon(|x-y|)(u(x)-u(y))^2 \\
&= \frac{1}{\sigma_\eta \varepsilon^2}\int_U\int_U \eta_\varepsilon\left(|T_\delta(x) - T_\delta(y)|\right)\left(u_\delta(x) - u_\delta(y)\right)^2 \rho_\delta(x)\rho_\delta(y)\,dy\,dx \\
&\geq \frac{1}{\sigma_\eta \varepsilon^2}\int_U\int_U \eta_\varepsilon\left(|x-y| + 2\delta\right)\left(u_\delta(x) - u_\delta(y)\right)^2 \rho_\delta(x)\rho_\delta(y)\,dy\,dx,
\end{aligned}$$

due to (5.91) and the assumptions that $t \mapsto \eta(t)$ is nonincreasing. Since $|\rho - \rho_\delta| \leq C(\delta + \lambda)$ and $\rho \geq \alpha > 0$, we have that

$$\rho_\delta \geq \rho - C(\delta + \lambda) \geq \rho(1 - C\alpha^{-1}(\delta + \lambda)),$$

for $\lambda + \delta$ sufficiently small so that $C\alpha^{-1}(\delta + \lambda) \leq 1/2$. Therefore

$$\mathcal{E}_{n,\varepsilon}(u) \geq (1 - C\alpha^{-1}(\delta + \lambda))I_{\varepsilon,\delta}(u_\delta),$$

which, upon rearrangement, completes the proof. □

Notes and references: This material in this section follows [14] closely. The idea of using transportation maps to compare discrete and continuum measures originally appeared in [61], and have been used in many subsequent works by the same authors. The method in [61] looks for a transportation map T_δ that pushes ρ directly onto μ_n; that is, we take $\rho = \rho_\delta$ in Lemma 5.31. It is far more challenging to prove existence of such a transportation map, and is related to hard problems in optimal matching in probability. The analogous result to Lemma 5.31 with $\rho = \rho_\delta$ can be found in [60]. The convergence rates for discrete to continuum can be made slightly sharper using [60, Theorem 1.1] in place of Lemma 5.31, although in dimension $n = 2$, the length scale restriction becomes $n\varepsilon^2 \gg \log(n)^{3/2}$, which is slightly suboptimal.

5.5.3 Nonlocal to local estimates

We must now estimate the local energy $I(u)$ in terms of the nonlocal energy $I_{\varepsilon,\delta}(u)$, in a way that does not require regularity of u. In general this is impossible, since u may be discontinuous and so $I(u) = \infty$ while $I_{\varepsilon,\delta}(u)$ is finite. However, if we apply a small amount of smoothing to u in just the right way, we can construct a smoothed version of u for which $I(u)$ is appropriately bounded.

Let us define

$$\psi_{1,t}(\tau) = \frac{1}{\sigma_{\eta,t}} \int_\tau^\infty \eta(s+2t)s\,ds, \tag{5.94}$$

where $\sigma_{\eta,t} = \int_{\mathbb{R}^d} |z_1|^2\eta(|z|+2t)\,dz$, and $\psi_{\varepsilon,\delta}(\tau) = \frac{1}{\varepsilon^d}\psi_{1,\delta/\varepsilon}(\tau/\varepsilon)$. Notice that

$$\psi_{\varepsilon,\delta}(\tau) = \frac{1}{\sigma_{\eta,\delta/\varepsilon}\varepsilon^2} \int_\tau^\infty \eta_\varepsilon(s+2\delta)s\,ds.$$

We could have taken the above as a definition of $\psi_{\varepsilon,\delta}$. We also note that $\sigma_\eta = \sigma_{\eta,0}$ and

$$\sigma_\eta - \frac{2\alpha(n)\mathrm{Lip}(\eta)}{d+2}t \le \sigma_{\eta,t} \le \sigma_\eta, \tag{5.95}$$

since η is Lipschitz and nonincreasing. Also, if $0 \le t \le 1/4$ we have

$$\sigma_{\eta,t} \ge \int_{\mathbb{R}^d} |z_1|^2\eta(|z|+\tfrac{1}{2})\,dz \ge \int_{B(0,1/2)} |z_1|^2\,dz = \frac{\alpha(n)}{2^{d+2}(d+2)}. \tag{5.96}$$

since $\eta \ge 1$ on $[0,1/2]$.

We now define the smoothing operator $\Lambda_{\varepsilon,\delta} : L^2(U) \to C^\infty(U)$ by

$$\Lambda_{\varepsilon,\delta}u(x) = \frac{1}{\theta_{\varepsilon,\delta}(x)} \int_U \psi_{\varepsilon,\delta}\left(|x-y|\right)u(y)\,dy, \tag{5.97}$$

where

$$\theta_{\varepsilon,\delta}(x) = \int_U \psi_{\varepsilon,\delta}(|x-y|)\,dy. \tag{5.98}$$

We first have some preliminary propositions

Proposition 5.33. *For all $\varepsilon,\delta > 0$ with $\varepsilon > 2\delta$ and $x \in U$ with dist$(x,\partial U) \ge \varepsilon - 2\delta$ we have $\theta_{\varepsilon,\delta}(x) = 1$.*

Proof. We simply compute

$$\begin{aligned}\theta_{\varepsilon,\delta}(x) = \int_U \psi_{\varepsilon,\delta}(|x-y|)\,dy &= \int_{B(0,1)} \psi_{1,\delta/\varepsilon}(|z|)\,dz \\ &= \int_0^1 n\alpha(n)\tau^{n-1}\psi_{1,\delta/\varepsilon}(\tau)\,d\tau \\ &= \frac{1}{\sigma_{\eta,t}}\int_0^1 n\alpha(n)\tau^{n-1}\int_\tau^1 \eta(s+\tfrac{2\delta}{\varepsilon})s\,ds\,d\tau \\ &= \frac{1}{\sigma_{\eta,t}}\int_0^1 \alpha(n)\eta(s+\tfrac{2\delta}{\varepsilon})s\int_0^s n\tau^{n-1}\,d\tau\,ds \\ &= \frac{1}{\sigma_{\eta,t}}\int_0^1 \alpha(n)s^{n+1}\eta(s+\tfrac{2\delta}{\varepsilon})ds \\ &= \frac{1}{n\sigma_{\eta,t}}\int_{B(0,1)} \eta(|z|+\tfrac{2\delta}{\varepsilon})|z|^2\,dz = 1. \qquad \square\end{aligned}$$

To prove nonlocal to local convergence, we define a more general nonlocal operator

$$I_{\varepsilon,\delta}(u;V) = \frac{1}{\sigma_\eta\varepsilon^2}\int_V\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)(u(x)-u(y))^2\rho(x)\rho(y)\,dx\,dy, \tag{5.99}$$

for $V \subset U$, which will simplify notation. As before, we write $I_{\varepsilon,\delta}(u) = I_{\varepsilon,\delta}(u;U)$. We first prove basic estimates on $\Lambda_{\varepsilon,\delta}$.

Proposition 5.34. *Let $u \in L^2(U)$ and $\varepsilon,\delta > 0$ with $\varepsilon \geq 4\delta$. Then*

$$\|\Lambda_{\varepsilon,\delta}u - u\|^2_{L^2(U_{\varepsilon-2\delta})} \leq C\sigma_\eta I_{\varepsilon,\delta}(u)\varepsilon^2. \tag{5.100}$$

Proof. First, we note that by monotonicity of η we have

$$\psi_{1,t}(\tau) = \frac{1}{\sigma_{\eta,t}}\int_\tau^{1-2t}\eta(s+2t)s\,ds \leq \frac{1}{\sigma_{\eta,t}}\int_0^1 \eta(\tau+2t)\,d\tau = \frac{1}{\sigma_{\eta,t}}\eta(\tau+2t),$$

and so

$$\psi_{\varepsilon,\delta}(\tau) = \frac{1}{\varepsilon^d}\psi_{1,\delta/\varepsilon}(\tau/\varepsilon) \leq \frac{1}{\sigma_{\eta,\delta/\varepsilon}\varepsilon^d}\eta\left(\tfrac{\tau}{\varepsilon}+2\tfrac{\delta}{\varepsilon}\right) = \frac{1}{\sigma_{\eta,\delta/\varepsilon}}\eta_\varepsilon(\tau+2\delta).$$

By (5.96) we have $\sigma_{\eta,\delta/\varepsilon} \geq c$, where $c > 0$ depends only on d, since $\varepsilon \geq 4\delta$. Therefore, for any $x \in U$ we have

$$\begin{aligned}|\Lambda_{\varepsilon,\delta}u(x) - u(x)|^2 &\leq \left(\frac{1}{\theta_{\varepsilon,\delta}(x)}\int_U \psi_{\varepsilon,\delta}(|x-y|)|u(y)-u(x)|\,dy\right)^2 \\ &\leq \frac{1}{\theta_{\varepsilon,\delta}(x)}\int_U \psi_{\varepsilon,\delta}(|x-y|)|u(y)-u(x)|^2\,dy \\ &\leq \frac{1}{c\theta_{\varepsilon,\delta}(x)}\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)(u(y)-u(x))^2\,dx.\end{aligned}$$

Therefore

$$\|\theta_{\varepsilon,\delta}(\Lambda_{\varepsilon,\delta}u - u)\|^2_{L^2(U)} \leq \frac{1}{c}\int_U\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)(u(y)-u(x))^2\,dxdy \leq \frac{\sigma_\eta}{c}I_{\varepsilon,\delta}(u)\varepsilon^2.$$

Applying Proposition 5.33 completes the proof. □

We now turn to comparing the local energy $I(u)$ to the nonlocal energy $I_{\varepsilon,\delta}(u)$.

Lemma 5.35. *There exists $C > 0$ such that for every $u \in L^2(U)$ and $\varepsilon, \delta > 0$ with $\varepsilon \geq 4\delta$ we have*

$$\int_U |\nabla\Lambda_{\varepsilon,\delta}u|^2\rho^2\theta^2_{\varepsilon,\delta}\,dx \leq \frac{\sigma_\eta(1+C\varepsilon)}{\sigma_{\eta,\delta/\varepsilon}}(I_{\varepsilon,\delta}(u)+I_{\varepsilon,\delta}(u;\partial_{\varepsilon-2\delta}U))+\frac{C}{\varepsilon^2}\|\Lambda_{\varepsilon,\delta}u-u\|^2_{L^2(\partial_{\varepsilon-2\delta}U)}.$$

Proof. Let $w = \Lambda_{\varepsilon,\delta}u$ and note that

$$\nabla w(x) = \frac{1}{\theta_{\varepsilon,\delta}(x)}\nabla v(x) - \frac{w(x)}{\theta_{\varepsilon,\delta}(x)}\nabla\theta_{\varepsilon,\delta}(x),$$

where

$$v(x) = \int_U \psi_{\varepsilon,\delta}\left(|x-y|\right)u(y)\,dy.$$

We now compute

$$\begin{aligned}\nabla v(x) &= \frac{1}{\varepsilon^d}\int_U \psi'_{1,\delta/\varepsilon}\left(\frac{|x-y|}{\varepsilon}\right)\frac{(x-y)}{\varepsilon|x-y|}u(y)\,dy\\ &= \frac{1}{\sigma_{\eta,\delta/\varepsilon}\varepsilon^{d+2}}\int_U \eta\left(\frac{|x-y|+2\delta}{\varepsilon}\right)(y-x)u(y)\,dy\\ &= \frac{1}{\sigma_{\eta,\delta/\varepsilon}\varepsilon^2}\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)(y-x)u(y)\,dy,\end{aligned}$$

and by a similar computation

$$\nabla\theta_{\varepsilon,\delta}(x) = \frac{1}{\sigma_{\eta,\delta/\varepsilon}\varepsilon^2}\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)(y-x)\,dy.$$

Therefore

(5.101)

$$\begin{aligned}\nabla w(x) = \frac{1}{\sigma_{\eta,\delta/\varepsilon}\theta_{\varepsilon,\delta}(x)\varepsilon^2}\Bigg[&\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)(y-x)(u(y)-u(x))\,dy\\ &+(u(x)-w(x))\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)(y-x)\,dy\Bigg].\end{aligned}$$

Let $\xi \in \mathbb{R}^d$ with $|\xi| = 1$ so that $\nabla w(x) \cdot \xi = |\nabla w(x)|$. Then we have

(5.102)
$$\begin{aligned}|\nabla w(x)| = \frac{1}{\sigma_{\eta,\delta/\varepsilon}\theta_{\varepsilon,\delta}(x)\varepsilon^2}\Bigg[\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)((y-x)\cdot\xi)(u(y)-u(x))\,dy \\ + (u(x)-w(x))\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)(y-x)\cdot\xi\,dy\Bigg].\end{aligned}$$

By the Cauchy-Schwarz inequality we have

$$\begin{aligned}&\left|\frac{1}{\sigma_{\eta,\delta/\varepsilon}\theta_{\varepsilon,\delta}(x)\varepsilon^2}\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)((y-x)\cdot\xi)(u(y)-u(x))\,dy\right| \\ &\qquad\le \frac{1}{\sigma_{\eta,\delta/\varepsilon}\theta_{\varepsilon,\delta}(x)\varepsilon^2}\left(\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)|(x-y)\cdot\xi|^2\,dy\right)^{1/2} \\ &\qquad\qquad\times\left(\int_U \eta_\varepsilon(|x-y|+2\delta)(u(y)-u(x))^2\,dy\right)^{1/2} \\ &\qquad\le \frac{1}{\sqrt{\sigma_{\eta,\delta/\varepsilon}}\theta_{\varepsilon,\delta}(x)\varepsilon}\left(\int_U \eta_\varepsilon(|x-y|+2\delta)(u(y)-u(x))^2\,dy\right)^{1/2}.\end{aligned}$$

Similarly, we have

$$\left|\frac{1}{\sigma_{\eta,\delta/\varepsilon}\theta_{\varepsilon,\delta}(x)\varepsilon^2}\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)(y-x)\cdot\xi\,dy\right| \le \frac{C}{\varepsilon}\mathbb{1}_{x\in\partial_{\varepsilon-2\delta}U},$$

since

$$\int_U \eta_\varepsilon\left(|x-y|+2\delta\right)(y-x)\cdot\xi\,dy = 0$$

if $\operatorname{dist}(x,\partial U) \ge \varepsilon - 2\delta$. Therefore

$$\begin{aligned}|\nabla w(x)| \le \frac{1}{\sqrt{\sigma_{\eta,\delta/\varepsilon}}\theta_{\varepsilon,\delta}(x)\varepsilon}\left(\int_U \eta_\varepsilon(|x-y|+2\delta)(u(y)-u(x))^2\,dy\right)^{1/2} \\ + \frac{C}{\varepsilon}|w(x)-u(x)|\mathbb{1}_{\partial_{\varepsilon-2\delta}U(x)}.\end{aligned}$$

Squaring and integrating against $\rho^2\theta_{\varepsilon,\delta}^2$ over U we have

$$\begin{aligned}\int_U |\nabla w|^2\rho^2\theta_{\varepsilon,\delta}^2\,dx \le \frac{1}{\sigma_{\eta,\delta/\varepsilon}\varepsilon^2}\int_U\int_U \eta_\varepsilon(|x-y|+2\delta)(u(y)-u(x))^2\rho(x)^2\,dxdy \\ + \frac{1}{\sigma_{\eta,\delta/\varepsilon}\varepsilon^2}\int_{\partial_{\varepsilon-2\delta}U}\int_U \eta_\varepsilon(|x-y|+2\delta)(u(y)-u(x))^2\rho(x)^2\,dydx \\ + \frac{C}{\varepsilon^2}\int_{\partial_{\varepsilon-2\delta}U}(w(x)-u(x))^2\,dx,\end{aligned}$$

where we applied Cauchy's inequality $2ab \leq a^2 + b^2$ to the cross terms. The proof is completed by noting that ρ is Lipschitz continuous and $\rho \geq \alpha > 0$, and so $\rho(x) \leq \rho(y)(1 + C\varepsilon)$ for $y \in B(x, \varepsilon)$. □

We can sharpen Lemma 5.35 if we have some information about the regularity of u near the boundary.

Theorem 5.36 (Nonlocal to local)**.** *Suppose that $u \in L^2(U)$ satisfies*

$$|u(x) - u(y)| \leq C\varepsilon \tag{5.103}$$

for all $x \in \partial_{\varepsilon-2\delta}U$ and $y \in B(x, \varepsilon - 2\delta) \cap U$. Then for $\varepsilon > 0$ and $\delta \leq c\varepsilon$, with $0 < c \leq 1/4$ depending only on η and n, we have

$$I(\Lambda_{\varepsilon,\delta}u) \leq \left(1 + C\left(\varepsilon + \tfrac{\delta}{\varepsilon}\right)\right) I_{\varepsilon,\delta}(u) + C\left(\varepsilon + \tfrac{\delta}{\varepsilon}\right). \tag{5.104}$$

Proof. For any $x \in \partial_{\varepsilon-2\delta}U$ we have by (5.103) that

$$\begin{aligned}|(\Lambda_{\varepsilon,\delta}u)(x) - u(x)| &\leq \frac{1}{\theta_{\varepsilon,\delta}(x)} \int_{B(x,\varepsilon-2\delta)\cap U} \psi_{\varepsilon,\delta}(|x-y|)|u(y) - u(x)|\,dx \\ &\leq \frac{C\varepsilon}{\theta_{\varepsilon,\delta}(x)} \int_{B(x,\varepsilon-2\delta)\cap U} \psi_{\varepsilon,\delta}(|x-y|)\,dx = C\varepsilon.\end{aligned}$$

Since $2\delta < \varepsilon$ we have

$$\|\Lambda_{\varepsilon,\delta}u - u\|^2_{L^2(\partial_{\varepsilon-2\delta}U)} \leq C\varepsilon^2|\partial_\varepsilon U| \leq C\varepsilon^3. \tag{5.105}$$

We again use (5.103) to obtain

$$\begin{aligned}I_{\varepsilon,\delta}(u; \partial_{\varepsilon-2\delta}U) &= \frac{1}{\sigma_\eta\varepsilon^2} \int_{\partial_{\varepsilon-2\delta}U} \int_{B(x,\varepsilon-2\delta)\cap U} \eta_\varepsilon(|x-y| + 2\delta)(u(x) - u(y))^2\rho(x)\rho(y)\,dxdy \\ &\leq C \int_{\partial_{\varepsilon-2\delta}U} \int_{B(x,\varepsilon-2\delta)\cap U} \eta_\varepsilon(|x-y| + 2\delta)\,dxdy \leq C|\partial_\varepsilon U| \leq C\varepsilon.\end{aligned}$$

By Proposition 5.33 we have $\theta_{\varepsilon,\delta} = 1$ on $U_{\varepsilon-2\delta}$. Combining these facts with (5.105) and Lemma 5.35 we have

$$\int_{U_{\varepsilon-2\delta}} |\nabla\Lambda_{\varepsilon,\delta}u|^2\rho^2\,dx \leq \frac{\sigma_\eta(1 + C\varepsilon)}{\sigma_{\eta,\delta/\varepsilon}}(I_{\varepsilon,\delta}(u) + C\varepsilon) + C\varepsilon.$$

Invoking (5.95), for $\delta \leq c\varepsilon$, with $0 < c \leq 1/4$ depending only on η and n, we have

$$\int_{U_{\varepsilon-2\delta}} |\nabla\Lambda_{\varepsilon,\delta}u|^2\rho^2\,dx \leq \left(1 + C\left(\varepsilon + \tfrac{\delta}{\varepsilon}\right)\right)(I_{\varepsilon,\delta}(u) + C\varepsilon) + C\varepsilon.$$

By (5.102) and (5.103) we have $|\nabla\Lambda_{\varepsilon,\delta}u(x)| \leq C$ for $x \in \partial_{\varepsilon-2\delta}U$, and so

$$\int_{\partial_{\varepsilon-2\delta}U} |\nabla\Lambda_{\varepsilon,\delta}u|^2\rho^2\,dx \leq C|\partial_\varepsilon U| \leq C\varepsilon,$$

which completes the proof. □

We also need estimates in the opposite direction, bounding the nonlocal energy $I_\varepsilon(u)$ by the local energy $I(u)$. While a similar result was already established in Lemma 5.29 under the assumption $u \in C^2(U)$, we need a result that holds for Lipschitz u to prove discrete to continuum convergence.

Lemma 5.37 (Local to nonlocal)**.** *There exists $C > 0$ such that for every Lipschitz continuous $u : U \to \mathbb{R}$ and $\varepsilon > 0$ we have*

$$I_\varepsilon(u) \leq (1 + C\varepsilon)I(u) + C Lip(u)^2\varepsilon. \tag{5.106}$$

Proof. We write

$$I_\varepsilon(u) = I_\varepsilon(u; U_\varepsilon) + I_\varepsilon(u; \partial_\varepsilon U)$$

Note that

$$(u(x) - u(y))^2 = \left(\int_0^1 \frac{d}{dt} u(x + t(y - x))\, dt\right)^2 \leq \int_0^1 |\nabla u(x + t(y - x)) \cdot (y - x)|^2\, dt,$$

where we used Jensen's inequality in the final step. Since ρ is Lipschitz and bounded below, we have $\rho(y) \leq \rho(x)(1 + C\varepsilon)$. Plugging these into the definition of $I_\varepsilon(u; U_\varepsilon)$ we have

$$I_\varepsilon(u; U_\varepsilon) \leq \frac{1 + C\varepsilon}{\sigma_\eta \varepsilon^2} \int_0^1 \int_{U_\varepsilon} \int_{B(x,\varepsilon)} \eta_\varepsilon(|x - y|)|\nabla u(x + t(y - x)) \cdot (y - x)|^2\, dy\, \rho(x)^2 dx\, dt.$$

Now make the change of variables $z = y - x$ in the inner integral to find that

$$\begin{aligned}
I_\varepsilon(u; U_\varepsilon) &\leq \frac{1 + C\varepsilon}{\sigma_\eta \varepsilon^2} \int_0^1 \int_{U_\varepsilon} \int_{B(0,\varepsilon)} \eta_\varepsilon(|z|)|\nabla u(x + tz) \cdot z|^2\, dz\, \rho(x)^2 dx\, dt \\
&= \frac{1 + C\varepsilon}{\sigma_\eta \varepsilon^2} \int_0^1 \int_{B(0,\varepsilon)} \eta_\varepsilon(|z|) \int_{U_\varepsilon} |\nabla u(x + tz) \cdot z|^2\, \rho(x)^2 dx\, dz\, dt \\
&\leq \frac{1 + C\varepsilon}{\sigma_\eta \varepsilon^2} \int_0^1 \int_{B(0,\varepsilon)} \eta_\varepsilon(|z|) \int_U |\nabla u(x) \cdot z|^2\, \rho(x)^2 dx\, dz\, dt \\
&= \frac{1 + C\varepsilon}{\sigma_\eta \varepsilon^2} \int_U \int_{B(0,\varepsilon)} \eta_\varepsilon(|z|)|\nabla u(x) \cdot z|^2\ dz\, \rho(x)^2 dx \\
&= (1 + C\varepsilon) \int_U |\nabla u|^2 \rho(x)^2 dx = (1 + C\varepsilon)I(u).
\end{aligned}$$

On the other hand we have

$$\begin{aligned}
I_\varepsilon(u; \partial_\varepsilon U) &= \frac{1}{\sigma_\eta \varepsilon^2} \int_{\partial_\varepsilon U} \int_U \eta_\varepsilon(|x - y|)(u(x) - u(y))^2 \rho(x)\rho(y)\, dxdy \\
&\leq C\mathrm{Lip}(u)^2 \int_{\partial_\varepsilon U} \int_U \eta_\varepsilon(|x - y|)\rho(x)\rho(y)\, dx\, dy \\
&\leq C\mathrm{Lip}(u)^2 \varepsilon.
\end{aligned}$$

This completes the proof. □

Notes and references: The material in this section roughly follows [58], though we have made some modifications to simplify the presentation. In particular, our smoothing operator $\Lambda_{\varepsilon,\delta}$ is different than the one in [58], which essentially uses $\Lambda_{\varepsilon,0}$ everywhere. Some of the core arguments in [58] appeared earlier in [10], which considered the non-random setting with constant kernel η. The spectral convergence rates from [10, 58] were recently improved in [14] by incorporating pointwise consistency of graph Laplacians.

5.5.4 Discrete to continuum

We now use the results from the previous sections to prove discrete to continuum convergence with rates for Laplacian learning via variational tools.

Let $g : U \to \mathbb{R}$ be Lipschitz continuous and define

$$A_{n,\varepsilon} = \{u \in \ell^2(\mathcal{X}_n) \, : \, u(x) = g(x) \text{ for all } x \in \partial_{2\varepsilon} U\}, \tag{5.107}$$

and

$$A = \{u \in H^1(U) \, : \, u = g \text{ on } \partial U\}. \tag{5.108}$$

We consider the discrete variational problem

$$\min_{u \in \mathcal{A}_{n,\varepsilon}} \mathcal{E}_{n,\varepsilon}(u) \tag{5.109}$$

and its continuum counterpart

$$\min_{u \in \mathcal{A}} I(u), \tag{5.110}$$

where we recall $\mathcal{E}_{n,\varepsilon}$ is defined in (5.53) and $I(u) = \int_U |\nabla u|^2 \rho^2 \, dx$.

To prove discrete to continuum convergence with rates, we require stability for the limiting problem.

Proposition 5.38 (Stability)**.** *Let $u \in C^2(\overline{U})$ such that $div(\rho^2 \nabla u) = 0$ in U. There exists $C > 0$, depending only on U, such that for all $w \in H^1(U)$*

$$\|u - w\|^2_{H^1(U)} \leq C \left(I(w) - I(u) + \|u\|_{C^1(U)} \|u - w\|_{L^2(\partial U)} + \|u - w\|^2_{L^2(\partial U)} \right). \tag{5.111}$$

Proof. Write $R = I(w) - I(u)$ so that $I(w) = I(u) + R$. Then we compute

$$\begin{aligned}
\alpha^2 \int_U |\nabla u - \nabla w|^2\,dx &\leq \int_U |\nabla u - \nabla w|^2 \rho^2\,dx \\
&\leq \int_U |\nabla u|^2\rho^2 - 2\rho^2 \nabla w \cdot \nabla u + |\nabla w|^2 \rho^2\,dx \\
&= I(u) + I(w) - 2\int_U \rho^2 \nabla w \cdot \nabla u\,dx \\
&= 2I(u) - 2\int_U \rho^2 \nabla w \cdot \nabla u\,dx + R \\
&= 2\int_U \rho^2 \nabla u \cdot \nabla(u - w)\,dx + R \\
&= 2\int_{\partial U} (u - w)\rho^2 \frac{\partial u}{\partial \nu}\,dS + R \\
&\leq C\alpha^{-2}\|u\|_{C^1(U)}\|u - w\|_{L^2(\partial U)} + R,
\end{aligned}$$

where we used integration by parts in the second to last line. We now use the Poincaré inequality proved in Exercise 4.18 to obtain

$$\int_U (u - w)^2\,dx \leq C\left(\int_U |\nabla u - \nabla w|^2\,dx + \int_{\partial U} (u - w)^2\,dx\right).$$

Combining this with the bound above completes the proof. □

Remark 5.39. We call Proposition 5.38 a *stability* estimate, since it shows that if u is a minimizer of I, and w is an approximate minimizer, so that $I(w) - I(u)$ is small, then u and w are close in $H^1(U)$, provided they are close on the boundary ∂U.

We also prove stability for the discrete graph problem. For the reader's reference, we refer to Section 5.3 for definitions of gradients and divergence operators on graphs.

Proposition 5.40 (Stability)**.** *Let $u \in \ell^2(\mathcal{X}_n)$ such that $\mathcal{L}_{n,\varepsilon}u(x) = 0$ for all $x \in \mathcal{X}_n \setminus \Gamma$, where $\Gamma \subset \mathcal{X}_n$. For all $w \in \ell^2(\mathcal{X}_n)$ with $w(x) = u(x)$ for all $x \in \Gamma$ we have*

$$\|\nabla_{n,\varepsilon} u - \nabla_{n,\varepsilon} w\|^2_{\ell^2(\mathcal{X}_n^2)} = \mathcal{E}_{n,\varepsilon}(w) - \mathcal{E}_{n,\varepsilon}(u). \tag{5.112}$$

Proof. The proof is a discrete analog of Proposition 5.38. Write $R = \mathcal{E}_{n,\varepsilon}(w) - \mathcal{E}_{n,\varepsilon}(u)$ and compute

$$\begin{aligned}
\|\nabla_{n,\varepsilon} u - \nabla_{n,\varepsilon} w\|^2_{\ell^2(\mathcal{X}_n^2)} &= \|\nabla_{n,\varepsilon} u\|^2_{\ell^2(\mathcal{X}_n^2)} - 2(\nabla_{n,\varepsilon} u, \nabla_{n,\varepsilon} w)_{\ell^2(\mathcal{X}_n^2)} + \|\nabla_{n,\varepsilon} w\|^2_{\ell^2(\mathcal{X}_n^2)} \\
&= 2(\|\nabla_{n,\varepsilon} u\|^2_{\ell^2(\mathcal{X}_n^2)} - (\nabla_{n,\varepsilon} u, \nabla_{n,\varepsilon} w)_{\ell^2(\mathcal{X}_n^2)}) + R \\
&= 2(\nabla_{n,\varepsilon} u, \nabla_{n,\varepsilon} u - \nabla_{n,\varepsilon} w)_{\ell^2(\mathcal{X}_n^2)} + R = 0 \\
&= -2(\mathcal{L}_{n,\varepsilon} u, u - w)_{\ell^2(\mathcal{X}_n)} + R = R,
\end{aligned}$$

since $\mathcal{L}_{n,\varepsilon} u = 0$ on $\mathcal{X}_n \setminus \Gamma$ and $u = w$ on Γ. □

We can now prove discrete to continuum convergence.

Theorem 5.41 (Discrete to continuum)**.** *Let $u \in \mathcal{A}$ be the solution of (5.110) and let $u_{n,\varepsilon} \in \mathcal{A}_{n,\varepsilon}$ be the solution of (5.109). There exists $C, c > 0$ such that*

$$\|u - u_{n,\varepsilon}\|^2_{H^1(\mathcal{X}_n)} \le C(\|u\|_{C^2(U)} + 1)(\varepsilon + \tfrac{\delta}{\varepsilon} + \lambda) \tag{5.113}$$

holds with probability at least $1 - Cn\exp\left(-cn\delta^d\lambda^2\right)$ for $\delta \le c\varepsilon$ and $\lambda > 0$.

Proof. The proof is split into several steps.

1. First, note that by elliptic regularity $u \in C^3(\overline{U})$, since $\rho \in C^2(\overline{U})$ and ∂U is smooth. Let $d(x) = \operatorname{dist}(x, \partial U)$ and define

$$\varphi(t) = \begin{cases} 1, & \text{if } t \le 1 \\ 2 - t, & \text{if } 1 < t \le 2 \\ 0, & \text{if } t > 2, \end{cases}$$

and

$$u_\varepsilon(x) = u(x) + \varphi\left(\tfrac{d(x)}{2\varepsilon}\right)(g(x) - u(x)).$$

Since d and φ are Lipschitz continuous, u_ε is Lipschitz. Since u and g are Lipschitz and $u = g$ on ∂U, we also have $|g(x) - u(x)| \le C\varepsilon$ for $x \in \partial_{4\varepsilon}U$. Recalling $|\nabla d| = 1$ a.e., we have

$$\begin{aligned}
|\nabla u_\varepsilon(x)| &= \left|\nabla u(x) + \varphi'\left(\tfrac{d(x)}{2\varepsilon}\right)\frac{\nabla d(x)}{2\varepsilon}(g(x) - u(x)) + \varphi\left(\tfrac{d(x)}{2\varepsilon}\right)(\nabla g(x) - \nabla u(x)\right| \\
&\le |\nabla u(x)| + \frac{1}{2\varepsilon}\varphi'\left(\tfrac{d(x)}{2\varepsilon}\right)|g(x) - u(x)| + |\nabla g(x) - \nabla u(x)| \\
&\le 2|\nabla u(x)| + |\nabla g(x)| + \frac{1}{2\varepsilon}\|g - u\|_{L^\infty(\partial_{4\varepsilon}U)} \\
&\le 2|\nabla u(x)| + |\nabla g(x)| + \frac{1}{2}C.
\end{aligned}$$

Therefore $|\nabla u_\varepsilon|$ is bounded. Since $u_\varepsilon = u$ on $U_{4\varepsilon}$ we have

$$I(u_\varepsilon) - I(u) = \int_U |\nabla u_\varepsilon|^2\rho^2\,dx - \int_U |\nabla u|^2\rho^2\,dx \le \int_{\partial_{4\varepsilon}U} |\nabla u_\varepsilon|^2\rho^2\,dx \le C\varepsilon. \tag{5.114}$$

Now, since $|\nabla u_\varepsilon| \le C$ a.e., u_ε is Lipschitz continuous with $\operatorname{Lip}(u)$ bounded independently of ε. By Lemma 5.28 we have that

$$|\mathcal{E}_{n,\varepsilon}(u_\varepsilon) - I_\varepsilon(u_\varepsilon)| \le C\left(\frac{1}{n} + \lambda\right)$$

with probability at least $1 - 2\exp\left(-cn\varepsilon^d\lambda^2\right)$. By Lemma 5.37 and (5.114) we have

$$I_\varepsilon(u_\varepsilon) \le (1 + C\varepsilon)I(u_\varepsilon) + C\varepsilon \le (1 + C\varepsilon)I(u) + C\varepsilon.$$

Since $u_\varepsilon = g$ on $\partial_{2\varepsilon} U$, the restriction of u_ε to the graph belongs to $A_{n,\varepsilon}$ and so $\mathcal{E}_{n,\varepsilon}(u_{n,\varepsilon}) \leq \mathcal{E}_{n,\varepsilon}(u_\varepsilon)$. This yields

$$\mathcal{E}_{n,\varepsilon}(u_{n,\varepsilon}) \leq (1 + C\varepsilon) I(u) + C(\varepsilon + \lambda) \tag{5.115}$$

with probability at least $1 - 2\exp\left(-cn\varepsilon^d\lambda^2\right)$, where we used $n\varepsilon^d \geq 1$ so $1/n \leq \varepsilon$.

2. We now establish a bound in the opposite direction to (5.115). Let $\delta > 0$ and let $E_\delta : \ell^2(\mathcal{X}_n) \to L^2(U)$ be the extension map, and $T_\delta : U \to \mathcal{X}_n$ the corresponding transportation map, provided by Lemma 5.31 and the discussion thereafter. Recall that $E_\delta u = u \circ T_\delta$. Let $x \in \partial_{\varepsilon - 2\delta} U$ and $y \in B(x, \varepsilon - 2\delta)$. Since $|T_\delta(x) - x| \leq \delta$ we have that $T_\delta(x) \in \partial_{\varepsilon-\delta} U$ and $T_\delta(y) \in B(x, \varepsilon - \delta)$. Since $u_{n,\varepsilon} = g$ on $\mathcal{X}_n \cap \partial_{2\varepsilon} U$

$$\begin{aligned}
|E_\delta u_{n,\varepsilon}(x) - E_\delta u_{n,\varepsilon}(y)| &= |g(T_\delta(x)) - g(T_\delta(y))| \\
&\leq \mathrm{Lip}(g)|T_\delta(x) - T_\delta(y)| \\
&\leq \mathrm{Lip}(g)(|T_\delta(x) - x| + |x - y| + |T_\delta(y) - y|) \\
&\leq \mathrm{Lip}(g)(\delta + \varepsilon + \delta) \leq 3\mathrm{Lip}(g)\varepsilon,
\end{aligned}$$

since $\delta \leq \varepsilon$. Thus, we can invoke Theorem 5.36 to obtain

$$I(\Lambda_{\varepsilon,\delta} E_\delta u_{n,\varepsilon}) \leq \left(1 + C\left(\varepsilon + \tfrac{\delta}{\varepsilon}\right)\right) I_{\varepsilon,\delta}(E_\delta u_{n,\varepsilon}) + C\left(\varepsilon + \tfrac{\delta}{\varepsilon}\right).$$

By Lemma 5.32 we have

$$I_{\varepsilon,\delta}(E_\delta u_{n,\varepsilon}) \leq (1 + C(\delta + \lambda))\mathcal{E}_{n,\varepsilon}(u_{n,\varepsilon}), \tag{5.116}$$

with probability at least $1 - Cn\exp\left(-cn\delta^d\lambda^2\right)$. Combining these two inequalities yields

$$I(\Lambda_{\varepsilon,\delta} E_\delta u_{n,\varepsilon}) \leq (1 + C(\varepsilon + \tfrac{\delta}{\varepsilon} + \lambda))\mathcal{E}_{n,\varepsilon}(u_{n,\varepsilon}) + C\left(\varepsilon + \tfrac{\delta}{\varepsilon}\right). \tag{5.117}$$

3. Combining (5.115) and (5.117) we have

$$I(\Lambda_{\varepsilon,\delta} E_\delta u_{n,\varepsilon}) - I(u) \leq C(\varepsilon + \tfrac{\delta}{\varepsilon} + \lambda), \tag{5.118}$$

with probability at least $1 - Cn\exp\left(-cn\delta^d\lambda^2\right)$, where $\delta \leq c\varepsilon$. Let $x \in \partial U$. As in the proof of Theorem 5.36, we have

$$|(\Lambda_{\varepsilon,\delta} E_\delta u_{n,\varepsilon})(x) - E_\delta u_{n,\varepsilon}(x)| \leq C\varepsilon.$$

Since $u_{n,\varepsilon} = g$ on $\mathcal{X}_n \cap \partial_{2\varepsilon} U$

$$|E_\delta u_{n,\varepsilon}(x) - g(x)| = |u_{n,\varepsilon}(T_\delta(x)) - g(x)| = |g(T_\delta(x)) - g(x)| \leq C\delta.$$

Since $g(x) = u(x)$ and $\delta \leq \varepsilon$ we have

$$|(\Lambda_{\varepsilon,\delta} E_\delta u_{n,\varepsilon})(x) - u(x)| \leq C\varepsilon$$

for all $x \in \partial U$. Therefore, we can apply Proposition 5.38 to obtain

$$\|u - \Lambda_{\varepsilon,\delta} E_\delta u_{n,\varepsilon}\|_{H^1(U)}^2 \leq C(\varepsilon + \tfrac{\delta}{\varepsilon} + \lambda). \tag{5.119}$$

with probability at least $1 - Cn \exp\left(-cn\delta^d\lambda^2\right)$.

4. We now extend the rate to the graph $\mathcal{X}_n$. By Proposition 5.34 we have

$$\|\Lambda_{\varepsilon,\delta} E_\delta u_{n,\varepsilon} - E_\delta u_{n,\varepsilon}\|_{L^2(U_\varepsilon)}^2 \leq C I_{\varepsilon,\delta}(E_\delta u_{n,\varepsilon})\varepsilon^2 \leq C\varepsilon^2,$$

since $I_{\varepsilon,\delta}(E_\delta u_{n,\varepsilon}) \leq C$ due to (5.116) and (5.115). Therefore

$$\begin{aligned}\|u - E_\delta u_{n,\varepsilon}\|_{L^2(U_\varepsilon)}^2 &\leq 2\|u - \Lambda_{\varepsilon,\delta} E_\delta u_{n,\varepsilon}\|_{L^2(U_\varepsilon)}^2 + 2\|\Lambda_{\varepsilon,\delta} E_\delta u_{n,\varepsilon} - E_\delta u_{n,\varepsilon}\|_{L^2(U_\varepsilon)}^2 \\ &= C(\varepsilon + \tfrac{\delta}{\varepsilon} + \lambda) + C\varepsilon^2.\end{aligned}$$

Since u is Lipschitz, we have $|E_\delta u - u| \leq C\delta \leq C\varepsilon$, and so

$$\begin{aligned}\|u - u_{n,\varepsilon}\|_{\ell^2(\mathcal{X}_n)}^2 &= \frac{1}{n}\sum_{x \in \mathcal{X}_n} (u(x) - u_{n,\varepsilon}(x))^2 \\ &= \int_U (u(T_\delta(x)) - u_{n,\varepsilon}(T_\delta(x)))^2 \, dx \\ &= \|E_\delta u - E_\delta u_{n,\varepsilon}\|_{L^2(U)}^2 \\ &\leq 2\|E_\delta u - u\|_{L^2(U)}^2 + 2\|u - E_\delta u_{n,\varepsilon}\|_{L^2(U)}^2 \\ &\leq C(\varepsilon + \tfrac{\delta}{\varepsilon} + \lambda) + 2\|u - E_\delta u_{n,\varepsilon}\|_{L^2(\partial_\varepsilon U)}^2.\end{aligned}$$

Using an argument similar to part 3, we have

$$\begin{aligned}\|u - E_\delta u_{n,\varepsilon}\|_{L^2(\partial_\varepsilon U)} &= \|u - E_\delta g\|_{L^2(\partial_\varepsilon U)} \\ &\leq \|g - E_\delta g\|_{L^2(\partial_\varepsilon U)} + \|u - g\|_{L^2(\partial_\varepsilon U)} \\ &\leq C(\delta + \varepsilon^2).\end{aligned}$$

Therefore

$$\|u - u_{n,\varepsilon}\|_{\ell^2(\mathcal{X}_n)}^2 \leq C(\varepsilon + \tfrac{\delta}{\varepsilon} + \lambda). \tag{5.120}$$

5. Define

$$w_\varepsilon(x) = \Lambda_{\varepsilon,\delta} E_\delta u_{n,\varepsilon}(x) + \varphi\left(\tfrac{4d(x)}{\varepsilon}\right)(g(x) - \Lambda_{\varepsilon,\delta} E_\delta u_{n,\varepsilon}(x)).$$

As in the proof of Theorem 5.36, and part 3, we have that $|\nabla \Lambda_{\varepsilon,\delta} E_\delta u_{n,\varepsilon}(x)| \leq C$ and

$$|\Lambda_{\varepsilon,\delta} E_\delta u_{n,\varepsilon}(x) - u(x)| \leq C\varepsilon$$

for $x \in \partial_{\varepsilon-2\delta}U$. Therefore $|\nabla w_\varepsilon| \leq C$ for $x \in \partial_{\varepsilon-2\delta}U$. Since $w_\varepsilon = g$ on ∂U we have $w_\varepsilon \in \mathcal{A}$ and so $I(u) \leq I(w_\varepsilon)$. Since $w_\varepsilon = \Lambda_{\varepsilon,\delta}E_\delta u_{n,\varepsilon}$ on $U_{\varepsilon/2}$ and $\varepsilon - 2\delta \geq \varepsilon/2$, since $\varepsilon \geq 4\delta$, we have that

$$\begin{aligned} I(u) - I(\Lambda_{\varepsilon,\delta}E_\delta u_{n,\varepsilon}) &\leq I(w_\varepsilon) - I(\Lambda_{\varepsilon,\delta}E_\delta u_{n,\varepsilon}) \\ &= \int_{\partial_{\varepsilon/2}U} (|\nabla w_\varepsilon|^2 - |\nabla \Lambda_{\varepsilon,\delta}E_\delta u_{n,\varepsilon}|^2)\rho^2\,dx \\ &\leq C|\partial_{\varepsilon/2}U| \leq C\varepsilon. \end{aligned}$$

Combining this with (5.117) we have

$$I(u) \leq (1 + C(\varepsilon + \tfrac{\delta}{\varepsilon} + \lambda))\mathcal{E}_{n,\varepsilon}(u_{n,\varepsilon}) + C\left(\varepsilon + \tfrac{\delta}{\varepsilon}\right)$$

with probability at least $1 - Cn\exp\left(-cn\delta^d\lambda^2\right)$. By Remark (5.30) we have

$$|\mathcal{E}_{n,\varepsilon}(u) - I(u)| \leq C(\|u\|_{C^2(U)} + 1)(\varepsilon + \lambda)$$

with probability at least $1 - 2\exp\left(cn\varepsilon^d\lambda^2\right)$. Therefore

$$\mathcal{E}_{n,\varepsilon}(u) - \mathcal{E}_{n,\varepsilon}(u_{n,\varepsilon}) \leq C(\|u\|_{C^2(U)} + 1)\left(\varepsilon + \tfrac{\delta}{\varepsilon} + \lambda\right).$$

Applying Proposition 5.40 we have

$$\|\nabla_{n,\varepsilon}u - \nabla_{n,\varepsilon}u_{n,\varepsilon}\|^2_{\ell^2(\mathcal{X}_n^2)} \leq C(\|u\|_{C^2(U)} + 1)\left(\varepsilon + \tfrac{\delta}{\varepsilon} + \lambda\right).$$

Combining this with (5.120) completes the proof. □

Appendix A

Mathematical preliminaries

We review here some basic mathematical preliminaries.

A.1 Inequalities

For $x \in \mathbb{R}^d$ the norm of x is

$$|x| := \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

When $n = 2$ or $n = 3$, $|x - y|$ is the usual Euclidean distance between x and y. The dot product between $x, y \in \mathbb{R}^d$ is

$$x \cdot y = \sum_{i=1}^{d} x_i y_i.$$

Notice that

$$|x|^2 = x \cdot x.$$

Simple inequalities, when used in a clever manner, are very powerful tools in the study of partial differential equations. We give a brief overview of some commonly used inequalities here.

The *Cauchy-Schwarz inequality* states that

$$|x \cdot y| \le |x||y|.$$

To prove the Cauchy-Schwarz inequality find the value of t that minimizes

$$h(t) := |x + ty|^2.$$

For $x, y \in \mathbb{R}^d$

$$|x + y|^2 = (x + y) \cdot (x + y) = x \cdot x + x \cdot y + y \cdot x + y \cdot y.$$

Therefore

$$|x+y|^2 = |x|^2 + 2x\cdot y + |y|^2.$$

Using the Cauchy-Schwarz inequality we have

$$|x+y|^2 \le |x|^2 + 2|x||y| + |y|^2 = (|x|+|y|)^2.$$

Taking square roots of both sides we have the *triangle inequality*

$$|x+y| \le |x| + |y|.$$

For $x, y \in \mathbb{R}^d$ the triangle inequality yields

$$|x| = |x-y+y| \le |x-y| + |y|.$$

Rearranging we obtain the *reverse triangle inequality*

$$|x-y| \ge |x| - |y|.$$

For real numbers a, b we have

$$0 \le (a-b)^2 = a^2 - 2ab + b^2.$$

Therefore

$$ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2. \tag{A.1}$$

This is called *Cauchy's inequality.*

Another useful inequality is *Young's inequality*: If $1 < p, q < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$ then

$$ab \le \frac{a^p}{p} + \frac{b^q}{q} \quad \text{for all } a, b > 0. \tag{A.2}$$

To see this, we use convexity of the mapping $x \mapsto e^x$ to obtain

$$ab = e^{\log(a)+\log(b)} = e^{\frac{1}{p}\log(a^p)+\frac{1}{q}\log(b^q)} \le \frac{1}{p}e^{\log(a^p)} + \frac{1}{q}e^{\log(b^q)} = \frac{a^p}{p} + \frac{b^q}{q}.$$

Young's inequality is a generalization of Cauchy's inequality, which is obtained by taking $p = q = 2$.

A.2 Topology

We will have to make use of basic point-set topology. We define the open ball of radius $r > 0$ centered at $x_0 \in \mathbb{R}^d$ by

$$B^0(x_0, r) := \{x \in \mathbb{R}^d \,:\, |x - x_0| < r\}.$$

The closed ball is defined as

$$B(x_0, r) := \{x \in \mathbb{R}^d \,:\, |x - x_0| \leq r\}.$$

Definition A.1. A set $U \subset \mathbb{R}^d$ is called *open* if for each $x \in U$ there exists $r > 0$ such that $B(x, r) \subset U$.

Exercise A.2. Let $U, V \subset \mathbb{R}^d$ be open. Show that

$$W := U \cup V := \{x \in \mathbb{R}^d \,:\, x \in U \text{ or } x \in V\}$$

is open. △

Definition A.3. We say that a sequence $\{x_k\}_{k=1}^{\infty}$ in $\mathbb{R}^d$ *converges* to $x \in \mathbb{R}^d$, written $x_k \to x$, if

$$\lim_{k \to \infty} |x_k - x| = 0.$$

Definition A.4. The *closure* of a set $U \subset \mathbb{R}^d$, denotes $\overline{U}$, is defined as

$$\overline{U} := \{x \in \mathbb{R}^d \,:\, \text{there exists a sequence } x_k \in U \text{ such that } x_k \to x\}.$$

The closure is the set of points that can be reached as limits of sequences belonging to U.

Definition A.5. We say that a set $U \subset \mathbb{R}^d$ is *closed* if $\overline{U} = U$.

Exercise A.6. Another definition of *closed* is: A set $U \subset \mathbb{R}^d$ is *closed* if the complement

$$\mathbb{R}^d \setminus U := \{x \in \mathbb{R}^d \,:\, x \notin U\}$$

is open. Verify that the two definitions are equivalent [This is not a trivial exercise]. △

Definition A.7. We define the *boundary* of an open set $U \subset \mathbb{R}^d$, denoted ∂U, as

$$\partial U := \overline{U} \setminus U.$$

Example A.1. The open ball $B^0(x, r)$ is open, and its closure is the closed ball $B(x, r)$. The boundary of the open ball $B^0(x, r)$ is

$$\partial B^0(x_0, r) := \{x \in \mathbb{R}^d \,:\, |x - x_0| = r\}.$$

It is a good idea to verify each of these facts from the definitions. △

We defined the boundary only for open sets, but is can be defined for any set.

Definition A.8. The *interior* of a set $U \subset \mathbb{R}^d$, denoted $\mathrm{int}(U)$, is defined as

$$\mathrm{int}(U) := \{x \in U \,:\, B(x,r) \subset U \text{ for small enough } r > 0\}.$$

Exercise A.9. Show that $U \subset \mathbb{R}^d$ is open if and only if $\mathrm{int}(U) = U$. △

We can now define the boundary of an arbitrary set $U \subset \mathbb{R}^d$.

Definition A.10. We define the *boundary* of a set $U \subset \mathbb{R}^d$, denoted ∂U, as

$$\partial U := \overline{U} \setminus \mathrm{int}(U).$$

Exercise A.11. Verify that

$$\partial B(x,r) = \partial B^0(x,r).$$

△

Definition A.12. We say a set $U \subset \mathbb{R}^d$ is *bounded* if there exists $M > 0$ such that $|x| \leq M$ for all $x \in U$.

Definition A.13. We say a set $U \subset \mathbb{R}^d$ is *compact* if U is closed and bounded.

Definition A.14. For open sets $V \subset U \subset \mathbb{R}^d$ we say that V is *compactly contained* in U if $\overline{V}$ is compact and $\overline{V} \subset U$. If V is compactly contained in U we write $V \subset\subset U$.

A.3 Differentiation

A.3.1 Partial derivatives

The *partial derivative* of a function $u = u(x_1, x_2, \dots, x_n)$ in the x_i variable is defined as

$$\frac{\partial u}{\partial x_i}(x) := \lim_{h \to 0} \frac{u(x + he_i) - u(x)}{h},$$

provided the limit exists. Here $e_1, e_2, \dots, e_n$ are the standard basis vectors in $\mathbb{R}^d$, so $e_i = (0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{R}^d$ has a one in the i^{th} entry. For simplicity of notation we will write

$$u_{x_i} = \frac{\partial u}{\partial x_i}.$$

The *gradient* of a function $u : \mathbb{R}^d \to \mathbb{R}$ is the vector of partial derivatives

$$\nabla u(x) := (u_{x_1}(x), u_{x_2}(x), \dots, u_{x_n}(x)).$$

We will treat the gradient as a column vector for matrix-vector multiplication.

Higher derivatives are defined iteratively. The second derivatives of u are defined as

$$\frac{\partial^2 u}{\partial x_i x_j} := \frac{\partial}{\partial x_i}\left(\frac{\partial u}{\partial x_j}\right).$$

This means that

$$\frac{\partial^2 u}{\partial x_i x_j}(x) = \lim_{h\to 0}\frac{1}{h}\left(\frac{\partial u}{\partial x_j}(x+he_i) - \frac{\partial u}{\partial x_j}(x)\right),$$

provided the limit exists. As before, we write

$$u_{x_ix_j} = \frac{\partial^2 u}{\partial x_i x_j}$$

for notational simplicity. When $u_{x_ix_j}$ and $u_{x_jx_i}$ exist and are continuous we have

$$u_{x_ix_j} = u_{x_jx_i},$$

that is the second derivatives are the same regardless of which order we take them in. We will generally always assume our functions are smooth (infinitely differentiable), so equality of mixed partials is always assumed to hold.

The *Hessian* of $u : \mathbb{R}^d \to \mathbb{R}$ is the matrix of all second partial derivatives

$$\nabla^2 u(x) := (u_{x_ix_j})_{i,j=1}^d = \begin{bmatrix} u_{x_1x_1} & u_{x_1x_2} & u_{x_1x_3} & \cdots & u_{x_1x_n} \\ u_{x_2x_1} & u_{x_2x_2} & u_{x_2x_3} & \cdots & u_{x_2x_n} \\ u_{x_3x_1} & u_{x_3x_2} & u_{x_3x_3} & \cdots & u_{x_3x_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{x_nx_1} & u_{x_nx_2} & u_{x_nx_3} & \cdots & u_{x_nx_n} \end{bmatrix}$$

Since we have equality of mixed partials, the Hessian is a symmetric matrix, i.e., $(\nabla^2 u)^T = \nabla^2 u$. Since we treat the gradient ∇u as a column vector, the product $\nabla^2 u(x)\nabla u(x)$ denotes the Hessian matrix multiplied by the gradient vector. That is,

$$[\nabla^2 u(x)\nabla u(x)]_j = \sum_{i=1}^d u_{x_ix_j}u_{x_i}.$$

Given a vector field $F : \mathbb{R}^d \to \mathbb{R}^d$ where $F(x) = (F^1(x), F^2(x), \ldots, F^d(x))$, the *divergence* of F is defined as

$$\text{div}F(x) := \sum_{i=1}^d F^i_{x_i}(x).$$

The *Laplacian* of a function $u : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\Delta u := \text{div}(\nabla u) = \sum_{i=1}^d u_{x_ix_i}.$$

A.3.2 Rules for differentiation

Most of the rules for differentiation from single variable calculus carry over to multivariable calculus.

Chain rule: If $v(t) = (v_1(t), v_2(t), \ldots, v_n(t))$ is a function $v : \mathbb{R} \to \mathbb{R}^d$, and $u : \mathbb{R}^d \to \mathbb{R}$, then

$$\text{(A.3)} \qquad \frac{d}{dt}u(v(t)) = \nabla u(v(t)) \cdot v'(t) = \sum_{i=1}^{d} u_{x_i}(v(t))v_i'(t).$$

Here $v'(t) = (v_1'(t), v_2'(t), \ldots, v_n'(t))$.

If $F(x) = (F^1(x), F^2(x), \ldots, F^d(x)$ is a function $F : \mathbb{R}^d \to \mathbb{R}^d$ then

$$\frac{\partial}{\partial x_j}u(F(x)) = \nabla u(F(x)) \cdot F_{x_j}(x) = \sum_{i=1}^{d} u_{x_i}(F(x))F^i_{x_j}(x),$$

where $F_{x_j} = (F^1_{x_j}, F^2_{x_j}, \ldots, F^d_{x_j})$. This is a special case of (A.3) with $t = x_j$.

Product rule: Given two functions $u, v : \mathbb{R}^d \to \mathbb{R}$, we have

$$\nabla(uv) = u\nabla v + v\nabla u.$$

This is entry-wise the usual product rule for single variable calculus.

Given a vector field $F : \mathbb{R}^d \to \mathbb{R}^d$ and a function $u : \mathbb{R}^d \to \mathbb{R}$ we have

$$\frac{\partial}{\partial x_i}(uF^i) = u_{x_i}F^i + uF^i_{x_i}.$$

Therefore

$$\operatorname{div}(uF) = \nabla u \cdot F + u\operatorname{div}F.$$

Exercise A.15. Let $|x| = \sqrt{x_1^2 + \cdots + x_n^2}$.

(a) Show that for $x \neq 0$

$$\frac{\partial}{\partial x_i}|x| = \frac{x_i}{|x|}.$$

(b) Show that for $x \neq 0$

$$\frac{\partial^2}{\partial x_i \partial x_j}|x| = \frac{\delta_{ij}}{|x|} - \frac{x_i x_j}{|x|^3},$$

where δ_{ij} is the Kronecker delta defined by

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j. \end{cases}$$

(c) Show that for $x \neq 0$

$$\Delta|x| = \frac{n-1}{|x|}.$$

△

Exercise A.16. Find all real numbers α for which $u(x) = |x|^\alpha$ is a solution of Laplace's equation

$$\Delta u(x) = 0 \quad \text{for } x \neq 0.$$

△

Exercise A.17. Let $1 \leq p \leq \infty$. The p-Laplacian is defined by

$$\Delta_p u := \operatorname{div}\left(|\nabla u|^{p-2}\nabla u\right)$$

for $1 \leq p < \infty$, and

$$\Delta_\infty u := \frac{1}{|\nabla u|^2}\sum_{i=1}^d \sum_{j=1}^d u_{x_i x_j} u_{x_i} u_{x_j}.$$

Notice that $\Delta_2 u = \Delta u$. A function u is called *p-harmonic* if $\Delta_p u = 0$.

(a) Show that

$$\Delta_p u = |\nabla u|^{p-2}\left(\Delta u + (p-2)\Delta_\infty u\right).$$

(b) Show that

$$\Delta_\infty u = \lim_{p\to\infty} \frac{1}{p}|\nabla u|^{2-p}\Delta_p u.$$

△

Exercise A.18. Let $1 \leq p \leq \infty$. Find all real numbers α for which the function $u(x) = |x|^\alpha$ is p-harmonic away from $x = 0$. △

A.4 Taylor series

A.4.1 One dimension

Let $u : \mathbb{R} \to \mathbb{R}$ be twice continuously differentiable. Then by the fundamental theorem of calculus

$$u(y) - u(x) = \int_x^y u'(t)\,dt = \int_x^y u'(x) + u'(t) - u'(x)\,dt.$$

Since $\int_x^y u'(x)\,dt = u'(x)(y-x)$, it follows that

$$u(y) = u(x) + u'(x)(y-x) + R_2(x,y), \tag{A.4}$$

where R_2 is the remainder given by

$$R_2(x,y) = \int_x^y u'(t) - u'(x)\,dt.$$

Applying the fundamental theorem again we have

$$R_2(x,y) = \int_x^y \int_x^t u''(s)\,ds\,dt. \tag{A.5}$$

Let $C > 0$ denote the maximum value of $|u''(s)|$. Assuming, without loss of generality, that $y > x$ we have

$$|R_2(x,y)| \leq \left| \int_x^y C|t-x|\,dt \right| = \frac{C}{2}|y-x|^2.$$

Exercise A.19. Verify the final equality above. △

When $|g(y)| \leq C|y|^k$ we write $g \in O(|y|^k)$. Thus $R_2(x,y) \in O(|y-x|^2)$ and we have deduced the first order Taylor series

$$u(y) = u(x) + u'(x)(y-x) + O(|y-x|^2). \tag{A.6}$$

A Taylor series expresses the fact that a sufficiently smooth function can be well-approximated locally by its tangent line. It is important to keep in mind that the constant C hidden in the $O((y-x)^2)$ term depends on how large $|u''|$ is. Also note that we can choose $C > 0$ to be the maximum of $|u''(s)|$ for s between x and y, which may be much smaller than the maximum of $|u'(s)|$ over all s (which may not exist).

It is useful sometimes to continue the Taylor series to higher order terms. For this, suppose u is three times continuously differentiable. We first write the Taylor series with remainder for $u'(t)$

$$u'(t) = u'(x) + u''(x)(t-x) + \int_x^t \int_x^\tau u'''(s)\,ds\,d\tau.$$

Proceeding as before, we use the fundamental theorem of calculus to find

$$\begin{aligned} u(y) &= u(x) + \int_x^y u'(t)\,dt \\ &= u(x) + \int_x^y u'(x) + u''(x)(t-x) + \int_x^t \int_x^\tau u'''(s)\,ds\,d\tau\,dt \\ &= u(x) + u'(x)(y-x) + \frac{1}{2}u''(x)(y-x)^2 + R_3(x,y), \end{aligned}$$

where

$$R_3(x,y) = \int_x^y \int_x^t \int_x^\tau u'''(s)\,ds\,d\tau\,dt.$$

As before, let $C > 0$ denote the maximum value of $|u'''(s)|$. Then

$$|R_3(x,y)| \leq \frac{C}{6}|y-x|^3.$$

Exercise A.20. Verify the inequality above. △

Therefore $R_3 \in O(|y-x|^3)$ and we have the second order Taylor expansion

$$u(y) = u(x) + u'(x)(y-x) + \frac{1}{2}u''(x)(y-x)^2 + O(|y-x|^3). \tag{A.7}$$

The second order Taylor series says that a sufficiently smooth function can be approximated up to $O((y-x)^3)$ accuracy with a parabola. Again, we note that the constant $C > 0$ hidden in $O((y-x)^3)$ depends on the size of $|u'''(s)|$, and $C > 0$ may be chosen as the maximum of $|u'''(s)|$ over s between x and y.

A.4.2 Higher dimensions

Taylor series expansions for functions $u : \mathbb{R}^d \to \mathbb{R}$ follow directly from the one dimensional case and the chain rule. Suppose u is twice continuously differentiable and fix $x, y \in \mathbb{R}^d$. For $t \in \mathbb{R}$ define

$$\varphi(t) = u(x + (y-x)t).$$

Since φ is a function of one variable t, we can use the one dimensional Taylor series to obtain

$$\varphi(t) = \varphi(0) + \varphi'(0)t + O(|t|^2). \tag{A.8}$$

The constant in the $O(|t|^2)$ term depends on the maximum of $|\varphi''(t)|$. All that remains is to compute the derivatives of φ. By the chain rule

$$\varphi'(t) = \sum_{i=1}^d u_{x_i}(x + (y-x)t)(y_i - x_i), \tag{A.9}$$

and

$$\begin{aligned}
\varphi''(t) &= \frac{d}{dt}\sum_{i=1}^d u_{x_i}(x + (y-x)t)(y_i - x_i) \\
&= \sum_{i=1}^d \frac{d}{dt} u_{x_i}(x + (y-x)t)(y_i - x_i) \\
&= \sum_{i=1}^d \sum_{j=1}^d u_{x_i x_j}(x + (y-x)t)(y_i - x_i)(y_j - x_j). \qquad \text{(A.10)}
\end{aligned}$$

In particular

$$\varphi'(0) = \sum_{i=1}^d u_{x_i}(x)(y_i - x_i) = \nabla u(x) \cdot (y-x),$$

and so (A.8) with $t = 1$ becomes

$$u(y) = u(x) + \nabla u(x) \cdot (y - x) + R_2(x, y),$$

where $R_2(x, y)$ satisfies $|R_2(x, y)| \le \frac{1}{2} \max_t |\varphi''(t)|$. Let $C > 0$ denote the maximum value of $|u_{x_i x_j}(z)|$ over all z, i and j. Then by (A.10)

$$|\varphi''(t)| \le C \sum_{i=1}^{d} \sum_{j=1}^{d} |y_i - x_i||y_j - x_j| \le C n^2 |x - y|^2.$$

It follows that $|R_2(x, y)| \le \frac{C}{2} n^2 |x - y|^2$, hence $R_2(x, y) \in O(|x - y|^2)$ and we arrive at the first order Taylor series

$$u(y) = u(x) + \nabla u(x) \cdot (y - x) + O(|x - y|^2). \tag{A.11}$$

This says that u can be locally approximated near x to order $O(|x-y|^2)$ by the affine function

$$L(y) = u(x) + \nabla u(x) \cdot (y - x).$$

We can continue this way to obtain the second order Taylor expansion. We assume now that u is three times continuously differentiable. Using the one dimensional second order Taylor expansion we have

$$\varphi(t) = \varphi(0) + \varphi'(0)t + \frac{1}{2}\varphi''(0)t^2 + O(|t|^3). \tag{A.12}$$

The constant in the $O(|t|^3)$ term depends on the maximum of $|\varphi'''(t)|$. Notice also that

$$\varphi''(0) = \sum_{i=1}^{d} \sum_{j=1}^{d} u_{x_i x_j}(x)(y_i - x_i)(y_j - x_j) = (y - x) \cdot \nabla^2 u(x)(y - x),$$

where $\nabla^2 u(x)$ is the Hessian matrix. Plugging this into (A.12) with $t = 1$ yields

$$u(y) = u(x) + \nabla u(x) \cdot (y - x) + \frac{1}{2}(y - x) \cdot \nabla^2 u(x)(y - x) + R_3(x, y),$$

where $R_3(x, y)$ satisfies $|R_3(x, y)| \le \frac{1}{6} \max_t |\varphi'''(t)|$. We compute

$$\begin{aligned} \varphi'''(t) &= \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{d}{dt} u_{x_i x_j}(x + (y - x)t)(y_i - x_i)(y_j - x_j) \\ &= \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} u_{x_i x_j x_k}(x + (y - x)t)(y_i - x_i)(y_j - x_j)(y_k - x_k). \end{aligned}$$

Let $C > 0$ denote the maximum value of $|u_{x_i x_j x_k}(z)|$ over all z, i, j, and k. Then we have

$$|\varphi'''(t)| \leq C \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} |y_i - x_i||y_j - x_j||y_k - x_k| \leq Cn^3|x-y|^3.$$

Therefore $|R_3(x,y)| \leq \frac{C}{6}n^3|x-y|^3$ and so $R_3 \in O(|x-y|^3)$. Finally we arrive at the second order Taylor expansion

$$u(y) = u(x) + \nabla u(x) \cdot (y-x) + \frac{1}{2}(y-x) \cdot \nabla^2 u(x)(y-x) + O(|x-y|^3). \tag{A.13}$$

This says that u can be locally approximated near x to order $O(|x-y|^3)$ by the quadratic function

$$L(y) = u(x) + \nabla u(x) \cdot (y-x) + \frac{1}{2}(y-x) \cdot \nabla^2 u(x)(y-x).$$

A.5 Function spaces

For an open set $U \subset \mathbb{R}^d$ we define

$$C^k(\overline{U}) := \left\{\text{Functions } u : \overline{U} \to \mathbb{R} \text{ that are } k\text{-times continuously differentiable on } \overline{U}\right\}.$$

The terminology k-times continuously differentiable means that all k^{th}-order partial derivatives of u exist and are continuous on $\overline{U}$. We write $C^0(\overline{U}) = C(\overline{U})$ for the space of functions that are continuous on $\overline{U}$.

Exercise A.21. Show that the function $u(x) = x^2$ for $x > 0$ and $u(x) = -x^2$ for $x \leq 0$ belongs to $C^1(\mathbb{R})$ but not to $C^2(\mathbb{R})$. △

We also define

$$C^\infty(\overline{U}) := \bigcap_{k=1}^{\infty} C^k(\overline{U})$$

to be the space of infinitely differentiable functions. Functions $u \in C^\infty(\overline{U})$ are called *smooth*.

Definition A.22. The *support* of a function $u : \overline{U} \to \mathbb{R}$ is defined as

$$\text{supp}(u) := \{x \in \overline{U} \;:\; u(x) \neq 0\}.$$

Definition A.23. We say that $u : \overline{U} \to \mathbb{R}$ is *compactly supported* in U if $\text{supp}(u) \subset\subset U$.

A function u is compactly supported in U if u vanishes near the boundary ∂U. Finally for $k \in \mathbb{N} \cup \{\infty\}$ we write

$$C_c^k(U) := \{u \in C^k(\overline{U}) \, : \, u \text{ is compactly supported in } U\}.$$

For a function $u : U \to \mathbb{R}$ we define the L^2-norm of u to be

$$\|u\|_{L^2(U)} := \left(\int_U u^2 \, dx \right)^{\frac{1}{2}}.$$

For two functions $u, v : U \to \mathbb{R}$ we define the L^2-inner product of u and v to be

$$(u, v)_{L^2(U)} := \int_U u \, v \, dx.$$

Notice that

$$\|u\|_{L^2(U)}^2 = (u, u)_{L^2(U)}.$$

We also define

$$L^2(U) := \Big\{ \text{Functions } u : U \to \mathbb{R} \text{ for which } \|u\|_{L^2(U)} < \infty \Big\}.$$

$L^2(U)$ is a Hilbert space (a complete inner-product space). We will often write $\|u\|$ in place of $\|u\|_{L^2(U)}$ and (u, v) in place of $(u, v)_{L^2(U)}$ when it is clear from the context that the L^2 norm is intended.

As before, we have the Cauchy-Schwarz inequality

$$(u, v)_{L^2(U)} \leq \|u\|_{L^2(U)} \|v\|_{L^2(U)}.$$

We also have

$$\|u + v\|_{L^2(U)}^2 = \|u\|_{L^2(U)}^2 + 2(u, v)_{L^2(U)} + \|v\|_{L^2(U)}^2.$$

Applying the Cauchy-Schwarz inequality we get the *triangle inequality*

$$\|u + v\|_{L^2(U)} \leq \|u\|_{L^2(U)} + \|v\|_{L^2(U)},$$

and the *reverse triangle inequality*

$$\|u - v\|_{L^2(U)} \geq \|u\|_{L^2(U)} - \|v\|_{L^2(U)}.$$

A.6 Analysis

A.6.1 The Riemann integral

Many students are accustomed to using different notation for integration in different dimensions. For example, integration along the real line in $\mathbb{R}$ is usually written

$$\int_a^b u(x) \, dx,$$

while integration over a region $U \subset \mathbb{R}^2$ is written

$$\iint_U u(x,y)\,dxdy \quad \text{or} \quad \iint_U u(\mathbf{x})\,d\mathbf{x},$$

where $\mathbf{x} = (x,y)$. Integration over a volume $U \subset \mathbb{R}^3$ is then written as

$$\iiint_U u(x,y,z)\,dxdydz \quad \text{or} \quad \iiint_U u(\mathbf{x})\,d\mathbf{x}.$$

This becomes cumbersome when we consider problems in an arbitrary number of dimensions n. In these notes, we use x (or y or z) for a point in $\mathbb{R}^d$, so $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^d$. We write

$$u(x) = u(x_1, x_2, \ldots, x_n)$$

for a function $u : \mathbb{R}^d \to \mathbb{R}$. The integration of u over a domain $U \subset \mathbb{R}^d$ is then written

$$\int_U u(x)\,dx \quad \text{or just} \quad \int_U u\,dx,$$

where $dx = dx_1 dx_2 \cdots dx_n$. This notation has the advantage of being far more compact without losing the meaning.

Let us recall the interpretation of the integral $\int_U u\,dx$ in the Riemann sense. We partition the domain into M rectangles and approximate the integral by a Riemann sum

$$\int_U u\,dx \approx \sum_{k=1}^{M} u(x_k)\Delta x_k,$$

where $x_k \in \mathbb{R}^d$ is a point in the k^{th} rectangle, and $\Delta x_k := \Delta x_{k,1}\Delta x_{k,2}\cdots\Delta x_{k,n}$ is the n-dimensional volume (or measure) of the k^{th} rectangle ($\Delta_{k,i}$ for $i = 1, \ldots, n$ are the side lengths of the k^{th} rectangle). Then the Riemann integral is defined by taking the limit as the largest side length in the partition tends to zero (provided the limit exists and does not depend on the choice of partition or points x_k). Notice here that $x_k = (x_{k,1}, \ldots, x_{k,n}) \in \mathbb{R}^d$ is a point in $\mathbb{R}^d$, and not the k^{th} entry of x. There is a slight abuse of notation here; the reader will have to discern from the context which is implied.

If $S \subset \mathbb{R}^d$ is an $n-1$ dimensional (or possibly lower dimensional) surface, we write the surface integral of u over S as

$$\int_S u(x)\,dS(x).$$

Here, $dS(x)$ is the surface area element at $x \in S$.

A.6.2 The Lebesgue integral

In these notes, and most rigorous analysis books, we interpret the integral in the Lebesgue sense (see [51] for definitions). The Lebesgue integral is more powerful for analysis and facilitates passing to limits within the integral with ease. The Lebesgue integral can be applied to *Lebesgue measurable functions* $u : \mathbb{R}^d \to \mathbb{R}$. The notion of measurability excludes certain pathological classes of functions u for which $\int_{\mathbb{R}^d} u\, dx$ cannot be assigned a sensible value. We again refer to [51] for notions of Lebesgue measure and measurable functions. Nearly all functions you encounter via simple constructions are measurable, and the reader not familiar with measure theory can ignore the issue and still understand the main ideas in these notes. We will always abbreviate *Lebesgue measurable* with *measurable*, and we write the Lebesgue measure of a set $A \subset \mathbb{R}^d$ by $|A|$.

Definition A.24. We say a measurable function $u : \mathbb{R}^d \to \mathbb{R}$ is *summable* if

$$\int_{\mathbb{R}^d} |u|\, dx < \infty.$$

Some properties will hold on all of $\mathbb{R}^d$ except for a set with measure zero. In this case we will say the property holds *almost everywhere* or abbreviated "a.e.".

The following theorems are the most useful for passing to limits within integrals.

Lemma A.25 (Fatou's Lemma)**.** *If $u_k : \mathbb{R}^d \to [0, \infty)$ is a sequence of nonnegative measureable functions then*

$$\int_{\mathbb{R}^d} \liminf_{k\to\infty} u_k(x)\, dx \le \liminf_{k\to\infty} \int_{\mathbb{R}^d} u_k(x)\, dx. \tag{A.14}$$

Remark A.26 (Reverse Fatou Lemma)**.** A simple consequence of Fatou's Lemma, called the *Reverse Fatou Lemma*, is often useful. Let $u_k : \mathbb{R}^d \to \mathbb{R}$ be a sequence of measureable functions, and assume there exists a summable function g such that $u_k \le g$ almost everywhere. Then

$$\limsup_{k\to\infty} \int_{\mathbb{R}^d} u_k(x)\, dx \le \int_{\mathbb{R}^d} \limsup_{k\to\infty} u_k(x)\, dx. \tag{A.15}$$

To see this, we apply Fatou's Lemma to the nonnegative sequence $g - u_k$ to obtain

$$\int_{\mathbb{R}^d} \liminf_{k\to\infty} (g(x) - u_k(x))\, dx \le \liminf_{k\to\infty} \int_{\mathbb{R}^d} g(x) - u_k(x)\, dx.$$

Noting that $\liminf(-u_k) = -\limsup(u_k)$, this can be rearranged to obtain (A.15), provided g is summable.

The following two theorems are more or less direct applications of Fatou's Lemma and the Reverse Fatou Lemma, but present the results in a way that is very useful in practice.

Theorem A.27 (Monotone Convergence Theorem)**.** *Let $u_k : \mathbb{R}^d \to \mathbb{R}$ be a sequence of measureable functions satisfying*

$$0 \leq u_1 \leq u_2 \leq u_3 \leq \cdots \leq u_k \leq u_{k+1} \leq \cdots \quad a.e. \tag{A.16}$$

If $\lim_{k\to\infty} u_k(x)$ exists almost everywhere then

$$\lim_{k\to\infty} \int_{\mathbb{R}^d} u_k(x)\,dx = \int_{\mathbb{R}^d} \lim_{k\to\infty} u_k(x)\,dx. \tag{A.17}$$

Proof. Let $u(x) = \lim_{k\to\infty} u_k(x)$. By Fatou's lemma we have

$$\int_{\mathbb{R}^d} u\,dx \leq \liminf_{k\to\infty} \int_{\mathbb{R}^d} u_k(x)\,dx. \tag{A.18}$$

Since $u_k(x) \leq u(x)$ for almost every $x \in U$, we clearly have $\int_{\mathbb{R}^d} u_k\,dx \leq \int_{\mathbb{R}^d} u\,dx$, and so

$$\limsup_{k\to\infty} \int_{\mathbb{R}^d} u_k(x)\,dx \leq \int_{\mathbb{R}^d} u\,dx. \tag{A.19}$$

Combining (A.18) and (A.19) completes the proof. □

Theorem A.28 (Dominated Convergence Theorem)**.** *Let $u_k : \mathbb{R}^d \to \mathbb{R}$ be a sequence of measureable functions, and assume there exists a summable function g such that*

$$|u_k| \leq g \quad a.e. \tag{A.20}$$

If $\lim_{k\to\infty} u_k(x)$ exists almost everywhere then

$$\lim_{k\to\infty} \int_{\mathbb{R}^d} u_k(x)\,dx = \int_{\mathbb{R}^d} \lim_{k\to\infty} u_k(x)\,dx. \tag{A.21}$$

Proof. Let $u(x) = \lim_{k\to\infty} u_k(x)$. Note that $|u_k - u| \leq |u_k| + |u| \leq 2g$. Therefore, we can apply the Reverse Fatou Lemma to find that

$$\limsup_{k\to\infty} \int_{\mathbb{R}^d} |u - u_k|\,dx \leq \int_{\mathbb{R}^d} \lim_{k\to\infty} |u_k - u|\,dx = 0.$$

Therefore $\lim_{k\to\infty} \int_U |u - u_k|\,dx = 0$ and we have

$$\left|\int_{\mathbb{R}^d} u\,dx - \int_{\mathbb{R}^d} u_k\,dx\right| = \left|\int_{\mathbb{R}^d} u - u_k\,dx\right| \leq \int_{\mathbb{R}^d} |u_k - u|\,dx \longrightarrow 0$$

as $k \to \infty$, which completes the proof. □

Finally, we recall Egoroff's Theorem.

Theorem A.29 (Egoroff's Theorem)**.** *Let $u_k : \mathbb{R}^d \to \mathbb{R}$ be a sequence of measurable functions, and assume there is a measurable function $u : \mathbb{R}^d \to \mathbb{R}$ such that*

$$\lim_{k\to\infty} u_k(x) = u(x) \quad a.e. \text{ in } A,$$

where $A \subset \mathbb{R}^d$ is measurable and $|A| < \infty$. Then for each $\varepsilon > 0$ there exists a measurable subset $E \subset A$ such that $|A - E| \leq \varepsilon$ and $u_k \to u$ uniformly on E.

A.7 Integration by parts

All of the sets $U \subset \mathbb{R}^d$ that we work with will be assumed to be open and bounded with a smooth boundary ∂U. A set $U \subset \mathbb{R}^d$ has a smooth boundary if at each point $x \in \partial U$ we can make an orthogonal change of coordinates so that for some $r > 0$, $\partial U \cap B(0,r)$ is the graph of a smooth function $u : \mathbb{R}^{n-1} \to \mathbb{R}$. If ∂U is smooth, we can define an outward normal vector $\nu = \nu(x)$ at each point $x \in \partial U$, and ν varies smoothly with x. Here, $\nu = (\nu_1, \dots, \nu_n) \in \mathbb{R}^d$ and ν is a unit vector so

$$|\nu| = \sqrt{\nu_1^2 + \cdots + \nu_n^2} = 1.$$

The *normal derivative* of $u \in C^1(\overline{U})$ at $x \in \partial U$ is

$$\frac{\partial u}{\partial \nu}(x) := \nabla u(x) \cdot \nu(x).$$

Integration by parts in $\mathbb{R}^d$ is based on the Gauss-Green Theorem.

Theorem A.30 (Gauss-Green Theorem)**.** *Let $U \subset \mathbb{R}^d$ be an open and bounded set with a smooth boundary ∂U. If $u \in C^1(\overline{U})$ then*

$$\int_U u_{x_i}\, dx = \int_{\partial U} u\nu_i \, dS.$$

The Gauss-Green Theorem is the natural extension of the fundamental theorem of calculus to dimensions $n \geq 2$. A proof of the Gauss-Green Theorem is outside the scope of this course.

We can derive a great many important integration by parts formulas from the Gauss-Green Theorem. These identities are often referred to as Green's identities or simply integration by parts.

Theorem A.31 (Integration by parts)**.** *Let $U \subset \mathbb{R}^d$ be an open and bounded set with a smooth boundary ∂U. If $u, v \in C^2(\overline{U})$ then*

(i) $\displaystyle\int_U u\Delta v\, dx = \int_{\partial U} u\frac{\partial v}{\partial \nu}\, dS - \int_U \nabla u \cdot \nabla v\, dx,$

(ii) $\displaystyle\int_U u\Delta v - v\Delta u\, dx = \int_{\partial U} u\frac{\partial v}{\partial \nu} - v\frac{\partial u}{\partial \nu}\, dS,$ *and*

(iii) $\displaystyle\int_U \Delta v\, dx = \int_{\partial U} \frac{\partial v}{\partial \nu}\, dS.$

Proof. (i) Notice that

$$\partial_{x_i}(uv_{x_i}) = u_{x_i}v_{x_i} + uv_{x_ix_i}.$$

Applying the Gauss-Green Theorem to uv_{x_i} we have

$$\int_{\partial U} uv_{x_i}\nu_i \, dS = \int_U u_{x_i}v_{x_i} + uv_{x_ix_i} \, dx.$$

Summing over i we have

$$\int_{\partial U} u\frac{\partial v}{\partial \nu} \, dS = \int_U \nabla u \cdot \nabla v + u\Delta v \, dx,$$

which is equivalent to (i).

(ii) Swap the roles of u and v in (i) and subtract the resulting identities to prove (ii).

(iii) Take $u = 1$ in (i). □

It will also be useful to prove the following version of the divergence theorem. Recall that for a vector field $F(x) = (F^1(x), \ldots, F^d(x))$ the divergence of F is

$$\text{div}(F) = F^1_{x_1} + F^2_{x_2} + \cdots + F^d_{x_n}.$$

Theorem A.32 (Divergence theorem)**.** *Let $U \subset \mathbb{R}^d$ be an open and bounded set with a smooth boundary ∂U. If $u \in C^1(\overline{U})$ and F is a C^1 vector field (i.e., $F^i \in C^1(\overline{U})$ for all i) then*

$$\int_U u \, div(F) \, dx = \int_{\partial U} u \, F \cdot \nu \, dS - \int_U \nabla u \cdot F \, dx.$$

Proof. The proof is similar to Theorem A.31 (i). Notice that

$$(uF^i)_{x_i} = u_{x_i}F^i + uF^i_{x_i},$$

and apply the Gauss-Green Theorem to find that

$$\int_{\partial U} uF^i\nu_i \, dS = \int_U u_{x_i}F^i + uF^i_{x_i} \, dx.$$

Summing over i we have

$$\int_{\partial U} u \, F \cdot \nu \, dS = \int_U \nabla u \cdot F + u \, \text{div}(F) \, dx,$$

which is equivalent to the desired result. □

Notice that when $u = 1$ Theorem A.32 reduces to

$$\int_U \text{div}(F) \, dx = \int_{\partial U} F \cdot \nu \, dS,$$

which is the usual divergence theorem. If we take $F = \nabla v$ for $v \in C^2(\overline{U})$, then we recover Theorem A.31 (i).

Exercise A.33. Let $u, w \in C^2(\overline{U})$ where $U \subset \mathbb{R}^d$ is open and bounded. Show that for $1 \leq p < \infty$

$$\int_U u \, \Delta_p w \, dx = \int_{\partial U} u \, |\nabla w|^{p-2}\frac{\partial w}{\partial \nu} \, dS - \int_U |\nabla w|^{p-2}\nabla u \cdot \nabla w \, dx.$$

The p-Laplacian Δ_p was defined in Exercise A.17. △

A.8 Convex functions

Here, we review some basic theory of convex functions.

Definition A.34. A function $u : \mathbb{R}^d \to \mathbb{R}$ is *convex* if

$$u(\lambda x + (1-\lambda)y) \le \lambda u(x) + (1-\lambda)u(y) \tag{A.22}$$

for all $x, y \in \mathbb{R}^d$ and $\lambda \in (0,1)$.

Definition A.35 (Strongly convex)**.** Let $\theta \ge 0$. A function $u : \mathbb{R}^d \to \mathbb{R}$ is *θ-strongly convex* if $u - \frac{\theta}{2}|x|^2$ is convex.

Exercise A.36. Show that u is θ-strongly convex for $\theta \ge 0$ if and only if

$$u(\lambda x + (1-\lambda)y) + \frac{\theta}{2}\lambda(1-\lambda)|x-y|^2 \le \lambda u(x) + (1-\lambda)u(y) \tag{A.23}$$

for all $x, y \in \mathbb{R}^d$ and $\lambda \in (0,1)$. The statement in (A.23) is often given as the definition of strong convexity. △

Lemma A.37. *Let $u : \mathbb{R} \to \mathbb{R}$ be twice continuously differentiable and $\theta \ge 0$. The following are equivalent*

(i) u is θ-strongly convex.

(ii) $u''(x) \ge \theta$ for all $x \in \mathbb{R}$.

(iii) $u(y) \ge u(x) + u'(x)(y-x) + \frac{\theta}{2}(y-x)^2$ for all $x, y \in \mathbb{R}$.

(iv) $(u'(x) - u'(y))(x-y) \ge \theta(x-y)^2$ for all $x, y \in \mathbb{R}$.

Proof. It is enough to prove the result for $\theta = 0$. Then if any statement holds for $\theta > 0$, we can define $v(x) = u(x) - \frac{\theta}{2}|x|^2$ and use the results for v with $\theta = 0$.

The proof is split into three parts.

1. (i) $\implies$ (ii): Assume u is convex. Let $x_0 \in \mathbb{R}$ and set $\lambda = \frac{1}{2}$, $x = x_0 - h$, and $y = x_0 + h$ for a real number h. Then

$$\lambda x + (1-\lambda)y = \frac{1}{2}(x_0 - h) + \frac{1}{2}(x_0 + h) = x_0,$$

and the convexity condition (A.22) yields

$$u(x_0) \le \frac{1}{2}u(x_0 - h) + \frac{1}{2}u(x_0 + h).$$

Therefore

$$u(x_0 - h) - 2u(x_0) + u(x_0 + h) \ge 0$$

for all h, and so

$$u''(x_0) = \lim_{h\to 0} \frac{u(x_0-h) - 2u(x_0) + u(x_0+h)}{h^2} \geq 0.$$

2. (ii) $\implies$ (iii): Assume that $u''(x) \geq 0$ for all x. Then recalling (A.4) and (A.5) we have

$$u(y) = u(x) + u'(x)(y-x) + R_2(x,y)$$

where $R_2(x,y) \geq 0$ for all x, y, which completes the proof.

3. (iii) $\implies$ (iv): Assume (iii) holds. Then we have

$$u(y) \geq u(x) + u'(x)(y-x)$$

and

$$u(x) \geq u(y) + u'(y)(x-y).$$

Subtracting the equations above proves that (iv) holds.

4. (iv) $\implies$ (i): Assume (iv) holds. Then u' is nondecreasing and so $u''(x) \geq 0$ for all $x \in \mathbb{R}$, hence (ii) holds, and so does (iii). Let $x, y \in \mathbb{R}^d$ and $\lambda \in (0,1)$, and set $x_0 = \lambda x + (1-\lambda)y$. Define

$$L(z) = u(x_0) + u'(x_0)(z - x_0).$$

By (iii) we have $u(z) \geq L(z)$ for all z. Therefore

$$u(\lambda x + (1-\lambda)y) = u(x_0) = \lambda L(x) + (1-\lambda)L(y) \leq \lambda u(x) + (1-\lambda)u(y),$$

and so u is convex. □

Lemma A.37 has a natural higher dimensional analog, but we first need some new notation. For a symmetric real-valued $n \times n$ matrix $A = (a_{ij})_{i,j=1}^d$, we write

$$A \geq 0 \iff v \cdot Av = \sum_{i,j=1}^d a_{ij} v_i v_j \geq 0, \ \forall v \in \mathbb{R}^d. \tag{A.24}$$

For symmetric real valued matrices A, B, we write $A \geq B$ if $A - B \geq 0$.

Theorem A.38. *Let $u : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable. The following are equivalent*

(i) u is θ-strongly convex.

(ii) $\nabla^2 u \geq \theta I$ for all $x \in \mathbb{R}^d$.

(iii) $u(y) \geq u(x) + \nabla u(x) \cdot (y - x) + \frac{\theta}{2}|y-x|^2$ for all $x, y \in \mathbb{R}^d$.

(iv) $(\nabla u(x) - \nabla u(y)) \cdot (x - y) \geq \theta |x-y|^2$ for all $x, y \in \mathbb{R}^d$.

Proof. Again, we may prove the result just for $\theta = 0$. The proof follows mostly from Lemma A.37, with some additional observations.

1. (i) $\implies$ (ii): Assume u is convex. Since convexity is defined along lines, we see that $g(t) = u(x + tv)$ is convex for all $x, v \in \mathbb{R}^d$, and by Lemma A.37 $g''(t) \geq 0$ for all t. By (A.10) we have

$$g''(t) = \frac{d^2}{dt^2} u(x + tv) = \sum_{i=1}^{d} \sum_{j=1}^{d} u_{x_i x_j}(x) v_i v_j = v \cdot \nabla^2 u(x) v, \tag{A.25}$$

and so $\nabla^2 u(x) \geq 0$ for all $x \in \mathbb{R}^d$.

2. (ii) $\implies$ (iii): Assume (ii) holds and let $g(t) = u(x + tv)$ for $x, v \in \mathbb{R}^d$. Let $y \in \mathbb{R}^d$. Then by (A.25) we have $g''(t) \geq 0$ for all t, and so by Lemma A.37

$$g(t) \geq g(s) + g'(s)(t - s)$$

for all s, t. Set $v = y - x$, $t = 1$ and $s = 0$ to obtain

$$u(y) \geq u(x) + \nabla u(x) \cdot (y - x),$$

where we used the fact that

$$g'(0) = \frac{d}{dt}\Big|_{t=0} u(x + tv) = \nabla u \cdot v.$$

3. (iii) $\implies$ (iv): The proof is similar to Lemma A.37.

4. (iv) $\implies$ (i): Assume (iv) holds, and define $g(t) = u(x + tv)$ for $x, v \in \mathbb{R}^d$. Then we have

$$(g'(t) - g'(s))(t - s) = (\nabla u(x + tv) - \nabla u(x + sv)) \cdot v(t - s) \geq 0$$

for all t, s. By Lemma A.37 we have that g is convex for al $x, v \in \mathbb{R}^d$, from which it easily follows that u is convex. □

A.9 Probability

Here, we give a brief overview of basic probability. For more details we refer the reader to [25].

A.9.1 Basic definitions

A *probability space* is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is a σ-algebra of measurable subsets of Ω and $\mathbb{P}$ is a nonegative measure on $\mathcal{F}$ with $\mathbb{P}(\Omega) = 1$ (i.e., a probability measure). Each $A \subset \Omega$ with $A \in \mathcal{F}$ is an event, with probability $\mathbb{P}(A)$. We think of each $\omega \in \Omega$ as a trial and if $\omega \in A$ then event A occured. For two events $A, B \in \mathcal{F}$

the union $A \cup B$ is the event that A or B occured, and the intersection $A \cap B$ is the event that both A and B occured. By subadditivity of measures we have

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B),$$

which is called the *union bound*.

Example A.2. Consider rolling a 6-sided die. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F}$ consists of all subsets of Ω, and $\mathbb{P}(A) = \#A/6$. If we roll the die twice, then $\Omega = \{1, 2, 3, 4, 5, 6\}^2$ and $\mathbb{P}(A) = \#A/36$. △

Example A.3. Consider drawing a number uniformly at random in the interval $[0, 1]$. Here, $\Omega = [0, 1]$, $\mathcal{F}$ is all Lebesgue measureable subsets of $[0, 1]$, and $\mathbb{P}(A)$ is the Lebesgue measure of $A \in \mathcal{F}$. △

We will from now on omit the σ-algebra $\mathcal{F}$ when referring to probability spaces.

Let $(\Omega, \mathbb{P})$ be a probability space. A *random variable* is a measurable function $X : \Omega \to \mathbb{R}^d$. That is, to each trial $\omega \in \Omega$ we associate the value $X(\omega)$.

Example A.4. In Example (A.2), suppose we win 10 times the number on the die in dollars. Then the random variable $X(\omega) = 10\omega$ describes our winnings. △

The image of Ω under X, denoted $\Omega^X = \{X(\omega) : \omega \in \Omega\} \subset \mathbb{R}^d$ is the *sample space* of X, and we often say X is a random variable on Ω^X. The random variable $X : \Omega \to \Omega^X$ defines a measure on Ω^X which we denote by $\mathbb{P}_X$. Indeed, for any $B \subset \Omega^X$, the probability that X lies in B, written $\mathbb{P}_X(X \in B)$ is

$$\mathbb{P}_X(X \in B) := \mathbb{P}(X^{-1}(B)).$$

With this new notation we can write

$$\mathbb{P}_X(X \in B) = \int_B d\,\mathbb{P}_X(x).$$

We say that X has a *density* if there exists a nonnegative Lebesgue measurable $\rho : \Omega^X \to \mathbb{R}$ such that

$$\mathbb{P}_X(X \in B) = \int_B \rho(x)\,dx.$$

Let $g : \Omega^X \to \mathbb{R}^m$. Then $Y = g(X)$ is a random variable. We define the *expectation* $\mathbb{E}_X[g(X)]$ to be

$$\mathbb{E}_X[g(X)] = \int_{\Omega^X} g(x)\,d\,\mathbb{P}_X(x) = \int_\Omega g(X(\omega))\,d\,\mathbb{P}(\omega).$$

In particular

$$\mathbb{E}_X[X] = \int_{\Omega^X} x\,d\,\mathbb{P}_X(x) = \int_\Omega X(\omega)\,d\,\mathbb{P}(\omega).$$

If X has a density then

$$\mathbb{E}_X[g(X)] = \int_{\Omega^X} g(x)\rho(x)\,dx.$$

We note that the expectation is clearly linear, so that

$$\mathbb{E}_X[f(X) + g(X)] = \mathbb{E}_X[f(X)] + \mathbb{E}_X[g(X)],$$

due to linearity of the integral.

A.9.2 Markov and Chebyshev inequalities

We introduce here basic estimates for bounding probabilities of random variables. An important result is Markov's inequality.

Proposition A.39 (Markov's inequality)**.** *Let $(\Omega, \mathbb{P})$ be a probability space and $X : \Omega \to [0, \infty)$ be a nonnegative random variable. Then for any $t > 0$*

$$\mathbb{P}_X(X \geq t) \leq \frac{\mathbb{E}_X[X]}{t}. \tag{A.26}$$

Proof. By definition we have

$$\mathbb{P}_X(X \geq t) = \int_t^\infty d\,\mathbb{P}_X(x) \leq \int_t^\infty \frac{x}{t}\,d\,\mathbb{P}_X(x) = \frac{1}{t}\int_0^\infty x\,d\,\mathbb{P}_X(x) = \frac{\mathbb{E}_X[X]}{t}. \qquad \square$$

Markov's inequality can be improved if we have information about the variance of X. We define the *variance* of a random variable X as

$$\mathrm{Var}(X) = \mathbb{E}_X[(X - \mathbb{E}_X[X])^2]. \tag{A.27}$$

Proposition A.40 (Chebyshev's inequality)**.** *Let $(\Omega, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable with finite mean $\mathbb{E}_X[X]$ and variance Var(X). Then for any $t > 0$*

$$\mathbb{P}_X(|X - \mathbb{E}_X[X]| \geq t) \leq \frac{Var(X)}{t^2}. \tag{A.28}$$

Proof. Let $Y = (X - \mathbb{E}_X[X])^2$. Then Y is a nonegative random variable and by Markov's inequality (A.26) we have

$$\mathbb{P}_X(|X - \mathbb{E}_X[X]| \geq t) = \mathbb{P}_X(Y \geq t^2) \leq \frac{\mathbb{E}_X[Y]}{t^2} = \frac{\mathrm{Var}\,(X)}{t^2}. \qquad \square$$

A.9.3 Sequences of indepedendent random variables

Let $(\Omega, \mathbb{P})$ be a probability space, and $X : \Omega \to \mathbb{R}^d$ be a random variable. We often want to construct other *independent* copies of the random variable X. For example, if we roll a die several times, then we have many instances of the same random variable. We clearly cannot use the same probability space for each roll of the die, otherwise all the rolls would always produce the same value (and would not be indepedent).

To construct an independent copy of X, we consider the product probability space $(\Omega \times \Omega, \mathbb{P} \times \mathbb{P})$ with the product probability measure $\mathbb{P} \times \mathbb{P}$. The product measure is the unique measure satisfying

$$(\mathbb{P} \times \mathbb{P})(A \times B) = \mathbb{P}(A)\mathbb{P}(B)$$

for all measurable $A, B \subset \Omega$. On the product probability space $\Omega^2 = \Omega \times \Omega$ the two independent copies of X are constructed via the random variable

$$(\omega_1, \omega_2) \mapsto (X(\omega_1), X(\omega_2)).$$

We normally give the random variables different names, so that $X_1(\omega_1, \omega_2) := X(\omega_1)$ and $X_2(\omega_1, \omega_2) := X(\omega_2)$. Then X_1 and X_2 are themselves random variables (now on Ω^2), and we say X_1 and X_2 are independent random variables with the same distribution as X, or *independent and identically distributed* random variables.

An important property concerns the expectation of products of independent random variables. If X_1 and X_2 are independent and identically distributed random variables with the same distribution as X (as above) then

$$\mathbb{E}_{(X_1,X_2)}[f(X_1)g(X_2)] = \mathbb{E}_X[f(X)]\mathbb{E}_X[g(X)]. \tag{A.29}$$

Indeed, we have

$$\begin{aligned}\mathbb{E}_{(X_1,X_2)}[f(X_1)g(X_2)] &= \int_\Omega \int_\Omega f(x)g(y)\, d\,\mathbb{P}_X(x)\, d\,\mathbb{P}_X(y) \\ &= \int_\Omega f(x)\, d\,\mathbb{P}_X(x) \int_\Omega g(y)\, d\,\mathbb{P}_X(y) \\ &= \mathbb{E}_X[f(X)]\mathbb{E}_X[g(X)].\end{aligned}$$

We also notice that

$$\mathbb{E}_X[f(X)] = \mathbb{E}_{(X_1,X_2)}[f(X_1)],$$

since

$$\mathbb{E}_{(X_1,X_2)}[f(X_1)] = \int_\Omega \int_\Omega f(x)\, d\,\mathbb{P}_X(x)\, d\,\mathbb{P}_X(y) = \int_\Omega f(x)\, d\,\mathbb{P}_X(x) = \mathbb{E}_X[f(x)].$$

We can continue constructing as many independent and identically distributed copies of X as we like. The construction is as follows. Let $n \geq 1$ and consider the product probability space $(\Omega^n, \mathbb{P}^n)$ with product measure

$$\mathbb{P}^n = \underbrace{\mathbb{P} \times \mathbb{P} \times \cdots \times \mathbb{P}}_{n \text{ times}}.$$

For $i = 1, \dots, n$ we define the random variable $X_i : \Omega^n \to \mathbb{R}^d$ by

$$X_i(\omega_1, \omega_2, \dots, \omega_n) = X(\omega_i).$$

We say that $X_1, X_2, \dots, X_n$ is a sequence of n *independent and identically distributed* (*i.i.d.*) random variables. It is important to note how all X_i for $i = 1, \dots, n$ are defined on the same probability space, which allows us to compute probabilities involving all the n random variables. As above, we have the product of expectations formula

$$\mathbb{E}_{(X_1,X_2,\dots,X_n)}[f_1(X_1)f_2(X_2)\cdots f_n(X_n)] = \prod_{i=1}^n \mathbb{E}_X[f_i(X)]. \tag{A.30}$$

We leave it to the reader to verify (A.30). In applications of probability theory, we will not burden the notation and will write $\mathbb{P}$ in place of $\mathbb{P}^n$ and $\mathbb{E}$ in place of $\mathbb{E}_X$ and $\mathbb{E}_{(X_1,X_2,\dots,X_n)}$. It will almost always be clear from context which probability measures and expectations are being used, and when it is not clear we will specifically denote the dependence. As above we have

$$\mathbb{E}_X[f(X)] = \mathbb{E}_{(X_1,X_2,\dots,X_n)}[f(X_i)],$$

for any i, so the choice of which expectation to use is irrelevant. Since we do not wish to always specify the base random variable X on which the sequence is constructed, we often write X_1 or X_i in place of X.

A.9.4 Law of large numbers

To get some practice using probability, we give a proof of the weak law of large numbers, using only the tools from Sections A.9.2 and A.9.3.

Theorem A.41 (Weak law of large numbers)**.** *Let $X_1, \dots, X_n$ be a sequence of independent and identically distributed random variables with finite mean $\mu := \mathbb{E}[X_i]$ and variance $\sigma^2 := Var(X_i)$. Let $S_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then for every $\varepsilon > 0$ we have*

$$\lim_{n\to\infty} \mathbb{P}(|S_n - \mu| \geq \varepsilon) = 0. \tag{A.31}$$

Remark A.42. The limit in (A.31) shows that $S_n \to \mu$ in probability as $n \to \infty$, which is known as the *weak law of large numbers*. In fact, inspecting the proof below, we have proved the slightly stronger statement

$$\lim_{n\to\infty} \mathbb{P}(|S_n - \mu| \geq \varepsilon n^{-\alpha}) = 0,$$

for any $\alpha \in (0, \frac{1}{2})$.

Proof. Note that $\mathbb{E}[S_n] = \mu$ and compute

$$\begin{aligned}\mathrm{Var}\,(S_n) &= \mathbb{E}[(S_n-\mu)^2] \\ &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n X_i - \mu\right)^2\right] \\ &= \frac{1}{n^2}\mathbb{E}\left[\sum_{i,j=1}^n (X_i-\mu)(X_j-\mu)\right] \\ &= \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}[(X_i-\mu)(X_j-\mu)].\end{aligned}$$

If $i \neq j$, then due to (A.30) we have $\mathbb{E}[(X_i-\mu)(X_j-\mu)] = 0$ and so

$$\mathrm{Var}\,(S_n) = \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}[(X_i-\mu)^2] = \frac{\sigma^2}{n}.$$

By Chebyshev's inequality (Proposition A.40) we have

$$\mathbb{P}(|S_n-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

for all $\varepsilon > 0$, which completes the proof. □

A.10 Miscellaneous results

A.10.1 Vanishing lemma

Lemma A.43. *Let $U \subset \mathbb{R}^d$ be open and bounded and let $u \in C(U)$. If*

$$\int_U u(x)\varphi(x)\,dx = 0 \quad \textit{for all } \varphi \in C_c^\infty(U)$$

then $u(x) = 0$ for all $x \in U$.

Proof. Let us sketch the proof. Assume to the contrary that $u(x_0) \neq 0$ at some $x_0 \in U$. We may assume, without loss of generality that $\varepsilon := u(x_0) > 0$. Since u is continuous, there exists $\delta > 0$ such that

$$u(x) \geq \frac{\varepsilon}{2} \quad \text{whenever } |x-x_0| < \delta.$$

Now let $\varphi \in C_c^\infty(U)$ be a test function satisfying $\varphi(x) > 0$ for $|x-x_0| < \delta$ and $\varphi(x) = 0$ for $|x-x_0| \geq \delta$. Then

$$0 = \int_U u(x)\varphi(x)\,dx = \int_{B(x_0,\delta)} u(x)\varphi(x)\,dx \geq \frac{\varepsilon}{2}\int_{B(x_0,\delta)} \varphi(x)\,dx > 0,$$

which is a contradiction. □

A test function satisfying the requirements in the proof of Lemma A.43 is given by

$$\varphi(x) = \begin{cases} \exp\left(-\frac{1}{\delta^2 - |x-x_0|^2}\right), & \text{if } |x-x_0| < \delta \\ 0, & \text{if } |x-x_0| \geq \delta. \end{cases}$$

A.10.2 Total variation of characteristic function is perimeter

Let $U \subset \mathbb{R}^d$ be an open and bounded set with a smooth boundary, and let

$$\chi_U(x) = \begin{cases} 1, & \text{if } x \in U \\ 0, & \text{otherwise.} \end{cases} \tag{A.32}$$

We show here that the length, also called perimeter, of ∂U, denoted $\mathrm{Per}(\partial U)$, is given by

$$\mathrm{Per}(\partial U) = \int_{\mathbb{R}^d} |\nabla \chi_U| \, dx. \tag{A.33}$$

We first need to understand how to interpret the total variation $\int_{\mathbb{R}^d} |\nabla u| \, dx$ for a non-smooth function $u : \mathbb{R}^d \to \mathbb{R}$. For this, we first note that when $u \in C^\infty(\mathbb{R}^d)$ with compact support we have

$$\int_{\mathbb{R}^d} |\nabla u| \, dx = \sup_{\substack{\varphi \in C^\infty(\mathbb{R}^d, \mathbb{R}^d) \\ |\varphi| \leq 1}} \int_{\mathbb{R}^d} \nabla u \cdot \varphi \, dx.$$

Indeed, the equality is immediate, since for any φ we have $\nabla u \cdot \varphi \leq |\nabla u|$ due to the condition $|\varphi| \leq 1$, and we can take $\varphi = \frac{\nabla u}{|\nabla u|}$ to saturate the inequality. Since u has compact support in $\mathbb{R}^d$, we can integrate by parts to obtain the identity

$$\int_{\mathbb{R}^d} |\nabla u| \, dx = \sup_{\substack{\varphi \in C^\infty(\mathbb{R}^d, \mathbb{R}^d) \\ |\varphi| \leq 1}} \int_{\mathbb{R}^d} u \, \mathrm{div} \varphi \, dx. \tag{A.34}$$

The identity (A.34) is taken to be the definition of the total variation $\int_{\mathbb{R}^d} |\nabla u| \, dx$ when u is not differentiable.

Therefore we have

$$\begin{aligned} \int_{\mathbb{R}^d} |\nabla \chi_U| \, dx &= \sup_{\substack{\varphi \in C^\infty(\mathbb{R}^d, \mathbb{R}^d) \\ |\varphi| \leq 1}} \int_{\mathbb{R}^d} \chi_U(x) \, \mathrm{div} \varphi(x) \, dx \\ &= \sup_{\substack{\varphi \in C^\infty(\mathbb{R}^d, \mathbb{R}^d) \\ |\varphi| \leq 1}} \int_U \mathrm{div} \varphi(x) \, dx \\ &= \sup_{\substack{\varphi \in C^\infty(\mathbb{R}^d, \mathbb{R}^d) \\ |\varphi| \leq 1}} \int_{\partial U} \varphi \cdot \nu \, dS. \end{aligned}$$

Since $\varphi \cdot \nu \leq 1$ and the choice of $\varphi = \nu$ for any smooth extension of ν to $\mathbb{R}^d$ yields $\varphi \cdot \nu = 1$, we have

$$\int_{\mathbb{R}^d} |\nabla \chi_U| \, dx = \int_{\partial U} dS = \mathrm{Per}(\partial U).$$

Bibliography

[1] R. K. Ando and T. Zhang. Learning on graph with Laplacian regularization. In *Advances in Neural Information Processing Systems*, pages 25–32, 2007.

[2] K. Andreev and H. Racke. Balanced graph partitioning. *Theory of Computing Systems*, 39(6):929–939, 2006.

[3] H. Attouch, X. Goudou, and P. Redont. The heavy ball with friction method, I. The continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. *Communications in Contemporary Mathematics*, 2(01):1–34, 2000.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, pages 585–591, 2002.

[5] M. Belkin and P. Niyogi. Semi-supervised learning on riemannian manifolds. *Machine learning*, 56(1-3):209–239, 2004.

[6] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

[7] M. Benyamin, J. Calder, G. Sundaramoorthi, and A. Yezzi. Accelerated PDE's for efficient solution of regularized inversion problems. *Journal of Mathematical Imaging and Vision*, 62(1), 2020.

[8] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

[9] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[10] D. Burago, S. Ivanov, and Y. Kurylev. A graph discretization of the Laplace-Beltrami operator. *Journal of Spectral Theory*, 4(4):675–714, 2014.

[11] J. Calder. The game theoretic p-Laplacian and semi-supervised learning with few labels. *Nonlinearity*, 32(1), 2018.

[12] J. Calder. Lecture notes on viscosity solutions. *Lecture notes*, 2018.

[13] J. Calder. Consistency of Lipschitz learning with infinite unlabeled data and finite labeled data. *SIAM Journal on Mathematics of Data Science*, 1(4):780–812, 2019.

[14] J. Calder and N. García Trillos. Improved spectral convergence rates for graph Laplacians on ε-graphs and k-NN graphs. *arXiv preprint*, 2019.

[15] J. Calder and D. Slepčev. Properly-weighted graph Laplacian for semi-supervised learning. *Applied Mathematics and Optimization: Special Issue on Optimization in Data Science*, 2019.

[16] J. Calder, D. Slepčev, and M. Thorpe. Rates of convergence for Laplacian semi-supervised learning with low labelling rates. *In preparation*, 2019.

[17] J. Calder and A. Yezzi. PDE Acceleration: A convergence rate analysis and applications to obstacle problems. *Research in the Mathematical Sciences*, 6(35), 2019.

[18] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20(1-2):89–97, 2004.

[19] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[20] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277, 2001.

[21] O. Chapelle, B. Scholkopf, and A. Zien. *Semi-supervised learning.* MIT, 2006.

[22] B. Dacorogna. *Direct methods in the calculus of variations*, volume 78. Springer Science & Business Media, 2007.

[23] E. de Giorgi. Sulla differenziabilita e l'analiticita delle estremali degli integrali multipli regolari. *Mem. Accad. Sci. Torino. Cl. Sci. Fis. Mat. Nat*, 3(3):25–43, 1957.

[24] P. K. Diederik, M. Welling, et al. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

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

[26] A. El Alaoui, X. Cheng, A. Ramdas, M. J. Wainwright, and M. I. Jordan. Asymptotic behavior of ℓ_p-based Laplacian regularization in semi-supervised learning. In *Conference on Learning Theory*, pages 879–906, 2016.

[27] S. Esedog, Y.-H. R. Tsai, et al. Threshold dynamics for the piecewise constant mumford–shah functional. *Journal of Computational Physics*, 211(1):367–384, 2006.

[28] L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. AMS, Providence, Rhode Island, 1998.

[29] M. Flores, J. Calder, and G. Lerman. Algorithms for Lp-based semi-supervised learning on graphs. *arXiv:1901.05031*, 2019.

[30] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*. springer, 2015.

[31] T. Goldstein and S. Osher. The split bregman method for l1-regularized problems. *SIAM journal on imaging sciences*, 2(2):323–343, 2009.

[32] J. He, M. Li, H.-J. Zhang, H. Tong, and C. Zhang. Manifold-ranking based image retrieval. In *Proceedings of the 12th annual ACM International Conference on Multimedia*, pages 9–16. ACM, 2004.

[33] J. He, M. Li, H.-J. Zhang, H. Tong, and C. Zhang. Generalized manifold-ranking-based image retrieval. *IEEE Transactions on Image Processing*, 15(10):3170–3177, 2006.

[34] M. Hein, J.-Y. Audibert, and U. v. Luxburg. Graph Laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8(Jun):1325–1368, 2007.

[35] M. Hein, J.-Y. Audibert, and U. Von Luxburg. From graphs to manifolds–weak and strong pointwise consistency of graph Laplacians. In *International Conference on Computational Learning Theory*, pages 470–485. Springer, 2005.

[36] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[38] R. Kyng, A. Rao, S. Sachdeva, and D. A. Spielman. Algorithms for Lipschitz learning on graphs. In *Conference on Learning Theory*, pages 1190–1223, 2015.

[39] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[40] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[41] B. Merriman, J. K. Bence, and S. Osher. Diffusion generated motion by mean curvature. In J. Taylor, editor, *Computational Crystal Growers Workshop*, pages 73–83. Sel. Lectures Math, AMS, Providence, RI, 1992.

[42] C. B. Morrey. On the solutions of quasi-linear elliptic partial differential equations. *Transactions of the American Mathematical Society*, 43(1):126–166, 1938.

[43] J. Moser. A new proof of de Giorgi's theorem concerning the regularity problem for elliptic differential equations. *Communications on Pure and Applied Mathematics*, 13(3):457–468, 1960.

[44] J. Nash. Continuity of solutions of parabolic and elliptic equations. *American Journal of Mathematics*, 80(4):931–954, 1958.

[45] Y. Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[46] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2002.

[47] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489, 2005.

[48] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[49] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[50] W. Rudin. *Functional analysis*. McGraw-hill, 1991.

[51] W. Rudin. *Real and complex analysis*. Tata McGraw-hill education, 2006.

[52] J. Shi and J. Malik. Normalized cuts and image segmentation. *Departmental Papers (CIS)*, page 107, 2000.

[53] Z. Shi, S. Osher, and W. Zhu. Weighted nonlocal Laplacian on interpolation from sparse data. *Journal of Scientific Computing*, 73(2-3):1164–1177, 2017.

[54] Z. Shi, B. Wang, and S. J. Osher. Error estimation of weighted nonlocal Laplacian on random point cloud. *arXiv preprint arXiv:1809.08622*, 2018.

[55] D. Slepčev and M. Thorpe. Analysis of p-Laplacian regularization in semisupervised learning. *SIAM Journal on Mathematical Analysis*, 51(3):2085–2120, 2019.

[56] G. Sundaramoorthi and A. Yezzi. Variational PDE's for acceleration on manifolds and applications to diffeomorphisms. *Neural Information Processing Systems*, 2018.

[57] D. Ting, L. Huang, and M. Jordan. An analysis of the convergence of graph Laplacians. *arXiv:1101.5435*, 2011.

[58] N. G. Trillos, M. Gerlach, M. Hein, and D. Slepcev. Error estimates for spectral convergence of the graph laplacian on random geometric graphs towards the laplace–beltrami operator. *arXiv preprint arXiv:1801.10108*, 2018.

[59] N. G. Trillos and R. Murray. A maximum principle argument for the uniform convergence of graph laplacian regressors. *arXiv preprint arXiv:1901.10089*, 2019.

[60] N. G. Trillos and D. Slepčev. On the rate of convergence of empirical measures in ∞-transportation distance. *Canadian Journal of Mathematics*, 67(6):1358–1383, 2015.

[61] N. G. Trillos and D. Slepčev. Continuum limit of total variation on point clouds. *Archive for rational mechanics and analysis*, 220(1):193–241, 2016.

[62] Y. Wang, M. A. Cheema, X. Lin, and Q. Zhang. Multi-manifold ranking: Using multiple features for better image retrieval. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 449–460. Springer, 2013.

[63] A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[64] B. Xu, J. Bu, C. Chen, D. Cai, X. He, W. Liu, and J. Luo. Efficient manifold ranking for image retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 525–534. ACM, 2011.

[65] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3166–3173, 2013.

[66] A. Yezzi, G. Sundaramoorthi, and M. Benyamin. PDE acceleration for active contours. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[67] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Semi-supervised learning by maximizing smoothness. *J. of Mach. Learn. Research*, 2004.

[68] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, pages 321–328, 2004.

[69] D. Zhou, J. Huang, and B. Schölkopf. Learning from labeled and unlabeled data on a directed graph. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 1036–1043. ACM, 2005.

[70] D. Zhou and B. Schölkopf. Regularization on discrete spaces. In *27th DAGM Conference on Pattern Recognition*, pages 361–368, 2005.

[71] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. In *Advances in Neural Information Processing Systems*, pages 169–176, 2004.

[72] X. Zhou, M. Belkin, and N. Srebro. An iterated graph Laplacian approach for ranking on manifolds. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 877–885. ACM, 2011.

[73] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine learning (ICML-03)*, pages 912–919, 2003.