Variational Methods For Geometric Statistical Inference

by

Matthew Thorpe

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Mathematics Institute

July 2015

THE UNIVERSITY OF WARWICK

Contents

Acknowledgments

Firstly I would like to thank my supervisors Florian Theil, Adam Johansen and Neil Cade. I am very grateful for their help, spotting my mistakes and stimulating discussions. Without their guidance this thesis would not have been possible.

I would like to thank EPSRC and Selex ES Ltd. for funding my placement through a CASE studentship in the MASDOC program at the University of Warwick.

My thanks go to my friends Louise Mason, Tom Hull, Tom Wardell and Dan Wheble for their company over the last years.

I am grateful to everyone I have shared an office with (which for fear of missing someone I will not attempt to name) for answering both my sensible and naive questions. You have all helped me mathematically and distracted me when I needed it. In a related thanks I feel very lucky to have been part of MASDOC which has been an excellent community and is never short of people willing to drink wine on a Friday night.

Thanks also go to my ever increasing family and in particular my parents and grandparents.

I am grateful to both examiners for their comments and feedback and also for both of them being very accommodating when coming to schedule the examination.

Finally I would also like to thank everyone at Warwick for creating an environment that has made the last four years very enjoyable.

Declarations

This thesis contains original and collaborative research carried out during the author's study in MASDOC (mathematics and statistics doctoral training centre) at the University of Warwick. The research was funded by an EPSRC CASE studentship with Selex ES Ltd. This thesis has been composed by myself and has not been submitted for any other degree or professional qualification. The work is my own, except where I have indicated to the contrary within the thesis. More specifically:

i) In Chapter 1, Section 1.1 is edited from a blog written with Neil Cade (Selex ES Ltd.) on the Smith Institute webpage [149].

ii) Chapter 2 is preliminary work in preparation for later chapters and contains no new results.

iii) Chapter 3 was joint work with Neil Cade, Adam Johansen (University of Warwick) and Florian Theil (University of Warwick) and a version has been submitted for publication [154]. Corrections were made on the advise of two anonymous referees.

iv) Chapters 4 and 5 are joint work with Adam Johansen and versions have been submitted for publication [150, 151].

v) Chapters 6 and 7 are joint work with Florian Theil and versions are currently being prepared to be submitted for publication [152, 153].

Abstract

Estimating multiple geometric shapes such as tracks or surfaces creates significant mathematical challenges particularly in the presence of unknown data association. In particular, problems of this type have two major challenges. The first is typically the object of interest is infinite dimensional whilst data is finite dimensional. As a result the inverse problem is ill-posed without regularization. The second is the data association makes the likelihood function highly oscillatory.

The focus of this thesis is on techniques to validate approaches to estimating problems in geometric statistical inference. We use convergence of the large data limit as an indicator of robustness of the methodology. One particular advantage of our approach is that we can prove convergence under modest conditions on the data generating process. This allows one to apply the theory where very little is known about the data. This indicates a robustness in applications to real world problems.

The results of this thesis therefore concern the asymptotics for a selection of statistical inference problems. We construct our estimates as the minimizer of an appropriate functional and look at what happens in the large data limit. In each case we will show our estimates converge to a minimizer of a limiting functional. In certain cases we also add rates of convergence.

The emphasis is on problems which contain a data association or classification component. More precisely we study a generalized version of the k-means method which is suitable for estimating multiple trajectories from unlabeled data which combines data association with spline smoothing. Another problem considered is a graphical approach to estimating the labeling of data points. Our approach uses minimizers of the Ginzburg-Landau functional on a suitably defined graph.

In order to study these problems we use variational techniques and in particular Γ-convergence. This is the natural framework to use for studying sequences of minimization problems. A key advantage of this approach is that it allows us to deal with infinite dimensional and highly oscillatory functionals.

Chapter 1

Introduction

"In theory, there is no difference between theory and practice. But in practice, there is."

- Yogi Berra

1.1 Motivation

Statistical estimators are used to extract information from data sets. Sometimes there may be some *true* value $\mu^\dagger$ which one hopes to recover in the large data limit, e.g. estimating the trajectory of a moving target from space-time measurements. Such a trajectory would form part of the data generating process. In other situation there may not be a true value of the parameter, e.g. deciding how to advertise products to a potential customer based on any available information, such as internet browsing history.

For the first type of problem it is natural to consider data $\{\xi_i = (t_i, z_i)\}_{i=1}^n$ of the form

$$z_i = F(\mu^\dagger, e_i, t_i) \quad i = 1, 2, \dots, n$$

where e_i is a random variable to account for noise, z_i is an observation and t_i is an input parameter. For example in the estimating trajectory problem one simple model is

$$z_i = F(\mu^\dagger, e_i, t_i) := \mu^\dagger(t_i) + e_i.$$

In many applications the function F will have no inverse. Hence one cannot in general use the inverse of F to reconstruct $\mu^\dagger$. One solution is to adopt a Bayesian-like approach.[1] In general one constructs (the maximum-a-posterior) estimate $\mu^{(n)}$ of $\mu^\dagger$ based on data $\{(t_i, z_i)\}_{i=1}^n$ by solving

$$\mu^{(n)} = \operatorname*{argmin}_{\mu} \sum_{i=1}^n |z_i - F(\mu, 0, t_i)|^2 + \lambda_n R(\mu) \tag{1.1}$$

[1]Technically without assumptions such as Gaussian noise and prior and with $\lambda_n = \lambda$ constant one could not use Bayes rule to write the maximum-a-posteriori estimate in the form (1.1). But for the purpose of this discussion we ignore the exact condition one needs on the distributions and with an abuse of notation assume the estimator is the maximum-a-posteriori estimate.

where R is a regularization term and λ_n is some appropriate scaling. In a fully Bayesian approach one should choose $\lambda_n = \lambda$ to be constant and then (under Gaussian assumptions) one can interpret $\lambda R(\mu)$ as the covariance of the prior. This is not always possible as the above minimization problem may become ill-posed, a fact well known in the Bayesian inverse community [4] and in the spline fitting community [46, 117]. In such cases one may be able take $\lambda_n \to \infty$ and still show $\mu^{(n)} \to \mu^\dagger$. There is a very active community that work on results of this type, see for example [4, 55, 72, 134, 139, 162], and references therein.

For the second class of problems there is no $\mu^\dagger$ so it does not make sense to look for F as before. One can see that F allows one to compare estimates with the data and therefore in some sense encodes the data generating model. Therefore without F one has to produce estimates without reference to any such model. As an example let us consider the k-means method which will also be the subject of Chapter 3 and Chapter 4. Given a data set $\{z_i\}_{i=1}^n \subset \mathbb{R}^\kappa$ it is the objective of the k-means method to partition the data into k clusters. This is done by minimizing the functional

$$f_n(\mu) = \sum_{i=1}^n \min_{j=1,\dots,k} |z_i - \mu_j| \, .$$

Minimizers $\mu^{(n)} = (\mu_1^{(n)}, \dots, \mu_k^{(n)}) \in \mathbb{R}^{k\times\kappa}$ of f_n are called cluster centers and the partitioning is defined by associating each data point to the closest center. One can see that f_n does not depend on the data generating model.

The problems we consider in this thesis involve estimators which do not directly use the data generating process (like the k-means method) and may also be ill-posed, requiring regularization as we saw at the start of this section. For example Chapter 3 and Chapter 4 use the k-means method where cluster centers are trajectories and data is space-time observations $\{\xi_i = (t_i, z_i)\}_{i=1}^n$. One can write the estimator as a minimizer of f_n where

$$f_n(\mu) = \frac{1}{n} \sum_{i=1}^n \min_{j=1,\dots,k} |\mu_j(t_i) - z_i|^2 + \lambda \sum_{j=1}^k \|\nabla^2 \mu_j\|_{L^2}^2.$$

Our results concern the asymptotic behavior of such estimators. The convergence in the large data limit is a measure of stability. A lack of convergence indicates ill-posedness. In particular there are two important questions one should consider:

(P1) Do our estimators converge?

(P2) Is there a limiting (large data) problem?

As we are looking at situation without truth then the second question is very important because establishing a limiting problem can provide justification for a choice of estimator. Furthermore, understanding the limit can help one design the finite dimensional problem so that features of the practitioners choice become important. For a sequence of minimization problems numbered according to the number of data points the convergence of the empirical distribution motivates a ‘limiting problem’ that we can understand as having an infinite amount of data.

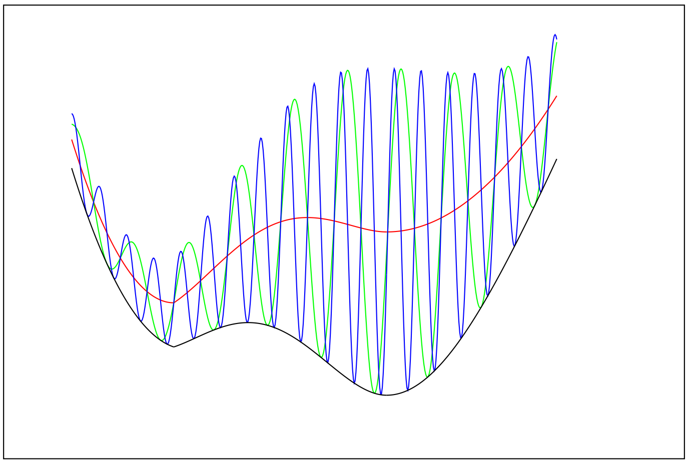
Crudely speaking in the k-means problem we expect

$$\frac{1}{n}\sum_{i=1}^{n}\min_{j=1,\dots,k}|\mu_j(t_i)-z_i|^2\approx\int\min_{j=1,\dots,k}|\mu_j(t)-z|^2\ P(\mathrm{d}(t,z))$$

where $(t_i,z_i)\overset{\text{iid}}{\sim}P$. So then what is the natural notion of convergence for sequence of minimization problems?

If we consider a sequence of oscillating functionals as in Figure 1.1 we see that although the minimum and minimizers are well behaved the function is not (in the sense that there is no strong limit). And whilst the weak limit exists it clearly does not capture the behavior of either minimizer or minimum. Approximately speaking the Γ-limit is the limiting lower semi-continuous envelope. We see in this case the Γ-limit completely captures the behavior of both the minimum and minimizers. Whilst the example in the figure shows functionals acting on a 1 dimensional space the same reasoning carries through to infinite dimensional spaces.

Figure 1.1: The Weak Limit Versus the Γ-Limit



The blue and green curves show two instances of a minimization problem that becomes increasingly oscillatory as the number of data points goes to infinity. The red curve gives the weak limit (i.e. the average over the oscillations) and the black curve is the Γ-limit. Clearly there is no strong limit.

Let us now be a little more precise regarding the application of Γ-convergence to minimization problems. There are two criteria we must show to infer the convergence of minimizers. The first is to find the Γ-limit. Often this will rely on the (almost sure) weak convergence of the empirical measure. The second is to show that minimizers are compact. If minimizers are not compact then we have sequences of estimators that do not converge (nor will any subsequence). Under these two conditions the Γ-convergence framework implies that the minimums converge and that (up to subsequences) minimizers are also convergent. Furthermore the limit of any subsequence of minimizers will minimize the Γ-limit. Hence if the Γ-limit has a unique minimizer then the entire sequence will converge (without the recourse to subsequences). Let us also emphasize that knowledge of the Γ-limit helps one to understand what features one should expect for estimators based on large (but finite) data sets. This is important criteria one can use in order to justify a choice of estimator. In this thesis the Γ-convergence methodology will be our framework to investigate the convergence properties for several examples of estimators.

1.2 Overview of Thesis

In this thesis we investigate questions (P1-P2) stated in the previous section for three problems. We give an overview of each type of problem below.

The k-Means Minimization Problem (Chapter 3 and Chapter 4). The first problem we consider is a k-means type problem where we generalize the k-means framework [105] to allow for cluster centers in different spaces to the data. This is motivated by the following smoothing-data association problem. We are given data $\{(t_i, z_i)\}_{i=1}^n$ sampled from k unknown curves μ_j, and in particular the association of data point to curve is unknown. The problem is then to recover the set of curves $(\mu_1, \dots, \mu_k)$ from $\{(t_i, z_i)\}_{i=1}^n$. By treating the unknown curves as cluster centers one can use the k-means method as estimators.

Our setting is we have data $\xi_i \in X$ and cluster centers $\mu_j \in Y$. The cost function $d : X \times Y \to [0, \infty)$ measures the similarity between a data point and a cluster center. In order for the problem to be well posed we use a regularization term $r : Y^k \to [0, \infty)$ scaled by λ. The object of interest is the optimal cluster centers, that is functions $\mu^{(n)} \in Y^k$ that minimize

$$f_n(\mu) = \frac{1}{n} \sum_{i=1}^n \min_{j=1,\dots,k} d(\xi_i, \mu_j) + \lambda r(\mu).$$

When $X = Y$ we show in Section 3.2 regularization is unnecessary and we can let $\lambda = 0$.

We prove asymptotics concerning the general case in Chapter 3 before we investigate the smoothing data-association further in Chapter 4 where we also prove a rate of convergence. We define f_∞ by

$$f_\infty(\mu) = \int_X \min_{j=1,\dots,k} d(x, \mu_j) \; P(\mathrm{d}x) + \lambda r(\mu)$$

and where $\xi_i \overset{\text{iid}}{\sim} P$. Formally, in Chapter 3 we show for sequences of minimizers $\mu^{(n)}$ and a minimizer $\mu^{(\infty)}$ of f_∞ that:

$$\text{for } X = Y \text{ and } \lambda = 0 \text{ then} \qquad \min_{\mu \in X^k} f_n(\mu) \to \min_{\mu \in X^k} f_\infty(\mu) \qquad \mu^{(n)} \to \mu^{(\infty)}$$

$$\text{for } X \neq Y \text{ and } \lambda > 0 \text{ then} \qquad \min_{\mu \in Y^k} f_n(\mu) \to \min_{\mu \in Y^k} f_\infty(\mu) \qquad \mu^{(n)} \to \mu^{(\infty)}$$

with probability one. And in Chapter 4 we show that:

$$\sum_{j=1}^k \left\| \mu_j^{(n)} - \mu_j^{(\infty)} \right\|_{L^2}^2 = O\left(\frac{1}{n}\right).$$

The earliest results regarding the asymptotics of the k-means method considered the application to Euclidean data sets, i.e. $X = Y = \mathbb{R}^\kappa$ and $d(x, y) = |x - y|^2$ where $|\cdot|$ is the Euclidean norm. Under the assumption that the limiting functional f_∞ has a unique minimizer $\mu^{(\infty)}$ then $\mu^{(n)} \to \mu^{(\infty)}$ with probability one [80,124]. When there is no unique minimizer these results do not hold [19]. With more generality the convergence of the minimum and minimizers for $X = Y$ a reflexive and separable Banach space has been studied in [95, 103]. And an analogous result in [96] for metric spaces. Convergence results for $X \neq Y$ are, as far as the author is aware, new.

The first result known to the author regarding the rate of convergence is a central limit theorem result proved for $X = Y = \mathbb{R}^\kappa$ and $d(x, y) = |x - y|^2$ [126], that is there exists a covariance matrix Σ such that

$$n^{\frac{1}{2}} \left(\mu^{(n)} - \mu^{(\infty)} \right) \to N(0, \Sigma)$$

where the convergence is in distribution. A simple application of this shows [42] that the minimum behaves as

$$\left| f_\infty(\mu^{(n)}) - f_\infty(\mu^{(n)}) \right| = O_p\left(\frac{1}{n} \right).$$

When considering convergence in expectation one has

$$\left| \mathbb{E} f_\infty(\mu^{(n)}) - f_\infty(\mu^{(\infty)}) \right| = O\left(\frac{1}{\sqrt{n}} \right), \tag{1.2}$$

see for example [10,102,104]. When $k = 1$ standard results imply $\left| \mathbb{E} f_\infty(\mu^{(n)}) - f_\infty(\mu^{(\infty)}) \right| = O\left(\frac{1}{n} \right)$ however when $k \geq 3$ it is known [18, 102] that there exists a constant $C > 0$ such that

$$\left| \mathbb{E} f_\infty(\mu^{(n)}) - f_\infty(\mu^{(\infty)}) \right| \geq \frac{C}{\sqrt{n}}$$

which in particular shows (1.2) is sharp. More generally (1.2) has been shown for $X = Y$ a Hilbert space when $d(x, y) = \|x - y\|^2$ and $\| \cdot \|$ is the norm on X and for $X = Y$ a separable and reflexive Banach space with $d(x, y) = \|x - y\|$, see [21, 95] respectively.

General Spline Smoothing (Chapter 5). The second problem, in Chapter 5, looks at the general spline problem. This is similar to the above except we remove the data association problem (i.e. $k = 1$) and assume there is a true data generating curve. In this case we can scale the regularization $\lambda_n \to 0$ and recover the 'truth' in the data rich limit.

Let $\mathcal{H}$ be a Hilbert space with norm $\|\cdot\|$ and inner product $(\cdot, \cdot)$. We consider the problem of recovering $\mu^\dagger \in \mathcal{H}$ from observations $\{(L_i, y_i)\}_{i=1}^n \subset \mathcal{H}^* \times \mathbb{R}$ and the model:

$$y_i = L_i \mu^\dagger + \epsilon_i$$

where ϵ_i is noise. We refer to this as the general spline problem.

A particular case of much interest is when $\mathcal{H} = H^m$ (the Sobolev space of degree m) and observation operators are of the form $L_i \mu = \mu(t_i)$. We call this the special spline problem.

We assume that there exists exists a decomposition $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $\mathcal{H}_i$ are Hilbert spaces with norms $\| \cdot \|_i$. The estimator $\mu^{(n)}$ of $\mu^\dagger$ is defined to be the minimizer of $f_n : \mathcal{H} \to [0, \infty)$ defined by

$$f_n(\mu) = \frac{1}{n} \sum_{i=1}^n |y_i - L_i \mu|^2 + \lambda_n \|\chi_1 \mu\|_1$$

where $\chi_1 : \mathcal{H} \to \mathcal{H}_1$ is the orthogonal projection. Under suitable conditions one can interpret $\mu^{(n)}$ as a maximum a-posteriori estimator [89]. We assume that $\mathcal{H}_0$ is finite dimensional and $\mathcal{H}_1$ is infinite dimensional. It is typically not possible to show $\|\mu^{(n)} - \mu\| \to 0$ which leaves one with two natural options. The first is to look for convergence in a weaker norm, e.g. instead of H^m we look at L^2, and the second option is to look for weak convergence.

Assume that $\|\cdot\|_1 = \|\mathcal{C}^{-1}\cdot\|_{L^2}$ where $\mathcal{C}$ is the covariance operator and the existence of a compact, positive semi-definite and self-adjoint operator U which satisfies

$$\frac{1}{n}\sum_{i=1}^{n} L_i^* L_i \to U$$

for some suitable notion of convergence. From the assumptions one has the existence of a eigenbasis $\{\psi_i\}$ of $\mathcal{H}$ satisfying

$$(\psi_i, U\psi_j) = \delta_{ij} \qquad \text{and} \qquad (\psi_i, \mathcal{C}^{-1}\psi_j) = \gamma_i \delta_{ij}.$$

One then constructs the Hilbert scale by defining the norm

$$\|\mu\|_\rho = \left(\sum_{i=1}^{\infty}(1+\gamma_i)^\rho (\mu, U\psi_i)^2\right)^{\frac{1}{2}}$$

and sets $\mathcal{H}_\rho^0 \{\mu \in \mathcal{H} : \|\mu\|_\rho < \infty\}$. One takes $\mathcal{H}_\rho$ to be the completion of $\mathcal{H}_\rho^0$ under $\|\cdot\|_\rho$. The main results of [46, 117] imply that

$$\mathbb{E}\left\|\mu^{(n)} - \mu\right\|_\rho^2 \lesssim \min\{1, \lambda_n^{\beta-\rho}\}\|\mu^\dagger\|_\beta^2 + \frac{1}{n}\left(C(\lambda_n, \rho) + m\right)$$

for any constant β with $\rho < \beta < \rho + 2$, where $\dim(\mathcal{H}_0) = m$ and

$$C(\lambda, \rho) = \sum_{j>m} \gamma_j^\rho (1 + \lambda\gamma_j)^{-2}.$$

For special splines it has been shown [131, 158] that for uniformly spaced observations the estimate $\mu^{(n)}$ of $\mu^\dagger$ satisfies the following bound:

$$\mathbb{E}\left\|\frac{\mathrm{d}^j}{\mathrm{d}t^j}(\mu^\dagger - \mu^{(n)})\right\|_{L^2}^2 \leq A_1\lambda\left\|\frac{\mathrm{d}^j}{\mathrm{d}t^j}\mu^\dagger\right\|_{L^2}^2 + \frac{A_2}{n\lambda^{\frac{2j+1}{2m}}}$$

where A_1, A_2 are constants and $0 \leq j < m$. In particular the optimal rate of convergence is for $\lambda_n \asymp n^{-\frac{2m}{2m+1}}$ in which case

$$\mathbb{E}\left\|\frac{\mathrm{d}^j}{\mathrm{d}t^j}(\mu^\dagger - \mu^{(n)})\right\|_{L^2}^2 = O\left(n^{-\frac{2(m-j)}{2m+1}}\right).$$

The results also generalize to non-uniform observations under assumptions on the ratio of the largest to smallest gap in observation times [131].

The result of Chapter 5 is to use weak convergence rather than strong convergence in Hilbert scales. However we are able to reuse a lot of the ideas used to prove strong convergence.

In particular the result of Chapter 5 is for any $F \in \mathcal{H}^*$ and any $\epsilon > 0$ we have

$$\mathbb{P}\left(\left|F(\mu^{(n)}) - F(\mu^\dagger)\right| \geq \epsilon\right) \to 0$$

as $n \to \infty$ when $\lambda_n = O\left(\frac{1}{\sqrt{n}}\right)$, i.e. $\mu^{(n)}$ converges weakly and in probability to $\mu^\dagger$. As in the strong convergence case we make use of the approximation $U \approx \frac{1}{n}\sum_{i=1}^n L_i^* L_i$ in order to prove boundedness.

An advantage of our result is that it negates the need for Hilbert scales which can be quite abstruse; by which we mean the spaces $\mathcal{H}_\rho$ can be difficult to identify. Even for Sobolev spaces understanding $\mathcal{H}_\rho$ is in general very difficult although for some values of ρ one can make informative statements such as identifying $\mathcal{H}_\rho$ with another Sobolev space with boundary conditions, see [46, Section 3]. The cost of our approach is that if one wants strong convergence then we are dependent on embedding theorems. Such embedding theorems exist for Hilbert scales but one gets a better rate of convergence (i.e. can scale $\lambda_n \to 0$ faster) if one proves the result directly for Hilbert scales rather than proving weak convergence first.

Our results show that for weak convergence one cannot scale $\lambda_n \to 0$ faster that $\frac{1}{\sqrt{n}}$. This is natural when one considers weak convergence as a finite dimensional projection and assumes a central limit theorem holds. Hence the results of Chapter 5 are optimal and in particular one cannot hope to recover the rates of convergence one has for strong convergence.

A Graphical Approach to Estimating the Data Association (Chapter 6 and Chapter 7). Chapters 6 and 7 look at the third problem where we use a graphical representation of the data in order to define an estimate to the data association problem, i.e. an estimate of $\mu : \{0, \dots, n\} \to \{0, 1\}$ where for simplicity we assume there are two classes (note the slight change of notation, μ is now estimating the data association only). We allow for a soft classification so that $\mu(j) \in \mathbb{R}$. We use minimizers of the Ginzburg-Landau functional which has two terms: the first penalizes soft assignments in order that $\mu(j) \approx \{0, 1\}$ and the second penalizes jumps between adjacent data points so $\mu(j) \approx \mu(j + 1)$. We use the structure of the graph to determine what data points are adjacent.

To be more precise we look for a function $\mu \in L^1(\Psi_n)$ where $\Psi_n = \{\xi_i\}_{i=1}^n \subset \mathbb{R}^d$ is the data and for convenience we write $L^1(\Psi_n)$ as the set of functions from Ψ_n to $\mathbb{R}$. The graph is constructed by weighting edges between points ξ_i and ξ_j by

$$W_{ij} = \eta_\epsilon(\xi_i - \xi_j)$$

where

$$\eta_\epsilon(x) = \frac{1}{\epsilon_n^d}\eta\left(\frac{x}{\epsilon}\right)$$

is the interaction potential that we scale by $\epsilon = \epsilon_n$ so that the graph remains sparse. We discuss the advantages of this in Chapter 6 . For a function $V : \mathbb{R} \to [0, \infty)$ we define the Ginzburg-

Landau functional $\mathcal{E}_n : L^1(\Psi_n) \to [0,\infty]$ by

$$\mathcal{E}_n(\mu) = \frac{1}{\epsilon_n} \sum_{i=1}^n V(\mu(\xi_i)) + \frac{1}{\epsilon_n}\frac{1}{n^2} \sum_{i,j} W_{ij} \left|\mu(\xi_i) - \mu(\xi_j)\right|.$$

We assume that $V(t) = 0 \Leftrightarrow t \in \{0,1\}$, e.g. $V(t) = t^2(1-t)^2$ so that the first term penalizes states not taking the values zero or one. The second term is defined as the graph total variation, i.e.

$$GTV_n(\mu) = \frac{1}{\epsilon_n}\frac{1}{n^2} \sum_{i,j} W_{ij} \left|\mu(\xi_i) - \mu(\xi_j)\right|.$$

Estimates of the data partition are given by minimizers of $\mathcal{E}_n$.

The asymptotics of the classical Ginzburg-Landau functional (in a continuous setting),

$$F_\epsilon(\mu) = \frac{1}{\epsilon}\int_X V(\mu(x))\,\mathrm{d}x + \frac{1}{\epsilon}\int_{X^2} \eta_\epsilon(x-y)\left|\mu(x) - \mu(y)\right|^2\,\mathrm{d}x\,\mathrm{d}y,$$

are well known, e.g. [5, 114]. These results show there exists some F_0 such that

$$F_0 = \Gamma\text{-}\lim_{\epsilon\to 0} F_\epsilon$$

and for any sequence $\mu^{(n)} \in L^1(X)$ and $\epsilon_n \to 0$ such that $\sup_{n\in\mathbb{N}} F_{\epsilon_n}(\mu^{(n)}) < \infty$ then $\{\mu^{(n)}\}_{n\in\mathbb{N}}$ is precompact in L^1. These results allow one to infer the convergence of the constrained minimization problem where the constraints respect the Γ-convergence, see Section 2.2.

More recently these results have been extended to discrete settings where $\{\xi_i\}_{i=1}^n$ form a regular graph [163]. These results apply when the data is deterministic. The appropriate notion of convergence of $\mu^{(n)} \to \mu$ where $\mu^{(n)} \in L^1(\Psi_n)$ and $\mu \in L^1(\mathbb{R}^d)$ is to define a piecewise constant approximation of $\mu^{(n)}$ on $L^1(\mathbb{R}^d)$, the details are left to Section 2.5.

For random data points it has been shown in [69] that the Γ-limit of GTV_n is a total variation distance (when η is isotropic). A consequence of our results shows this is also true when η is anisotropic. The compactness property for GTV_n requires the sequence $\mu^{(n)}$ be bounded in L^1 and GTV_n in order for compactness in L^1 which is easily seen as GTV_n is invariant under $\mu \mapsto \mu + c$.

Our results in Chapter 6 extend [5,163] to show the existence of a surface integral (where we leave the definition until Chapter 6) $\mathcal{E}_\infty$ such that $\mathcal{E}_\infty = \Gamma\text{-}\lim_{n\to\infty} \mathcal{E}_n$ and the compactness property holds when $\xi_i \overset{\text{iid}}{\sim} P$ for a probability measure P. To do so we use the methodology of [69].

The minimization problem, $\min_\mu \mathcal{E}_n(\mu)$, admits trivial minimizer $\mu \equiv 0$ or $\mu \equiv 1$. In order to obtain 'more interesting' minimizers one should impose constraints such as the mass constraint, adding a data fidelity term, or boundary conditions. In Chapter 7 we use the results of Chapter 6 to prove convergence results for each of the constrained minimization problems described. To do so one must show that the constraint respects the Γ-convergence. There is also a discussion on the results for more than 2 classes. In Chapters 6 and 7 the data is in $\mathbb{R}^d$, however we take some time in Chapter 8 to discuss the infinite dimensional case.

Chapter 2

Preliminary Material

2.1 Notation

The set of probability measures on X is denoted $\mathcal{P}(X)$ and the Borel σ-algebra by $\mathcal{B}(X)$. The problems which we address involve random observations usually denoted $\xi_i : \Omega \to X$ where we assume throughout the existence of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, rich enough to support a countably infinite sequence of such observations, $\{\xi_i^{(\omega)}\}_{i=1}^{\infty}$. All random elements are defined upon this common probability space and all stochastic quantifiers are to be understood as acting with respect to $\mathbb{P}$ unless otherwise stated. Where appropriate, to emphasize the randomness of the functionals f_n, we will write $f_n^{(\omega)}$ to indicate the functional associated with the particular observation sequence $\xi_1^{(\omega)}, \dots, \xi_n^{(\omega)}$ and we allow $P_n^{(\omega)}$ to denote the associated empirical measure.

For clarity we often write integrals using operator notation. Specifically, for a measure P (which is usually a probability distribution) we write

$$Ph = \int h(x)\; P(\mathrm{d}x).$$

A sequence of probability distributions P_n on a Polish space is said to converge weakly to a probability measure P, and we write $P_n \Rightarrow P$, if

$$P_n h \to Ph \qquad \text{for all } h \in C_b$$

where C_b is the space of continuous and bounded functions. A fact we will make use of is the almost sure weak convergence of the empirical measure, e.g. [56, Theorem 11.4.1]. For a sequence $\xi_i^{(\omega)}$ of random variables and each bounded and continuous function, h, it's possible to define the sequence of random variables $P_n^{(\omega)}(h)$ which converges almost surely to Ph by the strong law of large numbers. However this does not immediately imply the almost sure weak convergence of the empirical measure since taking the intersection over the uncountable set C_b does not necessarily have probability one, i.e. we must be careful when concluding the set

$$\Omega' = \bigcap_{h \in C_b} \left\{ \omega \in \Omega : P_n^{(\omega)} h \to Ph \right\}$$

has probability one. However, when the space X is a separable metric space then one can find a countable dense subset of C_b on which to apply the strong law of large numbers. By a continuity argument this extends to the whole of C_b.

With a slight abuse of notation we will sometimes write $P(U) := P\mathbb{I}_U$ for a measurable set U. We denote the support of a probability measure P by supp(P), i.e.

$$\text{supp}(P) = \inf \left\{ X' : X' \subset X, X' \text{ is closed, and } \int_{X \setminus X'} P(\mathrm{d}x) = 0 \right\}.$$

Throughout this thesis we say that a sequence of parameter estimators is consistent if, for any value of the "parameters", they converge with respect to the underlying topology in probability (with respect to the data-generating mechanism) to the true value.

The space of functions from Z onto Y that are L^p-integrable are denoted by $L^p(Z;Y)$ (for $1 \leq p \leq \infty$). Usually either $Y = \{0,1\}$ or $Y = \mathbb{R}$. If $Y = \mathbb{R}$ then we write $L^p(Z)$ instead of $L^p(Z;\mathbb{R})$. When we use the L^p norm with respect to a measure P the Y dependence is suppressed and we write $L^p(X;P)$. It will be obvious from the context what is meant.

We define the Sobolev spaces $W^{s,p}(I)$ on $I \subseteq \mathbb{R}$ by

$$W^{s,p} = W^{s,p}(I) = \left\{ f : I \to \mathbb{R} \text{ s.t. } \nabla^i f \in L^p(I) \text{ for } i = 0, \dots, s \right\}$$

where we use ∇ for the weak derivative, i.e. $g = \nabla f$ if for all $\phi \in C_c^\infty(I)$ (the space of smooth functions with compact support)

$$\int_I f(x) \frac{\mathrm{d}\phi}{\mathrm{d}x}(x) \, \mathrm{d}x = - \int_I g(x)\phi(x) \, \mathrm{d}x.$$

In particular, we will use the special case when $p = 2$ and we write $H^s = W^{s,2}$. This is a Hilbert space with norm:

$$\|f\|_{H^s}^2 = \sum_{i=0}^{s} \|\nabla^i f\|_{L^2}^2.$$

For a Banach space A one can define the dual space A^* to be the space of all bounded and linear maps over A into $\mathbb{R}$ equipped with the norm $\|F\|_{A^*} = \sup_{x \in A} |F(x)|$. Similarly one can define the second dual A^{**} as the space of all bounded and linear maps over A^* into $\mathbb{R}$. Reflexive spaces are defined to be spaces A such that A is isometrically isomorphic to A^{**}. These have the useful property that closed and bounded sets are weakly compact. For example any L^p space (with $1 < p < \infty$) is reflexive, as is any Hilbert space (by application of the Riesz Representation Theorem).

A sequence $x_n \in A$ is said to weakly convergence to $x \in A$ if $F(x_n) \to F(x)$ for all $F \in A^*$. We write $x_n \rightharpoonup x$. We say a functional $G : A \to \mathbb{R}$ is weakly continuous if $G(x_n) \to G(x)$ whenever $x_n \rightharpoonup x$ and strongly continuous if $G(x_n) \to G(x)$ whenever $\|x_n - x\|_A \to 0$. Note that weak continuity implies strong continuity. Similarly a functional G is weakly lower semi-continuous if $\liminf_{n\to\infty} G(x_n) \geq G(x)$ whenever $x_n \rightharpoonup x$.

For a space A and a set $K \subset A$ we write K^c for the complement of K in A, i.e. $K^c = A \setminus K$.

For an operator $U : \mathcal{H} \to \mathcal{H}$ we will use $\mathrm{Ran}(U)$ to denote the range of U, i.e.

$$\mathrm{Ran}(U) = \{\mu \in \mathcal{H} : \exists \nu \in \mathcal{H} \text{ s.t. } U\nu = \mu\}.$$

When U is linear and $(\mathcal{H}, \|\cdot\|)$ is a Banach space the operator norm is defined by

$$\|U\|_{\mathcal{L}(\mathcal{H},\mathcal{H})} := \sup_{\|\mu\|\leq 1} \|U\mu\|.$$

The Euclidean norm is given by $|\cdot|$ and with a small abuse of notation the dimension is inferred from the argument. The ball centered at x and with radius r in $\mathbb{R}^d$ is given as

$$B(x,r) = \left\{ y \in \mathbb{R}^d : |x-y| < r \right\}.$$

When the ball is centered at the origin we write $B(0,r)$.

For two real-valued and positive sequences a_n and r_n we write $a_n \lesssim r_n$ if $\frac{a_n}{r_n}$ is bounded. If $a_n \lesssim r_n$ and $r_n \lesssim a_n$ then we write $a_n \asymp r_n$. Alternatively we may sometimes write $a_n = O(r_n)$ if $\frac{a_n}{r_n}$ is bounded where a_n and r_n are two real valued deterministic sequences and r_n is positive. If $\frac{a_n}{r_n} \to 0$ as $n \to \infty$ we write $a_n = o(r_n)$. For random sequences a_n and r_n, where r_n are positive and real valued, we write $a_n = O_p(r_n)$ if $\frac{a_n}{r_n}$ is bounded in probability: for all $\epsilon > 0$ there exist deterministic constants M_ϵ, N_ϵ such that

$$\mathbb{P}\left(\left| \frac{a_n}{r_n} \right| \geq M_\epsilon \right) \leq \epsilon \qquad \forall n \geq N_\epsilon.$$

If $\frac{a_n}{r_n} \to 0$ in probability: for all $\epsilon > 0$

$$\mathbb{P}\left(\left| \frac{a_n}{r_n} \right| \geq \epsilon \right) \to 0 \qquad \text{as } n \to \infty$$

we write $a_n = o_p(r_n)$.

2.2 Γ-Convergence

Γ-convergence was introduced in the 1970's by De Giorgi as a tool for studying oscillatory objects. We are particularly motivated by using the Γ-limit to design our minimization problems so that our classifiers have certain properties. A key contribution of this thesis is to identify the limiting minimization problem associated with a variety of statistical inference problems. Knowledge of the limit aids the practitioner in designing the finite data problem, i.e. allows one to pick out important features of the data. In this sense Γ-convergence is used as a data analysis tool. See, for example [28, 50], for an introduction to Γ-convergence.

We have the following definition of Γ-convergence, see also Figure 1.1 in Chapter 1 for an illustration of Γ-convergence.

Definition 2.2.1 (Γ-convergence)**.** *Let (X, τ) be a topological space. A sequence $f_n : X \to \mathbb{R} \cup \{\pm\infty\}$ is said to Γ-converge on the domain X to $f_\infty : X \to \mathbb{R} \cup \{\pm\infty\}$ with respect to the topology τ on X, and we write $f_\infty = \Gamma\text{-}\lim_n f_n$, if for all $\zeta \in X$ we have* *(i) (lim inf inequality) for every sequence $(\zeta^{(n)})$ converging to ζ*

$$f_\infty(\zeta) \leq \liminf_{n\to\infty} f_n(\zeta^{(n)});$$

(ii) (recovery sequence) there exists a sequence $(\zeta^{(n)})$ converging to ζ such that

$$f_\infty(\zeta) \geq \limsup_{n\to\infty} f_n(\zeta^{(n)}).$$

We give the above definition of Γ-convergence in terms of a general topological space. In this thesis the topology will either be the topology of weak convergence or strong convergence.

When it exists the Γ-limit is always lower semi-continuous [27, Proposition 1.31], and hence there exists minimizers over compact intervals. The following result justifies the use of Γ-convergence as a variational type of convergence.

Theorem 2.2.1 (Convergence of Minimizers). *Let (X, τ) be a topological space and $f_n : X \to [0,\infty]$ be a sequence of functionals. Let $\mu^{(n)}$ be a sequence of almost minimizers of f_n. If $\mu^{(n)}$ are precompact and $f_\infty = \Gamma\text{-}\lim_n f_n$ where $f_\infty : X \to [0,\infty]$ is not identically $+\infty$ then*

$$\min_X f_\infty = \lim_{n\to\infty} \inf_X f_n.$$

Furthermore any cluster point of $\mu^{(n)}$ minimizes f_∞.

A simple consequence of the above is if one can show that the Γ-limit has a unique minimizer then any sequence of almost minimizers converges (without the recourse to subsequences).

Corollary 2.2.2. *If in addition to the assumptions of Theorem 2.2.1 the minimizer of the Γ-limit is unique then any sequence of almost minimizers $\mu^{(n)}$ of f_n converges weakly to the minimizer of f_∞.*

For the Γ-convergence results to carry through to functions on domains $f_n : \Theta_n \to [0,\infty]$ we require that Θ_n are compatible in the following sense.

Definition 2.2.2. *Assume that f_n Γ-converges to f_∞ on a topological space (X, τ). Let Θ_n, Θ be subsets of X. Then we say that $(\Theta_n, \Theta, f_n, f_\infty)$ are compatible with respect to Γ-convergence if*

1. *Θ is closed,*
2. *there exists $\zeta \in \Theta$ such that $f_\infty(\zeta) < \infty$,*
3. *if $\zeta^{(n)} \in \Theta_n$ and $\zeta^{(n)} \to \zeta$ then $\zeta \in \Theta$ and*
4. *for all $\mu \in \Theta$ there exists a sequence $\mu^{(n)} \in \Theta_n$ such that $\mu^{(n)} \to \mu$ and*

$$\limsup_{n\to\infty} f_n(\mu^{(n)}) \leq f_\infty(\mu).$$

We immediately see that if Θ_n, Θ are compatible with respect to Γ-convergence then f_n Γ-converges to f_∞ on Θ. Theorem 2.2.1 holds when restricting f_n and f_∞ to compatible subsets.

Corollary 2.2.3. *Let (X, τ) be a topological space and $f_n : X \to [0, \infty]$ be a sequence of functionals Γ-converging to $f_\infty : X \to [0, \infty]$. Assume Θ_n, Θ are compatible with respect to Γ-convergence. If any sequence $\mu^{(n)}$ of almost minimizers is precompact then*

$$\min_{\Theta} f_\infty = \lim_{n\to\infty} \inf_{\Theta_n} f_n.$$

Furthermore any cluster point of $\mu^{(n)}$ minimizes f_∞ in Θ.

Another property of Γ-convergence we will exploit is it stability under continuous perturbations. We say g_n converges continuously to g_∞ if $g_n(\zeta^{(n)}) \to g_\infty(\zeta)$ whenever $\zeta^{(n)} \to \zeta$. We have the following proposition.

Proposition 2.2.4. *If f_n Γ-converges to f_∞ on a topological space (X, τ) and g_n continuously converges to g_∞ then*

$$f_\infty + g_\infty = \Gamma\text{-}\lim_{n\to\infty} f_n + g_n.$$

If $g_n \geq 0$ then any compactness of f_n will also carry through to $f_n + g_n$. Hence convergence of minimizers/convergence of minima results for f_n carry through to $f_n + g_n$.

2.3 The Gâteaux Derivative

We very quickly recap Gâteaux derivatives (also known as directional derivatives) and remind the reader that Taylor's theorem holds in the multi-dimensional case.

Definition 2.3.1. *We say that $f : \mathcal{H} \to \mathbb{R}$ is Gâteaux differentiable at $\mu \in \mathcal{H}$ in direction $\nu \in \mathcal{H}$ if the limit*

$$\partial f(\mu; \nu) = \lim_{r\to 0} \frac{f(\mu + r\nu) - f(\mu)}{r}$$

exists. We may define second order derivatives by

$$\partial^2 f(\mu; \nu, \nu') = \lim_{r\to 0} \frac{\partial f(\mu + r\nu'; \nu) - \partial f(\mu; \nu)}{r}$$

for $\mu, \nu, \nu' \in \mathcal{H}$. Similarly for higher order derivatives. To simplify notation, when it is clear, we write

$$\partial^s f(\mu; \nu) := \partial^s f(\mu; \nu, \ldots, \nu).$$

Theorem 2.3.1 (Taylor's Theorem). *If $f : \mathcal{H} \to \mathbb{R}$ is m times continuously Gâteaux differentiable on a convex subset $K \subset \mathcal{H}$ then, for $\mu, \nu \in K$:*

$$\begin{aligned} f(\nu) = f(\mu) + \partial f(\mu; \nu - \mu) + \frac{1}{2!}\partial^2 f(\mu; \nu - \mu, \nu - \mu) + \ldots \\ + \frac{1}{(m-1)!}\partial^{m-1} f(\mu; \nu - \mu, \ldots, \nu - \mu) + R_m \end{aligned}$$

where

$$R_m(\mu, \nu - \mu) = \frac{1}{(m-1)!} \int_0^1 (1-t)^{m-1} \partial^m f((1-t)\mu + t\nu; \nu - \mu)\, dt.$$

2.4 Total Variation Distance

For the convenience of the reader we define the weighted total variation distance and recall some well known results. We start by defining the total variation distance.

Definition 2.4.1. *For a domain $X \subset \mathbb{R}^d$ the weighted total variation $TV(\cdot; \rho, \eta)$ of a function $\mu \in L^1(X)$ with respect to a density ρ and potential η is defined by*

$$TV(\mu; \rho, \eta) = \sup \left\{ \int_X \mu(x) \operatorname{div}(\phi(x))\, \mathrm{d}x \; : \; \phi \in C_c^\infty(X; \mathbb{R}^d), \right. \\ \left. \sup_{x \in X} \sigma^* \left(-\rho^{-2}(x)\phi(x) \right) < \infty \right\}, \tag{2.1}$$

$$\sigma^*(\phi) = \sup \left\{ \nu \cdot \phi - \sigma(\nu) \; : \; \nu \in \mathbb{R}^d \right\} \in \{0, \infty\}, \tag{2.2}$$

$$\sigma(\nu) = \int_{\mathbb{R}^d} \eta(x) |x \cdot \nu|\, \mathrm{d}x. \tag{2.3}$$

The space of functions with finite weighted total variation is denoted by $BV(X; \rho, \eta)$. The standard total variation distance on X is defined by

$$\widehat{TV}(\mu) = \sup \left\{ \int_X \mu(x) \operatorname{div}(\phi)\, \mathrm{d}x \; : \; \phi \in C_c^\infty(X), \|\phi\|_{L^\infty(X)} \leq 1 \right\}.$$

The standard bounded variation space $\widehat{BV}(X)$ is the set of functions such that $\widehat{TV}(\mu) < \infty$.

When $\mu \in L^1(X; \{0,1\})$ then one can write the total variation distance as a surface integral. In particular one can write:

$$TV(\mu; \rho, \eta) = \int_{\partial\{\mu = 1\}} \sigma(n(x)) \rho^2(x)\, \mathrm{d}\mathcal{H}^{d-1}(x)$$

where $n(x)$ is the outward unit normal for the set $\partial\{\mu = 1\}$, $\mathcal{H}^{d-1}$ is the $d-1$ dimensional Hausdorff measure. The equivalence when $\mu \in L^1(X; \{0,1\})$ can be seen from the simplification of $TV(\cdot; \rho, \eta)$ when $\mu \in C^1$:

$$TV(\mu; \rho, \eta) = \int_X \sigma(\nabla\mu(x)) \rho^2(x)\, \mathrm{d}x = \int_X \int_{\mathbb{R}^d} \eta(y) |y \cdot \nabla\mu(x)| \rho^2(x)\, \mathrm{d}y\, \mathrm{d}x.$$

One may also write

$$TV(\mu; \rho, \eta) = \int_{\mathbb{R}^d} \eta(z) TV_z(\mu; \rho)\, \mathrm{d}z$$

where $TV_z(\cdot;\rho)$ is defined by

$$TV_z(\mu;\rho) = \sup\left\{ \int_X \mu(x)\,\mathrm{div}(\phi(x))\,\mathrm{d}x \,:\, \phi \in C_c^\infty(X;\mathbb{R}^d), \right.$$
$$\left. -\nu\cdot\phi(x) \leq |z\cdot\nu|\rho^2(x)\,\forall \nu, x\in\mathbb{R}^d \right\} \tag{2.4}$$

The following proposition is a slight generalization of a well known result regarding the convergence of difference quotients to the total variation semi-norm. The proof is omitted but it is a trivial adaptation of, for example, [97, Theorem 13.48].

Proposition 2.4.1. *Assume $\mu^{(n)} \to \mu$ in L^1. For a sequence $\epsilon_n \to 0$ and a function $\rho : X \to [0,\infty)$ define $f_n : X \to [0,\infty)$ by*

$$f_n(z) = \frac{1}{\epsilon_n}\int_X \left|\mu^{(n)}(x+\epsilon_n z) - \mu^{(n)}(x)\right| \rho^2(x)\,\mathrm{d}x.$$

Then

$$\liminf_{n\to\infty} f_n(z) \geq TV_z(\mu;\rho).$$

For each $\mu \in BV(X;\rho,\eta)$ the following theorem gives the existence of a measure that one can understand as the weak derivative of μ. See for example [15, 60] for more details.

Theorem 2.4.2. *For every $\mu \in BV(X;\rho,\eta)$ there exists a Radon measure $\lambda_{\rho,\eta}$ on X and a $\lambda_{\rho,\eta}$-measurable function $\alpha : X \to \mathbb{R}$ such that $\alpha(x) = 1$ for $\lambda_{\rho,\eta}$-almost every $x \in X$ and*

$$\int_X \mu(x)\mathrm{div}\phi(x)\,\mathrm{d}x = -\int_X \frac{\phi(x)\cdot x}{\rho^2(x)\sigma(x)}\alpha(x)\,\lambda_{\rho,\eta}(\mathrm{d}x)$$

for all $\phi \in C_c^1(X;\mathbb{R}^d)$. In particular,

$$\lambda_{\rho,\eta}(X) = TV(\mu;\rho,\eta).$$

For the standard total variation distance we write $\hat{\lambda}$ and have the following relationship:

$$\lambda_{\rho,\eta}(\mathrm{d}x) = \rho^2(x)\sigma(x)\,\hat{\lambda}(\mathrm{d}x).$$

In particular

$$TV(\mu;\rho,\eta) = \int_X \rho^2(x)\sigma(x)\,\hat{\lambda}(\mathrm{d}x).$$

A useful approximation result we will make use of is for all $\mu \in BV(X;\rho,\eta)$ there exists a sequence $\mu^{(n)} \in BV(X;\rho,\eta)\cap C^\infty(\mathbb{R}^d)$ such that

$$\mu^{(n)} \to \mu \text{ in } L^1(X) \quad \text{and} \quad TV(\mu^{(n)};\rho,\eta) \to TV(\mu;\rho,\eta)$$

or equivalently $\lambda_{\rho,\eta}^{(n)}(X) \to \lambda_{\rho,\eta}(X)$ (where $\lambda_{\rho,\eta}^{(n)}$ is the measure given by Theorem 2.4.2 and induced by $\mu^{(n)}$), see e.g. [60, Theorem 2, Section 5.2.2].

The Rellich-Kondrachov theorem implies that any bounded set in BV is relatively compact in L^1. In particular if a sequence $\mu^{(n)}$ can be bounded in BV then one can infer the existence of a subsequence converging in L^1.

2.5 Transportation Theory

In Chapters 6 and 7 we will look for convergence of the data association function $\mu : \Psi_n \to \{1, \dots, k\}$ (where $\Psi_n = \{\xi_i\}_{i=1}^n$). This requires a notion of convergence suitable for comparing functions on different domains. We wish to define a map $T_n : X \to \Psi_n$ that will allow us to extend functions $\mu^{(n)}$ on Ψ_n to functions $\tilde{\mu}^{(n)}$ on X, i.e. $\tilde{\mu}^{(n)} = \mu^{(n)} \circ T_n$. The challenge is to define T_n optimally in the sense that as little mass as possible is moved. We start by defining the p-OT distance.

Definition 2.5.1. *If $1 \leq p < \infty$ then the p-OT distance between $P, Q \in \mathcal{P}(X)$ is defined by*

$$d_p(P,Q) = \min \left\{ \left(\int_{X^2} |x-y|^p \, \pi(\mathrm{d}x, \mathrm{d}y) \right)^{\frac{1}{p}} : \pi \in \Gamma(P,Q) \right\} \tag{2.5}$$

where $\Gamma(P,Q)$ is the set of couplings between P and Q, i.e. the set of probability measures on $X \times X$ such that the first marginal is P and the second marginal is Q.

If $p = \infty$ then the ∞-OT distance between $P, Q \in \mathcal{P}(X)$ is defined by

$$d_\infty(P,Q) = \min \left\{ \operatorname*{ess\,sup}_{\pi} \{ |x-y| : (x,y) \in X \times X \} : \pi \in \Gamma(P,Q) \right\}. \tag{2.6}$$

The minimization problem in (2.5) and (2.6) is known as Kantorovich's optimal transportation problem. The minimization is convex and and therefore the minimum is achieved [40, 168]. One can also show that d_p defines a metric. Elements $\pi \in \Gamma(P,Q)$ are called transference plans. The distance d_2 is also known as the Wasserstein metric and d_∞ the ∞-transportation distance. For bounded $X \subset \mathbb{R}^d$ convergence in d_p (for $1 \leq p < \infty$) is equivalent to the weak convergence of probability measures [168] and therefore with probability one, $d_p(P_n, P) \to 0$ where P_n is the empirical measure.

When P has density with respect to the Lebesgue measure then the Kantorovich minimization problem is equivalent to the Monge optimal transportation problem [67]:

$$\text{Minimize} \int_X |x - T(x)|^p \, P(\mathrm{d}x) \quad \text{over all measurable maps } T \text{ such that } T_\# P = Q$$

where the push forward measure is defined by

$$T_\# P(A) = P(T^{-1}(A))$$

for any $A \in \mathcal{B}(X)$. If $Q = T_\# P$ then we call T a transportation map between P and Q.

Let P_n be the empirical measure and $1 \leq p < \infty$ then from $d_p(P_n, P) \to 0$ (almost surely) one can immediately infer the existence of a sequence of transport plans such that

$$\|\mathrm{Id} - T_n\|_{L^p(X;P)}^p = \int_X |x - T_n(x)|^p \, P(\mathrm{d}x) \to 0 \tag{2.7}$$

as $n \to \infty$. We call any sequence of transportation maps $\{T_n\}$ that satisfy (2.7) stagnating.

In the next definition we use stagnating transport maps to define piecewise constant approximations of functions on Ψ_n in order to define a suitable notion of convergence.

Definition 2.5.2. *Let $\mu^{(n)} \in L^p(\Psi_n) = L^p(X; P_n)$ and $\mu \in L^p(X; P)$ We say $\mu^{(n)} \to \mu$ in $TL^p(X)$ if*

$$\|\mu^{(n)} \circ T_n - \mu\|_{L^p}^p = \int_X |\mu^{(n)}(T_n(x)) - \mu(x)|^p \, P(dx) \to 0 \tag{2.8}$$

for any sequence of stagnating transportation maps $T_n : X \to \Psi_n$. Similarly $\mu^{(n)}$ is bounded in TL^p if $\|\mu^{(n)} \circ T_n\|_{L^p}$ is bounded and $\mu^{(n)}$ is precompact in TL^p if $\mu^{(n)} \circ T_n$ is precompact in L^p.

One can show that if (2.8) holds for one sequence of stagnating transport maps then it holds for any sequence of stagnating transport maps [69, Lemma 3.5].

In Chapters 6 and 7 we will assume that P has density ρ which is bounded above and below by positive constants then (2.7) is equivalent to $\|\mathrm{Id} - T_n\|_{L^p(X)} \to 0$ and $L^p(X; P) = L^p(X)$. We will mostly consider the case $p = 1$ however generalizations to $1 < p < \infty$ are straightforward.

Now consider an arbitrary $T : X \to X$ and a measurable $\varphi : X \to [0, \infty)$. Recall that

$$\int_X \varphi(x) \, T_\# P(\mathrm{d}x) := \sup \left\{ \int_X s(x) T_\# P(\mathrm{d}x) : 0 \leq s \leq \varphi \text{ and } s \text{ is simple} \right\}.$$

If $s(x) = \sum_{i=1}^N a_i \delta_{U_i}(x)$ where $a_i = s(x)$ for any $x \in U_i$ then

$$\int_X s(x) T_\# P(\mathrm{d}x) = \sum_{i=1}^N a_i T_\# P(U_i) = \sum_{i=1}^N a_i P(V_i)$$

for $V_i = T^{-1}(U_i)$. Note that $a_i = s(x)$ for any $x \in T(U_i)$. From this it is not hard to see the following change of variables formula:

$$\int_X \varphi(x) \, T_\# P(\mathrm{d}x) = \int_X \varphi(T(x)) \, P(\mathrm{d}x). \tag{2.9}$$

A particularly useful version of this will be when $T_\# P(\mathrm{d}x) = P_n(\mathrm{d}x)$ where P_n is the empirical measure. In which case (2.9) implies

$$\frac{1}{n} \sum_{i=1}^n \varphi(\xi_i) = \int_X \varphi(T(x)) \, P(\mathrm{d}x).$$

As an aside one can generalize the TL^p norm to functions $\mu \in L^p(X; P)$ and $\zeta \in$

$L^p(Y;Q)$ where P, Q are arbitrary measures on X and Y respectively. Let us define

$$d_{TL^p}((P,\mu),(Q,\zeta))$$
$$= \inf_{\pi\in\Gamma(P,Q)} \left\{ \left(\int_{X\times Y} |x-y|^p \, \pi(\mathrm{d}x,\mathrm{d}y) \right)^{\frac{1}{p}} + \left(\int_{X\times Y} |\mu(x)-\zeta(y)|^p \, \pi(\mathrm{d}x,\mathrm{d}y) \right)^{\frac{1}{p}} \right\}.$$

Let P have density with respect to the Lebesgue measure and take a sequence of measures P_n defined on a common space X (where we do not assume that P_n is the empirical measure). Then $(P_n, f_n) \to (P, f)$ in TL^p is equivalent to weak convergence of measures (due to the first term) and $\|\mu^{(n)} \circ T_n - \mu\|_{L^p(X;P)} \to 0$ (due to the second term), see [69, Proposition 3.6]. Since we we will always work with the empirical measure then with probability one P_n converges weakly to P. Hence the first term plays no role in this thesis and so is not included.

We recall the following theorem which will be useful later.

Theorem 2.5.1. *[70] Let $X \subset \mathbb{R}^d$ with $d \geq 2$ be open, connected and bounded with Lipschitz boundary. Let P be a probability measure on X with density (with respect to Lebesgue) ρ which is bounded above and below by positive constants. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with distribution P and let P_n be the empirical measure. Then there exists a constant $C > 0$ such that almost surely there exists a sequence of transportation maps $\{T_n\}_{n=1}^\infty$ from P to P_n such that*

$$\text{if } d = 2 \text{ then} \qquad \limsup_{n\to\infty} \frac{\sqrt{n}\|\mathrm{Id} - T_n\|_{L^\infty(X)}}{(\log n)^{\frac{3}{4}}} \leq C$$
$$\text{and if } d \geq 3 \text{ then} \qquad \limsup_{n\to\infty} \frac{n^{\frac{1}{d}}\|\mathrm{Id} - T_n\|_{L^\infty(X)}}{(\log n)^{\frac{1}{d}}} \leq C.$$

Chapter 3

Convergence of the k-Means Minimization Problem in a General Setting

Abstract

The k-means method is an iterative clustering algorithm which associates each observation with one of k clusters. It traditionally employs cluster centers in the same space as the observed data. By relaxing this requirement, it is possible to apply the k-means method to infinite dimensional problems, for example multiple target tracking and smoothing problems in the presence of unknown data association. Via a Γ-convergence argument, the associated optimization problem is shown to converge in the sense that both the k-means minimum and minimizers converge in the large data limit to quantities which depend upon the observed data only through its distribution. The theory is supplemented with two examples to demonstrate the range of problems now accessible by the k-means method. The first example combines a non-parametric smoothing problem with unknown data association. The second addresses tracking using sparse data from a network of passive sensors.

3.1 Introduction

The k-means algorithm [105] is a technique for assigning each of a collection of observed data to exactly one of k clusters, each of which has a unique center, in such a way that each observation is assigned to the cluster whose center is closest to that observation in an appropriate sense.

The k-means method has traditionally been used with limited scope. Its usual application has been in Euclidean spaces which restricts its application to finite dimensional problems. There are relatively few theoretical results using the k-means methodology in infinite dimensions of which [21, 34, 49, 95, 96, 103, 148] are the only papers known to the author. In the right framework, post-hoc track estimation in multiple target scenarios with unknown data association can be viewed as a clustering problem and therefore accessible to the k-means method.

In such problems one typically has finite-dimensional data, but would wish to estimate infinite dimensional tracks with the added complication of unresolved data association. It is our aim to propose and characterize a framework for the k-means method which can deal with this problem.

A natural question to ask of any clustering technique is whether the estimated clustering stabilizes as more data becomes available. More precisely, we ask whether certain estimates converge, in an appropriate sense, in the large data limit. In order to answer this question in our particular context we first establish a related optimization problem and make precise the notion of convergence.

Consistency of estimators for ill-posed inverse problems has been well studied, for example [52, 118], but without the data association problem. In contrast to standard statistical consistency results, we do not assume that there exists a structural relationship between the optimization problem and the data-generating process in order to establish convergence to true parameter values in the large data limit; rather, we demonstrate convergence to the solution of a related limiting problem.

This chapter shows the convergence of the minimization problem associated with the k-means method in a framework that is general enough to include examples where the cluster centers are not necessarily in the same space as the data points. In particular we are motivated by the application to infinite dimensional problems, e.g. the smoothing-data association problem. The smoothing-data association problem is the problem of associating data points $\{(t_i, z_i)\}_{i=1}^n \subset [0,1] \times \mathbb{R}^\kappa$ to unknown trajectories $\mu_j : [0,1] \to \mathbb{R}^\kappa$ for $j = 1, 2, \dots, k$. By treating the trajectories μ_j as the cluster centers one may approach this problem using the k-means methodology. The comparison of data points to cluster centers is a pointwise distance: $d((t_i, z_i), \mu_j) = |\mu_j(t_i) - z_i|^2$ (where $|\cdot|$ is the Euclidean norm on $\mathbb{R}^\kappa$). To ensure the problem is well-posed some regularization is also necessary. For $k = 1$ the problem reduces to smoothing and coincides with the limiting problem studied in [79]. We will discuss the smoothing-data association problem more in Section 3.3.3.

Let us now introduce the notation for our variational approach. The k-means method is a strategy for partitioning a data set $\Psi_n = \{\xi_i\}_{i=1}^n \subset X$ into k clusters where each cluster has center μ_j for $j = 1, 2, \dots, k$. First let us consider the special case when $\mu_j \in X$. The data partition is defined by associating each data point with the cluster center closest to it which is measured by a cost function $d : X \times X \to [0, \infty)$. Traditionally the k-means method considers Euclidean spaces $X = \mathbb{R}^\kappa$, where typically we choose $d(x, y) = |x - y|^2 = \sum_{i=1}^\kappa (x_i - y_i)^2$. We define the energy for a choice of cluster centers given data by

$$f_n : X^k \to \mathbb{R} \qquad\qquad f_n(\mu|\Psi_n) = \frac{1}{n} \sum_{i=1}^n \bigwedge_{j=1}^k d(\xi_i, \mu_j),$$

where for any k variables, $a_1, a_2, \dots, a_k$, $\bigwedge_{j=1}^k a_j := \min\{a_1, \dots, a_k\}$. The optimal choice of μ is that which minimizes $f_n(\cdot|\Psi_n)$. We define

$$\hat{\theta}_n = \min_{\mu \in X^k} f_n(\mu|\Psi_n) \in \mathbb{R}.$$

An associated "limiting problem" can be defined

$$\theta = \min_{\mu \in X^k} f_\infty(\mu)$$

where we assume that $\xi_i \overset{\text{iid}}{\sim} P$ for some suitable probability distribution, P, and define

$$f_\infty(\mu) = \int \bigwedge_{j=1}^{k} d(x, \mu_j) P(\mathrm{d}x).$$

In Section 3.2 we validate the formulation by first showing that, under regularity conditions and with probability one, the minimum energy converges: $\hat{\theta}_n \to \theta$. And secondly by showing that (up to a subsequence) the minimizers converge: $\mu^{(n)} \to \mu^{(\infty)}$ where $\mu^{(n)}$ minimizes f_n and $\mu^{(\infty)}$ minimizes f_∞ (again with probability one).

In a more sophisticated version of the k-means method the requirement that $\mu_j \in X$ can be relaxed. We instead allow $\mu = (\mu_1, \mu_2, \dots, \mu_k) \in Y^k$ for some other Banach space, Y, and define d appropriately. This leads to interesting statistical questions. When Y is infinite dimensional even establishing whether or not a minimizer exists is non-trivial.

When the cluster center is in a different space to the data, bounding the set of minimizers becomes less natural. For example, consider the smoothing problem in which one wishes to fit a continuous function to a set of data points. The natural choice of cost function is a pointwise distance of the data to the curve. The optimal solution is for the cluster center to interpolate the data points: in the limit the cluster center may no longer be well defined. In particular we cannot hope to have converging sequences of minimizers.

In the smoothing literature this problem is prevented by using a regularization term $r : Y^k \to \mathbb{R}$. For a cost function $d : X \times Y \to [0, \infty)$ the energies $f_n(\cdot|\Psi_n), f_\infty(\cdot) : Y^k \to \mathbb{R}$ are redefined

$$f_n(\mu|\Psi_n) = \frac{1}{n} \sum_{i=1}^{n} \bigwedge_{j=1}^{k} d(\xi_i, \mu_j) + \lambda_n r(\mu)$$
$$f_\infty(\mu) = \int \bigwedge_{j=1}^{k} d(x, \mu_j) P(\mathrm{d}x) + \lambda r(\mu).$$

Adding regularization changes the nature of the problem so we commit time in Section 3.3 to justifying our approach. Particularly we motivate treating $\lambda_n = \lambda$ as a constant independent of n. We are able to repeat the analysis from Section 3.2; that is to establish that the minimum and a subsequence of minimizers still converge.

Early results assumed $Y = X$ were Euclidean spaces and showed the convergence of minimizers to the appropriate limit [80, 124]. The motivation for the early work in this area was to show consistency of the methodology. In particular this requires there to be an underlying 'truth'. This requires the assumption that there exists a unique minimizer to the limiting energy. These results do not hold when the limiting energy has more than one minimizer [19]. In this chapter we discuss only the convergence of the method and as such require no assumption as to the existence or uniqueness of a minimizer to the limiting problem. Consistency has been strengthened to a central limit theorem in [126] also assuming a unique minimizer to the limiting energy. Other rates of convergence have been shown in [10, 18, 42, 104]. In Hilbert spaces there exist convergence results and rates of convergence for the minimum. In [21] the authors show that $|f_n(\mu^{(n)}) - f_\infty(\mu^{(\infty)})|$ is of order $\frac{1}{\sqrt{n}}$, however, there are no results for the convergence of minimizers. Results exist for $k \to \infty$, see for example [34] (which are also valid for $Y \neq X$).

Assuming that $Y = X$, the convergence of the minimization problem in a reflexive and separable Banach space has been proved in [103] and a similar result in metric spaces in [96]. In [95], the existence of a weakly converging subsequence was inferred using the results of [103].

The chapter is structured as follows. In Section 3.2 we consider convergence in the special case when the cluster centers are in the same space as the data points, i.e. $Y = X$. In this case we don't have an issue with well-posedness as the data has the same dimension as the cluster centers. For this reason we use energies defined without regularization. Theorem 3.2.5 shows that the minimum converges, i.e. $\hat{\theta}_n \to \theta$ as $n \to \infty$, for almost every sequence of observations and furthermore we have a subsequence $\mu^{(n_m)}$ of minimizers of f_{n_m} which weakly converge to some $\mu^{(\infty)}$ which minimizes f_∞.

This result is generalized in Section 3.3 to an arbitrary X and Y. The analogous result to Theorem 3.2.5 is Theorem 3.3.6. We first motivate the problem and in particular our choice of scaling in the regularization in Section 3.3.1 before proceeding to the results in Section 3.3.2. Verifying the conditions on the cost function d and regularization term r is non-trivial and so we show an application to the smoothing-data association problem in Section 3.3.3.

To demonstrate the generality of the results in this chapter, two applications are considered in Section 3.4. The first is the data association and smoothing problem. We show the minimum converging as the data size increases. We also numerically investigate the use of the k-means energy to determine whether two targets have crossed tracks. The second example uses measured times of arrival and amplitudes of signals from moving sources that are received across a network of three sensors. The cluster centers are the source trajectories in $\mathbb{R}^2$.

3.2 Convergence when $Y = X$

We assume we are given data points $\xi_i \in X$ for $i = 1, 2, \ldots$ where X is a reflexive and separable Banach space with norm $\|\cdot\|_X$ and Borel σ-algebra $\mathcal{X}$. These data points realize a sequence of $\mathcal{X}$-measurable random elements on $(\Omega, \mathcal{F}, \mathbb{P})$ which will also be denoted, with a slight abuse of notation, ξ_i.

We define

$$f_n^{(\omega)} : X^k \to \mathbb{R}, \quad f_n^{(\omega)}(\mu) = P_n^{(\omega)} g_\mu = \frac{1}{n} \sum_{i=1}^{n} \bigwedge_{j=1}^{k} d(\xi_i^{(\omega)}, \mu_j) \tag{3.1}$$

$$f_\infty : X^k \to \mathbb{R}, \quad f_\infty(\mu) = P g_\mu = \int_X \bigwedge_{j=1}^{k} d(x, \mu_j) P(\mathrm{d}x) \tag{3.2}$$

where

$$g_\mu(x) = \bigwedge_{j=1}^{k} d(x, \mu_j),$$

P is a probability measure on $(X, \mathcal{X})$ and the empirical measure $P_n^{(\omega)}$ associated with $\{\xi_i^{(\omega)}\}_{i=1}^n$ is defined by

$$P_n^{(\omega)} h = \frac{1}{n} \sum_{i=1}^{n} h(\xi_i^{(\omega)})$$

for any $\mathcal{X}$-measurable function $h : X \to \mathbb{R}$. We assume ξ_i are iid according to P with $P = \mathbb{P} \circ \xi_i^{-1}$.

We wish to show

$$\hat{\theta}_n^{(\omega)} \to \theta \quad \text{for almost every } \omega \text{ as } n \to \infty \tag{3.3}$$

where

$$\hat{\theta}_n^{(\omega)} = \inf_{\mu \in X^k} f_n^{(\omega)}(\mu)$$
$$\theta = \inf_{\mu \in X^k} f_\infty(\mu).$$

We define $\| \cdot \|_k : X^k \to [0, \infty)$ by

$$\|\mu\|_k := \max_j \|\mu_j\|_X \quad \text{for } \mu = (\mu_1, \mu_2, \ldots, \mu_k) \in X^k. \tag{3.4}$$

The reflexivity of $(X, \| \cdot \|_X)$ carries through to $(X^k, \| \cdot \|_k)$.

Our strategy is similar to that of [124] but we embed the methodology into the Γ-convergence framework. We show that (3.2) is the Γ-limit in Theorem 3.2.2 and that minimizers are bounded in Proposition 3.2.3. We may then apply Theorem 2.2.1 to infer (3.3) and the existence of a weakly converging subsequence of minimizers.

The key assumptions on d and P are given in Assumptions 1.1-1.4. The first assumption can be understood as a 'closeness' condition for the space X with respect to d. If we let $d(x, y) = 1$ for $x \neq y$ and $d(x, x) = 0$ then our cost function d does not carry any information on how far apart two points are. Assume there exists a probability density for P which has unbounded support. Then $f_n^{(\omega)}(\mu) \geq \frac{n-k}{n}$ (for almost every ω), with equality when we choose $\mu_j \in \{\xi_i^{(\omega)}\}_{i=1}^n$. I.e. any set of k unique data points will minimize $f_n^{(\omega)}$. Since our data points are unbounded we may find a sequence $\|\xi_{i_n}^{(\omega)}\|_X \to \infty$. Now we choose $\mu_1^{(n)} = \xi_{i_n}^{(\omega)}$ and clearly our cluster center is unbounded. We see that this choice of d violates the first assumption. We also add a moment condition to the upper bound to ensure integrability. Note that this also implies that $Pd(\cdot, 0) \leq \int_X M(\|x\|) \, P(\mathrm{d}x) < \infty$ so $f_\infty(0) < \infty$ and, in particular, that f_∞ is not identically infinity.

The second assumption is slightly stronger condition on d than a weak lower semi-continuity condition in the first variable and strong continuity in the second variable. The condition allows the application of Fatou's lemma for weakly converging probabilities, see [64].

The third assumption allows us to view $d(\xi_i, y)$ as a collection of random variables. The fourth implies that we have at least k open balls (where k is known) with positive probability and therefore we are not overfitting clusters to data.

Assumptions 1. *We have the following assumptions on* $d : X \times X \to [0, \infty)$ *and* P*.*

1.1 There exist continuous, strictly increasing functions $m, M : [0, \infty) \to [0, \infty)$ *such that*

$$m(\|x - y\|_X) \leq d(x, y) \leq M(\|x - y\|_X) \quad \text{for all } x, y \in X$$

with $\lim_{r\to\infty} m(r) = \infty$*,* $M(0) = 0$*, there exists* $\gamma < \infty$ *such that* $M(\|x + y\|_X) \leq \gamma M(\|x\|_X) + \gamma M(\|y\|_X)$ *and finally* $\int_X M(\|x\|_X)\ P(dx) < \infty$ *(and* M *is measurable).*

1.2. For each $x, y \in X$ *we have that if* $x_m \to x$ *and* $y_n \rightharpoonup y$ *as* $n, m \to \infty$ *then*

$$\liminf_{n,m\to\infty} d(x_m, y_n) \geq d(x, y) \quad \text{and} \quad \lim_{m\to\infty} d(x_m, y) = d(x, y).$$

1.3. For each $y \in X$ *we have that* $d(\cdot, y)$ *is* $\mathcal{X}$*-measurable.*

1.4. There exist k *different centers* $\mu_j^\dagger \in X$*,* $j = 1, 2, \ldots, k$ *such that for all* $\delta > 0$

$$P(B(\mu_j^\dagger, \delta)) > 0 \qquad \forall\, j = 1, 2, \ldots, k$$

where $B(\mu, \delta) := \{x \in X : \|\mu - x\|_X < \delta\}$*.*

We now show that for a particular common choice of cost function, d, Assumptions 1.1 to 1.3 hold.

Remark 3.2.1. *For any* $p > 0$ *let* $d(x, y) = \|x - y\|_X^p$ *then* d *satisfies Assumptions 1.1 to 1.3.*

Proof. Taking $m(r) = M(r) = r^p$ we can bound $m(\|x - y\|_X) \leq d(x, y) \leq M(\|x - y\|_X)$ and m, M clearly satisfy $m(r) \to \infty$, $M(0) = 0$, are strictly increasing and continuous. One can also show that

$$M(\|x + y\|_X) \leq 2^{p-1} \left(\|x\|_X^p + \|y\|_X^p\right)$$

hence Assumption 1.1 is satisfied.

Let $x_m \to x$ and $y_n \rightharpoonup y$. Then

$$\begin{aligned}
\liminf_{n,m\to\infty} d(x_m, y_n)^{\frac{1}{p}} &= \liminf_{n,m\to\infty} \|x_m - y_m\|_X \\
&\geq \liminf_{n,m\to\infty} (\|y_n - x\|_X - \|x_m - x\|_X) \\
&= \liminf_{n\to\infty} \|y_n - x\|_X \quad \text{since } x_m \to x \\
&\geq \|y - x\|_X
\end{aligned}$$

where the last inequality follows as a consequence of the Hahn-Banach Theorem and the fact that $y_n - x \rightharpoonup y - x$ which implies $\liminf_{n\to\infty} \|y_n - x\|_X \geq \|y - x\|_X$. Clearly $d(x_m, y) \to d(x, y)$ and so Assumption 1.2 holds.

The third assumption holds by the Borel measurability of metrics on complete separable metric spaces. □

We now state the first result of the chapter which formalizes the understanding that f_∞ is the limit of $f_n^{(\omega)}$.

Theorem 3.2.2. *Let $(X, \|\cdot\|_X)$ be a reflexive and separable Banach space with Borel σ-algebra, $\mathcal{X}$; let $\{\xi_i\}_{i\in\mathbb{N}}$ be a sequence of independent X-valued random elements with common law P. Assume $d : X \times X \to [0,\infty)$ and that P satisfies the conditions in Assumptions 1. Define $f_n^{(\omega)} : X^k \to \mathbb{R}$ and $f_\infty : X^k \to \mathbb{R}$ by* (3.1) *and* (3.2) *respectively. Then*

$$f_\infty = \Gamma\text{-}\lim_n f_n^{(\omega)}$$

for $\mathbb{P}$-almost every ω.

Proof. Define Ω' as the intersection of three events:

$$\Omega' = \left\{\omega \in \Omega : P_n^{(\omega)} \Rightarrow P\right\} \cap \left\{\omega \in \Omega : P_n^{(\omega)}(B(0,q)^c) \to P(B(0,q)^c)\ \forall q \in \mathbb{N}\right\}$$
$$\cap \left\{\omega \in \Omega : \int_X \mathbb{I}_{B(0,q)^c}(x) M(\|x\|_X)\ P_n^{(\omega)}(\mathrm{d}x) \to \int_X \mathbb{I}_{B(0,q)^c}(x) M(\|x\|_X)\ P(\mathrm{d}x)\ \forall q \in \mathbb{N}\right\}.$$

By the almost sure weak convergence of the empirical measure [56] the first of these events has probability one, the second and third are characterized by the convergence of a countable collection of empirical averages to their population average and, by the strong law of large numbers, each has probability one. Hence $\mathbb{P}(\Omega') = 1$.

Fix $\omega \in \Omega'$: we will show that the lim inf inequality holds and a recovery sequence exists for this ω and hence for every $\omega \in \Omega'$. We start by showing the lim inf inequality, allowing $\{\mu^{(n)}\}_{n=1}^\infty \in X^k$ to denote any sequence which converges weakly to $\mu \in X^k$. We are required to show:

$$\liminf_{n\to\infty} f_n^{(\omega)}(\mu^{(n)}) \geq f_\infty(\mu).$$

By Theorem 1.1 in [64] we have

$$\int_X \liminf_{n\to\infty, x'\to x} g_{\mu^{(n)}}(x')\ P(\mathrm{d}x) \leq \liminf_{n\to\infty} \int_X g_{\mu^{(n)}}(x)\ P_n^{(\omega)}(\mathrm{d}x) = \liminf_{n\to\infty} P_n^{(\omega)} g_{\mu^{(n)}}.$$

For each $x \in X$, we have by Assumption 1.2 that

$$\liminf_{x'\to x, n\to\infty} d(x', \mu_j^{(n)}) \geq d(x, \mu_j).$$

By taking the minimum over j we have

$$\liminf_{x'\to x, n\to\infty} g_{\mu^{(n)}}(x') = \bigwedge_{j=1}^k \liminf_{x'\to x, n\to\infty} d(x', \mu_j^{(n)}) \geq \bigwedge_{j=1}^k d(x, \mu_j) = g_\mu(x).$$

Hence

$$\liminf_{n\to\infty} f_n^{(\omega)}(\mu^{(n)}) = \liminf_{n\to\infty} P_n^{(\omega)} g_{\mu^{(n)}} \geq \int_X g_\mu(x)\ P(\mathrm{d}x) = f_\infty(\mu)$$

as required.

We now establish the existence of a recovery sequence for every $\omega \in \Omega'$ and every $\mu \in X^k$. Let $\mu^{(n)} = \mu \in X^k$. Let ζ_q be a $C^\infty(X)$ sequence of functions such that $0 \leq \zeta_q(x) \leq 1$ for all $x \in X$, $\zeta_q(x) = 1$ for $x \in B(0, q-1)$ and $\zeta_q(x) = 0$ for $x \notin B(0, q)$. Then the function $\zeta_q(x) g_\mu(x)$ is continuous in x (and with respect to convergence in $\|\cdot\|_X$) for all q. We also have

$$\begin{aligned}
\zeta_q(x) g_\mu(x) &\leq \zeta_q(x) d(x, \mu_1) \\
&\leq \zeta_q(x) M(\|x - \mu_1\|_X) \\
&\leq \zeta_q(x) M(\|x\|_X + \|\mu_1\|_X) \\
&\leq M(q + \|\mu_1\|_X)
\end{aligned}$$

so $\zeta_q g_\mu$ is a continuous and bounded function, hence by the weak convergence of $P_n^{(\omega)}$ to P we have

$$P_n^{(\omega)} \zeta_q g_\mu \to P \zeta_q g_\mu$$

as $n \to \infty$ for all $q \in \mathbb{N}$. For all $q \in \mathbb{N}$ we have

$$\begin{aligned}
\limsup_{n\to\infty} |P_n^{(\omega)} g_\mu - P g_\mu| &\leq \limsup_{n\to\infty} |P_n^{(\omega)} g_\mu - P_n^{(\omega)} \zeta_q g_\mu| + \limsup_{n\to\infty} |P_n^{(\omega)} \zeta_q g_\mu - P \zeta_q g_\mu| \\
&\quad + \limsup_{n\to\infty} |P \zeta_q g_\mu - P g_\mu| \\
&= \limsup_{n\to\infty} |P_n^{(\omega)} g_\mu - P_n^{(\omega)} \zeta_q g_\mu| + |P \zeta_q g_\mu - P g_\mu|.
\end{aligned}$$

Therefore,

$$\limsup_{n\to\infty} |P_n^{(\omega)} g_\mu - P g_\mu| \leq \limsup_{q\to\infty} \limsup_{n\to\infty} |P_n^{(\omega)} g_\mu - P_n^{(\omega)} \zeta_q g_\mu|$$

by the dominated convergence theorem. We now show that the right hand side of the above expression is equal to zero. We have

$$\begin{aligned}
|P_n^{(\omega)} g_\mu - P_n^{(\omega)} \zeta_q g_\mu| &\leq P_n^{(\omega)} \mathbb{I}_{(B(0,q-1))^c} g_\mu \\
&\leq P_n^{(\omega)} \mathbb{I}_{(B(0,q-1))^c} d(\cdot, \mu_1) \\
&\leq P_n^{(\omega)} \mathbb{I}_{(B(0,q-1))^c} M(\|\cdot - \mu_1\|_X) \\
&\leq \gamma \left(P_n^{(\omega)} \mathbb{I}_{(B(0,q-1))^c} M(\|\cdot\|_X) + M(\|\mu_1\|_X) P_n^{(\omega)} \mathbb{I}_{(B(0,q-1))^c} \right) \\
&\to \gamma \left(P \mathbb{I}_{(B(0,q-1))^c} M(\|\cdot\|_X) + M(\|\mu_1\|_X) P \mathbb{I}_{(B(0,q-1))^c} \right) \quad \text{as } n \to \infty \\
&\to 0 \quad \text{as } q \to \infty
\end{aligned}$$

where the last limit follows by the monotone convergence theorem. We have shown

$$\lim_{n\to\infty} |P_n^{(\omega)} g_\mu - P g_\mu| = 0.$$

Hence

$$f_n^{(\omega)}(\mu) \to f_\infty(\mu)$$

as required. ☐

Now we have established almost sure Γ-convergence we establish the boundedness condition in Proposition 3.2.3 so we can apply Theorem 2.2.1.

Proposition 3.2.3. *Assuming the conditions of Theorem 3.2.2 and define* $\|\cdot\|_k$ *by* (3.4)*, there exists* $R > 0$ *such that*

$$\inf_{\mu\in X^k} f_n^{(\omega)}(\mu) = \inf_{\|\mu\|_k\leq R} f_n^{(\omega)}(\mu) \quad \forall n \text{ sufficiently large}$$

for $\mathbb{P}$*-almost every* ω*. In particular* R *is independent of* n*.*

Proof. The structure of the proof is similar to [96, Lemma 2.1]. We argue by contradiction. In particular we argue that if a cluster center is unbounded then in the limit the minimum is achieved over the remaining $k-1$ cluster centers. We then use Assumption 1.4 to imply that adding an extra cluster center will strictly decrease the minimum, and hence we have a contradiction.

We define Ω'' to be

$$\Omega'' = \cap_{\delta\in\mathbb{Q}\cap(0,\infty),l=1,2,\dots,k} \left\{\omega \in \Omega' : P_n^{(\omega)}(B(\mu_l^\dagger,\delta)) \to P(B(\mu_l^\dagger,\delta))\right\}.$$

As Ω'' is the countable intersection of sets of probability one, we have $\mathbb{P}(\Omega'') = 1$. Fix $\omega \in \Omega''$ and assume that the cluster centers $\mu^{(n)} \in X^k$ are almost minimizers, i.e.

$$f_n^{(\omega)}(\mu^{(n)}) \leq \inf_{\mu\in X^k} f_n^{(\omega)}(\mu) + \varepsilon_n$$

for some sequence $\varepsilon_n > 0$ such that

$$\lim_{n\to\infty} \varepsilon_n = 0. \tag{3.5}$$

Assume that $\lim_{n\to\infty} \|\mu^{(n)}\|_k = \infty$. There exists $l_n \in \{1,\dots,k\}$ with $\lim_{n\to\infty} \|\mu_{l_n}^{(n)}\|_X = \infty$. Fix $x \in X$ then

$$d(x,\mu_{l_n}^{(n)}) \geq m(\|\mu_{l_n}^{(n)} - x\|_X) \to \infty.$$

Therefore, for each $x \in X$,

$$\lim_{n\to\infty} \left(\bigwedge_{j=1}^k d(x,\mu_j^{(n)}) - \bigwedge_{j\neq l_n} d(x,\mu_j^{(n)}) \right) = 0.$$

Let $\delta > 0$ then there exists N such that for $n \geq N$

$$\bigwedge_{j=1}^k d(x,\mu_j^{(n)}) - \bigwedge_{j\neq l_n} d(x,\mu_j^{(n)}) \geq -\delta.$$

Hence

$$\liminf_{n\to\infty}\int\left(\bigwedge_{j=1}^{k} d(x,\mu_j^{(n)}) - \bigwedge_{j\neq l_n} d(x,\mu_j^{(n)})\right)\; P_n^{(\omega)}(\mathrm{d}x)\geq -\delta.$$

Letting $\delta\to 0$ we have

$$\liminf_{n\to\infty}\int\left(\bigwedge_{j=1}^{k} d(x,\mu_j^{(n)}) - \bigwedge_{j\neq l_n} d(x,\mu_j^{(n)})\right)\; P_n^{(\omega)}(\mathrm{d}x)\geq 0$$

and moreover

$$\liminf_{n\to\infty}\left(f_n^{(\omega)}\left(\mu^{(n)}\right) - f_n^{(\omega)}\left((\mu_j^{(n)})_{j\neq l_n}\right)\right)\geq 0, \tag{3.6}$$

where we interpret $f_n^{(\omega)}$ accordingly. It suffices to demonstrate that

$$\liminf_{n\to\infty}\left(\inf_{\mu\in X^k} f_n^{(\omega)}(\mu) - \inf_{\mu\in X^{k-1}} f_n^{(\omega)}(\mu)\right) < 0. \tag{3.7}$$

Indeed, if (3.7) holds, then

$$\begin{aligned}
&\liminf_{n\to\infty}\left(f_n^{(\omega)}\left(\mu^{(n)}\right) - f_n^{(\omega)}\left((\mu_j^{(n)})_{j\neq l_n}\right)\right)\\
&=\lim_{n\to\infty}\underbrace{\left(f_n^{(\omega)}\left(\mu^{(n)}\right) - \inf_{\mu\in X^k} f_n^{(\omega)}(\mu)\right)}_{\leq\varepsilon_n} + \liminf_{n\to\infty}\left(\inf_{\mu\in X^k} f_n^{(\omega)}(\mu) - f_n^{(\omega)}\left((\mu_j^{(n)})_{j\neq l_n}\right)\right)\\
&<0 \quad \text{by (3.5) and (3.7)},
\end{aligned}$$

but this contradicts (3.6).

We now establish (3.7). By Assumption 1.4 there exists k centers $\mu_j^\dagger\in X$ and $\delta_1>0$ such that $\min_{j\neq l}\|\mu_j^\dagger-\mu_l^\dagger\|_X\geq\delta_1$. Hence for any $\mu\in X^{k-1}$ there exists $l\in\{1,2,\dots,k\}$ such that we have

$$\|\mu_l^\dagger-\mu_j\|_X\geq\frac{\delta_1}{2}\qquad\text{for } j=1,2,\dots,k-1.$$

Proceeding with this choice of l, for $x\in B(\mu_l^\dagger,\delta_2)$ (for any $\delta_2\in(0,\delta_1/2)$) we have

$$\|\mu_j-x\|_X\geq\frac{\delta_1}{2}-\delta_2$$

and therefore $d(\mu_j,x)\geq m(\frac{\delta_1}{2}-\delta_2)$ for all $j=1,2,\dots,k-1$. Also

$$D_l(\mu):=\min_{j=1,2,\dots,k-1} d(x,\mu_j)-d(x,\mu_l^\dagger)\geq m(\frac{\delta_1}{2}-\delta_2)-M(\delta_2). \tag{3.8}$$

So for δ_2 sufficiently small there exists $\epsilon>0$ such that

$$D_l(\mu)\geq\epsilon.$$

Since the right hand side is independent of $\mu \in X^{k-1}$,

$$\inf_{\mu \in X^{k-1}} \max_l D_l(\mu) \geq \epsilon.$$

Define the characteristic function

$$\chi_\mu(\xi) = \begin{cases} 1 & \text{if } \|\xi - \mu^\dagger_{l(\mu)}\|_X < \delta_2 \\ 0 & \text{otherwise,} \end{cases}$$

where $l(\mu)$ is the maximizer in (3.8). For each $\omega \in \Omega''$ one obtains

$$\begin{aligned}
\inf_{\mu \in X^{k-1}} f_n^{(\omega)}(\mu) &= \inf_{\mu \in X^{k-1}} \frac{1}{n} \sum_{i=1}^{n} \bigwedge_{j=1}^{k-1} d(\xi_i, \mu_j) \\
&\geq \inf_{\mu \in X^{k-1}} \frac{1}{n} \sum_{i=1}^{n} \left[\bigwedge_{j=1}^{k-1} d(\xi_i, \mu_j) \left(1 - \chi_\mu(\xi_i)\right) + \left(d(\xi_i, \mu^\dagger_{l(\mu)}) + \epsilon \right) \chi_\mu(\xi_i) \right] \\
&\geq \inf_{\mu \in X^k} f_n^{(\omega)}(\mu) + \epsilon \min_{l=1,2,\ldots,k} P_n^{(\omega)}(B(\mu^\dagger_l, \delta_2)).
\end{aligned}$$

Then since $P_n^{(\omega)}(B(\mu^\dagger_l, \delta_2)) \to P(B(\mu^\dagger_l, \delta_2)) > 0$ by Assumption 1.4 (for $\delta_2 \in \mathbb{Q} \cap (0, \infty)$) we can conclude (3.7) holds. □

Remark 3.2.4. *One can easily show that Assumption 1.2 implies that d is weakly lower semi-continuous in its second argument which carries through to $f_n^{(\omega)}$. It follows that on any bounded (or equivalently as X is reflexive: weakly compact) set the infimum of $f_n^{(\omega)}$ is achieved. Hence the infimum in Proposition 3.2.3 is actually a minimum.*

We now easily prove convergence by application of Theorem 2.2.1.

Theorem 3.2.5. *Assuming the conditions of Theorem 3.2.2 and Proposition 3.2.3 the minimization problem associated with the k-means method converges. I.e. for $\mathbb{P}$-almost every ω:*

$$\min_{\mu \in X^k} f_\infty(\mu) = \lim_{n \to \infty} \min_{\mu \in X^k} f_n^{(\omega)}(\mu).$$

Furthermore any sequence of minimizers $\mu^{(n)}$ of $f_n^{(\omega)}$ is almost surely weakly precompact and any weak limit point minimizes f_∞.

Remark 3.2.6. *If in addition to the conditions in the above theorem the Γ-limit f_∞ has a unique minimizer then it follows (see Corollary 2.2.2) that the entire sequence $\mu^{(n)}$ of minimizers of f_n converge weakly to a minimizer of f_∞ without the recourse to subsequences. The same reasoning applies to Theorem 3.3.6 in the $X \neq Y$ case.*

3.3 The Case of General Y

In the previous section the data, ξ_i, and cluster centers, μ_j, took their values in a common space, X. We now remove this restriction and let $\xi_i : \Omega \to X$ and $\mu_j \in Y$. We may want to use this framework to deal with finite dimensional data and infinite dimensional cluster centers, which can lead to the variational problem having uninformative minimizers.

In the previous section the cost function d was assumed to scale with the underlying norm. This is no longer appropriate when $d : X \times Y \to [0, \infty)$. In particular if we consider the smoothing-data association problem then the natural choice of d is a pointwise distance which will lead to the optimal cluster centers interpolating data points. Hence, in any H^s norm with $s \geq 1$, the optimal cluster centers "blow up".

One possible solution would be to weaken the space to L^2 and allow this type of behavior. This is undesirable from both modeling and mathematical perspectives: If we first consider the modeling point of view then we do not expect our estimate to perfectly fit the data which is observed in the presence of noise. It is natural that the cluster centers are smoother than the data alone would suggest. It is desirable that the optimal clusters should reflect reality. From the mathematical point of view, restricting ourselves to only very weak spaces gives no hope of obtaining a strongly convergent subsequence.

An alternative approach is, as is common in the smoothing literature, to use a regularization term. This approach is also standard when dealing with ill-posed inverse problems. This changes the nature of the problem and so requires some justification. In particular the scaling of the regularization with the data is of fundamental importance. In the following section we argue that scaling motivated by a simple Bayesian interpretation of the problem is not strong enough (unsurprisingly, countable collections of finite dimensional observations do not carry enough information to provide consistency when dealing with infinite dimensional parameters). In the form of a simple example we show that the optimal cluster center is unbounded in the large data limit when the regularization goes to zero sufficiently quickly. The natural scaling in this example is for the regularization to vary with the number of observations as n^p for $p \in [-\frac{4}{5}, 0]$. We consider the case $p = 0$ in Section 3.3.2. This type of regularization is understood as penalized likelihood estimation [75].

Although it may seem undesirable for the limiting problem to depend upon the regularization it is unavoidable in ill-posed problems such as this one: there is not sufficient information, in even countably infinite collections of observations to recover the unknown cluster centers and exploiting known (or expected) regularity in these solutions provides one way to combine observations with qualitative prior beliefs about the cluster centers in a principled manner. There are many precedents for this approach, including [79] in which the consistency of penalized splines is studied using, what in this thesis we call, the Γ-limit. In that paper a fixed regularization was used to define the limiting problem in order to derive an estimator. Naturally, regularization strong enough to alter the limiting problem influences the solution and we cannot hope to obtain consistent estimation in this setting, even in settings in which the cost function can be interpreted as the log likelihood of the data generating process. In the setting of [79], the regularization is finally scaled to zero whereupon under assumptions the estimator converges to the truth but such a step is not feasible in the more complicated settings considered here.

When more structure is available it may be desirable to further investigate the regularization. For example with $k = 1$ the non-parametric regression model is equivalent to the white noise model [32] for which optimal scaling of the regularization is known [4, 185]. It is the subject of further work to extend these results to $k > 1$.

With our redefined k-means type problem we can replicate the results of the previous section, and do so in Theorem 3.3.6. That is, we prove that the k-means method converges where Y is a general separable and reflexive Banach space and in particular need not be equal to X.

This section is split into three subsections. In the first we motivate the regularization term. The second contains the convergence theory in a general setting. Establishing that the assumptions of this subsection hold is non-trivial and so, in the third subsection, we show an application to the smoothing-data association problem.

3.3.1 Regularization

In this section we use a toy, $k = 1$, smoothing problem to motivate an approach to regularization which is adopted in what follows. We assume that the cluster centers are periodic with equally spaced observations so we may use a Fourier argument. In particular we work on the space of 1-periodic functions in H^2,

$$Y = \left\{ \mu : [0,1] \to \mathbb{R} \text{ s.t. } \mu(0) = \mu(1) \text{ and } \mu \in H^2 \right\}. \tag{3.9}$$

For arbitrary sequences (a_n), (b_n) and data $\Psi_n = \{(t_j, z_j)\}_{j=1}^n \subset [0,1] \times \mathbb{R}^d$ we define the functional

$$f_n^{(\omega)}(\mu) = a_n \sum_{j=0}^{n-1} |\mu(t_j) - z_j|^2 + b_n \|\nabla^2 \mu\|_{L^2}^2. \tag{3.10}$$

Data are points in space-time: $[0,1] \times \mathbb{R}$. The regularization is chosen so that it penalizes the L^2 norm of the second derivative. For simplicity, we employ deterministic measurement times t_j in the following proposition although this lies outside the formal framework which we consider subsequently. Another simplification we make is to use convergence in expectation rather than almost sure convergence. This simplifies our arguments. We stress that this section is the motivation for the problem studied in Section 3.3.2. We will give conditions on the scaling of a_n and b_n that determine whether $\mathbb{E} \min f_n^{(\omega)}$ and $\mathbb{E}\mu^{(n)}$ stay bounded where $\mu^{(n)}$ is the minimizer of $f_n^{(\omega)}$.

Proposition 3.3.1. *Let data be given by $\Psi_n = \{(t_j, z_j)\}_{j=1}^n$ with $t_j = \frac{j}{n}$ under the assumption $z_j = \mu^\dagger(t_j) + \epsilon_j$ for ϵ_j iid noise with finite variance and $\mu^\dagger \in L^2$ and define Y by* (3.9)*. Then $\inf_{\mu \in Y} f_n^{(\omega)}(\mu)$ defined by* (3.10) *stays bounded (in expectation) if $a_n = O(\frac{1}{n})$ for any positive sequence b_n.*

Proof. Assume n is odd. Both μ and z are 1-periodic so we can write

$$\mu(t) = \frac{1}{n} \sum_{l=-\frac{n-1}{2}}^{\frac{n-1}{2}} \hat{\mu}_l e^{2\pi i l t} \qquad \text{and} \qquad z_j = \frac{1}{n} \sum_{l=-\frac{n-1}{2}}^{\frac{n-1}{2}} \hat{z}_l e^{\frac{2\pi i l j}{n}}$$

with

$$\hat{\mu}_l = \sum_{j=0}^{n-1} \mu(t_j) e^{-\frac{2\pi i l j}{n}} \qquad \text{and} \qquad \hat{z}_l = \sum_{j=0}^{n-1} z_j e^{-\frac{2\pi i l j}{n}}.$$

We will continue to use the notation that $\hat{\mu}_l$ is the Fourier transform of μ. We write

$$\hat{\mu} := \left(\hat{\mu}_{-\frac{n-1}{2}}, \hat{\mu}_{-\frac{n-1}{2}+1}, \dots, \hat{\mu}_{\frac{n-1}{2}}\right).$$

Similarly for z.

Substituting the Fourier expansion of μ and z into $f_n^{(\omega)}$ implies

$$f_n^{(\omega)}(\mu) = \frac{a_n}{n}\left(\langle\hat{\mu},\hat{\mu}\rangle - 2\langle\hat{\mu},\hat{z}\rangle + \langle\hat{z},\hat{z}\rangle + \frac{\gamma_n}{n}\langle l^4\hat{\mu},\hat{\mu}\rangle\right)$$

where $\gamma_n = \frac{16\pi^4 b_n}{a_n}$ and $\langle\hat{x},\hat{z}\rangle = \sum_l \hat{x}_l\overline{\hat{z}}_l$. The Gateaux derivative $\partial f_n^{(\omega)}(\mu;\nu)$ of $f_n^{(\omega)}$ at μ in the direction ν is

$$\partial f_n^{(\omega)}(\mu;\nu) = \frac{2a_n}{n}\left\langle \hat{\mu} - \hat{z} + \frac{\gamma_n l^4}{n}\hat{\mu}, \hat{\nu}\right\rangle.$$

Which implies the minimizer $\mu^{(n)}$ of $f_n^{(\omega)}$ is (in terms of its Fourier expansion)

$$\hat{\mu}_l^{(n)} = \left(1 + \frac{\gamma_n l^4}{n}\right)^{-1}\hat{z} := \left(\left(1 + \frac{\gamma_n l^4}{n}\right)^{-1}\hat{z}_l\right)_{l=-\frac{n-1}{2}}^{\frac{n-1}{2}}.$$

It follows that the minimum is

$$\begin{aligned}\mathbb{E}\left(f_n^{(\omega)}(\mu^{(n)})\right) &= \frac{a_n}{n}\mathbb{E}\left(\left\langle\left(1+\frac{n}{\gamma_n l^4}\right)^{-1}\hat{z},\hat{z}\right\rangle\right)\\ &\leq a_n\sum_{j=0}^{n-1}\mathbb{E}z_j^2\\ &\lesssim 2a_n n\left(\|\mu^\dagger\|_{L^2}^2 + \mathrm{Var}(\epsilon)\right).\end{aligned}$$

Similar expressions can be obtained for the case of even n. □

Clearly the natural choice for a_n is

$$a_n = \frac{1}{n}$$

which we use from here. We let $b_n = \lambda n^p$ and therefore $\gamma_n = 16\pi^4\lambda n^{p+1}$. From Proposition 3.3.1 we immediately have $\mathbb{E}\min f_n^{(\omega)}$ is bounded for any choice of p. In our next proposition we show that for $p \in [-\frac{4}{5}, 0]$ our minimizer is bounded in H^2 whilst outside this window the norm either blows up or the second derivative converges to zero. For simplicity in the calculations we impose the further condition that $\mu^\dagger(t) = 0$.

Proposition 3.3.2. *In addition to the assumptions of Proposition 3.3.1 let $a_n = \frac{1}{n}$, $b_n = \lambda n^p$, $\epsilon_j \overset{\text{iid}}{\sim} N(0,\sigma^2)$ and assume that $\mu^{(n)}$ is the minimizer of $f_n^{(\omega)}$.*

1. *For n sufficiently large there exists $M_1 > 0$ such that for all p and n the L^2 norm is bounded:*

$$\mathbb{E}\|\mu^{(n)}\|_{L^2}^2 \leq M_1.$$

2. *If $p > 0$ then*

$$\mathbb{E}\|\nabla^2 \mu^{(n)}\|_{L^2}^2 \to 0 \quad \text{as } n \to \infty.$$

If we further assume that $\mu^\dagger(t) = 0$, then the following statements are true.

3. *For all $p \in [-\frac{4}{5}, 0]$ there exists $M_2 > 0$ such that*

$$\mathbb{E}\|\nabla^2 \mu^{(n)}\|_{L^2}^2 \leq M_2.$$

4. *If $p < -\frac{4}{5}$ then*

$$\mathbb{E}\|\nabla^2 \mu^{(n)}\|_{L^2}^2 \to \infty \quad \text{as } n \to \infty.$$

Proof. The first two statements follow from

$$\mathbb{E}\|\mu^{(n)}\|_{L^2}^2 \lesssim 2\left(\|\mu^\dagger\|_{L^2}^2 + \mathrm{Var}(\epsilon)\right)$$
$$\mathbb{E}\|\nabla^2 \mu^{(n)}\|_{L^2}^2 \lesssim \frac{8\pi^4 n}{\gamma_n}\left(\|\mu^\dagger\|_{L^2}^2 + \mathrm{Var}(\epsilon)\right)$$

which are easily shown. Statement 3 is shown after statement 4.

Following the calculation in the proof of Proposition 3.3.1, and assuming that $\mu^\dagger(t) = 0$, it is easily shown that

$$\mathbb{E}\|\nabla^2 \mu^{(n)}\|_{L^2}^2 = \frac{16\pi^4\sigma^2}{n} \sum_{l=-\frac{n-1}{2}}^{\frac{n-1}{2}} \frac{l^4}{(1 + 16\pi^4\lambda n^p l^4)^2} =: S(n) \tag{3.11}$$

since $\mathbb{E}|\hat{z}_l|^2 = \sigma^2 n$. To show $S(n) \to \infty$ we will manipulate the Riemann sum approximation of

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{x^4}{(1 + 16\pi^4\lambda x^4)^2}\, \mathrm{d}x = C$$

where $0 < C < \infty$. We have

$$\begin{aligned}
\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{x^4}{(1 + 16\pi^4\lambda x^4)^2}\, \mathrm{d}x &= n^{1+\frac{p}{4}} \int_{-\frac{1}{2}n^{-1-\frac{p}{4}}}^{\frac{1}{2}n^{-1-\frac{p}{4}}} \frac{n^{4+p} w^4}{(1 + 16\pi^4\lambda n^{4+p} w^4)^2}\, \mathrm{d}w \quad \text{where } x = n^{1+\frac{p}{4}} w \\
&\approx n^{\frac{5p}{4}} \sum_{l=-\left\lfloor \frac{1}{2}n^{-\frac{p}{4}} \right\rfloor}^{\left\lfloor \frac{1}{2}n^{-\frac{p}{4}} \right\rfloor} \frac{l^4}{(1 + 16\pi^4\lambda n^p l^4)^2} =: R(n).
\end{aligned}$$

Therefore assuming $p > -4$ we have

$$S(n) \geq \frac{16\pi^4\sigma^2}{n^{1+\frac{5p}{4}}} R(n).$$

So for $1 + \frac{5p}{4} < 0$ we have $S(n) \to \infty$. Since $S(n)$ is monotonic in p then $S(n) \to \infty$ for all $p < -\frac{4}{5}$. This shows that statement 4 is true.

Finally we establish the third statement. If $p = -\frac{4}{5}$ then

$$S(n) = 16\pi^4\sigma^2 \left\{ R(n) + \frac{1}{n} \left(\sum_{l=-\frac{n-1}{2}}^{\lfloor \frac{n^{\frac{1}{5}}}{2} \rfloor - 1} \frac{l^4}{(1+16\pi^4\lambda n^p l^4)^2} + \sum_{l=\lfloor \frac{n^{\frac{1}{5}}}{2} \rfloor + 1}^{\frac{n-1}{2}} \frac{l^4}{(1+16\pi^4\lambda n^p l^4)^2} \right) \right\}$$
$$\leq 16\pi^4\sigma^2 R(n) + \frac{2\pi^4\sigma^2}{n^{\frac{1}{5}}(1+\pi^4\lambda)^2}.$$

The remaining cases $p \in [-\frac{4}{5}, 0]$ are a consequence of (3.11) which implies that $p \mapsto \mathbb{E}(\nabla^2\mu)$ is non-increasing. □

By the Poincaré inequality it follows that if $p \geq -\frac{4}{5}$ then the H^2 norm of our minimizer stays bounded as $n \to \infty$. Our final calculation in this section is to show that the regularization for $p \in [-\frac{4}{5}, 0]$ is not too strong. We have already shown that $\|\nabla^2\mu^{(n)}\|_{L^2}$ is bounded (in expectation) in this case but we wish to make sure that we don't have the stronger result that $\|\nabla^2\mu^{(n)}\|_{L^2} \to 0$.

Proposition 3.3.3. *With the assumptions of Proposition 3.3.1 and* $a_n = \frac{1}{n}$, $b_n = \lambda n^p$ *with* $p \in [-\frac{4}{5}, 0]$ *there exists a choice of* $\mu^\dagger$ *and a constant* $M > 0$ *such that if* $\mu^{(n)}$ *is the minimizer of* $f_n^{(\omega)}$ *then*

$$\mathbb{E}\|\nabla^2\mu^{(n)}\|_{L^2}^2 \geq M. \tag{3.12}$$

Proof. We only need to prove the proposition for $p = 0$ (the strongest regularization) and find one $\mu^\dagger$ such that (3.12) is true. Let $\mu^\dagger(t) = 2\cos(2\pi t) = e^{2\pi i t} + e^{-2\pi i t}$. Then the Fourier transform of $\mu^\dagger$ satisfies $\hat{\mu}_l^\dagger = 0$ for $l \neq \pm 1$ and $\hat{\mu}_l^\dagger = n$ for $l = \pm 1$. So,

$$\begin{aligned}
\mathbb{E}\|\nabla^2\mu^{(n)}\|_{L^2}^2 &= \frac{16\pi^4}{n^2} \sum_{l=-\frac{n-1}{2}}^{\frac{n-1}{2}} \frac{l^4}{(1+16\pi^4\lambda l^4)^2} \mathbb{E}|\hat{z}_l|^2 \\
&\gtrsim \frac{16\pi^4}{n^2} \sum_{l=-\frac{n-1}{2}}^{\frac{n-1}{2}} \frac{l^4}{(1+16\pi^4\lambda l^4)^2} |\hat{\mu}_l^\dagger|^2 \\
&= \frac{32\pi^4}{(1+16\pi^4\lambda)^2} > 0
\end{aligned}$$

which completes the proof. □

We have shown that the minimizer is bounded for any $p \geq -\frac{4}{5}$ and $\|\nabla^2\mu^{(n)}\|_{L^2} \to 0$ for $p > 0$. The case $p > 0$ is clearly undesirable as we would be restricting ourselves to straight lines. The natural scaling for this problem is in the range $p \in [-\frac{4}{5}, 0]$. In the remainder of this chapter we consider the case $p = 0$. This has the advantage that, not only $\mathbb{E}\|\nabla^2\mu^{(n)}\|_{L^2}$, but also $\mathbb{E}f_n^{(\omega)}(\mu^{(n)})$ is $O(1)$ as $n \to \infty$. In fact we will show that with this choice of regularization we do not need to choose k dependent on the data generating model. The regularization makes the methodology sufficiently robust to have convergence even for poor choices of k. For example, if there exists a data generating process which is formed of a $k^\dagger$-mixture model then for our method to be robust does not require us to choose $k = k^\dagger$. Of course with the 'wrong' choice of k the results may be physically meaningless and we should take care in how to interpret the results. The point to stress is that the methodology does not rely on a data generating model.

The disadvantage of this is to potentially increase the bias in the method. Since the k-means is already biased we believe the advantages of our approach outweigh the disadvantages. In particular we have in mind applications where only a coarse estimate is needed. For example the k-means method may be used to initialize some other algorithm. Another application could be part of a decision making process: in Section 3.4.1 we show the k-means methodology can be used to determine whether two tracks have crossed.

3.3.2 Convergence For General Y

Let $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$ be reflexive, separable Banach spaces We will also assume that the data points, $\Psi_n = \{\xi_i\}_{i=1}^n \subset X$ for $i = 1, 2, \ldots, n$ are iid random elements with common law P. As before $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ but now the cluster centers $\mu_j \in Y$ for each j. The cost function is $d : X \times Y \to [0, \infty)$.

The energy functions associated with the k-means algorithm in this setting are slightly different to those used previously:

$$g_\mu : X \to \mathbb{R}, \quad g_\mu(x) = \bigwedge_{j=1}^k d(x, \mu_j),$$

$$f_n^{(\omega)} : Y^k \to \mathbb{R}, \quad f_n^{(\omega)}(\mu) = P_n^{(\omega)} g_\mu + \lambda r(\mu), \tag{3.13}$$

$$f_\infty : Y^k \to \mathbb{R}, \quad f_\infty(\mu) = P g_\mu + \lambda r(\mu). \tag{3.14}$$

The aim of this section is to show the convergence result:

$$\hat{\theta}_n^{(\omega)} = \inf_{\mu \in Y^k} f_n^{(\omega)}(\mu) \to \inf_{\mu \in Y^k} f_\infty(\mu) = \theta \quad \text{and} \quad \text{as } n \to \infty \text{ for } \mathbb{P}\text{-almost every } \omega$$

and that minimizers converge (almost surely).

The key assumptions are given in Assumption 2; they imply that $f_n^{(\omega)}$ is weakly lower semi-continuous and coercive. In particular, Assumption 2.2 allows us to prove the lim inf inequality as we did for Theorem 3.2.2. Assumption 2.1 is likely to mean that our convergence results are limited to the case of bounded noise. In fact, when applying the problem to the smoothing-data association problem, it is necessary to bound the noise in order for Assumption 2.5 to hold. Assumption 2.5 implies that $f_n^{(\omega)}$ is (uniformly) coercive and hence allows us to easily bound the set of minimizers. In the next chapter we will remove the bounded noise assumption for the smoothing-data association problem. Assumption 2.3 is a measurability condition we require in order to integrate and the weak lower semi-continuity of r is needed for the to obtain the lim inf inequality in the Γ-convergence proof.

We note that, since $Pd(\cdot, \mu_1) \leq \sup_{x \in \text{supp}(P)} d(x, \mu_1) < \infty$, we have $f_\infty(\mu) < \infty$ for every $\mu \in Y^k$ (and since $r(\mu) < \infty$ for each $\mu \in Y^k$).

Assumptions 2. *We have the following assumptions on $d : X \times Y \to [0, \infty)$, $r : Y^k \to [0, \infty)$ and P.*

2.1. *For all* $y \in Y$ *we have* $\sup_{x \in \mathrm{supp}(P)} d(x,y) < \infty$ *where* $\mathrm{supp}(P) \subseteq X$ *is the support of* P.

2.2. *For each* $x \in X$ *and* $y \in Y$ *we have that if* $x_m \to x$ *and* $y_n \rightharpoonup y$ *as* $n, m \to \infty$ *then*

$$\liminf_{n,m\to\infty} d(x_m, y_n) \geq d(x,y) \quad \textit{and} \quad \lim_{m\to\infty} d(x_m, y) = d(x,y).$$

2.3. *For every* $y \in Y$ *we have that* $d(\cdot, y)$ *is* $\mathcal{X}$*-measurable.*

2.4. r *is weakly lower semi-continuous.*

2.5. r *is coercive.*

We will follow the structure of Section 3.2. We start by showing that under the above conditions $f_n^{(\omega)}$ Γ-converges to f_∞. We then show that the regularization term guarantees that the minimizers to $f_n^{(\omega)}$ lie in a bounded set. An application of Theorem 2.2.1 gives the desired convergence result. Since we were able to restrict our analysis to a weakly compact subset of Y we are easily able to deduce the existence of a weakly convergent subsequence.

Similarly to the previous section on the product space Y^k we use the analogous norm $\|\mu\|_k := \max_j \|\mu_j\|_Y$.

Theorem 3.3.4. *Let* $(X, \|\cdot\|_X)$ *and* $(Y, \|\cdot\|_Y)$ *be separable and reflexive Banach spaces. Assume* $r : Y^k \to [0,\infty)$, $d : X \times Y \to [0,\infty)$ *and the probability measure* P *on* $(X, \mathcal{X})$ *satisfy the conditions in Assumptions 2. For independent samples* $\{\xi_i^{(\omega)}\}_{i=1}^n$ *from* P *define* $P_n^{(\omega)}$ *to be the empirical measure and* $f_n^{(\omega)} : Y^k \to \mathbb{R}$ *and* $f_\infty : Y^k \to \mathbb{R}$ *by* (3.13) *and* (3.14) *respectively and where* $\lambda > 0$. *Then*

$$f_\infty = \Gamma\text{-}\lim_n f_n^{(\omega)}$$

for $\mathbb{P}$*-almost every* ω.

Proof. Define

$$\Omega' = \left\{\omega \in \Omega : P_n^{(\omega)} \Rightarrow P\right\} \cap \left\{\omega \in \Omega : \xi_i^{(\omega)} \in \mathrm{supp}(P) \ \forall i \in \mathbb{N}\right\}.$$

Then $\mathbb{P}(\Omega') = 1$. For the remainder of the proof we consider an arbitrary $\omega \in \Omega'$. We start with the lim inf inequality. Let $\mu^{(n)} \rightharpoonup \mu$ then

$$\liminf_{n\to\infty} f_n^{(\omega)}(\mu^{(n)}) \geq f_\infty(\mu)$$

follows (as in the proof of Theorem 3.2.2) by applying Theorem 1.1 in [64] and the fact that r is weakly lower semi-continuous.

We now establish the existence of a recovery sequence. Let $\mu \in Y^k$ and let $\mu^{(n)} = \mu$. We want to show

$$\lim_{n\to\infty} f_n^{(\omega)}(\mu) = \lim_{n\to\infty} P_n^{(\omega)} g_\mu + \lambda r(\mu) = P g_\mu + \lambda r(\mu) = f_\infty(\mu).$$

Clearly this is equivalent to showing that

$$\lim_{n\to\infty} P_n^{(\omega)} g_\mu = P g_\mu.$$

Now g_μ are continuous by assumption on d. Let $M = \sup_{x\in\text{supp}(P)} d(x,\mu_1) < \infty$ and note that $g_\mu(x) \leq M$ for all $x \in \text{supp}(P)$ and therefore bounded. Hence $P_n^{(\omega)} g_\mu \to P g_\mu$. □

Proposition 3.3.5. *Assuming the conditions of Theorem 3.3.4, then for $\mathbb{P}$-almost every ω there exists $N < \infty$ and $R > 0$ such that*

$$\min_{\mu\in Y^k} f_n^{(\omega)}(\mu) = \min_{\|\mu\|_k \leq R} f_n^{(\omega)}(\mu) < \inf_{\|\mu\|_k > R} f_n^{(\omega)}(\mu) \quad \forall n \geq N.$$

In particular R is independent of n.

Proof. Let

$$\Omega'' = \left\{\omega \in \Omega' : P_n^{(\omega)} \Rightarrow P\right\} \cap \left\{\omega \in \Omega' : P_n^{(\omega)} d(\cdot,0) \to P d(\cdot,0)\right\}.$$

Then, for every $\omega \in \Omega''$, $f_n^{(\omega)}(0) \to f_\infty(0) < \infty$ where with a slight abuse of notation we denote the zero element in both Y and Y^k by 0. Take N sufficiently large so that

$$f_n^{(\omega)}(0) \leq f_\infty(0) + 1 \qquad \text{for all } n \geq N.$$

Then $\min_{\mu\in Y^k} f_n^{(\omega)}(\mu) \leq f_\infty(0) + 1$ for all $n \geq N$. By coercivity of r there exists R such that if $\|\mu\|_k > R$ then $\lambda r(\mu) \geq f_\infty(0) + 1$. Therefore any such μ is not a minimizer and in particular any minimizer must be contained in the set $\left\{\mu \in Y^k : \|\mu\|_k \leq R\right\}$. □

The convergence results now follows by applying Theorem 3.3.4 and Proposition 3.3.5 to Theorem 2.2.1.

Theorem 3.3.6. *Assuming the conditions of Theorem 3.3.4 and Proposition 3.3.5 the minimization problem associated with the k-means method converges in the following sense:*

$$\min_{\mu\in Y^k} f_\infty(\mu) = \lim_{n\to\infty} \min_{\mu\in Y^k} f_n^{(\omega)}(\mu)$$

for $\mathbb{P}$-almost every ω. Furthermore any sequence of minimizers $\mu^{(n)}$ of $f_n^{(\omega)}$ is almost surely weakly precompact and any weak limit point minimizes f_∞.

It was not necessary to assume that cluster centers are in a common space. A trivial generalization would allow each $\mu_j \in Y^{(j)}$ with the cost and regularization terms appropriately defined; in this setting Theorem 3.3.6 holds.

3.3.3 Application to the Smoothing-Data Association Problem

In this section we give an application to the smoothing-data association problem and show the assumptions in the previous section are met. For $k = 1$ the smoothing-data association problem is the problem of fitting a curve to a data set (no data association). For $k > 1$ we couple the smoothing problem with a data association problem. Each data point is associated with an unknown member of a collection of k curves. Solving the problem involves simultaneously estimating both the data partition (i.e. the association of observations to curves) and the curve which best fits each subset of the data. By treating the curve of best fit as the cluster center we are able to approach this problem using the k-means methodology. The data points are points in space-time whilst cluster centers are functions from time to space.

We let the Euclidean norm on $\mathbb{R}^\kappa$ be given by $|\cdot|$. Let $X = \mathbb{R} \times \mathbb{R}^\kappa$ be the data space. We will subsequently assume that the support of P, the common law of our observations, is contained within $\tilde{X} = [0,T] \times X'$ where $X' \subseteq [-\tilde{N}, \tilde{N}]^\kappa$. We define the cluster center space to be $Y = H^2([0,T])$, the Sobolev space of functions from $[0,T]$ to $\mathbb{R}^\kappa$. Clearly X and Y are separable and reflexive. The cost function $d : X \times Y \to [0,\infty)$ is defined by

$$d(\xi, \mu_j) = |z - \mu_j(t)|^2 \tag{3.15}$$

where $\mu_j \in Y$ and $\xi = (t,z) \in X$. We introduce a regularization term that penalizes the second derivative. This is a common choice in the smoothing literature, e.g. [132]. The regularization term $r : Y^k \to [0,\infty)$ is given by

$$r(\mu) = \sum_{j=1}^k \|\nabla^2 \mu_j\|_{L^2}^2. \tag{3.16}$$

The k-means energy f_n for data points $\{\xi_i = (t_i, z_i)\}_{i=1}^n$ is therefore written

$$f_n(\mu) = \frac{1}{n}\sum_{i=1}^n \bigwedge_{j=1}^k d(\xi_i, \mu_j) + \lambda r(\mu) = \frac{1}{n}\sum_{i=1}^n \bigwedge_{j=1}^k |z_i - \mu_j(t_i)|^2 + \lambda \sum_{j=1}^k \|\nabla^2 \mu_j\|_{L^2}^2. \tag{3.17}$$

In most cases it is reasonable to assume that any minimizer of f_∞ must be uniformly bounded, i.e. there exists N (which will in general depend on P) such that if $\mu^{(\infty)}$ minimizes f_∞ then $|\mu^{(\infty)}(t)| \leq N$ for all $t \in [0,T]$. Under this assumption we redefine Y to be

$$Y = \{\mu_j \in H^2([0,T]) : |\mu_j(t)| \leq N \, \forall t \in [0,T]\}. \tag{3.18}$$

Since pointwise evaluation is a bounded linear functional in H^s (for $s \geq 1$) this space is weakly closed. We now minimize f_n over Y^k. Note that we are not immediately guaranteed that minimizers of f_n over $(H^s)^k$ are contained in Y^k. However when we apply Theorem 3.3.6 we can conclude that minimizers $\mu^{(n)}$ of f_n over Y_k are weakly compact in $(H^s)^k$ and any limit point is a minimizer of f_∞ in Y^k. And therefore any limit point is a minimizer of f_∞ over $(H^s)^k$.

If no such N exists then our results in Theorem 3.3.6 are still valid however the minimum of f_∞ over $(H^s)^k$ is not necessarily equal to the minimum of f_∞ over Y^k.

Our results show that the Γ-limit for $\mathbb{P}$-almost every ω is

$$f_\infty(\mu) = \int_X \bigwedge_{j=1}^k d(x,\mu_j) P(\mathrm{d}x) + \lambda r(\mu) = \int_X \bigwedge_{j=1}^k |z-\mu_j(t)|^2 P(\mathrm{d}x) + \lambda \sum_{j=1}^k \|\nabla^2 \mu_j\|_{L^2}^2. \tag{3.19}$$

We start with the key result for this section, that is the existence of a weakly converging subsequence of minimizers. Our result relies upon the regularity of Sobolev functions. For our result to be meaningful we require that the minimizer should at least be continuous. In fact every $g \in H^2([0,T])$ is in $C^s([0,T])$ for any $s < \frac{3}{2}$. The regularity in the space allows us to further deduce the existence of a strongly converging subsequence.

Theorem 3.3.7. *Let $X = [0,T] \times \mathbb{R}^\kappa$ and define Y by* (3.18)*. Define $d : X \times Y \to [0,\infty)$ by* (3.15) *and $r : Y^k \to [0,\infty)$ by* (3.16)*. For independent samples $\{\xi_i\}_{i=1}^n$ from P which has compact support $\tilde{X} \subset X$ define $f_n, f_\infty : Y^k \to \mathbb{R}$ by* (3.17) *and* (3.19) *respectively.*

Then (1) any sequence of minimizers $\mu^{(n)} \in Y^k$ of f_n is $\mathbb{P}$-almost surely weakly-precompact (in H^2) with any weak limit point of $\mu^{(n)}$ minimizes f_∞ and (2) if $\mu^{(n_m)} \rightharpoonup \mu$ is a weakly converging (in H^2) subsequence of minimizers then the convergence is uniform (in C^0).

To prove the first part of Theorem 3.3.7 we are required to check the boundedness and continuity assumptions on d (Proposition 3.3.8) and show that r is weakly lower semi-continuous and coercive (Proposition 3.3.9). This statement is then a straightforward application of Theorem 3.3.6. Note that we will have shown the result of Theorem 3.3.4 holds: $f_\infty = \Gamma\text{-}\lim_n f_n^{(\omega)}$.

In what follows we check that properties hold for any $x \in \tilde{X}$, which should be understood as implying that they hold for P-almost any $x \in X$; this is sufficient for our purposes as the collection of sequences $\xi_1, \ldots$ for which one or more observations lies in the complement of $\tilde{X}$ is $\mathbb{P}$-null and the support of P_n is $\mathbb{P}$-almost surely contained within $\tilde{X}$.

Proposition 3.3.8. *Let $\tilde{X} = [0,T] \times [-\tilde{N},\tilde{N}]^\kappa$ and define Y by* (3.18)*. Define $d : \tilde{X} \times Y \to [0,\infty)$ by* (3.15)*. Then (i) for all $y \in Y$ we have $\sup_{x\in\tilde{X}} d(x,y) < \infty$ and (ii) for any $x \in X$ and $y \in Y$ and any sequences $x_m \to x$ and $y_n \rightharpoonup y$ as $m,n \to \infty$ then we have $\liminf_{n,m\to\infty} d(x_m,y_n) = d(x,y)$.*

Proof. We start with (i). Let $y \in Y$ and $x = (t,z) \in [0,T] \times [-\tilde{N},\tilde{N}]^\kappa$, then

$$\begin{aligned} d(x,y) &= |z - y(t)|^2 \\ &\leq 2|z|^2 + 2|y(t)|^2 \\ &\leq 2\tilde{N}^2 + 2 \sup_{t\in[0,T]} |y(t)|^2. \end{aligned}$$

Since y is continuous then $\sup_{t\in[0,T]} |y(t)|^2 < \infty$ and moreover we can bound $d(x,y)$ independently of x which shows (i).

For (ii) we let $(t_m, z_m) = x_m \to x = (t, z)$ in $\mathbb{R}^{\kappa+1}$ and $y_n \rightharpoonup y$. Then

$$\begin{aligned} d(x_m, y_n) &= |z_m - y_n(t_m)|^2 \\ &= |z_m|^2 - 2z_m \cdot y_n(t_m) + |y_n(t_m)|^2. \end{aligned} \tag{3.20}$$

Clearly $|z_m|^2 \to |z|^2$ and we now show that $y_n(t_m) \to y(t)$ as $m, n \to \infty$.

We start by showing that the sequence $\|y_n\|_Y$ is bounded. Each y_n can be associated with $\Lambda_n \in Y^{**}$ by $\Lambda_n(\nu) = \nu(y_n)$ for $\nu \in Y^*$. As y_n is weakly convergent it is weakly bounded. So,

$$\sup_{n\in\mathbb{N}} |\Lambda_n(\nu)| = \sup_{n\in\mathbb{N}} |\nu(y_n)| \leq M_\nu$$

for some $M_\nu < \infty$. By the uniform boundedness principle [44]

$$\sup_{n\in\mathbb{N}} \|\Lambda_n\|_{Y^{**}} < \infty.$$

And so,

$$\sup_{n\in\mathbb{N}} \|y_n\|_Y = \sup_{n\in\mathbb{N}} \|\Lambda_n\|_{Y^{**}} < \infty.$$

Hence there exists $M > 0$ such that $\|y_n\|_Y \leq M$. Therefore

$$\begin{aligned} |y_n(r) - y_n(s)| = \left| \int_s^r \nabla y_n(t) \, \mathrm{d}t \right| \leq \int_s^r |\nabla y_n(t)| \, \mathrm{d}t = \int_0^T \mathbb{I}_{[s,r]}(t) \, |\nabla y_n(t)| \, \mathrm{d}t \\ \leq \|\mathbb{I}_{[s,r]}\|_{L^2} \, \|\nabla y_n(t)\|_{L^2} \leq M\sqrt{|r-s|}. \end{aligned}$$

Since y_n is uniformly bounded and equi-continuous then by the Arzelà–Ascoli theorem there exists a uniformly converging subsequence, say $y_{n_m} \to \hat{y}$. By uniqueness of the weak limit $\hat{y} = y$. But this implies that

$$y_n(t) \to y(t)$$

uniformly for $t \in [0, T]$. Now as

$$|y_n(t_m) - y(t)| \leq |y_n(t_m) - y(t_m)| + |y(t_m) - y(t)|$$

then $y_n(t_m) \to y(t)$ as $m, n \to \infty$. Therefore the second and third terms of (3.20) satisfies

$$\begin{aligned} 2z_m \cdot y_m(t_m) &\to 2z \cdot y(t) \\ |y_n(t_m)|^2 &\to |y(t)|^2 \end{aligned}$$

as $m, n \to \infty$. Hence

$$d(x_m, y_n) \to |z|^2 - 2z \cdot y(t) + |y(t)|^2 = |z - y(t)|^2 = d(x, y)$$

which completes the proof. □

Proposition 3.3.9. *Define* Y *by* (3.18) *and* $r : Y^k \to [0, \infty)$ *by* (3.16). *Then* r *is weakly lower semi-continuous and coercive.*

Proof. We start by showing r is weakly lower semi-continuous. For any weakly converging sequence $\mu_1^{(n)} \rightharpoonup \mu_1$ in H^2 we have that $\nabla^2\mu_1^{(n)} \rightharpoonup \nabla^2\mu_1$ weakly in L^2. Hence it follows that r is weakly lower semi-continuous.

To show r is coercive let $\hat{r}(\mu_1) = \|\nabla^2\mu_1\|_{L^2}^2$ for $\mu_1 \in Y$. We will show $\hat{r}$ is coercive. Let $\mu_1 \in Y$ and note that since $\mu_1 \in C^1$ the first derivative exists (strongly). Clearly we have $\|\mu_1\|_{L^2} \leq N\sqrt{T}$ and using a Poincaré inequality

$$\left\| \frac{\mathrm{d}\mu_1}{\mathrm{d}t} - \frac{1}{T}\int_0^T \frac{\mathrm{d}\mu_1}{\mathrm{d}t}\,\mathrm{d}t \right\|_{L^2} \leq C\|\nabla^2\mu_1\|_{L^2}$$

for some C independent of μ_1. Therefore

$$\left\| \frac{\mathrm{d}\mu_1}{\mathrm{d}t} \right\|_{L^2} \leq C\|\nabla^2\mu_1\|_{L^2} + \left| \frac{1}{T}\int_0^T \frac{\mathrm{d}\mu_1}{\mathrm{d}t}\,\mathrm{d}t \right| \leq C\|\nabla^2\mu_1\|_{L^2} + \frac{2N}{T}.$$

It follows that if $\|\mu_1\|_{H^2} \to \infty$ then $\|\nabla^2\mu_1\|_{L^2} \to \infty$, hence $\hat{r}$ is coercive. □

Finally, the existence of a strongly convergent subsequence in Theorem 3.3.7 follows from the fact that H^2 is compactly embedded into H^1. Hence the convergence is strong in H^1. By Morrey's inequality H^1 is embedded into a Hölder space ($C^{0,\frac{1}{2}}$) which is a subset of uniformly continuous functions. This implies the convergence is uniform in C^0.

3.4 Examples

In this section we give two exemplar applications of the methodology. In principle any cost function, d, and regularization, r, (that satisfy the conditions) could be used. For illustrative purposes we choose d and r to make the minimization simple to implement. In particular, in Example 1 our choices allow us to use smoothing splines.

3.4.1 Example 1: A Smoothing-Data Association Problem

We use the k-means method to solve a smoothing-data association problem. For each $j = 1, 2, \ldots, k$ we take functions $x^j : [0,T] \times \mathbb{R}$ for $j = 1, 2, \ldots, k$ as the "true" cluster centers, and for sample times t_i^j for $i = 1, 2, \ldots n_j$, uniformly distributed over $[0,T]$, we let

$$z_i^j = x^j(t_i^j) + \epsilon_i^j \tag{3.21}$$

where ϵ_i^j are iid noise terms.

The observations take the form $\xi_i = (t_i, z_i)$ for $i = 1, 2, \ldots, n = \sum_{j=1}^k n_j$ where we have relabeled the observations to remove the (unobserved) target reference. We model the observations with density (with respect to the Lebesgue measure)

$$p((t,z)) = \frac{1}{T}\mathbb{I}_{[0,T]}(t)\sum_{j=1}^k w_j p_\epsilon(z - x^j(t))$$

on $\mathbb{R} \times \mathbb{R}$ where p_ϵ denotes the common density of the ϵ_i^j and w_j denotes the probability that an observation is generated by trajectory j. We let each cluster center be equally weighted: $w_j = \frac{1}{k}$. The cluster centers were fixed and in particular did not vary between numerical experiments.

When the noise is bounded this is precisely the problem described in Section 3.3.2 with $\kappa = 1$, hence the problem converges. We use a truncated Gaussian noise term.

In the theoretical analysis of the algorithm we have considered only the minimization problem associated with the k-means algorithm; of course minimizing complex functionals of the form of f_n is itself a challenging problem. Practically, we adopt the usual k-means strategy [105] of iteratively assigning data to the closest of a collection of k centers and then re-estimating each center by finding the center which minimizes the average regularized cost of the observations currently associated with that center. As the energy function is bounded below and monotonically decreasing over iterations, this algorithm converges to a local (but not necessarily global) minimum.

More precisely, in the particular example considered here we employ the following iterative procedure:

1. Initialize $\varphi^{(0)} : \{1, 2, \dots, n\} \to \{1, 2, \dots, k\}$ arbitrarily.

2. For a given data partition $\varphi^{(r)} : \{1, 2, \dots, n\} \to \{1, 2, \dots, k\}$ we independently find the cluster centers $\mu^{(r)} = (\mu_1^{(r)}, \mu_2^{(r)}, \dots, \mu_k^{(r)})$ where each $\mu_j^{(r)} \in H^2([0, T])$ by

$$\mu_j^{(r)} = \underset{\mu_j}{\operatorname{argmin}} \frac{1}{n} \sum_{i:\varphi^{(r)}(i)=j} |z_i - \mu_j(t_i)|^2 + \lambda \|\nabla^2 \mu_j\|_{L^2}^2 \quad \text{for } j = 1, 2, \dots, k.$$

   This is done using smoothing splines.

3. Data is repartitioned using the cluster centers $\mu^{(r)}$

$$\varphi^{(r+1)}(i) = \underset{j=1,2,\dots,k}{\operatorname{argmin}} |z_i - \mu_j^{(r)}(t_i)|.$$

4. If $\varphi^{(r+1)} \neq \varphi^{(r)}$ then return to Step 2. Else we terminate.

Let $\mu^{(n)} = (\mu_1^{(n)}, \dots, \mu_k^{(n)})$ be the output of the k-means algorithm from n data points. To evaluate the success of the methodology when dealing with a finite sample of n data points we look at how many iterations are required to reach convergence (defined as an assignment which is unchanged over the course of an algorithmic iteration), the number of data points correctly associated, the metric

$$\eta(n) = \frac{1}{k} \sqrt{\sum_{j=1}^{k} \|\mu_j^{(n)} - x^j\|_{L^2}^2}$$

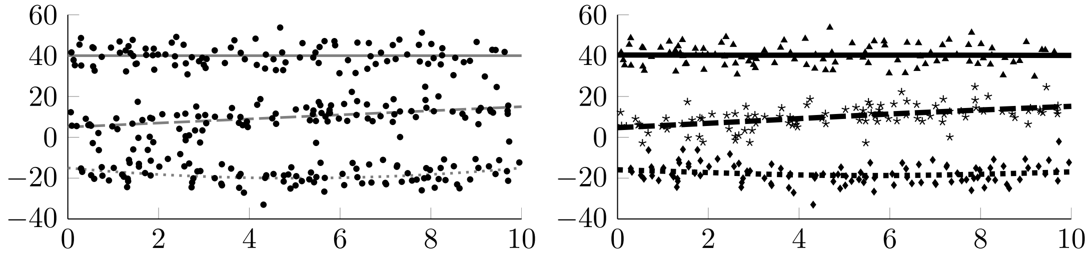
and the energy

$$\hat{\theta}_n = f_n(\mu^{(n)})$$

where

$$f_n(\mu) = \frac{1}{n} \sum_{i=1}^{n} \bigwedge_{j=1}^{k} |z_i - \mu_j(t_i)|^2 + \lambda \sum_{j=1}^{k} \|\nabla^2 \mu_j\|_{L^2}^2.$$

Figure 3.1: Smoothed data association trajectory results for the k-means method.



The figure on the left shows the raw data with the data generating model. That on the right shows the output of the k-means algorithm. The parameters used are: $k = 3, T = 10$, ϵ_i^j from a $N(0, 5)$ truncated at ± 100, $\lambda = 1$, $x^1(t) = -15 - 2t + 0.2t^2$, $x^2(t) = 5 + t$ and $x^3(t) = 40$.

Figure 3.1 shows the raw data and output of the k-means algorithm for one realization of the model. We run Monte Carlo trials for increasing numbers of data points; in particular we run 10^3 numerical trials independently for each $n = 300, 600, \ldots, 3000$ where we generate the data from (3.21) and cluster using the above algorithm. Each numerical experiment is independent.

Results, shown in Figure 3.2, illustrate that as measured by η the performance of the k-means method improves with the size of the available data set, as do the proportion of data points correctly assigned. The minimum energy stabilizes as the size of the data set increases, although the algorithm does take more iterations for the method to converge. We also note that the energy of the data generating functions is higher than the minimum energy.

Since the iterative k-means algorithm described above does not necessarily identify global minima, we tested the algorithm on two targets whose paths intersect as shown in Figure 3.3. The data association hypotheses corresponding to correct and incorrect associations, after the crossing point, correspond to two local minima. The observation window $[0, T]$ was expanded to investigate the convergence to the correct data association hypothesis. To enable this to be described in more detail we introduce the crossing and non-crossing energies:

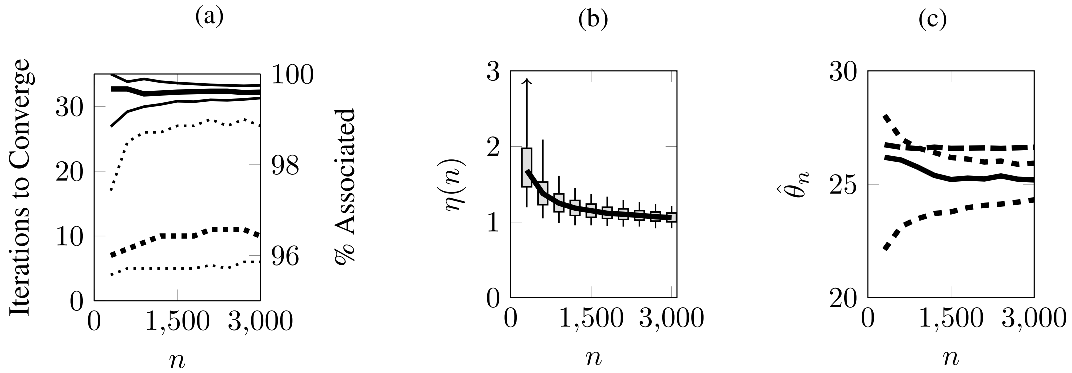
$$E_{\text{c}} = \frac{1}{T} f_n(\mu_{\text{c}})$$
$$E_{\text{nc}} = \frac{1}{T} f_n(\mu_{\text{nc}})$$

where μ_{c} and μ_{nc} are the k-means centers for the crossing (correct) and non-crossing (incorrect) solutions. To allow the association performance to be quantified, we therefore define the relative energy

$$\Delta E = E_{\text{c}} - E_{\text{nc}}.$$

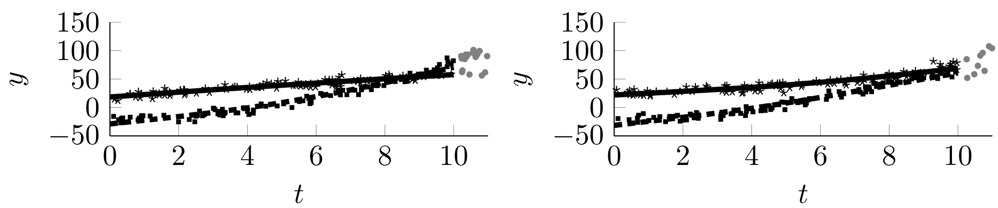
To determine how many numerical trials we should run in order to get a good number of simulations that produce crossing and non-crossing outputs we first ran the experiment until we achieved at least 100 tracks that crossed and at least 100 that did not. I.e. let N_t^{c} be the number of trials that output tracks that crossed and N_t^{nc} be the number of trials that output tracks that did not cross. We stop when $\min\{N_t^{\text{c}}, N_t^{\text{nc}}\} \geq 100$. Let $N_t = 10\,(N_t^{\text{c}} + N_t^{\text{c}})$. We then re-ran the experiment with N_t trials so we expect that we get 1000 tracks that do not cross and 1000 tracks that do cross at each time t.

Figure 3.2: Monte Carlo convergence results.



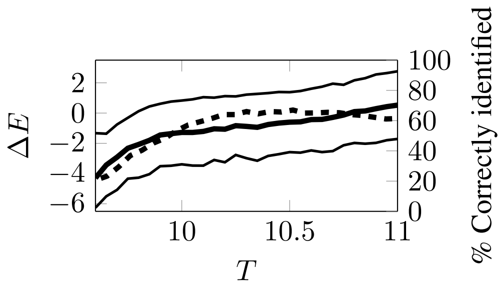
Convergence results for the parameters given in Figure 3.1. In (a) the thick dotted line corresponds to the median number of iterations taken for the method to converge and the thinner dotted lines are the 25% and 75% quantiles. The thick solid line corresponds to the median percentage of data points correctly identified and the thinner solid line are the 25% and 75% quantiles. (b) shows the median value of $\eta(n)$ (solid), interquartile range (box) and the interval between the 5% and 95% percentiles (whiskers). (c) shows the mean minimum energy $\hat{\theta}_n$ (solid) and the 10% and 90% quantiles (dashed). The energy associated with the data generating model is also shown (long dashes). In order to increase the chance of finding a global minimum for each Monte Carlo trial ten different initializations were tried and the one that had the smallest energy on termination was recorded.

Figure 3.3: Crossing tracks in the k-means method.



Typical data sets for times up to $T_{\max}$ with cluster centers, fitted up till time T, exhibiting crossing and non-crossing behavior. The parameters used are $k = 2$, $T_{\min} = 9.6 \leq T \leq 11 = T_{\max}$, $\epsilon_i^j \overset{\text{iid}}{\sim} N(0, 5)$, $x^1(t) = -20 + t^2$ and $x^2(t) = 20 + 4t$. There are $n = 220$ data points uniformly distributed over $[0, 11]$ with 110 observations for each track. The crossing occurs at approximately $t \approx 8.6$ but we wait a further time unit before investigating the decision making procedure.

Figure 3.4: Energy differences in the k-means method.



Mean results are shown for data obtained using the parameters given in Figure 3.3 for data up to time T (between T_{min} and T_{max}). The thick solid line shows the mean ΔE and the thinner lines one standard deviation either side of the mean. The dashed line shows the percentage of times we correctly identified the tracks as crossing.

The results in Figure 3.4 show that initially the better solution to the k-means minimization problem is the one that incorrectly partitions the tracks after the intersection. However, as time is run forward the k-means favors the partition that correctly associates tracks to targets. This is reflected in both an increase in ΔE and the percentage of outputs that correctly identify the switch. Our results show that for $T > 9.7$ the energy difference between the two minima grows linearly with time. However, when we look which minima the k-means algorithm finds our results suggest that after time $T \approx 10.25$ the probability of finding the correct minima stabilizes at approximately 64%. There is reasonably large variance in the energy difference. The mean plus standard deviation is positive for all T greater than 9.8, however it takes until $T = 10.8$ for the average energy difference to be positive.

3.4.2 Example 2: Passive Electromagnetic Source Tracking

In the previous example the data is simply a linear projection of the trajectories. In contrast, here we consider the more general case where the measurement X and model Y spaces are very different; being connected by a complicated mapping that results in a very non-linear cost function d. While the increased complexity of the cost function does lead to a (linear in data size) increase in computational cost, the problem is equally amenable to our approach.

In this example we consider the tracking of targets that periodically emit radio pulses as they travel on a two dimensional surface. These emissions are detected by an array of (three) sensors that characterize the detected emissions in terms of 'time of arrival', 'signal amplitude' and the 'identity of the sensor making the detection'.

Expressed in this way, the problem has a structure which does not fall directly within the framework which the theoretical results of previous sections cover. In particular, the observations are not independent (we have exactly one from each target in each measurement interval), they are not identically distributed and they do not admit an empirical measure which is weakly convergent in the large data limit.

This formulation could be refined so that the problem did fall precisely within the framework; but only at the expense of losing physical clarity. This is not done but as shall be seen below, even in the current formulation, good performance is obtained. This gives some confidence that k-means like strategies in general settings, at least when the qualitatively important features of the problem are close to those considered theoretically, and gives some heuristic justification for the lack of rigor.

Three sensors receive amplitude and time of arrival from each target with periodicity τ. Data at each sensor are points in $\mathbb{R}^2$ whilst the cluster centers (trajectories) are time-parameterized curves in a different $\mathbb{R}^2$ space.

In the generating model, for clarity we again index the targets in the observed amplitude and time of arrival. However, we again assume that this identifier is not observed and this notation is redefined (identities suppressed) when we apply the k-means method.

Let $x_j(t) \in \mathbb{R}^2$ be the position of target j for $j = 1, 2, \dots k$ at time $t \in [0, T]$. In every time frame of length τ each target emits a signal which is detected at three sensors. The time difference from the start of the time frame to when the target emits this signal is called the time offset. The time offset for each target is a constant which we call o_j for $j = 1, 2, \dots, k$. Target j therefore emits a signal at times

$$\tilde{t}_j(m) = m\tau + o_j$$

for $m \in \mathbb{N}$ such that $\tilde{t}_j(m) \leq T$. Note that this is not the time of arrival and we do not observe $\tilde{t}_j(m)$.

Sensor p at position z_p detects this signal some time later and measures the time of arrival $t_j^p(m) \in [0, T]$ and amplitude $a_j^p(m) \in \mathbb{R}$ from target j. The time of arrival is

$$t_j^p(m) = m\tau + o_j + \frac{|x_j(m) - z_p|}{c} + \epsilon_j^p(m) = \tilde{t}_j(m) + \frac{|x_j(m) - z_p|}{c} + \epsilon_j^p(m)$$

where c is the speed of the signal and $\epsilon_j^p(m)$ are iid noise terms with variance σ^2. The amplitude is

$$a_j^p(m) = \log\left(\frac{\alpha}{|x_j(m) - z_p|^2 + \beta}\right) + \delta_j^p(m)$$

where α and β are constants and $\delta_j^p(m)$ are iid noise terms with variance ν^2. We assume the parameters α, β, c, σ, τ, ν and z_p are known.

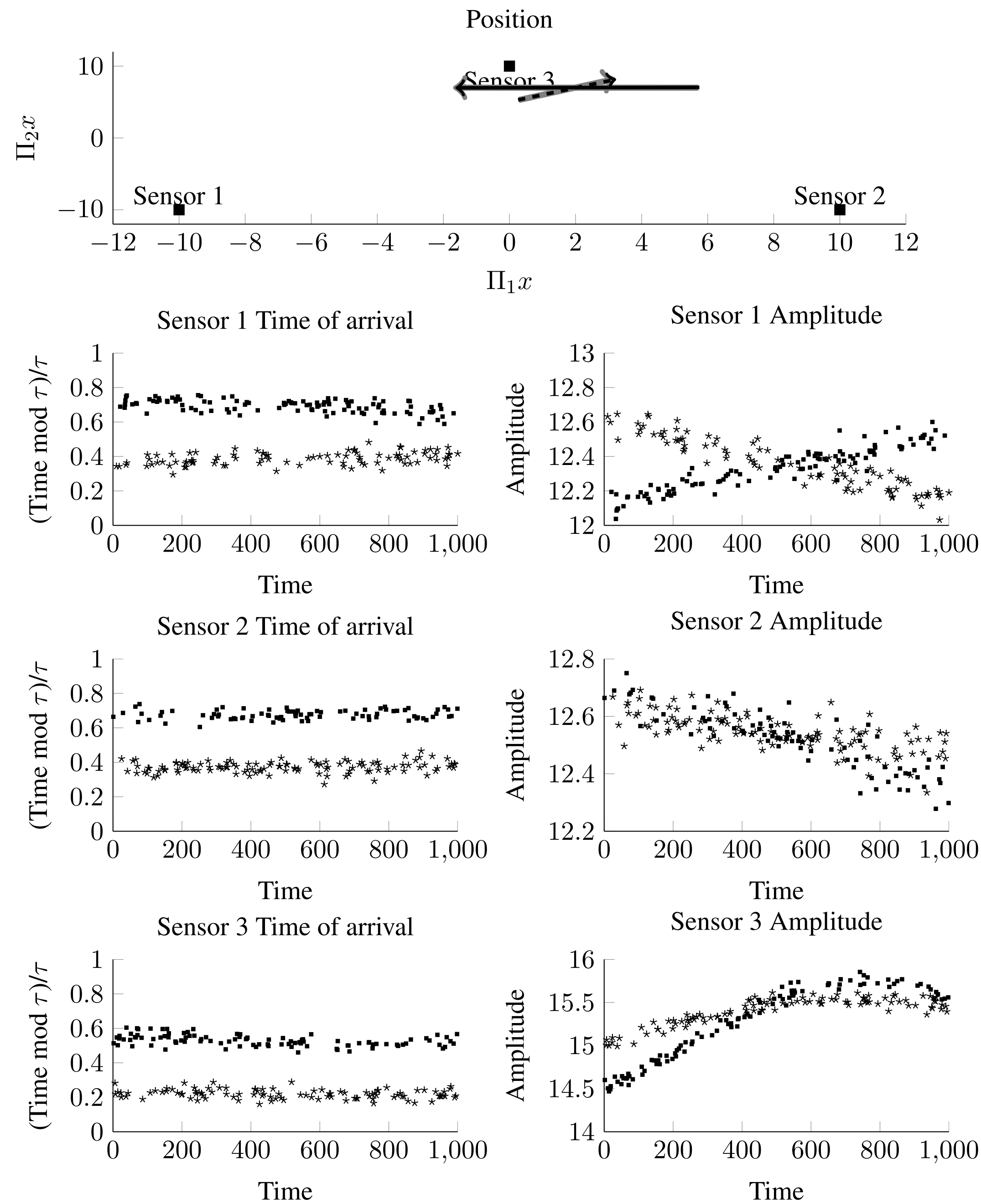
To simplify the notation $\Pi_q x : \mathbb{R}^2 \to \mathbb{R}$ is the projection of x onto it's q^{th} coordinate for $q = 1, 2$. I.e. the position of target j at time t can be written $x_j(t) = (\Pi_1 x_j(t), \Pi_2 x_j(t))$.

In practice we do not know to which target each observation corresponds. We use the k-means method to partition a set $\{\xi_i = (t_i, a_i, p_i)\}_{i=1}^n$ into the k targets. Note the relabeling of indices; $\xi_i = (t_i, a_i, p_i)$ is the time of arrival t_i, amplitude a_i and sensor p_i of the i^{th} detection. The cluster centers are in a function-parameter product space $\mu_j = (\hat{x}_j(t), \hat{o}_j) \in C^0([0, T]; \mathbb{R}^2) \times [0, \tau) \subset C^0([0, T]; \mathbb{R}^2) \times \mathbb{R}$ that estimates the j^{th} target's trajectory and time offset. The k-means minimization problem is

$$\mu^{(n)} = \underset{\mu \in (C^0 \times [0,\tau))^k}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \bigwedge_{j=1}^{k} d(\xi_i, \mu_j)$$

for a choice of cost function d. If we look for cluster centers as straight trajectories then we can

Figure 3.5: Representative data and resulting tracks for the passive tracking example.



Representative data is shown for the parameters $k = 2$, $\tau = 1$, $T = 1000$, $c = 100$, $z_1 = (-10, -10)$, $z_2 = (10, -10)$, $z_3 = (0, 10)$, $\epsilon_j^p(m) \overset{\text{iid}}{\sim} N(0, 0.03^2)$, $\delta_j^p(m) \overset{\text{iid}}{\sim} N(0, 0.05^2)$, $\alpha = 10^8$, $\beta = 5$, $x_1(t) = \frac{\sqrt{2}t}{400}(1, 1) + (0, 5)$, $x_2(t) = (6, 7) - \frac{t}{125}(1, 0)$, $o_1 = 0.3$ and $o_2 = 0.6$, given the sensor configuration shown at the top of the figure. The k-means method was run until it converged, with the trajectory component of the resulting cluster centers plotted with the true trajectories at the top of the figure. Target one is the dashed line with starred data points, target two is the solid line and square data points.

restrict ourselves to functions of the form $x_j(t) = x_j(0) + v_j t$ and consider the cluster centers as finite dimensional objects. This allows us to redefine our minimization problem as

$$\mu^{(n)} = \underset{\mu \in (\mathbb{R}^4 \times [0,\tau))^k}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{j} \bigwedge_{j=1}^{k} d(\xi_i, \mu_j)$$

so that now $\mu_j = (x_j(0), v_j, o_j) \in \mathbb{R}^2 \times \mathbb{R}^2 \times [0, \tau)$. We note that in this finite dimensional formulation it is not necessary to include a regularization term; a feature already anticipated in the definition of the minimization problem.

For $\mu_j = (x_j, v_j, o_j)$ we define the cost function

$$d((t,a,p),\mu_j) = ((t,a) - \psi(\mu_j,p,m)) \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\nu^2} \end{pmatrix} \left(\begin{pmatrix} t \\ a \end{pmatrix} - \psi(\mu_j,p,m)^\top \right)$$

where $m = \max\{n \in \mathbb{N} : n\tau \le t\}$,

$$\psi(\mu_j,p,m) = \left(\frac{|x_j + m\tau v_j - z_p|}{c} + o_j + m\tau, \log\left(\frac{\alpha}{|x_j + m\tau v_j - z_p|^2 + \beta} \right) \right)$$

and superscript $\top$ denotes the transpose.

We initialize the partitions by choosing $\varphi^{(0)} : \{1,2,\ldots,n\} \to \{1,2,\ldots,k\}$ uniformly randomly. At the r^{th} iteration the k-means minimization problem is then partitioned into k independent problems

$$\mu_j^{(r)} = \underset{\mu_j}{\operatorname{argmin}} \sum_{i \in (\varphi^{(r-1)})^{-1}(j)} d((t_i,a_i,p_i),\mu_j) \quad \text{for } 1 \le j \le k.$$

A range of initializations for μ_j are used to increase the chance of the method converging to a global minimum.

For optimal centers conditioned on partition $\varphi^{(r-1)}$ we can define the partition $\varphi^{(r)}$ to be the optimal partition of $\{(t_i,a_i,p_i)\}_{i=1}^n$ conditioned on centers $(\mu_j^{(r)})$ by solving

$$\begin{aligned} \varphi^{(r)} : \{1,2,\ldots,n\} &\to \{1,2,\ldots,k\} \\ i &\mapsto \underset{j=1,2,\ldots,k}{\operatorname{argmin}} d((t_i,a_i,p_i),\mu_j^{(r)}). \end{aligned}$$

The method has converged when $\varphi^{(r)} = \varphi^{(r-1)}$ for some r. Typical simulated data and resulting trajectories are shown in Figure 3.5.

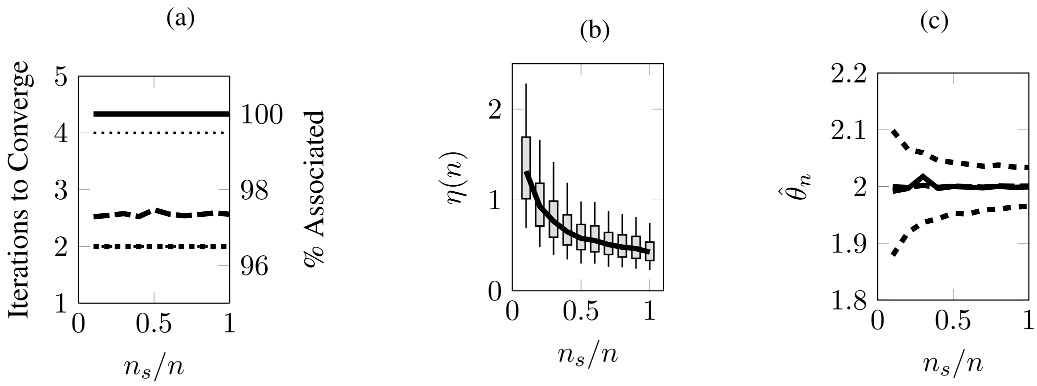
To illustrate the convergence result achieved above we performed a test on a set of data simulated from the same model as shown in Figure 3.5. We sample n_s observations from $\{(t_i,a_i,p_i)\}_{i=1}^n$ and compare our results as $n_s \to n$. Let $\hat{x}^{(n_s)}(t) = (\hat{x}_1^{(n_s)}(t), \ldots, \hat{x}_k^{(n_s)}(t))$ be the position output by the k-means method described above using n_s data points and $x(t) = (x_1(t), \ldots, x_k(t))$ be the true values of each cluster center. We use the metric

$$\eta(n_s) = \frac{1}{k} \sqrt{\sum_{j=1}^{k} \|\hat{x}_j^{(n_s)} - x_j\|_{L^2}^2}$$

to measure how close the estimated position is to the exact position. Note we do not use the estimated time offset given by the first model. The number of iterations required for the method to converge is also recorded. Results are shown in Figure 3.6.

In this example the data has enough separation that we are always able to recover the true data partition. We also see improvement in our estimated cluster centers and convergence of the minimum energy as we increase the size of the data. Finding global minima is difficult and although we run the k-means method from multiple starting points we sometimes only find local minima. For $\frac{n_s}{n} = 0.3$ we see the effect of finding local minima. In this case only one Monte Carlo trial produces a bad result, but the error η is so great (around 28 times greater than the average) that it can be seen in the mean result shown in Figure 3.6(c).

Figure 3.6: Monte Carlo convergence results.



Convergence results for 10^3 Monte Carlo trials with the parameters given in Figure 3.5; expressed with the notation used in Figure 3.2. In (a) we have also recorded the mean number of iterations to converge (long dashes). The 25% and 75% quantiles for the number of iterations to converge is 2 and 4 for all n respectively. The 25% and 75% quantiles for the percentage of data points correctly identified is 100% in both cases for all n. This is due to large separation in the data space. To increase the chance of finding a global minimum for each Monte Carlo trial, out of five different initializations, that which had the smallest energy on terminating was recorded.

Chapter 4

Rate of Convergence for a Smoothing Spline with Data Association Model

Abstract

The problem of estimating multiple trajectories from unlabeled data comprises two coupled problems. The first is a data association problem: how to map data points onto individual trajectories. The second is, given a solution to the data association problem, to estimate those trajectories. We construct estimators as a solution to a variational problem which uses smoothing splines under a k-means like framework and show that, as the number of data points increases, we have stability. More precisely, we show that these estimators converge weakly in an appropriate Sobolev space with probability one. Furthermore, we show that the estimators converge in probability with rate $\frac{1}{\sqrt{n}}$ in the L^2 norm (strongly).

4.1 Introduction

Given observations from multiple moving targets we face two (coupled) problems. The first is associating observations to targets: the data association problem. The second is estimating the trajectory of each target given the appropriate set of observations. When there is one target then the data association problem is trivial. However, when the number of targets is greater than one (even when the number of targets is known) the set of data association hypothesis grows combinatorially with the number of data points. Very quickly it becomes infeasible to check every possibility. Hence the value of fast approximate solutions.

In this chapter we interpret the problem of estimating multiple trajectories with unknown data association (see Figure 4.1) in such a way that the k-means method may be applied to find a solution. As a special case of the previous chapter this is a non-standard application of the k-means method where we generalize the notion of a 'cluster center' to partition finite dimensional data using infinite dimensional cluster centers. In this chapter the cluster centers are trajectories in some function space and the data are space-time observations.

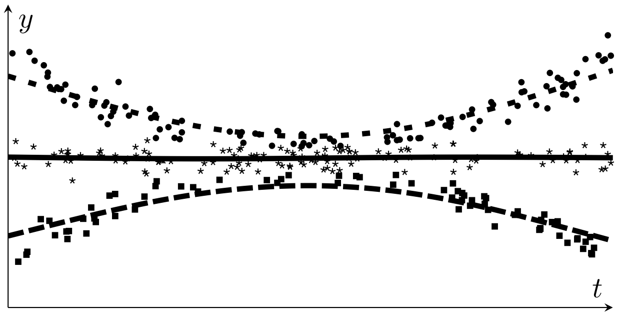


Figure 4.1: Unlabeled data is generated from three targets and using minimizers of (4.2) we can find a partitioning of the data set and non-parametrically estimate each trajectory using the k-means algorithm.

Let $\Theta \subset (H^s)^k$ where H^s is the Sobolev space of degree s. Given a data set

$$\{(t_i, y_i)\}_{i=1}^n \subset [0,1] \times \mathbb{R}^d$$

and a model for the observation process

$$y_i = \mu^\dagger_{\varphi(i)}(t_i) + \epsilon_i \tag{4.1}$$

where $\epsilon_i \overset{\text{iid}}{\sim} \phi_0$ and $t_i \overset{\text{iid}}{\sim} \phi_T$ for densities ϕ_0 and ϕ_T on $[0,1]$ and $\mathbb{R}^d$ respectively. We assume that the index of the cluster responsible for any given observation is an independent random variable with a categorical distribution of parameter vector $p = (p_1, \ldots, p_k)$, writing $\varphi(i) \sim \mathrm{Cat}(p)$ to mean $\mathbb{P}(\varphi(i) = j) = p_j$. This assumptions allow us to write the density of y given t, which we denote by $\phi_Y(y|t)$, as

$$\phi_Y(y|t) = \sum_{j=1}^k p_j \phi_0(\mu_j^\dagger(t) - y).$$

We can summarize the data generating process as follows. A cluster is selected at random: $\mathbb{P}(\varphi = j) = p_j$, the time and observation error are drawn independently from their respective distributions, $t \sim \phi_T$, and $\epsilon \sim \phi_0$; and we observe $(t, y = \mu^\dagger_\varphi(t) + \epsilon)$.

The aim is to estimate $\mu^\dagger = (\mu_1^\dagger, \ldots, \mu_k^\dagger) \in \Theta$. In particular the data association

$$\varphi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, k\}$$

is unknown. With a single trajectory ($k = 1$) the problem is precisely the spline smoothing problem, see Chapter 5. For $k > 1$ trajectories there is an additional data association problem coupled to the spline smoothing problem. We call this the smoothing-data association (SDA) problem.

We assume k is fixed and known. The aim of this chapter is to construct a sequence of estimators $\mu^{(n)}$ of $\mu^\dagger$ from the data $\{(t_i, y_i)\}_{i=1}^n$ and study the asymptotic behavior as $n \to \infty$.

For each n our estimate is given as the minimizer of $f_n : \Theta \to \mathbb{R}$ defined by

$$f_n(\mu) = \frac{1}{n} \sum_{i=1}^{n} \bigwedge_{j=1}^{k} |y_i - \mu_j(t_i)|^2 + \lambda \sum_{j=1}^{k} \|\nabla^s \mu_j\|_{L^2}^2 \tag{4.2}$$

where $|\cdot|$ is the Euclidean norm on $\mathbb{R}^d$, $\bigwedge_{j=1}^{k} z_j = \min\{z_1, \dots, z_k\}$ and λ is a positive constant. Penalizing the s^{th} derivative ensures that the problem is well posed. Optimizing this function can be interpreted as seeking a hard data association: given $\mu \in \Theta$ each observation (t_i, y_i) is associated with the trajectory closest to it so the corresponding data association solution is given by

$$\varphi^\mu(i) = \operatorname*{argmin}_{j=1,2,\dots,k} |\mu_j(t_i) - y_i|.$$

Here, we focus upon characterizing the solution of this problem rather than computational methods to obtain this solution. However, a variant of the k-means method would be readily applicable — for this reason we term the μ_j cluster centers.

The choice of regularization scheme and, in particular, of λ is not straightforward. For $k = 1$ there are many results in the spline literature on the selection of $\lambda = \lambda_n$ and the resulting asymptotic behavior as $n \to \infty$, see for example Chapter 5 and [3, 45–47, 98, 117, 131, 143, 144, 146, 156–158, 173]. In this case one has $\lambda_n \to 0$ and can expect $\mu^{(n)}$ to converge to $\mu^\dagger$. Convergence is either with respect to a Hilbert scale, e.g. L^2, or in the dual space, i.e. weak convergence. Using a Hilbert scale in effect measures the convergence in a norm weaker than H^s.

The approach we take is to penalize the s^{th} derivative where we assume s is known. Choosing s is also an interesting problem which we don't address in this thesis. By Taylor's Theorem we can write $H^s = \mathcal{H}_0 \oplus \mathcal{H}_1$ where

$$\mathcal{H}_0 = \text{span}\left\{\zeta_i(t) = \frac{t^i}{i!} : i = 0, 1, \dots, s-1\right\},$$
$$\mathcal{H}_1 = \left\{g \in \mathcal{H} : \nabla^i g(0) = 0 \text{ for all } i = 0, 1, \dots, s-1\right\}.$$

We use $\|\cdot\|_1 = \|\nabla^s \cdot\|_{L^2}$ as the norm on $\mathcal{H}_1$ and denote the $\mathcal{H}_0$ norm by $\|\cdot\|_0$. Since $\mathcal{H}_0$ is finite dimensional we are free to use any norm we choose without changing the topology. We can view $H^s = \mathcal{H}_0 \oplus \mathcal{H}_1$ as a multiscale decomposition of H^s. The polynomial component represents a coarse approximation. The regularization penalizes oscillations on the fine scale, i.e. in $\mathcal{H}_1$.

In the case $k = 1$, f_n is quadratic and one can find an explicit representation of $\mu^{(n)}$, i.e. there exists a random function $G_{n,\lambda}$ such that with probability one $\mu^{(n)} = G_{n,\lambda}\nu^{(n)}$ for some function $\nu^{(n)}$ which depends on the data. When $k > 1$ the problem is no longer convex and the methodology used in the $k = 1$ case fails. The authors know of no method which would allow $\lambda \to 0$ for $k > 1$ and therefore we treat λ as a constant in this chapter. A consequence of this regularization is that we cannot expect to recover the *true* cluster centroids, even in the large data limit.

The first result of this chapter (Theorem 4.2.1) is to show that there exists $\mu^{(\infty)} \in \Theta$

such that (up to subsequences) $\mu^{(n)} \rightharpoonup \mu^{(\infty)}$ a.s. in H^s and $\mu^{(\infty)}$ is a minimizer of f_∞ defined by

$$f_\infty(\mu) = \int_0^1 \int_{\mathbb{R}^d} \bigwedge_{j=1}^k |y - \mu_j(t)|^2 \, \phi_Y(y|t)\phi_T(t) \, \mathrm{d}y\mathrm{d}t + \lambda \sum_{j=1}^k \|\nabla^s \mu_j\|_{L^2}^2 \tag{4.3}$$

Considering the law of large numbers the limit f_∞ is natural. The functional f_∞ can be seen as a limit of f_n, the nature of which will be made rigorous in Section 4.2.

We recall that the motivation for the minimization problem (4.2) is to embed the problem into a framework that allows the application of the k-means method. Large data limits for the k-means have been studied extensively in finite dimensions, see for example [10, 18, 42, 80, 124–126]. There are fewer results for the infinite dimensional case, with Chapter 3 and [21, 95, 96, 103] the only results known to us. Of these, only Chapter 3 can be applied to finite dimensional data and infinite dimensional cluster centers but this required bounded noise. The first contribution of this chapter is to extend this convergence result to unbounded noise for the SDA problem.

By the compact embedding of H^s into L^2 we have that (upto subsequences) $\mu^{(n)} \to \mu^{(\infty)}$ a.s. in L^2. The second result of this chapter is to show in Theorem 4.3.1 that the rate of convergence in L^2 is of order $\frac{1}{\sqrt{n}}$ in probability. I.e.

$$\|\mu^{(n)} - \mu^{(\infty)}\|_{L^2} = O_p\left(\frac{1}{\sqrt{n}}\right).$$

This is closely related to the central limit theorem first proved for the k-means method by Pollard [126] for Euclidean data. We extend his methodology to cluster centers in H^s to prove our rate of convergence result.

Section 4.2 contains the convergence results. The rate of convergence results are in Section 4.3.

4.2 Convergence

To show convergence we apply Theorem 2.2.1. The following two subsections prove that the conditions required to apply this theorem, i.e. that f_∞ is the Γ-limit of f_n and that the minimizers $\mu^{(n)}$ are uniformly bounded, hold with probability one. Using the compact embedding of H^s into L^2 we are able to conclude that upto subsequences convergence is strong in L^2.

For a fixed $\delta > 0$ we define the set Θ to be the set of functions in $(H^s)^k$ which have minimum separation distance of δ:

$$\Theta = \left\{\mu \in (H^s)^k : |\mu_j(t) - \mu_l(t)| \geq \delta \; \forall t \in [0,1] \text{ and } j \neq l\right\}. \tag{4.4}$$

For $d = 1$ this is a strong assumption as we restrict ourselves to trajectories that do not intersect. However, for larger d the assumption is much less stringent.

First let us show that Θ is weakly closed in $(H^s)^k$. Take any sequence $\mu^{(n)} \in \Theta$ such that $\mu^{(n)} \rightharpoonup \mu \in (H^s)^k$. We have to show $\mu \in \Theta$. Pick $t \in [0,1]$, $j \neq l$ and define $F : \Theta \to \mathbb{R}^d$

by $F : \nu \to \nu_j(t) - \nu_l(t)$, note that F is in the dual space of $(H^s)^k$. Hence

$$\delta \leq |\mu_j^{(n)}(t) - \mu_l^{(n)}(t)| = |F(\mu^{(n)})| \to |F(\mu)| = |\mu_j(t) - \mu_l(t)|.$$

Therefore $\mu \in \Theta$. Furthermore we can show that f_n, f_∞ are weakly lower semi-continuous by Propositions 3.3.8 and 3.3.9 hence they obtain their minimizers on Θ.

The minimum separation distance implies that f_∞ is locally quadratic on Θ. This is a consequence of being able to define the association

$$j(t, y) = \operatorname*{argmin}_j |y - \mu_j(t)|$$

uniquely for (Lebesgue) almost every $y \in \mathbb{R}$ and every $t \in [0, 1]$. Modulo some technical difficulties at the boundary of each partition (which we address in Lemma 4.3.2) we can write f_∞ as the sum of k quadratic functionals and is therefore quadratic itself. This implies f_∞ is differentiable and in particular allows the application of Taylor's theorem in Section 4.3.

We now state our assumptions.

Assumptions 3. *We use the following assumptions on the data model.*

3.1 The data sequence (t_i, y_i) is independent and identically distributed in accordance with the model (4.1), with $\varphi(i) \sim$ Cat(p), $\epsilon_i \sim \phi_0$, $t_i \sim \phi_T$ and $\varphi(i), \epsilon_j, t_k$ are independent for all i, j and k. We assume ϕ_0 and ϕ_T are densities with respect to the Lebesgue measure on $\mathbb{R}^d$ and $[0, 1]$ respectively and use the same symbols to refer to these densities and to their associated measures.

3.2 The density ϕ_0 is centered and with finite second moment.

3.3 For all $\epsilon \in \mathbb{R}^d$ we assume $\phi_0(\epsilon) > 0$.

3.4 We can bound ϕ_T away from 0, i.e. $\inf_{t\in[0,1]} \phi_T(t) > 0$.

Observe that

$$\begin{aligned}
f_\infty(\mu^\dagger) &= \frac{1}{n}\sum_{i=1}^n \bigwedge_{j=1}^k |\mu_j^\dagger(t_i) - y_i|^2 + \lambda \sum_{j=1}^k \|\nabla^s \mu_j^\dagger\|_{L^2}^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n |y_i - \mu_{\varphi(i)}^\dagger(t_i)|^2 + \lambda \sum_{j=1}^k \|\nabla^s \mu_j^\dagger\|_{L^2}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 + \lambda \sum_{j=1}^k \|\nabla^s \mu_j^\dagger\|_{L^2}^2 \\
&\to \mathrm{Var}(\epsilon_i) + \lambda \sum_{j=1}^k \|\nabla^s \mu_j^\dagger\|_{L^2}^2 =: \alpha < \infty
\end{aligned}$$

where the convergence is almost surely by the strong law of large numbers. Hence Assumption 3.2 implies that there exists N such that $\min_{\mu\in\Theta} f_n(\mu) < \alpha + 1$ for $n \geq N$ and $N < \infty$ with probability one (although N could depend on the sequence $\{(t_i, y_i)\}_{i=1}^n$ and so we could have $\sup_{\omega\in\Omega} N = \infty$).

To simplify our proofs we use Assumption 3.3 although the results of this chapter can be proved without it. The assumption is used in bounding the minimizers of f_n. Clearly if ϕ_0 has bounded support then each y_i is uniformly bounded (a.s.) and one can show that $|\mu^{(n)}(t)|$ is bounded uniformly in n and t (a.s.). When the support is unbounded, but Assumption 3.3 does not hold, our proofs hold with some trivial but notationally messy modifications.

Assumption 3.4 will be used in the rate of convergence section. This is used to show that f_∞ is positive definite.

We now state the main result for this section. The proof is an application of Theorem 2.2.1 once we have shown the Γ-limit (Theorem 4.2.2) and the uniform bound on the set of minimizers (Theorem 4.2.4).

Theorem 4.2.1. *Define* $f_n, f_\infty : \Theta \to \mathbb{R}$ *by* (4.2) *and* (4.3)*, where* $\Theta \subset (H^s)^k$ *for* $s \geq 1$ *is given by* (4.4)*, respectively. Under Assumptions 3.1-3.3 any sequence of minimizers* $\mu^{(n)}$ *of* f_n *are, with probability one, weakly compact and any weak limit* $\mu^{(\infty)}$ *is a minimizer of* f_∞*. Furthermore if* $\mu^{(n_m)} \rightharpoonup \mu^{(\infty)}$ *in* H^s *then* $\mu^{(n_m)} \to \mu^{(\infty)}$ *in* L^2*.*

4.2.1 The Γ-Limit

We claim the Γ-limit of (f_n) is given by (4.3).

Theorem 4.2.2. *Define* $f_n, f_\infty : \Theta \to \mathbb{R}$ *by* (4.2) *and* (4.3) *respectively where* $\Theta \subset (H^s)^k$ *for* $s \geq 1$ *is given by* (4.4)*. Under Assumptions 3.1-3.2*

$$f_\infty = \Gamma\text{-}\lim_n f_n$$

for almost every sequence of observations $(t_1, y_1), (t_2, y_2), \ldots$.

Proof. We are required to show that the two inequalities in Definition 2.2.1 hold with probability 1. In order to do this we follow Chapter 3 and consider a subset of Ω of full measure, Ω', and show that both statements hold for every data sequence obtained from that set.

For clarity let $P(\mathrm{d}(t,y)) = \phi_Y(\mathrm{d}y|t)\phi_T(\mathrm{d}t)$. Define $g_\mu(t,y) = \bigwedge_{j=1}^k (y - \mu(t))^2$. Let $P_n^{(\omega)}$ be the associated empirical measure arising from the particular elementary event ω, which we define via it's action on any continuous bounded function $h : [0,1] \times \mathbb{R}^d \to \mathbb{R}$: $P_n^{(\omega)} h = \frac{1}{n}\sum_{i=1}^n h\left(t_i^{(\omega)}, y_i^{(\omega)}\right)$ where $\left(t_i^{(\omega)}, y_i^{(\omega)}\right)$ emphasizes that these are the observations associated with elementary event ω. To highlight the dependence of f_n on ω we write $f_n^{(\omega)}$. We can write

$$f_n^{(\omega)}(\mu) = P_n^{(\omega)} g_\mu + \lambda \sum_{j=1}^k \|\nabla^s \mu_j\|_{L^2}^2 \qquad \text{and} \qquad f_\infty = P g_\mu + \lambda \sum_{j=1}^k \|\nabla^s \mu_j\|_{L^2}^2.$$

We define

$$\Omega' = \left\{\omega \in \Omega : P_n^{(\omega)} \Rightarrow P\right\} \cap \left\{\omega \in \Omega : P_n^{(\omega)}(B(0,q))^c \to P(B(0,q))^c \ \forall q \in \mathbb{N}\right\}$$
$$\cap \left\{\omega \in \Omega : \int_{(B(0,q))^c} |y|^2 \, P_n^{(\omega)}(\mathrm{d}t, y) \to \int_{(B(0,q))^c} |y|^2 \, P(\mathrm{d}(t,y)) \ \forall q \in \mathbb{N}\right\}$$

then $\mathbb{P}(\Omega') = 1$ by the almost sure weak convergence of the empirical measure [56] and the strong law of large numbers.

Fix $\omega \in \Omega'$ and we start with the lim inf inequality. Let $\mu^{(n)} \rightharpoonup \mu$. By Theorem 1.1 in [64] we have

$$\int_{[0,1]\times\mathbb{R}^d} \liminf_{n\to\infty,(t',y')\to(t,y)} g_{\mu^{(n)}}((t',y'))\, P(\mathrm{d}(t,y)) \\ \leq \liminf_{n\to\infty} \int_{[0,1]\times\mathbb{R}^d} g_{\mu^{(n)}}(t,y)\, P_n^{(\omega)}(\mathrm{d}(t,y)).$$

By the same argument as in Proposition 3.3.8(ii) we have

$$\liminf_{n\to\infty,(t',y')\to(t,y)} \left(y' - \mu_j^{(n)}(t')\right)^2 \geq (y - \mu_j(t))^2 .$$

Taking the minimum over j we have

$$\liminf_{n\to\infty,(t',y')\to(t,y)} g_{\mu^{(n)}}(t',y') \geq g_\mu(t,y).$$

And as a consequence of the Hahn-Banach theorem $\liminf_{n\to\infty} \|\nabla^s \mu_j^{(n)}\|_{L^2}^2 \geq \|\nabla^2 \mu_j\|_{L^2}^2$. Therefore

$$\liminf_{n\to\infty} f_n^{(\omega)}(\mu^{(n)}) \geq f_\infty(\mu)$$

as required.

We now establish the existence of a recovery sequence for every $\omega \in \Omega'$ and every $\mu \in \Theta$. Let $\mu^{(n)} = \mu \in \Theta$. Let ζ_q be a $C^\infty(\mathbb{R}^{d+1})$ sequence of functions such that $0 \leq \zeta_q(t,y) \leq 1$ for all $(t,y) \in \mathbb{R}^{d+1}$, $\zeta_q(t,y) = 1$ for $(t,y) \in B(0,q-1)$ and $\zeta_q(t,y) = 0$ for $(t,y) \notin B(0,q)$. Then the function $\zeta_q(t,y)g_\mu(t,y)$ is continuous for all q. We also have, for any $(t,y) \in [0,1]\times\mathbb{R}^d$,

$$\begin{aligned} \zeta_q(t,y)g_\mu(t,y) &\leq \zeta_q(t,y)|y-\mu_1(t)|^2 \\ &\leq 2\zeta_q(t,y)\left(|y|^2 + |\mu_1(t)|^2\right) \\ &\leq 2\zeta_q(t,y)\left(|y|^2 + \|\mu_1\|_{L^\infty([0,1])}^2\right) \\ &\leq 2|q|^2 + 2\|\mu_1\|_{L^\infty([0,1])}^2 < \infty \end{aligned}$$

so $\zeta_q g_\mu$ is a continuous and bounded function, hence by the weak convergence of $P_n^{(\omega)}$ to P we have

$$P_n^{(\omega)}\zeta_q g_\mu \to P\zeta_q g_\mu$$

as $n\to\infty$ for all $q\in\mathbb{N}$. For all $q\in\mathbb{N}$ we have

$$\begin{aligned} \limsup_{n\to\infty} |P_n^{(\omega)}g_\mu - Pg_\mu| &\leq \limsup_{n\to\infty} |P_n^{(\omega)}g_\mu - P_n^{(\omega)}\zeta_q g_\mu| + \limsup_{n\to\infty} |P_n^{(\omega)}\zeta_q g_\mu - P\zeta_q g_\mu| \\ &\quad + \limsup_{n\to\infty} |P\zeta_q g_\mu - Pg_\mu| \\ &= \limsup_{n\to\infty} |P_n^{(\omega)}g_\mu - P_n^{(\omega)}\zeta_q g_\mu| + |P\zeta_q g_\mu - Pg_\mu|. \end{aligned}$$

Therefore,

$$\limsup_{n\to\infty} |P_n^{(\omega)} g_\mu - P g_\mu| \le \limsup_{q\to\infty} \limsup_{n\to\infty} |P_n^{(\omega)} g_\mu - P_n^{(\omega)} \zeta_q g_\mu|$$

by the dominated convergence theorem. We now show that the right hand side of the above expression is equal to zero. We have

$$\begin{aligned}
|P_n^{(\omega)} g_\mu - P_n^{(\omega)} \zeta_q g_\mu| &\le P_n^{(\omega)} \mathbb{I}_{(B(0,q-1))^c} g_\mu \\
&\le \int_{[0,1]\times\mathbb{R}^d} \mathbb{I}_{(B(0,q-1))^c}(t,y) |y-\mu_1(t)|^2 \, P_n^{(\omega)}(\mathrm{d}(t,y)) \\
&\le 2 \int_{[0,1]\times\mathbb{R}^d} \mathbb{I}_{(B(0,q-1))^c}(t,y) |y|^2 \, P_n^{(\omega)}(\mathrm{d}(t,y)) \\
&\qquad + 2\|\mu_1\|^2_{L^\infty([0,1])} \int_{[0,1]\times\mathbb{R}^d} \mathbb{I}_{(B(0,q-1))^c}(t,y) \, P_n^{(\omega)}(\mathrm{d}(t,y)) \\
&\overset{n\to\infty}{\to} 2 \int_{[0,1]\times\mathbb{R}^d} \mathbb{I}_{(B(0,q-1))^c}(t,y) |y|^2 \, P(\mathrm{d}(t,y)) \\
&\qquad + 2\|\mu_1\|^2_{L^\infty([0,1])} \int_{[0,1]\times\mathbb{R}^d} \mathbb{I}_{(B(0,q-1))^c}(t,y) \, P(\mathrm{d}(t,y)) \\
&\overset{q\to\infty}{\to} 0
\end{aligned}$$

where the last limit follows by the monotone convergence theorem and Assumption 3.2. We have shown

$$\lim_{n\to\infty} |P_n^{(\omega)} g_\mu - P g_\mu| = 0.$$

Hence

$$f_n^{(\omega)}(\mu) \to f_\infty(\mu)$$

as required. □

4.2.2 Boundedness

The aim of this subsection is to show that the minimizers of f_n are uniformly bounded in n for almost every sequence of observations. We divide this into two parts; bounding each of the $\mathcal{H}_0$ and $\mathcal{H}_1$ norms. The $\mathcal{H}_1$ bound follows easily from the regularization. For the $\mathcal{H}_0$ bound we exploit the equivalence of norms on finite-dimensional vector spaces to choose a convenient norm on $\mathcal{H}_0$.

From our assumptions we may infer the existence of a set $\hat{\Omega} \subseteq \Omega$ such that for all $\omega \in \hat{\Omega}$ we have

$$f_n^{(\omega)}(\mu^\dagger) = P_n^{(\omega)} g_{\mu^\dagger} + \lambda \sum_{j=1}^k \|\nabla^s \mu_j^\dagger\|^2_{L^2} \to P g_{\mu^\dagger} + \lambda \sum_{j=1}^k \|\nabla^s \mu_j^\dagger\|^2_{L^2} =: \alpha$$

and $\mathbb{P}(\hat{\Omega}) = 1$. Now we let $\mu^{(n)}$ be a sequence minimizers and note that for n sufficiently large we have

$$\lambda \|\mu^{(n)}\|_1^2 \le f_n(\mu^{(n)}) \le f_n(\mu^\dagger) \le \alpha + 1.$$

Therefore $\|\mu^{(n)}\|_1$ is bounded almost surely. We are left to show the corresponding result for $\|\mu^{(n)}\|_0$.

The following lemma will be used to establish the main result of this subsection, Theorem 4.2.4. It shows that, if for some sequence $\nu^{(n)} \in H^s$ with $\|\nabla^s \nu^{(n)}\|_{L^2} \leq \sqrt{\alpha}$ and $\|\nu^{(n)}\|_0 \to \infty$, then we have that $|\nu^{(n)}(t)| \to \infty$ with the exception of at most finitely many $t \in [0,1]$. When applied to $\mu_j^{(n)}$ this will be used to show that in the limit, if any center is unbounded, then the minimization can be achieved over $k-1$ clusters — and hence to provide a contradiction.

Lemma 4.2.3. *Let $\nu^{(n)} \in H^s$ satisfy $\|\nabla^s \nu^{(n)}\|_{L^2} \leq \sqrt{\alpha}$ and $\|\nu^{(n)}\|_0 \to \infty$. Then, with the exception of at most finitely many $t \in [0,1]$ we have $|\nu^{(n)}(t)| \to \infty$. Furthermore for each $t \in (0,1)$ with $|\nu^{(n)}(t)| \to \infty$ there exists $c > 0$ such that $|\nu^{(n)}(r)| \to \infty$ uniformly for $r \in [t-c, t+c]$.*

Proof. Let the norm on $\mathcal{H}_0$ be given by

$$\|\nu\|_0 := \sum_{i=0}^{s-1} \frac{|\nabla^i \nu(0)|}{i!}. \tag{4.5}$$

By Taylor's theorem and the bound on $\|\nabla^s \nu^{(n)}\|_{L^2}$ we have

$$\left| \nu^{(n)}(t) - \sum_{i=0}^{s-1} \frac{\nabla^i \nu^{(n)}(0)}{i!} t^i \right| \leq \sqrt{\alpha}.$$

Now let $Q_n(t) = \sum_{i=0}^{s-1} \frac{\nabla^i \nu^{(n)}(0)}{i!} t^i$. If $\|\nu^{(n)}\|_0 \to \infty$ then at least one of $|\frac{\nabla^i \nu^{(n)}(0)}{i!}| \to \infty$. If all but one of the terms stay bounded then it is easy to see that for $t > 0$ we have $|Q_n(t)| \to \infty$ and the convergence is uniform over all intervals $[t,1]$. Now if $\frac{\nabla^{i_1} \nu^{(n)}(0)}{i_1!} \to \infty$ and $\frac{\nabla^{i_2} \nu^{(n)}(0)}{i_2!} \to -\infty$ and all other terms are bounded by M, i.e. $\left| \frac{\nabla^i \nu^{(n)}(0)}{i!} t^i \right| \leq M$ for all $i \neq i_1, i_2$, then

$$\gamma_n(t) = \frac{\nabla^{i_1} \nu^{(n)}(0)}{i_1!} t^{i_1} + \frac{\nabla^{i_2} \nu^{(n)}(0)}{i_2!} t^{i_2}$$

can remain bounded for at most two values of t. Assume two values exist for which γ_n is bounded which we denote by $t_1^* = 0$ and $t_2^* > 0$. Without loss of generality we assume that $\gamma_n(t) \to \infty$ for $t > t_2^*$ and $\gamma_n(t) \to -\infty$ for $0 < t < t_2^*$ (equivalently we assume that $i_1 > i_2$). One can show (by differentiating γ_n) that the stationary points $\tau_i^{(n)}$ of γ_n satisfy

$$\tau_1^{(n)} = 0 \qquad \text{and} \qquad \tau_2^{(n)} \to \frac{i_2}{i_1} t_2^* < t_2^*.$$

Pick $t \in (t_2^*, 1)$, we immediately have that $|\nu^{(n)}(t)| \to \infty$ from

$$|\nu^{(n)}(t) - \gamma_n(t)| \leq (s-2)M + \sqrt{\alpha}.$$

Furthermore we choose c such that $t - c > t_2^*$ then since γ_n is eventually increasing on $[t_2^*, 1]$

for n sufficiently large we have

$$\nu^{(n)}(r) \geq \gamma_n(t-c) - (s-2)M - \sqrt{\alpha}$$

for all $r \in [t-c, 1]$. Hence $\nu^{(n)}(r) \to \infty$ uniformly on $[t-c, 1]$.

Now we pick $t \in (0, t_2^*)$ and find c such that $[t-c, t+c] \subset (0, t_2^*)$. The same argument implies

$$\nu^{(n)}(r) \leq \min\{\gamma_n(t-c), \gamma_n(t+c)\} + (s-2)M + \sqrt{\alpha}.$$

And therefore $\nu^{(n)}(r) \to -\infty$ uniformly on $[t-c, t+c]$.

Similar arguments hold if γ_n has one or zero bounded values.

An analogous argument can be employed if there are three unbounded terms in (4.5), in which case we consider γ_n of the form:

$$\gamma_n(t) = \frac{\nabla^{i_1}\nu^{(n)}(0)}{i_1!}t^{i_1} + \frac{\nabla^{i_2}\nu^{(n)}(0)}{i_2!}t^{i_2} + \frac{\nabla^{i_3}\nu^{(n)}(0)}{i_3!}t^{i_3}.$$

This sequence of functions, can be bounded at at most three values of the argument, t_1^*, t_2^*, t_3^* and we repeat the previous argument. Such a process can be continued iteratively until we have considered the case where γ_n has s unbounded terms. □

We proceed to the main result of this subsection.

Theorem 4.2.4. *Define $f_n, f_\infty : \Theta \to \mathbb{R}$, where $\Theta \subset (H^s)^k$ for $s \geq 1$ is given by* (4.4)*, by* (4.2) *and* (4.3)*, respectively. Let $\mu^{(n)}$ be the minimizer of f_n then, under Assumptions 3.1-3.3, for almost every sequence of observations there exists a constant $M < \infty$ such that $\|\mu^{(n)}\|_{H^s} \leq M$ for all n.*

Proof. As in the proof of Theorem 4.2.2 we let $\omega \in \Omega''$ where

$$\Omega'' = \left\{\omega \in \Omega' : \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 \to \text{Var}(\epsilon_1)\right\}$$
$$\cap \left(\cap_{c\in\mathbb{Q}} \left\{\omega \in \Omega' : P_n^{(\omega)}\left(B(c, \frac{\delta}{4})\right) \to P\left(B(x, \frac{\delta}{4})\right)\right\}\right)$$

where Ω' is defined in the proof of Theorem 4.2.2. We have $\mathbb{P}(\Omega'') = 1$. For the remainder of the proof we assume $\omega \in \Omega''$. Then there exists $N^{(\omega)} < \infty$ such that $f_n^{(\omega)}(\mu^{(n)}) \leq \alpha + 1$ for all $n \geq N^{(\omega)}$. Hence

$$\lambda\|\mu^{(n)}\|_1^2 \leq f_n^{(\omega)}(\mu^{(n)}) \leq \alpha + 1.$$

It remains to show the $\mathcal{H}_0$ bound. The structure of the proof is similar to [96, Lemma 2.1]. We will argue by contradiction. In particular we argue that if a cluster center is unbounded then in the limit the minimum is achieved over the remaining $k-1$ cluster centers.

Step 1: *The minimization is achieved over $k-1$ cluster centers.* Assume there exists j^* such that $\|\mu_{j^*}^{(n)}\|_0 \to \infty$, then by Lemma 4.2.3 $|\mu_{j^*}^{(n)}(t)| \to \infty$ for all but finitely many $t \in [0,1]$ and for each t with $|\mu_{j^*}^{(n)}(t)| \to \infty$ there exists c such that $|\mu_{j^*}^{(n)}(r)| \to \infty$ uniformly for $r \in [t-c, t+c]$.

Pick t such that $|\mu_{j^*}^{(n)}(t)| \to \infty$ and find c. Let $t_m \to t$, then there exists an $M > 0$ such that for all $m > M$ we have $|t_m - t| < c$. It follows that

$$\lim_{m\to\infty, n\to\infty} |\mu_{j^*}^{(n)}(t_m)| = \infty.$$

This easily implies

$$\lim_{n\to\infty,(t',y')\to(t,y)} \left|\mu_{j^*}^{(n)}(t') - y'\right|^2 = \infty$$

for any $y \in \mathbb{R}^d$. Therefore

$$\liminf_{n\to\infty,(t',y')\to(t,y)} \left(\bigwedge_{j=1}^{k} \left|\mu_j^{(n)}(t') - y'\right|^2 - \bigwedge_{j\neq j^*} \left|\mu_j^{(n)}(t') - y'\right|^2 \right) = 0.$$

Note that the above expression holds for P-almost every $(t, y) \in [0, 1] \times \mathbb{R}^d$. By Theorem 1.1 in [64] and the above we have

$$\liminf_{n\to\infty} \left(\int_{[0,1]\times\mathbb{R}^d} \bigwedge_{j=1}^{k} |\mu_j^{(n)}(t) - y|^2 - \bigwedge_{j\neq j^*} |\mu_j^{(n)}(t) - y|^2 \; P_n^{(\omega)}(\mathrm{d}t, \mathrm{d}y) \right) \geq 0.$$

Hence

$$\liminf_{n\to\infty} \left(f_n^{(\omega)}(\mu^{(n)}) - f_n^{(\omega)}((\mu_j^{(n)})_{j\neq j^*}) - \lambda \|\nabla^s \mu_{j^*}^{(n)}\|_{L^2}^2 \right) \geq 0$$

where we interpret $f_n^{(\omega)}((\mu_j^{(n)})_{j\neq j^*})$ accordingly. So,

$$\liminf_{n\to\infty} \left(f_n^{(\omega)}(\mu^{(n)}) - f_n^{(\omega)}((\mu_j^{(n)})_{j\neq j^*}) \right) \geq 0.$$

Step 2: *The contradiction.* If we can show that there exists $\epsilon > 0$ such that the following holds (i.e. we can do strictly better by fitting k centers than fitting $k-1$ centers) then we can conclude:

$$\liminf_{n\to\infty} \left(f_n^{(\omega)}(\mu^{(n)}) - f_n^{(\omega)}((\mu_j^{(n)})_{j\neq j^*}) \right) \geq -\epsilon.$$

Now,

$$f_n^{(\omega)}(\mu^{(n)}) \leq f_n^{(\omega)}(\hat{\mu}^{(n)}) = \frac{1}{n} \sum_{i=1}^{n} \bigwedge_{j=1}^{k} |\hat{\mu}_j^{(n)}(t_i) - y_i|^2 + \lambda \sum_{j\neq j^*} \|\nabla^s \hat{\mu}_j^{(n)}\|_{L^2}^2,$$

where

$$\hat{\mu}_j^{(n)}(t) = \begin{cases} \mu_j^{(n)}(t) & \text{for } j \neq j^* \\ c_n & \text{for } j = j^* \end{cases}$$

for a constant c_n. Now each $\hat{\mu}_j^{(n)}$ must have a minimum separation distance of δ. For now we assume that we can choose c_n such that this criterion is fulfilled. So if $|y_i - c_n| \leq \frac{\delta}{4}$ then

$$|y_i - c_n| + \frac{\delta}{4} \leq |\mu_j^{(n)}(t_i) - y_i|$$

for all $j \neq j^*$. And therefore $|y_i - c_n|^2 + \frac{\delta^2}{16} \leq |\mu_j^{(n)}(t_i) - y_i|^2$ which implies

$$\begin{aligned}
f_n^{(\omega)}((\mu_j^{(n)})_{j\neq j^*}) &= \frac{1}{n}\sum_{i=1}^n \bigwedge_{j\neq j^*} |\mu_j^{(n)}(t_i) - y_i|^2 + \lambda \sum_{j\neq j^*} \|\nabla^s \mu_j^{(n)}\|_{L^2}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \bigwedge_{j\neq j^*} |\mu_j^{(n)}(t_i) - y_i|^2 \mathbb{I}_{(t_i,y_i)\nsim j^*} \\
&\quad\quad + \frac{1}{n}\sum_{i=1}^n \bigwedge_{j\neq j^*} |\mu_j^{(n)}(t_i) - y_i|^2 \mathbb{I}_{(t_i,y_i)\sim_n j^*} + \lambda \sum_{j\neq j^*} \|\nabla^s \mu_j^{(n)}\|_{L^2}^2 \\
&\geq \frac{1}{n}\sum_{i=1}^n \bigwedge_{j\neq j^*} |\mu_j^{(n)}(t_i) - y_i|^2 \mathbb{I}_{(t_i,y_i)\nsim j^*} + \frac{1}{n}\sum_{i=1}^n |c_n - y_i|^2 \mathbb{I}_{(t_i,y_i)\sim_n j^*} \\
&\quad\quad + \frac{\delta^2}{16} P_n^{(\omega)}\left(B\left(c_n, \frac{\delta}{4}\right)\right) + \lambda \sum_{j\neq j^*} \|\nabla^s \mu_j^{(n)}\|_{L^2}^2 \\
&= f_n^{(\omega)}(\hat{\mu}^{(n)}) + \frac{\delta^2}{16} P_n^{(\omega)}\left(B\left(c_n, \frac{\delta}{4}\right)\right).
\end{aligned}$$

Where $(t_i, y_i) \sim_n j$ means coordinate (t_i, y_i) is associated to center $\hat{\mu}_j^{(n)}$ in the sense that $(t, y) \sim_n j \Leftrightarrow j = \mathrm{argmin}_{i=1,\dots,k} |y - \hat{\mu}_i^{(n)}(t)|$ (and if the minimum is not uniquely achieved then we take the smallest j such that $j \in \mathrm{argmin}_{i=1,\dots,k} |y - \hat{\mu}_i^{(n)}(t)|$). If we can show that $P_n^{(\omega)}\left(B\left(c_n, \frac{\delta}{4}\right)\right)$ is bounded away from zero, then the result follows.

Since we assumed ϵ_1 has unbounded support on $\mathbb{R}$ if we can show that $|c_n| \leq M$ for a constant M and n sufficiently large (a.s.) then

$$\begin{aligned}
P_n^{(\omega)}\left(B\left(c_n, \frac{\delta}{4}\right)\right) &\geq \inf_{c\in[-M,M]\cap\mathbb{Q}} P_n^{(\omega)}\left(B(c, \frac{\delta}{4}\right) \\
&\to \inf_{c\in[-M,M]} P\left(B(c, \frac{\delta}{4}\right) \\
&= \inf_{c\in[-M,M]} \int_0^1 \int_{\mathbb{R}^d} \mathbb{I}_{|y-c|\leq\frac{\delta}{4}} \phi_Y(y|t)\phi_T(t)\, \mathrm{d}y\mathrm{d}t.
\end{aligned}$$

By Assumption 3.3, there exists $\epsilon' > 0$ such that $\phi_Y(y|t) \geq \epsilon'$ for all $y \in [-M, M]$ and $t \in [0, 1]$. Hence we may bound the final expression above by

$$\inf_{c\in[-M,M]} \int_0^1 \int_{\mathbb{R}^d} \mathbb{I}_{|y-c|\leq\frac{\delta}{4}} \phi_Y(y|t)\phi_T(t)\, \mathrm{d}y\mathrm{d}t \geq \frac{\delta}{4}\epsilon'.$$

We are left to show such an M exists. If there exists M_{k-1} such that for all $j \neq j^*$ we have $\|\mu_j^{(n)}\|_{H^s} \leq M_{k-1}$ then we know each center has size of range at most $2M_{k-1}$. So $k-1$ centers have size of range at most $2M_{k-1}(k-1) + \delta(k-2) =: C$. Hence there exists $c_n \in [0, C+\delta]^d$ such that $\hat{\mu}_{j^*}^{(n)}(t) = c_n$ and $\hat{\mu}^{(n)} \in \Theta$.

Now if no such M_{k-1} exists then there exists a second cluster such that $\|\mu_{j^{**}}^{(n)}\|_{H^s} \to \infty$ where $j^{**} \neq j^*$. By the same argument

$$\liminf_{n\to\infty} \left(f_n^{(\omega)}(\mu^{(n)}) - f_n^{(\omega)}((\mu_j^{(n)})_{j\neq j^*, j^{**}}) \right) \geq 0$$

and

$$f_n^{(\omega)}(\mu^{(n)}) - f_n^{(\omega)}((\mu_j^{(n)})_{j\neq j^*,j^{**}}) \leq -\frac{\delta^2}{16}P_n^{(\omega)}\left(B\left(c_n,\frac{\delta}{4}\right)\right) - \frac{\delta^2}{16}P_n^{(\omega)}\left(B\left(c_n',\frac{\delta}{4}\right)\right)$$

for a constant c_n'. By induction it is clear that we can find M_{k-l} such that $k-l$ cluster centers are bounded. The result then follows □

Remark 4.2.5. *Note that in the above theorem we did not need to assume a* correct *choice of k. If the true number of cluster centers is $k^\dagger$ and we incorrectly use $k \neq k^\dagger$, then the resulting cluster centers are still bounded. In fact for all the results of this chapter the correct choice of k is not necessary: although the minimizers of f_∞ may no longer make physical sense, the problem is still robust in that the conclusions of Theorems 4.2.1 and 4.3.1 hold.*

4.3 Rate of Convergence

In [126] a Central Limit Theorem was shown to hold for the k-means method in Euclidean spaces. We extend that argument to show that for converging sequences of minimizers, which exist by Theorem 4.2.1, the L^2 rate of convergence is of order $\frac{1}{n}$. We state the main result of this section now but leave the proof to the end.

Theorem 4.3.1. *Define $f_n, f_\infty : \Theta \to \mathbb{R}$, where Θ is given by* (4.4)*, by* (4.2) *and* (4.3)*, respectively. Let $\{\mu^{(n)}\}_{n\in\mathbb{N}} \subset \Theta$ where $\mu^{(n)}$ minimizes f_n. Let $\mu^{(n_m)}$ be any subsequence that converges to some $\mu^{(\infty)}$ then under Assumptions 3.1-3.4 we have:*

$$\|\mu^{(n_m)} - \mu^{(\infty)}\|_{L^2}^2 = O_p\left(\frac{1}{n_m}\right).$$

For clarity we will assume that the entire sequence $\mu^{(n)}$ converges in the remainder of this chapter and hence we may avoid writing subsequences.

We let $Y_n(\mu) = \sqrt{n}(f_n(\mu) - f_\infty(\mu))$ and then, by Taylor expanding around $\mu^{(\infty)}$, we have

$$Y_n(\mu^{(n)}) = Y_n(\mu^{(\infty)}) + \partial Y_n(\mu^{(\infty)}; \mu^{(n)} - \mu^{(\infty)}) + \text{h.o.t.}$$

In Lemma 4.3.5 using Chebyshev's inequality we bound the Gâteaux derivative of Y_n in probability. Similarly one can Taylor expand f_∞ around $\mu^{(\infty)}$. After some manipulation of the Taylor expansion one has

$$\frac{1}{2}\partial^2 f_\infty(\mu^{(\infty)}; \mu^{(n)} - \mu^{(\infty)}) = f_n(\mu^{(n)}) - f_n(\mu^{(\infty)}) + O_p\left(\frac{1}{\sqrt{n}}\|\mu^{(n)} - \mu^{(\infty)}\|_{L^2}\right)$$

where we leave the details until the proof of Theorem 4.3.1 at the end of the section. We note that $f_n(\mu^{(n)}) - f_n(\mu^{(\infty)}) \leq 0$. Therefore

$$\frac{\kappa}{2}\|\mu^{(n)} - \mu^{(\infty)}\|_{L^2}^2 \leq O_p(\frac{1}{\sqrt{n}}\|\mu^{(n)} - \mu^{(\infty)}\|_{L^2})$$

where $\kappa > 0$ exists as a consequence of f_∞ being positive definite (Lemma 4.3.4).

Lemmata 4.3.2 and 4.3.4 provide the first and second Gâteaux derivatives of f_∞. It is also shown that the second derivative is positive definite.

Lemma 4.3.2. *Define f_∞ by (4.3) and $\Theta \subset (H^s)^k$ for $s \geq 1$ by (4.4). Then for $\mu \in \Theta$, $\nu \in (H^s)^k$ we have that f_∞ is Gâteaux differentiable at μ in the direction ν with*

$$\begin{aligned}\partial f_\infty(\mu;\nu) = &-2\int_0^1\int_{\mathbb{R}} \left(y-\mu_{j(t,y)}(t)\right)\cdot \nu_{j(t,y)}(t)\phi_Y(y|t)\phi_T(t)\; dydt \\ &+ 2\lambda\sum_{j=1}^k (\nabla^s\nu_j, \nabla^s\mu_j)\end{aligned}$$

where $j(t,y)$ is chosen arbitrarily from the set

$$j(t,y) \in \operatorname*{argmin}_j |y-\mu_j(t)|. \tag{4.6}$$

Remark 4.3.3. *Since μ_j are continuous the boundary between each element of the resulting partition is itself continuous and has Lebesgue measure zero. The set on which $j(t,y)$ is not uniquely defined therefore has measure zero. Hence we will treat $j(t,y)$ as though it was uniquely defined.*

Proof of Lemma 4.3.2: Fix $\mu \in \Theta$, $\nu \in (H^s)^k$ and $r > 0$. Define

$$j_r(t,y) = \operatorname*{argmin}_j |y-\mu_j(t) - r\nu_j(t)|, \qquad j(t,y) = j_0(t,y).$$

Then for (t,y) in the interior of the partition associated with μ_j we have

$$j_r(t,y) = j(t,y) \qquad \text{for } r \text{ sufficiently small.}$$

More precisely at each t we have $j_r(t,y) = j(t,y)$ for all y with $\operatorname{dist}(y, B(t)) > r\max_i \|\nu_i\|_{L^\infty}$ where $B(t)$ is the set of boundary points:

$$B(t) = \left\{ y\in\mathbb{R}^d \;:\; j(t,y) \text{ is not uniquely defined}\right\}.$$

Let $Y(r,t) = \{y \;:\; \operatorname{dist}(y,B(t)) \leq r\max_i\|\nu_i\|_{L^\infty}\}$ then

$$\begin{aligned}&\int_{\mathbb{R}^d} \left|y-\mu_{j_r(t,y)}(t)\right|^2 - \left|y-\mu_{j(t,y)}(t)\right|^2\; \phi_Y(y|t)\,\mathrm{d}y \\ &\qquad = \int_{Y(r,t)} \left(2y-\mu_{j(t,y)}(t)-\mu_{j_r(t,y)}(t)\right)\cdot\left(\mu_{j(t,y)}(t)-\mu_{j_r(t,y)}(t)\right)\;\phi_Y(y|t)\,\mathrm{d}y \\ &\qquad \leq 2r\max_i\|\nu_i\|_{L^\infty}\int_{Y(r,t)}\left|\mu_{j(t,y)}(t)-\mu_{j_r(t,y)}(t)\right|\;\phi_Y(y|t)\,\mathrm{d}y \\ &\qquad = O(r^2)\end{aligned}$$

where the penultimate line follows from: if $j(t,y) \neq j_r(t,y)$ then $\frac{\mu_{j(t,y)}(t)+\mu_{j_r(t,y)}(t)}{2} \in B(t)$

(for r sufficiently small). Therefore

$$
\begin{aligned}
\partial f_\infty(\mu;\nu) &= \lim_{r\to 0} \frac{f_\infty(\mu + r\nu) - f_\infty(\mu)}{r} \\
&= \lim_{r\to 0} \frac{1}{r}\Bigg\{ \int_0^1 \int_{\mathbb{R}^d} \Bigg(|y - \mu_{j_r(t,y)}(t)|^2 - |y - \mu_{j(t,y)}(t)|^2 + r^2|\nu_{j_r(t,y)}(t)|^2 \\
&\qquad - 2r\left(y - \mu_{j_r(t,y)}(t)\right)\cdot \nu_{j_r(t,y)}(t)\Bigg)\phi_Y(y|t)\phi_T(t)\,\mathrm{d}y\mathrm{d}t \\
&\qquad + \lambda \sum_{j=1}^k \left(2r(\nabla^s \nu_j, \nabla^s \mu_j) + r^2\|\nabla^s\nu_j\|^2_{L^2}\right)\Bigg\} \\
&= -2\int_0^1 \int_{\mathbb{R}^d} \left(y - \mu_{j(t,y)}(t)\right)\cdot \nu_{j(t,y)}(t)\phi_Y(y|t)\phi_T(t)\,\mathrm{d}y\mathrm{d}t \\
&\qquad + 2\lambda\sum_{j=1}^k (\nabla^s\nu_j, \nabla^s\mu_j)
\end{aligned}
$$

which proves the result. □

Lemma 4.3.4. *Under the conditions of Lemma 4.3.2 and additionally Assumption 3.4, we have that f_∞ has a second Gâteaux derivative at $\mu \in \Theta$ with respect to $\nu, \eta \in (H^s([0,1]))^k$ given by*

$$
\partial^2 f_\infty(\mu;\nu,\eta) = 2\int_0^1 \int_{\mathbb{R}} \eta_{j(t,y)}(t)\cdot\nu_{j(t,y)}(t)\phi_Y(y|t)\phi_T(t)\,dydt + 2\lambda\sum_{j=1}^k (\nabla^s\nu_j, \nabla^s\eta_j)
$$

where $j(t,y)$ is defined by (4.6). *Furthermore, $\partial^2 f_\infty$ is positive definite at $\mu^{(\infty)}$, the minimizer of f_∞ in Θ. I.e. there exists $\kappa > 0$ such that for all $\nu \in (H^s([0,1]))^k$ we have*

$$
\partial^2 f_\infty(\mu^{(\infty)};\nu) \geq \kappa\|\nu\|^2_{L^2}.
$$

Proof. The proof of the first statement follows the same reasoning as that of Lemma 4.3.2. To establish the second, observe that:

$$
\begin{aligned}
\partial^2 f_\infty(\mu^{(\infty)};\nu) &= 2\sum_{j=1}^k \int_0^1 \int_{\mathbb{R}} |\nu_j(t)|^2 \mathbb{I}_{\{(t,y)\sim j\}}(t,y)\phi_Y(y|t)\phi_T(t)\,\mathrm{d}y\mathrm{d}t + 2\lambda\sum_{j=1}^k \|\nabla^s\nu_j\|^2_{L^2} \\
&\geq \hat{\kappa}\sum_{j=1}^k \int_0^1 |\nu_j(t)|^2\ \phi_T(t)\,\mathrm{d}t \\
&\geq \kappa\sum_{j=1}^k \|\nu_j\|^2_{L^2}
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{\kappa} &= 2\inf_{t\in[0,1]}\min_{j=1,2,\ldots,k}\int_{\mathbb{R}} \mathbb{I}_{\{(t,y)\sim j\}}(t,y)\phi_Y(y|t)\,\mathrm{d}y \\
\kappa &= \hat{\kappa}\inf_{t\in[0,1]}\phi_T(t)
\end{aligned}
$$

where we recall $(t, y) \sim j$ means coordinate (t, y) is associated with center $\mu_j^{(\infty)}$. It remains to show $\hat{\kappa} > 0$. Recall that there exists a uniform bound on $\mu^{(\infty)}$ so that $|\mu^{(\infty)}(t)| \leq M$ for all $t \in [0, 1]$. Define ϵ' by

$$\epsilon' = \inf_{t\in[0,1]} \min_{|y|\leq M} \phi_Y(y|t).$$

By Assumptions 3.1,3.3 and 3.4, we have $\epsilon' > 0$. So

$$\hat{\kappa} \geq 2 \inf_{t\in[0,1]} \min_{j=1,2,\ldots,k} \mathrm{Vol}(A_t^j)\epsilon' M$$

where A_t^j is the partition of $\mathbb{R}^d$ corresponding to center $\mu_j^{(\infty)}$ at time t. The minimum separation distance assumptions on Θ imply that $\mathrm{Vol}(A_t^j) \geq \mathrm{Vol}(B(0, \frac{\delta}{2}))$. Which implies that $\hat{\kappa} > 0$. The result then follows. □

We now consider Y_n. In particular we want to bound $\partial Y_n(\mu^{(\infty)}; \mu^{(n)} - \mu^{(\infty)})$.

Lemma 4.3.5. *Define $f_n, f_\infty : \Theta \to \mathbb{R}$ by* (4.2) *and* (4.3) *respectively where Θ is given by* (4.4). *Take Assumption 3.1 and define*

$$Y_n : \Theta \to \mathbb{R}, \qquad Y_n(\mu) = \sqrt{n}\left(f_n(\mu) - f_\infty(\mu)\right).$$

Then for $\mu \in \Theta$, $\nu \in (H^s)^k$ we have that Y_n is Gâteaux differentiable at μ in the direction ν with

$$\partial Y_n(\mu; \nu) = 2\sqrt{n}\left(\int_0^1 \int_{\mathbb{R}} \left(y - \mu_{j(t,y)}(t)\right) \cdot \nu_{j(t,y)}(t)\phi_Y(y|t)\phi_T(t)\, dydt \right.$$
$$\left. -\frac{1}{n}\sum_{i=1}^n \left(y_i - \mu_{j(t_i,y_i)}(t_i)\right) \cdot \nu_{j(t_i,y_i)}(t_i)\right)$$

where $j(t, y)$ is defined by (4.6). *Furthermore, for a sequence $\nu^{(n)}$ with*

$$\|\nu^{(n)}\|_{L^2} = o_p(1) \qquad \text{and} \qquad \|\nu^{(n)}\|_{H^s} = O_p(1)$$

we have $\partial Y_n(\mu; \nu^{(n)}) = O_p(\|\nu^{(n)}\|_{L^2})$.

Proof. Calculating the Gâteaux derivative is similar to Lemma 4.3.2 and is omitted. By linearity and continuity of ∂Y_n we can write

$$\partial Y_n\left(\mu; \frac{\nu^{(n)}}{\|\nu^{(n)}\|_{L^2}}\right) = \sum_m \frac{(\nu^{(n)}, e_m)}{\|\nu^{(n)}\|_{L^2}} \partial Y_n(\mu; e_m)$$

where e_m is the Fourier basis for $(L^2)^k$ (we assume $e_m = (\hat{e}_{m_1}, \ldots, \hat{e}_{m_k})$ where $\hat{e}_m$ is a Fourier basis for L^2). Let $V_m = \mathbb{E}\left(\partial Y_n(\mu; e_m)\right)^2$ and $Z_i = (y_i - \mu_{j(t_i,y_i)}(t_i)) \cdot \hat{e}_m$, then

$$\begin{aligned} V_m &= \frac{4}{n}\mathbb{E}\left(\sum_{i=1}^{n}(Z_i - \mathbb{E}Z_i)\right)^2 \\ &= 4\mathbb{E}\left(Z_1 - \mathbb{E}Z_1\right)^2 \\ &\leq 4\mathbb{E}\left|\epsilon_1\right|^2 =: C. \end{aligned}$$

Where C is finite by Assumption 3.2. Therefore, by Chebyshev's inequality for any $M > 0$

$$\mathbb{P}\left(\left|\partial Y_n(\mu; e_m)\right| \geq M\right) \leq \frac{C}{M^2}.$$

Hence

$$\begin{aligned} \mathbb{P}\left(\left|\partial Y_n\left(\mu; \frac{\nu^{(n)}}{\|\nu^{(n)}\|_{L^2}}\right)\right| \geq M\right) &\leq \sum_m \frac{|(\nu^{(n)}, e_m)|}{\|\nu^{(n)}\|_{L^2}}\mathbb{P}\left(\left|\partial Y_n(\mu; e_m)\right| \geq M\right) \\ &\leq \sum_m \frac{|(\nu^{(n)}, e_m)|}{\|\nu^{(n)}\|_{L^2}}\frac{C}{M^2} \\ &\leq \frac{C}{M^2}. \end{aligned}$$

Which implies $\partial Y_n\left(\mu; \frac{\nu^{(n)}}{\|\nu^{(n)}\|_{L^2}}\right) = O_p(1)$. □

We now have the necessary pieces in place to prove Theorem 4.3.1

Proof of Theorem 4.3.1. By Theorem 4.2.1 we have (up to subsequences) $\|\mu^{(n)} - \mu^{(\infty)}\|_{L^2} = o_p(1)$ and $\|\mu^{(n)}\|_{H^s} = O_p(1)$.

By Taylor's Theorem and the fact that f_∞ is quadratic we have

$$f_\infty(\mu^{(n)}) = f_\infty(\mu^{(\infty)}) + \partial f_\infty\left(\mu^{(\infty)}; \mu^{(n)} - \mu^{(\infty)}\right) + \frac{1}{2}\partial^2 f_\infty\left(\mu^{(\infty)}; \mu^{(n)} - \mu^{(\infty)}\right).$$

Since $\mu^{(\infty)}$ minimizes f_∞ the linear term above must be zero. Hence

$$f_\infty(\mu^{(n)}) = f_\infty(\mu^{(\infty)}) + \frac{1}{2}\partial^2 f_\infty\left(\mu^{(\infty)}; \mu^{(n)} - \mu^{(\infty)}\right).$$

Similarly, and using Lemma 4.3.5

$$\begin{aligned} Y_n(\mu^{(n)}) &= Y_n(\mu^{(\infty)}) + O_p\left(\partial Y_n\left(\mu^{(\infty)}; \mu^{(n)} - \mu^{(\infty)}\right)\right) \\ &= Y_n(\mu^{(\infty)}) + O_p\left(\|\mu^{(n)} - \mu^{(\infty)}\|_{L^2}\right). \end{aligned}$$

From the definition of Y_n we also have

$$f_n(\mu^{(n)}) = f_\infty(\mu^{(n)}) + \frac{1}{\sqrt{n}}Y_n(\mu^{(n)}).$$

Substituting into the above we obtain

$$\begin{aligned}
f_n(\mu^{(n)}) &= f_\infty(\mu^{(\infty)}) + \frac{1}{2}\partial^2 f_\infty(\mu^{(\infty)}; \mu^{(n)} - \mu^{(\infty)}) + \frac{1}{\sqrt{n}} Y_n(\mu^{(\infty)}) \\
&\quad + O_p\left(\frac{1}{\sqrt{n}}\|\mu^{(n)} - \mu^{(\infty)}\|_{L^2}\right) \\
&= f_n(\mu^{(\infty)}) + \frac{1}{2}\partial^2 f_\infty(\mu^{(\infty)}; \mu^{(n)} - \mu^{(\infty)}) + O_p\left(\frac{1}{\sqrt{n}}\|\mu^{(n)} - \mu^{(\infty)}\|_{L^2}\right).
\end{aligned}$$

Rearranging for $\partial^2 f_\infty$ and using $f_n(\mu^{(n)}) \leq f_n(\mu^{(\infty)})$ we have

$$\begin{aligned}
\partial^2 f_\infty(\mu^{(\infty)}; \mu^{(n)} - \mu^{(\infty)}) &= 2\left(f_n(\mu^{(n)}) - f_n(\mu^{(\infty)})\right) + O_p(\frac{1}{\sqrt{n}}\|\mu^{(n)} - \mu^{(\infty)}\|_{L^2}) \\
&\leq O_p(\frac{1}{\sqrt{n}}\|\mu^{(n)} - \mu^{(\infty)}\|_{L^2}).
\end{aligned}$$

Recall that by Lemma 4.3.4 we have that $\partial^2 f_\infty$ is positive definite at $\mu^{(\infty)}$. Therefore:

$$\kappa\|\mu^{(n)} - \mu^{(\infty)}\|_{L^2}^2 \leq O_p(\frac{1}{\sqrt{n}}\|\mu^{(n)} - \mu^{(\infty)}\|_{L^2}).$$

Which implies

$$\|\mu^{(n)} - \mu^{(\infty)}\|_{L^2} = O_p(\frac{1}{\sqrt{n}})$$

which completes the proof. □

Chapter 5

Weak Convergence For Generalized Spline Smoothing

Abstract

Establishing the convergence of splines can be cast as a variational problem which is amenable to a Γ-convergence approach. We consider the case in which the regularization coefficient scales with the number of observations, n, as $\lambda_n = n^{-p}$. Using standard theorems from the Γ-convergence literature, we prove that general splines are consistent in the sense that estimators converge weakly in probability if $p \leq \frac{1}{2}$. Without further assumptions this rate is sharp. This differs from rates for strong convergence using Hilbert scales where one can often choose $p > \frac{1}{2}$.

5.1 Introduction

Given a Hilbert space, $\mathcal{H}$, with dual $\mathcal{H}^*$, the general spline problem [90, 171] is to recover $\mu^\dagger \in \mathcal{H}$ from observations, $\{(L_i, y_i)\}_{i=1}^n \subseteq \mathcal{H}^* \times \mathbb{R}$, and the model

$$y_i = L_i \mu^\dagger + \epsilon_i, \tag{5.1}$$

where $\epsilon_i \in \mathbb{R}$ and $L_i \in \mathcal{H}^*$ are independent random variables. We assume that $\mathcal{H}$ can be decomposed into $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $(\mathcal{H}_l, \|\cdot\|_l)$ (for $l = 0, 1$) are themselves both Hilbert spaces. For example, one may apply the theory to the special spline problem where $\mathcal{H} = H^m([0,1])$ ($m \geq 1$) is the Sobolev space of degree m and the observation operators are of the form $L_i\mu = \mu(t_i)$ in which t_i is sampled from some distribution over $[0,1]$. Throughout this chapter we refer to (5.1) as the general spline model when $L_i \in \mathcal{H}^*$ and $\mathcal{H}$ is any Hilbert space, and the special spline model when L_i is the pointwise evaluation operator and $\mathcal{H} = H^m$.

We assume that dim$(\mathcal{H}_0) = m < \infty$ and dim$(\mathcal{H}_1) = \infty$. This can be seen as a multi-scale decomposition of $\mathcal{H}$. The projection of a function $\mu \in \mathcal{H}$ into the subspace $\mathcal{H}_0$ is a coarse approximation of that function. Continuing with the special spline example, one can write

$$\mu(t) = \sum_{i=0}^{m-1} \frac{\nabla^i \mu(0)}{i!} t^i + \int_0^t \frac{(t-u)^{m-1}}{(m-1)!} \nabla^m \mu(u) \, \mathrm{d}u$$

for any $\mu \in H^m$. The space $\mathcal{H}_0$ is then the space of polynomials of degree at most $m-1$. Hence $\dim(\mathcal{H}_0) = m$.

Imposing a penalty on the $\mathcal{H}_1$ space, we construct a sequence of estimators $\mu^{(n)}$ of $\mu^\dagger$ as the minimizers of

$$f_n(\mu) = \frac{1}{n} \sum_{i=1}^n |y_i - L_i \mu|^2 + \lambda_n \|\chi_1 \mu\|_1^2$$

where $\chi_i : \mathcal{H} \to \mathcal{H}_i$ $(i = 0, 1)$ is the projection of $\mathcal{H}$ onto $\mathcal{H}_i$. This chapter addresses the asymptotics (as $n \to \infty$) of the general spline problem and in particular how one should choose λ_n to ensure $\mu^{(n)}$ converges (weakly in probability) to $\mu^\dagger$.

There are two bodies of literature on the specification of λ_n. On the one hand there are methods which define λ_n as the minimizer of some loss function, for example average square error. This class of techniques includes cross-validation [172], generalized cross-validation [47] and penalized likelihood techniques [82, 85, 91, 109, 136, 170]. These methods provide a numerical value of λ_n for a given n and a given set of data. In the case of special splines there are many results on the asymptotic behavior of λ_n and $\mu^{(n)}$ for these methods, see for example [3, 45, 47, 98, 144, 156, 157, 173]. The alternative approach, and the one we take in this chapter, is to choose a sequence such that the estimates $\mu^{(n)}$ converge to $\mu^\dagger$ in an appropriate sense at the fastest possible rate. This strategy gives a scaling regime for λ_n, but it does not in general give specific numerical values of λ_n, i.e. it provides the optimal rate of convergence but not the associated multiplicative constant.

In considering the convergence of the sequence, $\mu^{(n)}$, one can look for convergence with respect to some norm or in the dual space. Many results in the literature demonstrate the strong convergence of $\mu^{(n)} \to \mu^\dagger$ via the use of Hilbert scales — see, for example, [46, 117, 131, 143, 146, 158]. It is not typically possible to obtain strong convergence with respect to the original norm and it is common to resort to the use of weaker norms; for example, in the special spline problem, one starts with the space H^s but looks for convergence in L^2. The alternative, which is pursued in this chapter, is to consider weak convergence in the original space, $\mathcal{H}$.

Note that for special splines strong convergence in a larger space is a weaker result than weak convergence in the original space: by the Sobolev embedding theorem, weak convergence in H^s implies strong convergence in L^2; however, the converse does not hold.

In this chapter we show that the estimators of the general spline problem weakly converge in probability in the large data limit, $\mu^{(n)} \rightharpoonup \mu^\dagger$, for regularization λ_n that scales to zero no faster than $n^{-\frac{1}{2}}$. In this scaling regime we say that the general spline problem is consistent. For insufficient regularization the spline estimators may in some sense 'blow up'. In particular for scaling outside this regime we construct (uniformly bounded) observation operators L_i such that $\mathbb{E}\left[\|\mu^{(n)}\|^2\right] \to \infty$. Hence without further assumptions our results are sharp.

If we are interested in estimating $\mu^\dagger$ at a point t then we let $F(\mu) = \mu(t)$ where $F \in \mathcal{H}^*$. In this setting weak convergence is the natural form to consider. However, if one is interested in a global approximation of $\mu^\dagger$, then convergence of $\mu^{(n)} - \mu^\dagger$ in an appropriate norm is the appropriate concept. The two formulations imply different scaling results for λ_n.

There are many results in the ill-posed inverse problems literature that may be applied to the strong convergence of the general spline problem, for brevity we only mention those most relevant to this work. In [170] two different methods of estimating λ_n were compared as $n \to \infty$ using the general spline formulation. The reproducing kernel Hilbert space setting was used in [89] which also discussed the probabilistic interpretation behind the estimator $\mu^{(n)}$. In [46, 117] the authors prove the strong convergence and optimal rates for the spline model using an approximation $\frac{1}{n}\sum_{i=1}^n L_i^* L_i \approx U$ where U is compact, positive definite, self-adjoint and with dense inverse. See also [35, 108] that consider ill-posed inverse problems without noise using similar methods. In these papers the scaling regime for λ_n is given in terms of the rate of decay of the eigenvalues of the inverse covariance (regularization) operator $\mathcal{C}^{-1}$ (where $\|\cdot\|_1 = \|\mathcal{C}^{-1}\cdot\|_{L^2}$).

There are many more recent results addressing the asymptotic properties of splines, including [43, 79, 88, 94, 100, 138, 174, 180, 182, 183]. Many of these recent results concern the asymptotics of penalized splines where one fixes the number of knot points as apposed to the smoothing spline case where the number of knots is equal to the number of data points. As far as we are aware there are no asymptotic results concerning the weak convergence of splines (either general or special).

It is known that the special spline problem is equivalent to a white noise problem [32]. Strong convergence and rates for the white noise problem have been well studied see, for example, [4, 23, 74] and references therein.

One advantage of our approach is that we gain intuition in what happens when $\lambda_n \to 0$ too quickly. Our results show a critical rate, with respect to the scaling of λ_n, at which the methodology is ill-posed below this rate and well-posed at or above this rate. The second advantage of our approach is that, by using the Γ-convergence framework, as long as we can show that minimizers are uniformly bounded the convergence follows easily (we also need to show the Γ-limit is unique, but for our problem this is not difficult). This is easier than showing, directly, that $\mu^{(n)} - \mu^\dagger$ converges to zero. We are consequently able to employ simpler assumptions than those required by more direct arguments.

The outline of this chapter is as follows. In the next section we remind the reader of the spline methodology and prove a known existence and uniqueness result. Section 5.3 contains the results for the convergence of the general spline model under appropriate conditions on the scaling in the regularization using the Γ-convergence framework. We discuss the special spline model in Section 5.4.

5.2 The Spline Framework

In this section we recap the spline methodology and find an explicit representation for our estimators. In particular we construct our estimate as a minimizer of a quadratic functional. We will show the existence and uniqueness of the minimizer.

We consider the separable Hilbert space $\mathcal{H}$ with inner product and norm given by $(\cdot,\cdot)$

and $\|\cdot\|$ respectively. We assume we can write $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $(\mathcal{H}_0, (\cdot,\cdot)_0, \|\cdot\|_0)$, $(\mathcal{H}_1, (\cdot,\cdot)_1, \|\cdot\|_1)$ are Hilbert spaces with $\dim(\mathcal{H}_0) = m$ and $\dim(\mathcal{H}_1) = \infty$. We may write

$$\|\mu\| = \|\mu\|_0 + \|\mu\|_1.$$

We wish to estimate $\mu^\dagger \in \mathcal{H}$ given observations of the form (L_i, y_i) and in particular L_i is random. For convenience we summarize the general spline model in the definition below. One can also see [171] for more details on the general spline model.

The General Spline Model. *The general spline model is given by* (5.1) *where $L_i \in \mathcal{H}^*$ are random variables and ϵ_i are iid random variables from a centered distribution ϕ_0 with variance σ^2. The L_i are assumed to be observed without noise and to be members of a family indexed by $\mathcal{I}$; we write L_t to mean the operator L which depends upon a parameter $t \in \mathcal{I}$. The 'randomness' of L is characterized by the distribution, ϕ_T, of a random index $t \in \mathcal{I}$. For a sample $t_i \sim \phi_T$ we write L_i as shorthand for L_{t_i}. The operator L_i is therefore interpreted as a realization of L_{t_i}. We assume that t_i, ϵ_i are independent and for convenience we define $\phi_{L_t\mu^\dagger}$ to be the distribution ϕ_0 shifted by $-L_t\mu^\dagger$. By the Riesz Representation Theorem there exists $\eta_i \in \mathcal{H}$ such that $L_i\mu = (\eta_i, \mu)$ for all $\mu \in \mathcal{H}$. The sequence of observed data points $(t_1, y_1), (t_2, y_2), \dots$ is a realization of a sequence of random elements on $(\Omega, \mathcal{F}, \mathbb{P})$. To mitigate the notational burden, we suppress the ω-dependence of t_i, y_i and L_i.*

For example in the case of special splines $L_i\mu^\dagger = \mu^\dagger(t_i)$ for some t_i a random variable distributed in $[0,1]$. Observing L_i without noise is equivalent here to observing t_i without noise. It will be convenient to introduce the natural filtration associated with the marginal sequence (L_i) and we define for $n \in \mathbb{N}$, $\mathcal{G}_n = \sigma(L_1, \dots, L_n)$, a sequence of sub-$\sigma$-algebras of $\mathcal{F}$. We use $\mathbb{E}[\cdot|\mathcal{G}_n]$ to denote a version of the associated conditional expectation.

We take our sequence of estimators $\mu^{(n)}$ of $\mu^\dagger$ as minimizers, which are subsequently shown to be unique, of $f_n^{(\omega)}$ where

$$f_n^{(\omega)}(\mu) = \frac{1}{n}\sum_{i=1}^n (y_i - L_i\mu)^2 + \lambda_n\|\mu\|_1^2. \tag{5.2}$$

By completing the square we can easily show $\mu^{(n)}$ is given implicitly by

$$G_{n,\lambda_n}\mu^{(n)} = \frac{1}{n}\sum_{i=1}^n y_i\eta_i$$

where

$$G_{n,\lambda} = \frac{1}{n}\sum_{i=1}^n \eta_i L_i + \lambda\chi_1 \tag{5.3}$$

and for clarity we also suppress the ω-dependence of $G_{n,\lambda}$ from the notation. It will be necessary in our proofs to bound $\|G_{n,\lambda_n}\|_{\mathcal{H}^*}$ in terms of λ_n (for almost every sequence of observations). We do this by imposing a bound on $\|L_t\|_{\mathcal{H}^*}$ or equivalently on $\|\eta_t\|$ for almost every $t \in \mathcal{I}$. See Section 5.4 for a discussion of the special spline problem and in particular how one can find η_i. In order to bound the $\mathcal{H}_0$ norm of $\mu^{(n)}$ we need conditions on our observation operators L_t.

In particular we will use the observation operators to define a norm on $\mathcal{H}_0$. Hence our proofs require a uniqueness assumption of L_t in $\mathcal{H}_0$ (Assumption 4.3 below). It is not enough that L_t are unique over $\mathcal{H}$ as this would not necessarily contain any information on the $\mathcal{H}_0$ projection of $\mu^{(n)}$, e.g. if $L_t\mu = L_t\chi_1\mu$ for all $\mu \in \mathcal{H}$. For clarity and future reference we now summarize the assumptions described in the previous paragraphs.

Assumptions 4. *We make the following assumptions on $f_n^{(\omega)} : \mathcal{H} \to \mathbb{R}$ defined by* (5.2) *and $\mathcal{H}$.*

4.1 Let $(\mathcal{H}, (\cdot,\cdot), \|\cdot\|)$ be a separable Hilbert spaces with $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $(\mathcal{H}_0, (\cdot,\cdot)_0, \|\cdot\|_0)$ and $(\mathcal{H}_1, (\cdot,\cdot)_1, \|\cdot\|_1)$ are Hilbert spaces. Assume dim$(\mathcal{H})$ = dim$(\mathcal{H}_1) = \infty$ and dim$(\mathcal{H}_0) = m < \infty$.

4.2 The distribution of $L_i := L_{t_i}$ is specified implicitly by that of $t_i \in \mathcal{I}$ and we assume $t_i \overset{\text{iid}}{\sim} \phi_T$.

4.3 We assume $|supp(\phi_T)| \geq m$ and that the L_t are unique in $\mathcal{H}_0$ in the sense that if $L_t\mu = L_r\mu$ for all $\mu \in \mathcal{H}_0$ then $t = r$.

4.4 There exists $\alpha > 0$ such that $\|\eta_t\| = \|L_t\|_{\mathcal{H}^*} \leq \alpha$ for ϕ_T-almost every $t \in \mathcal{I}$.

We have the following lemma which gives the existence of a unique minimizer to (5.2).

Lemma 5.2.1. *Define $f_n^{(\omega)} : \mathcal{H} \to \mathbb{R}$ by* (5.2) *and assume $\lambda_n > 0$. Under Assumptions 4.1-4.4 the operator $G_{n,\lambda_n} : \mathcal{H} \to \mathcal{H}$ defined by* (5.3) *has a well defined inverse G_{n,λ_n}^{-1} on span$\{\eta_1, \ldots, \eta_n\}$ for almost every $\omega \in \Omega$. In particular, there almost surely exists $N < \infty$ such that for all $n \geq N$ there exists a unique minimizer $\mu^{(n)} \in \mathcal{H}$ to $f_n^{(\omega)}$ which is given by*

$$\mu^{(n)} = \frac{1}{n}\sum_{i=1}^{n} y_i G_{n,\lambda_n}^{-1}\eta_i. \tag{5.4}$$

Proof. We claim that any minimizer of $f_n^{(\omega)}$ lies in the set $\mathcal{H}_0 \oplus \text{span}\{\chi_1\eta_1, \ldots, \chi_1\eta_n\} =: \mathcal{H}'_n$. If so, and we can show that G_{n,λ_n}^{-1} is well defined on $\mathcal{H}'_n$, then we can conclude the minimizer must be of the form (5.4).

Because $\mathcal{H}'_n$ is finite dimensional, we arrive at the same topology whichever norm we choose. We define $\Omega' \subset \Omega$ by

$$\Omega' := \{\omega \in \Omega : \text{the number of unique } t_j \text{ in } \{t_i\}_{i=1}^{\infty} \text{ is greater than } m \text{ and } \|L_i\|_{\mathcal{H}^*} \leq \alpha \,\forall i\}.$$

By Assumptions 4.3 and 4.4, $\mathbb{P}(\Omega') = 1$. Let $\omega \in \Omega'$ then there exists N such that for all $n \geq N$ we have that $\{L_i\}_{i=1}^N$ contains m distinct elements. Therefore $\|\mu\|_{\mathcal{H}'_n}^2 := \frac{1}{n}\sum_{i=1}^n (L_i\mu)^2 + \lambda_n\|\mu\|_1^2$ defines a norm on $\mathcal{H}'_n$ for any $n \geq N$.

We first show that any minimizer of $f_n^{(\omega)}$ lies in $\mathcal{H}'_n$. Let

$$\mu = \sum_{j=1}^{m} a_j\phi_j + \sum_{j=1}^{n} b_j\chi_1\eta_j + \rho$$

where ϕ_j are a basis for $\mathcal{H}_0$ and $\rho \perp \mathcal{H}'_n$. Then since $L_i\rho = (\eta_i, \rho) = 0$ we have:

$$f_n^{(\omega)}(\mu) = \frac{1}{n}\sum_{i=1}^n \left(y_i - L_i\chi_{\mathcal{H}'_n}\mu\right)^2 + \lambda_n \left\|\sum_{j=1}^n b_j\chi_1\eta_j\right\|_1^2 + \lambda_n\|\rho\|_1^2$$

where $\chi_{\mathcal{H}'_n}$ denotes the projection onto $\mathcal{H}'_n$. Trivially any minimizer of $f_n^{(\omega)}$ must have $\|\rho\|_1 = 0$ and since $\rho \in \mathcal{H}_1$ this implies $\rho = 0$. Hence minimizers of $f_n^{(\omega)}$ lie in $\mathcal{H}'_n$.

We now show that G_{n,λ_n} has a well defined inverse on $\mathcal{H}'_n$. For any $r \in \mathcal{H}'_n$ the weak formulation of $G_{n,\lambda_n}\mu^{(n)} = r$ is given by

$$B(\mu^{(n)}, \nu) = (r, \nu) \qquad \forall \nu \in \mathcal{H}'_n$$

where

$$B(\mu, \nu) = \frac{1}{n}\sum_{i=1}^n (L_i\mu)(L_i\nu) + \lambda_n(\chi_1\mu, \chi_1\nu).$$

Now we apply the Lax-Milgram lemma to imply the existence of a unique weak solution to $G_{n,\lambda_n}\mu^{(n)} = r$. Clearly $B : \mathcal{H}'_n \times \mathcal{H}'_n \to \mathbb{R}$ is a bilinear form. We will show it is also bounded and coercive. As $\omega \in \Omega'$, $\|L_i\|_{\mathcal{H}^*} \leq \alpha$ and for $\mu, \nu \in \mathcal{H}'_n$ we have

$$\begin{aligned}
|B(\mu, \nu)| &\leq \frac{1}{n}\sum_{i=1}^n |L_i\mu L_i\nu| + \lambda_n\|\mu\|_1\|\nu\|_1 \\
&\leq \alpha^2\|\mu\|\|\nu\| + \lambda_n\|\mu\|_1\|\nu\|_1 \\
&\leq \left(\alpha^2 + \lambda_n\right)\|\mu\|\|\nu\|.
\end{aligned}$$

Hence B is bounded. Similarly, for some constant c independent of μ,

$$B(\mu, \mu) = \frac{1}{n}\sum_{i=1}^n (L_i\mu)^2 + \lambda_n\|\mu\|_1^2 = \|\mu\|_{\mathcal{H}'_n}^2 \geq c\|\mu\|^2$$

where the inequality follows by the equivalence of norms on finite dimensional spaces. Hence B is coercive and by the Lax-Milgram Lemma there exists a unique weak solution. A strong solution follows from the equivalence of the strong and weak topology on finite dimensional spaces or alternatively from the following short calculation. We have

$$(r, \nu) = \left(\frac{1}{n}\sum_{i=1}^n (L_i\mu^{(n)})\eta_i, \nu\right) + \left(\lambda_n\chi_1\mu^{(n)}, \nu\right) \qquad \forall \nu \in \mathcal{H}'_n$$

Hence

$$\left(r - \frac{1}{n}\sum_{i=1}^n (L_i\mu^{(n)})\eta_i - \lambda_n\chi_1\mu^{(n)}, \nu\right) = 0 \qquad \forall \nu \in \mathcal{H}'_n.$$

So choosing $\nu = r - \frac{1}{n}\sum_{i=1}^n (L_i\mu^{(n)})\eta_i - \lambda_n\chi_1\mu^{(n)}$ implies

$$\|r - \frac{1}{n}\sum_{i=1}^n (L_i\mu^{(n)})\eta_i - \lambda_n\chi_1\mu^{(n)}\|^2 = 0$$

and therefore

$$r = \frac{1}{n}\sum_{i=1}^{n}(L_i\mu^{(n)})\eta_i - \lambda_n\chi_1\mu^{(n)} = G_{n,\lambda_n}\mu^{(n)}.$$

As this is true for all $r \in \mathcal{H}'_n$ we can infer the existence of an inverse operator $G^{-1}_{n,\lambda_n} : \mathcal{H}'_n \to \mathcal{H}'_n$ such that $G^{-1}_{n,\lambda_n} r = \mu^{(n)}$. One can verify that G^{-1}_{n,λ_n} is linear. As $\omega \in \Omega'$ was arbitrary, the result holds almost surely. □

5.3 Consistency

We demonstrate consistency by applying the Γ-convergence framework. This requires us to find the Γ-limit, and to show that the Γ-limit has a unique minimizer and that the minimizers of $f_n^{(\omega)}$ are uniformly bounded. The next three subsections demonstrate that each of these requirements is satisfied under the stated assumptions and allow the application of Corollary 2.2.2 to conclude the consistency of the spline model, as summarized in Theorem 5.3.1. We start by stating the remainder of the conditions employed.

Assumptions 4. *4.5 We have $\lambda_n = n^{-p}$ with $0 < p \leq \frac{1}{2}$.*

4.6 For $\nu \in \mathcal{H}$ the following relation holds:

$$\int_{\mathcal{I}}(L_t\nu)^2\phi_T(dt) = 0 \Leftrightarrow \nu = 0.$$

4.7 For each $\mu \in \mathcal{H}$ each $L_t\mu$ is continuous in t, i.e $\|L_s - L_t\|_{\mathcal{H}^*} \to 0$ as $s \to t$.

Assumption 4.5 gives the admissible scaling regime in λ_n. Clearly if $p \leq 0$ then $\lambda_n \not\to 0$ hence we expect the limit, if it even exists, to be biased. We are required to show that the minimizers are bounded in probability. To do so we show they are bounded in expectation. We will show in Theorem 5.3.3 that for $p > \frac{1}{2}$ we cannot bound minimizers in expectation; hence it is not possible to extend our proofs for $p \notin (0, \frac{1}{2}]$. The reason Theorem 5.3.1 holds in probability and not in expectation is that the Γ-convergence framework requires $\mu^{(n)}$ to be a minimizer and as such we cannot make conclusions about the average minimizer since $\mathbb{E}[\mu^{(n)}|\mathcal{G}_n]$ is not a minimizer.

We will show that the second derivative of f_∞ is given by $\int_{\mathcal{I}}(L_t\nu)^2\ \phi_T(\mathrm{d}t)$ for the direction ν. Assumption 4.6 is used to establish that f_∞ is strictly convex, and hence the minimizer is unique.

It will be necessary to show that

$$\frac{1}{n}\sum_{i=1}^{n}|L_i\mu| \to \int_{\mathcal{I}}|L_t\mu|\ \phi_T(\mathrm{d}t) \tag{5.5}$$

for all $\mu \in \mathcal{H}$ with probability one. We impose Assumption 4.7 (together with Assumption 4.4) to imply that $L_t\mu$ is continuous and bounded in t for all $\mu \in \mathcal{H}$ and therefore by the weak convergence of the empirical measure we infer that (5.5) holds for all $\mu \in \mathcal{H}$ and for almost every sequence $\{L_i\}_{i=1}^{\infty}$. In particular we can define a set $\Omega' \subset \Omega$ independent of μ, on which (5.5) holds, such that $\mathbb{P}(\Omega') = 1$.

Theorem 5.3.1. *Define $f_n^{(\omega)} : \mathcal{H} \to \mathbb{R}$ by* (5.2)*. Under Assumptions 4.1-4.7 the minimizer $\mu^{(n)}$ of $f_n^{(\omega)}$ converges in the following topological sense: for all $\epsilon, \delta > 0$ and $F \in \mathcal{H}^*$ there exists $N = N(\epsilon, \delta, F) \in \mathbb{N}$ such that*

$$\mathbb{P}\left(\left|F(\mu^{(n)}) - F(\mu^\dagger)\right| \geq \epsilon\right) \leq \delta \quad \textit{for } n \geq N.$$

Remark 5.3.2. *We view the convergence in the above theorem as the natural generalization of convergence in probability to the context. To fit with standard notation we have been careful not to say converges weakly in probability which could reasonably mislead the reader into interpreting the theorem as the convergence of $\mu^{(n)} \to \mu^\dagger$ is uniform over $F \in \mathcal{H}^*$ and not pointwise as we claim in the theorem.*

The following theorem shows that if $p > \frac{1}{2}$ then without imposing further assumptions it is always possible to construct observation functionals $\{L_t\}_{t\in\mathcal{I}}$ such that $\mathbb{E}\left[\|\mu^{(n)}\|^2\right] \to \infty$.

Theorem 5.3.3. *Define $f_n^{(\omega)} : \mathcal{H} \to \mathbb{R}$ by* (5.2)*, let $\mu^{(n)}$ be the minimizer of $f_n^{(\omega)}$ and take any $\alpha > 0$ and $p > \frac{1}{2}$. Take Assumptions 4.1-4.2 and assume that $\lambda = n^{-p}$. Then there exists a distribution ϕ_T on $\mathcal{I}$ such that $\|L_t\|_{\mathcal{H}^*} = \|\eta_t\| \leq \alpha$ for almost every $\omega \in \Omega$ (i.e. Assumption 4.4 holds) and $\mathbb{E}[\|\mu^{(n)}\|^2] \to \infty$.*

Essentially when considering weak convergence, one is restricting to finite dimensional projections. It is therefore not surprising that $n^{-\frac{1}{2}}$ is the best we can do. For $p > \frac{1}{2}$ and a sequence of real valued iid random variables X_i of finite variance (which are not identically zero) we have $n^{2p}\mathbb{E}(\frac{1}{n}\sum_{i=1}^n X_i)^2 \to \infty$. In light of this elementary observation Theorem 5.3.3 is not surprising. The proof is given in Section 5.3.4.

5.3.1 The Γ-Limit

We claim the Γ-limit of $f_n^{(\omega)}$, for almost every $\omega \in \Omega$, is given by

$$f_\infty(\mu) = \int_{\mathcal{I}} \int_{-\infty}^{\infty} |y - L_t\mu|^2 \ \phi_{L_t\mu^\dagger}(\mathrm{d}y) \ \phi_T(\mathrm{d}t). \tag{5.6}$$

Theorem 5.3.4. *Define $f_n^{(\omega)}, f_\infty : \mathcal{H} \to \mathbb{R}$ by* (5.2) *and* (5.6) *respectively. Under Assumptions 4.1-4.2, 4.5 and 4.7,*

$$f_\infty = \Gamma\text{-}\lim_n f_n^{(\omega)}$$

for almost every $\omega \in \Omega$.

Proof. We are required to show the two inequalities in Definition 2.2.1 hold with probability 1. In order to do this we consider a subset of Ω of full measure, Ω', and show that both statements hold for every data sequence obtained from that set.

Define $g_\mu(t, y) = (y - L_t\mu)^2$. For clarity let $P(\mathrm{d}(t, y)) = \phi_T(\mathrm{d}t)\phi_{L_t\mu^\dagger}(\mathrm{d}y)$ and P_n be the empirical measure associated with the observations, i.e. for any measurable $h : \mathcal{I} \times \mathbb{R} \to \mathbb{R}$ we define $P_n h = \frac{1}{n}\sum_{i=1}^n h(t_i, y_i)$. Further, let $P_n^{(\omega)}$ denote the measure arising from the particular realization ω. Defining:

$$\Omega' = \left\{\omega : P_n^{(\omega)} \Rightarrow P\right\} \cap \left\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^n \epsilon_i^2(\omega) \to \sigma^2 \text{ and } \frac{1}{n}\sum_{i=1}^n \epsilon_i(\omega) \to 0\right\},$$

then $\mathbb{P}(\Omega') = 1$ by the almost sure weak convergence of the empirical measure [56] and the strong law of large numbers. Let $\omega \in \Omega'$.

We start with the lim inf inequality. Pick $\nu \in \mathcal{H}$ and let $\nu^{(n)} \rightharpoonup \nu$. By Theorem 1.1 in [64] we have

$$\begin{aligned}\int_{\mathcal{I}}\int_{-\infty}^{\infty} \liminf_{n\to\infty,(t',y')\to(t,y)} g_{\nu^{(n)}}(t',y')\ P(\mathrm{d}(t,y)) &\leq \liminf_{n\to\infty} \int_{\mathcal{I}}\int_{-\infty}^{\infty} g_{\nu^{(n)}}(t,y)\ P_n^{(\omega)}(\mathrm{d}(t,y)) \\ &= \liminf_{n\to\infty} f_n^{(\omega)}(\nu^{(n)}).\end{aligned}$$

Now we show

$$\liminf_{n\to\infty,(t',y')\to(t,y)} g_{\nu^{(n)}}(t',y') \geq g_\nu(t,y) \tag{5.7}$$

which proves the lim inf inequality. Let $(t_m, y_m) \to (t,y)$ then

$$\begin{aligned}\left(g_{\nu^{(n)}}(t_m,y_m)\right)^{\frac{1}{2}} &= |y_m - L_{t_m}\nu^{(n)}| \\ &\geq |L_{t_m}\nu^{(n)} - y| - |y_m - y| \\ &\geq |y - L_t\nu^{(n)}| - |L_{t_m}\nu^{(n)} - L_t\nu^{(n)}| - |y_m - y| \\ &\geq |y - L_t\nu^{(n)}| - \|L_{t_m} - L_t\|_{\mathcal{H}^*}\|\nu^{(n)}\| - |y_m - y|.\end{aligned}$$

A consequence of the uniform boundedness principle is that any weakly convergent sequence is bounded, hence there exists some $C > 0$ such that $\|\nu^{(n)}\| \leq C$. It follows from the above, and Assumption 4.7, that

$$\liminf_{n\to\infty,m\to\infty} \left(g_{\nu^{(n)}}(t_m,y_m)\right)^{\frac{1}{2}} \geq |y - L_t\nu| = \left(g_\nu(t,y)\right)^{\frac{1}{2}}.$$

As our choice of sequence (t_m, y_m) was arbitrary we can conclude that (5.7) holds.

For the recovery sequence we choose $\nu \in \mathcal{H}$ and let $\nu^{(n)} = \nu$. We are required to show

$$Pg_\nu \geq \limsup_{n\to\infty}\left(P_n^{(\omega)}g_\nu + \lambda_n\|\mu\|_1^2\right) = \limsup_{n\to\infty} P_n^{(\omega)}g_\nu.$$

Since we can write

$$g_\nu(t_i,y_i) = (L_i\mu^\dagger)^2 + \epsilon_i^2 + (L_i\nu)^2 + 2\epsilon_iL_i\mu^\dagger - 2L_i\mu^\dagger L_i\nu - 2\epsilon_iL_i\nu$$

and each term is either a continuous and bounded functional, or its convergence is addressed directly by the construction of Ω', we have $P_n^{(\omega)}g_\nu \to Pg_\nu$ as required. As $\omega \in \Omega'$ was arbitrary, the result holds almost surely. □

Remark 5.3.5. *Note that in the above theorem we did not need a lower bound on the decay of λ_n (only that $\lambda_n \geq 0$). We only used that $\lambda_n = o(1)$.*

5.3.2 Uniqueness of the Γ-limit

To show the Γ-limit has a unique minimizer we show it is strictly convex. The following lemma gives the second Gâteaux derivative of f_∞. After which we conclude in Corollary 5.3.7 that the Γ-limit is unique.

Lemma 5.3.6. *Under Assumptions 4.1-4.2 define $f_\infty : \mathcal{H} \to \mathbb{R}$ by* (5.6)*. Then the first and second Gâteaux derivatives of f_∞ are given by*

$$\partial f_\infty(\mu;\nu) = 2\int_{\mathcal{I}}\int_{-\infty}^{\infty}(L_t\mu - y)L_t(\nu)\phi_{L_t\mu^\dagger}(dy)\phi_T(dt)$$

$$\partial^2 f_\infty(\mu;\nu,\zeta) = 2\int_{\mathcal{I}}(L_t\nu)(L_t\zeta)\phi_T(dt).$$

Proof. We first compute the first Gâteaux derivative.

$$\begin{aligned}\partial f_\infty(\mu;\nu) &= \lim_{r\to 0}\int_{\mathcal{I}}\int_{-\infty}^{\infty}\frac{(y-L_t(\mu+r\nu))^2-(y-L_t\mu)^2}{r}\,\phi_{L_t\mu^\dagger}(\mathrm{d}y)\phi_T(\mathrm{d}t)\\ &= 2\int_{\mathcal{I}}\int_{-\infty}^{\infty}(L_t\mu-y)L_t(\nu)\phi_{L_t\mu^\dagger}(\mathrm{d}y)\phi_T(\mathrm{d}t)\\ &\qquad + \lim_{r\to 0} r\int_{\mathcal{I}}\int_{-\infty}^{\infty}(L_t\nu)^2\phi_{L_t\mu^\dagger}(\mathrm{d}y)\phi_T(\mathrm{d}t)\\ &= 2\int_{\mathcal{I}}\int_{-\infty}^{\infty}(L_t\mu-y)L_t(\nu)\phi_{L_t\mu^\dagger}(\mathrm{d}y)\phi_T(\mathrm{d}t)\end{aligned}$$

The second Gâteaux derivative follows similarly:

$$\begin{aligned}\partial^2 f_\infty(\mu;\nu,\zeta) &= \lim_{r\to 0} 2\int_{\mathcal{I}}\int_{-\infty}^{\infty}\frac{(L_t(\mu+r\zeta)-y)L_t\nu-(L_t\mu-y)L_t\nu}{r}\phi_{L_t\mu^\dagger}(\mathrm{d}y)\phi_T(\mathrm{d}t)\\ &= 2\int_{\mathcal{I}}\int_{-\infty}^{\infty}(L_t\nu)(L_t\zeta)\phi_{L_t\mu^\dagger}(\mathrm{d}y)\phi_T(\mathrm{d}t)\\ &= 2\int_{\mathcal{I}}(L_t\nu)(L_t\zeta)\phi_T(\mathrm{d}t)\end{aligned}$$

which complete the proof. ☐

Corollary 5.3.7. *Under Assumptions 4.1,4.2 and 4.6, define $f_\infty : \mathcal{H} \to \mathbb{R}$ by* (5.6)*. Then f_∞ has a unique minimizer which is achieved for $\mu = \mu^\dagger$.*

Proof. It is easy to check that $\partial f_\infty(\mu^\dagger;\nu) = 0$ for all $\nu \in \mathcal{H}$. By Lemma 5.3.6 and Assumption 4.6 the second Gâteaux derivative satisfies $\partial^2 f_\infty(\mu;\nu) > 0$ for all $\nu \neq 0$. Then by Taylor's Theorem (and noting that f_∞ is quadratic), for $\mu \neq \mu^\dagger$,

$$f_\infty(\mu) = f_\infty(\mu^\dagger) + \frac{1}{2}\partial^2 f_\infty(\mu^\dagger;\mu-\mu^\dagger) > f_\infty(\mu^\dagger)$$

as required. ☐

5.3.3 Bound on Minimizers

In this subsection we show that $\|\mu^{(n)}\| = O_p(1)$. The bound in $\mathcal{H}_0$ can be obtained using fewer assumptions (than the bound in $\mathcal{H}$), which is natural considering $\mathcal{H}_0$ is finite dimensional. We may choose the norm on $\mathcal{H}_0$ without changing the topology (all norms are equivalent on finite dimensional spaces). We will use

$$\|\mu\|_0 = \int_{\mathcal{I}} |L_t \mu| \phi_T(\mathrm{d}t).$$

Loosely speaking we can then write $\|\mu^{(n)}\|_0 \lesssim f_n^{(\omega)}(\mu^{(n)})$. The bound in $\mathcal{H}_0$ then follows if $\min f_n^{(\omega)}$ is bounded. We make this argument rigorous in Lemma 5.3.8. After this result we concentrate on bounding $\mu^{(n)}$ in $\mathcal{H}$.

Lemma 5.3.8. *Define $f_n^{(\omega)} : \mathcal{H} \to \mathbb{R}$ by* (5.2)*. Under Assumptions 4.1-4.5 and 4.7 the minimizers $\mu^{(n)}$ of $f_n^{(\omega)}$ are, with probability one, eventually bounded in $\mathcal{H}_0$, i.e. for almost every $\omega \in \Omega$ there exist constants $C, N > 0$ such that $\|\mu^{(n)}\|_0 \le C$ for all $n \ge N$.*

Proof. We define P and $P_n^{(\omega)}$ as in the proof of Theorem 5.3.4, let

$$\Omega' = \left\{\omega \in \Omega : P_n^{(\omega)} \Rightarrow P\right\} \cap \left\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^n \epsilon_i^2(\omega) \to \sigma^2 \text{ and } \frac{1}{n}\sum_{i=1}^n |\epsilon_i(\omega)| \to P|\epsilon_1|\right\}$$

and $\mu^{(n)}$ be a minimizer of $f_n^{(\omega)}$. Assume $\omega \in \Omega'$. As

$$f_n^{(\omega)}(\mu^{(n)}) \le f_n^{(\omega)}(\mu^\dagger) \le \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 + \lambda_1 \|\mu^\dagger\|_1^2 \to \sigma^2 + \lambda_1\|\mu^\dagger\|_1^2,$$

there exists N such that $f_n^{(\omega)}(\mu^{(n)}) \le \sigma^2 + \lambda_1 \|\mu^\dagger\|_1^2 + 1$ for $n \ge N$.

Now

$$\begin{aligned} f_n^{(\omega)}(\mu) &= \frac{1}{n}\sum_{i=1}^n (y_i - L_i\mu)^2 + \lambda_n \|\mu\|_1^2 \\ &\ge \frac{1}{n}\sum_{i=1}^n (|L_i\mu| - |y_i| - 1) \\ &= \frac{1}{n}\sum_{i=1}^n |L_i\mu| - \frac{1}{n}\sum_{i=1}^n |y_i| - 1 \\ &\ge \frac{1}{n}\sum_{i=1}^n |L_i\mu| - \frac{1}{n}\sum_{i=1}^n |L_i\mu^\dagger| - \frac{1}{n}\sum_{i=1}^n |\epsilon_i| - 1 \\ &\to \int_{\mathcal{I}} |L_t\mu| \phi_T(\mathrm{d}t) - c \end{aligned}$$

where the convergence follows since $|L_t\mu|$ is a continuous and bounded functional in t and c is given by

$$\lim_{n\to\infty} \left(\frac{1}{n}\sum_{i=1}^n |L_i\mu^\dagger| + \frac{1}{n}\sum_{i=1}^n |\epsilon_i| + 1\right) \le \int_{\mathcal{I}} |L_t\mu^\dagger|\phi_T(\mathrm{d}t) + \sigma + 1 =: c.$$

We now show that $\int_{\mathcal{I}} |L_t\mu| \phi_T(\mathrm{d}t)$ is a norm on $\mathcal{H}_0$ and hence that the above constant, c, is finite. This will also show that $\|\mu\|_0 \leq f_n^{(\omega)}(\mu) + c$ which completes the proof.

The triangle inequality, absolute homogeneity and $\int_{\mathcal{I}} |L_t\mu| \phi_T(\mathrm{d}t) \geq 0$ are trivial to establish. If $\int_{\mathcal{I}} |L_t\mu| \phi_T(\mathrm{d}t) = 0$ then, by Assumption 4.3, we have at least m disjoint subsets of positive measure (with respect to ϕ_T) on $\mathcal{I}$. It follows that on each of these subsets $L_t\mu = 0$. As $\mathcal{H}_0$ is m-dimensional this determines μ, and hence $\mu = 0$.

As $\omega \in \Omega'$ was arbitrary and $\mathbb{P}(\Omega') = 1$, the result holds almost surely. □

Remark 5.3.9. *In the above lemma we did not need the lower bound on λ_n (only that $\lambda_n \geq 0$). The result holds for all $\lambda_n = O(1)$.*

Continuing with the bound in $\mathcal{H}$ we write

$$\mu^{(n)} = \frac{1}{n}\sum_{i=1}^n L_i\mu^\dagger G_{n,\lambda_n}^{-1}\eta_i + \frac{1}{n}\sum_{i=1}^n \epsilon_i G_{n,\lambda_n}^{-1}\eta_i = G_{n,\lambda_n}^{-1}U_n\mu^\dagger + \frac{1}{n}\sum_{i=1}^n \epsilon_i G_{n,\lambda_n}^{-1}\eta_i \tag{5.8}$$

where

$$U_n = \frac{1}{n}\sum_{i=1}^n \eta_i L_i. \tag{5.9}$$

We bound $\|G_{n,\lambda_n}^{-1}U_n\mu^\dagger\|$ in Lemma 5.3.11 and $\|\frac{1}{n}\sum_{i=1}^n \epsilon_i G_{n,\lambda_n}^{-1}\eta_i\|$ in Lemma 5.3.12.

In the proof of Lemma 5.3.11 we show that G_{n,λ_n}^{-1} maps from $\mathrm{Ran}(U_n)$ to $\mathrm{Ran}(U_n)$. Lemma 5.3.10 gives the conditions necessary to infer the existence of a orthonormal basis of eigenfunctions $\{\psi_j^{(n)}\}_{j=1}^\infty$ of $\mathrm{Ran}(U_n)$. Hence we can write

$$\|G_{n,\lambda_n}^{-1}U_n\mu\|^2 = \sum_{j=1}^\infty (G_{n,\lambda_n}^{-1}U_n\mu, \psi_j^{(n)}).$$

From here we exploit the fact that $\psi_j^{(n)}$ are eigenfunctions. We leave the details until the proof of Lemma 5.3.11.

Lemma 5.3.12 is a consequence of being able to bound $\|G_{n,\lambda_n}^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{H})}$ in terms of λ_n. One is then left to show $\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i\right)^2 = O(\frac{1}{n})$. We start by showing that U_n is compact, bounded, self-adjoint and positive semi-definite.

Lemma 5.3.10. *Define U_n by* (5.9)*. Under Assumptions 4.1 and 4.4, U_n is almost surely a bounded, self-adjoint, positive semi-definite and compact operator on $\mathcal{H}$.*

Proof. In this proof we consider $\omega \in \Omega'$ where $\Omega' = \{\omega : \|\eta_i(\omega)\| \leq \alpha \text{ for all } i\}$, noting that $\mathbb{P}(\Omega') = 1$ by Assumption 4.4.

Boundedness of U_n follows easily as

$$\|U_n\mu\| \leq \frac{1}{n}\sum_{i=1}^n \alpha^2\|\mu\| = \alpha^2\|\mu\|.$$

Let $(\cdot,\cdot)_{\mathbb{R}^n}$ be the inner product on $\mathbb{R}^n$ given by

$$(x,y)_{\mathbb{R}^n} = \frac{1}{n}\sum_{i=1}^n x_iy_i \qquad \forall x,y \in \mathbb{R}^n.$$

Now for $x \in \mathbb{R}$ and $\nu \in \mathcal{H}$ we have

$$(x, L_i\nu)_{\mathbb{R}^1} = xL_i\nu = x(\eta_i, \nu) = (x\eta_i, \nu)$$

which shows $L_i^* : \mathbb{R} \to \mathcal{H}$ is given by $L_i^* x = x\eta_i$. Now if we define $T_n = (L_1, \ldots, L_n) : \mathcal{H} \to \mathbb{R}^n$ then for $x \in \mathbb{R}^n$, $\nu \in \mathcal{H}$

$$(T_n\nu, x)_{\mathbb{R}^n} = \frac{1}{n}\sum_{i=1}^{n} L_i\nu x_i = (\frac{1}{n}\sum_{i=1}^{n} x_i\eta_i, \nu).$$

Hence $T_n^* x = \frac{1}{n}\sum_{i=1}^{n} x_i\eta_i$. We have shown $U_n = T_n^* T_n$, and is therefore self-adjoint.

To show U_n is positive semi-definite then we need

$$(U_n\nu, \nu) \geq 0$$

for all $\nu \in \mathcal{H}$. This follows easily as

$$(U_n\nu, \nu) = \frac{1}{n}\sum_{i=1}^{n}(L_i\nu)^2 \geq 0.$$

For compactness of U_n (for n fixed) let $\nu^{(m)}$ be a sequence with $\|\nu^{(m)}\| \leq 1$. Since $|L_i\nu^{(m)}| \leq \alpha$ for every $\omega \in \Omega'$, there exists a convergent subsequence m_p such that

$$L_i\nu^{(m_p)} \to \kappa_i \quad \forall i = 1, 2, \ldots, n \quad \text{say.}$$

So $U_n\nu^{(m_p)} \to \frac{1}{n}\sum_{i=1}^{n} \eta_i\kappa_i \in \mathcal{H}$ as $m_p \to \infty$. Therefore each U_n is compact. □

Using the basis whose existence is implied by the previous lemma, we can bound the first term on the RHS of (5.8).

Lemma 5.3.11. *Under Assumptions 4.1-4.4 define G_{n,λ_n} and U_n by* (5.3) *and* (5.9) *respectively. Then with probability one we have*

$$\|G_{n,\lambda_n}^{-1} U_n\|_{\mathcal{L}(\mathcal{H},\mathcal{H})} \leq 1$$

for all n.

Proof. First note that $\dim(\text{Ran}(U_n)) = \dim(\text{span}\{\eta_1, \ldots, \eta_n\}) \leq n$. Without loss of generality we will assume $\dim(\text{Ran}(U_n)) = n$ (else we can assume the dimension is m_n where $m_n \leq n$ is an increasing sequence). Clearly χ_1 is a self-adjoint, bounded and compact operator on $\text{Ran}(U_n)$ as is U_n by Lemma 5.3.10. Therefore there exists a simultaneous diagonalisation of U_n and χ_1 on $\text{Ran}(U_n)$. I.e. there exists $\beta_j^{(n)}, \gamma_j^{(n)}$ and $\phi_j^{(n)}$ such that

$$U_n\psi_j^{(n)} = \beta_j^{(n)}\psi_j^{(n)} \quad \text{and} \quad \chi_1\psi_j^{(n)} = \gamma_j^{(n)}\psi_j^{(n)}$$

for all $j = 1, 2, \ldots, n$. Furthermore $\psi_j^{(n)}$ form an orthonormal basis of $\text{Ran}(U_n)$. Since χ_1 and

U_n are both semi-positive definite then $\beta_j^{(n)}, \gamma_j^{(n)} \geq 0$. We have

$$G_{n,\lambda_n}\psi_j^{(n)} = U_n\psi_j^{(n)} + \lambda_n\chi_1\psi_j^{(n)} = \left(\beta_j^{(n)} + \lambda_n\gamma_j^{(n)}\right)\psi_j^{(n)}.$$

So,

$$G_{n,\lambda_n}^{-1}\psi_j^{(n)} = \frac{1}{\beta_j^{(n)} + \lambda_n\gamma_j^{(n)}}\psi_j^{(n)}.$$

In particular this shows that

$$G_{n,\lambda_n}^{-1}U_n : \mathcal{H} \to \text{Ran}(U_n).$$

Assume $\mu \in \mathcal{H}, \nu \in \text{Ran}(U_n)$, then

$$\mu = \sum_{i=1}^n (\mu, \psi_i^{(n)})\psi_i^{(n)} + \hat{\mu} \qquad \text{and} \qquad \nu = \sum_{i=1}^n (\nu, \psi_i^{(n)})\psi_i^{(n)}$$

where $\hat{\mu} \in \text{Ran}(U_n)^\perp$. Therefore,

$$(U_n\mu, \psi_j^{(n)}) = \sum_{i=1}^n (\mu, \psi_i^{(n)})(U_n\psi_i^{(n)}, \psi_j^{(n)}) = \beta_j^{(n)}(\mu, \psi_j^{(n)})$$
$$(G_{n,\lambda_n}^{-1}\nu, \psi_j^{(n)}) = \sum_{i=1}^n (\nu, \psi_i^{(n)})(G_{n,\lambda_n}^{-1}\psi_i^{(n)}, \psi_j^{(n)}) = \frac{1}{\beta_j^{(n)} + \lambda_n\gamma_j^{(n)}}(\nu, \psi_j^{(n)}).$$

Which implies

$$(G_{n,\lambda_n}^{-1}U_n\mu, \psi_j^{(n)}) = \frac{1}{\beta_j^{(n)} + \lambda_n\gamma_j^{(n)}}(U_n\mu, \psi_j^{(n)}) = \frac{\beta_j^{(n)}}{\beta_j^{(n)} + \lambda_n\gamma_j^{(n)}}(\mu, \psi_j^{(n)}).$$

Hence

$$\begin{aligned}
\|G_{n,\lambda_n}^{-1}U_n\mu\|^2 &= \sum_{j=1}^n (G_{n,\lambda_n}^{-1}U_n\mu, \psi_j^{(n)})^2 \\
&= \sum_{j=1}^n \left(\frac{\beta_j^{(n)}}{\beta_j^{(n)} + \lambda_n\gamma_j^{(n)}}\right)^2 (\mu, \psi_j^{(n)})^2 \\
&\leq \sum_{j=1}^n (\mu, \psi_j^{(n)})^2 \\
&\leq \|\mu\|^2.
\end{aligned}$$

This proves the lemma. □

We now focus on bounding $\|G_{n,\lambda_n}^{-1}\nu^{(n)}\|$ where $\nu^{(n)} = \frac{1}{n}\sum_{i=1}^n \epsilon_i\eta_i$.

Lemma 5.3.12. *Under Assumptions 4.1-4.5 define G_{n,λ_n} by* (5.3). *Then*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i G_{n,\lambda_n}^{-1}\eta_i\right\|^2 \middle| \mathcal{G}_n\right] = O(1) \quad \textit{almost surely.}$$

Proof. Recalling B from the proof of Lemma 5.2.1, we have

$$(G_{n,\lambda_n}\mu,\mu) = B(\mu,\mu) \geq \lambda_n \|\mu\|_1^2.$$

This implies $\|G_{n,\lambda_n}\mu\| \geq \lambda_n\|\mu\|_1$. By Lemma 5.2.1 there exists a well defined inverse of G_{n,λ_n} at η_i, hence we let $\mu = G_{n,\lambda_n}^{-1}\eta_i$ and we have

$$\|G_{n,\lambda_n}^{-1}\eta_i\|_1 \leq \frac{1}{\lambda_n}\|\eta_i\| \leq \frac{\alpha}{\lambda_n}.$$

Now, define $\nu^{(n)} = \frac{1}{n}\sum_{i=1}^n \epsilon_i\eta_i$ and

$$\begin{aligned}\mathbb{E}\left[\left\|G_{n,\lambda_n}^{-1}\nu^{(n)}\right\|_1^2 \middle| \mathcal{G}_n\right] &\overset{\text{a.s.}}{=} \frac{\sigma^2}{n^2}\sum_{i=1}^n \left\|G_{n,\lambda_n}^{-1}\eta_i\right\|_1^2 \\ &\leq \frac{\alpha^2\sigma^2}{n\lambda_n^2}.\end{aligned}$$

Combined with Lemma 5.3.8 (the $\mathcal{H}_0$ bound) this proves the lemma. □

Recalling (5.8) and via Lemmas 5.3.11 and 5.3.12 we obtain the following asymptotic bound on minimizers in $\mathcal{H}$.

Theorem 5.3.13. *Under Assumptions 4.1-4.5 we have*

$$\mathbb{E}\left[\|\mu^{(n)}\|^2|\mathcal{G}_n\right] = O(1) \quad \textit{almost surely.} \tag{5.10}$$

This is a stronger result than we needed; we were only required to show that $\|\mu^{(n)}\|$ is bounded in probability. Taking expectation of (5.10) one has

$$\mathbb{E}\|\mu^{(n)}\|^2 = O(1).$$

Hence applying Chebyshev's inequality we may conclude that $\|\mu^{(n)}\| = O_p(1)$.

Corollary 5.3.14. *Under Assumptions 4.1-4.5 we have* $\|\mu^{(n)}\| = O_p(1)$.

We conclude this section with a brief analysis of the rate of convergence. For any $F \in \mathcal{H}^*$, by the Riesz Representation Theorem, there exists $\xi \in \mathcal{H}$ such that $F(\mu) = (\mu,\xi)$ for all $\mu \in \mathcal{H}$. Hence

$$F(\mu^{(n)}) - F(\mu^\dagger) = ((G_{n,\lambda_n}^{-1}U_n - \mathrm{Id})\mu^\dagger + G_{n,\lambda_n}^{-1}\nu^{(n)},\xi)$$

where $\nu^{(n)} = \frac{1}{n}\sum_{i=1}^n \epsilon_i\eta_i$. Decomposing $\mathcal{H}$ into $\mathcal{H} = \overline{\mathrm{Ran}(U_n)} \oplus \mathrm{Ran}(U_n)^\perp$ one can write

$$\begin{aligned}F(\mu^{(n)}) - F(\mu^\dagger) &= \left(\left(G_{n,\lambda_n}^{-1}U_n - \chi_{\overline{\mathrm{Ran}(U_n)}}\right)\mu^\dagger,\xi\right) - \left(\chi_{\mathrm{Ran}(U_n)^\perp}\mu^\dagger,\xi\right) + \left(G_{n,\lambda_n}^{-1}\nu^{(n)},\xi\right) \\ &= \sum_{j=1}^n \frac{-\lambda_n}{\beta_j^{(n)}+\lambda_n}\left(\mu^\dagger,\psi_j^{(n)}\right)\left(\psi_j^{(n)},\xi\right) \\ &\qquad - \left(\chi_{\mathrm{Ran}(U_n)^\perp}\mu^\dagger,\xi\right) + \left(G_{n,\lambda_n}^{-1}\nu^{(n)},\xi\right)\end{aligned}$$

where $\chi_{\overline{\mathrm{Ran}(U_n)}}$ is the projection onto $\overline{\mathrm{Ran}(U_n)}$. The first term is of order 1 and the third term, by the proof of Lemma 5.3.12, is of order $\frac{1}{\sqrt{n}\lambda_n}$. The second term is independent of λ_n.

Theorem 5.3.15. *Under Assumptions 4.1-4.6, for $F \in \mathcal{H}^*$ take $\xi \in \mathcal{H}$ such that $F(\mu) = (\mu, \xi)$ and assume there exists $q > \frac{1}{2}$ such that*

$$\left|\left(\xi, \psi_j^{(j)}\right)\right| \lesssim j^{-q} \qquad \text{and} \qquad \left|\left(\mu^\dagger, \psi_j^{(j)}\right)\right| \lesssim j^{-q}$$

where $\{\psi_j^{(n)}\}_{j=1}^n$ are the set of orthonormal functions that span the range of U_n. Then, for all $0 < \delta < 1 - \frac{1}{2q}$, we have

$$\mathbb{E}\left[|F(\mu^{(n)}) - F(\mu^\dagger)|\,|\mathcal{G}_n\right] = O(1) + O\left(n^{-2q\delta}\right) + O\left(\frac{1}{\lambda_n\sqrt{n}}\right) \qquad \textit{almost surely.}$$

Proof. Let A be such that

$$\left|\left(\xi, \psi_j^{(j)}\right)\right|, \left|\left(\mu^\dagger, \psi_j^{(j)}\right)\right| \leq A j^{-q}.$$

We are left to show that

$$\left(\chi_{\mathrm{Ran}(U_n)^\perp}\mu^\dagger, \xi\right) = O\left(n^{-2q\delta}\right).$$

Now since $\dim(U_j) = j$ we can assume that $\psi_{j+1}^{(j+1)}$ is orthogonal to $\psi_j^{(j)}$. Hence the set $\{\psi_j^{(j)}\}_{j=1}^\infty$ forms an orthonormal basis for $\mathcal{H}$.

For any $1 > \delta > 0$ one has

$$n^{2q\delta}\left(\chi_{\mathrm{Ran}(U_n)^\perp}\mu^\dagger, \xi\right) = n^{2q\delta}\left|\sum_{j=n+1}^\infty \left(\mu^\dagger, \psi_j^{(j)}\right)\left(\psi_j^{(j)}, \xi\right)\right| \leq A^2 \sum_{j=1}^\infty \left(\frac{n^\delta}{j+n}\right)^{2q}.$$

A simple maximization argument shows

$$\frac{n^\delta}{j+n} \leq \left(\frac{\delta - 1}{\delta j}\right)^{1-\delta} \qquad \text{for all } j.$$

Hence

$$n^{2q\delta}\left(\chi_{\mathrm{Ran}(U_n)^\perp}\mu^\dagger, \xi\right) \leq A^2\left(\frac{\delta-1}{\delta}\right)^{2q(1-\delta)}\sum_{j=1}^\infty \frac{1}{j^{2q(1-\delta)}}$$

where the above sum is finite for $\delta < 1 - \frac{1}{2q}$. Which proves

$$\left|\left(\chi_{\mathrm{Ran}(U_n)^\perp}\mu^\dagger, \xi\right)\right| = O\left(n^{-2q\delta}\right)$$

for all $\delta < 1 - \frac{1}{2q}$. □

Remark 5.3.16. *Since $\mu^\dagger, \xi \in \mathcal{H}$ and $\{\psi_j^{(j)}\}_{j=1}^\infty$ form an orthonormal set in $\mathcal{H}$,*

$$\|\mu^\dagger\|^2 \geq \sum_{j=1}^\infty (\mu^\dagger, \psi_j^{(j)})^2$$

and similarly for ξ. It immediately follows that there exists $q \geq \frac{1}{2}$ and A such that

$$\left|\left(\mu^\dagger, \psi_j^{(j)}\right)\right| \leq Aj^{-q} \qquad \textit{and} \qquad \left|\left(\xi, \psi_j^{(j)}\right)\right| \leq Aj^{-q}.$$

The application of the above theorem requires the slightly stronger assumption that $q > \frac{1}{2}$.

5.3.4 Sharpness of the Scaling Regime - Proof of Theorem 5.3.3

Proof of Theorem 5.3.3. Fix any $\alpha > 0$ and without loss of generality we can choose $\{\eta_t\}_{t\in\mathcal{I}}$ such that $\|\eta_t\| = \alpha$ for all $t \in \mathcal{I}$. Define $L_t \in \mathcal{H}$ by $L_t = (\eta_i, \cdot)$.

In the proof of Lemma 5.2.1 we showed

$$|(G_{n,\lambda_n}\mu, \nu)| \leq (\alpha^2 + \lambda_n)\|\mu\|\|\nu\|.$$

Letting $\nu = G_{n,\lambda_n}\mu$, for $\mu \in \text{span}\{\eta_1, \dots, \eta_n\}$, one has

$$\|G_{n,\lambda_n}\mu\|^2 \leq (\alpha^2 + \lambda_n)\|\mu\|\|G_{n,\lambda_n}\mu\|.$$

And hence

$$\|G_{n,\lambda_n}\mu\| \leq (\alpha^2 + \lambda_n)\|\mu\|.$$

Which implies

$$\|G_{n,\lambda_n}^{-1}\mu\| \geq \frac{1}{\alpha^2 + \lambda_n}\|\mu\|.$$

Now, for $\nu^{(n)} = \frac{1}{n}\sum_{i=1}^n \epsilon_i\eta_i$, we consider

$$\begin{aligned}\mathbb{E}\left[\|G_{n,\lambda_n}^{-1}\nu^{(n)}\|^2\middle|\mathcal{G}_n\right] &\geq \frac{1}{(\alpha^2+\lambda_n)^2}\mathbb{E}\left[\|\nu^{(n)}\|^2|\mathcal{G}_n\right] \\ &\overset{\text{a.s.}}{=} \frac{\sigma^2\alpha^2}{\lambda_n^2 n(\alpha^2+\lambda_n)^2} \\ &\to \infty\end{aligned}$$

as $\lambda_n^2 n \to \infty$. Hence by taking expectations:

$$\mathbb{E}\left[\|G_{n,\lambda_n}^{-1}\nu^{(n)}\|^2\right] \to \infty.$$

By noting

$$\mathbb{E}\left[\|\mu^{(n)}\|^2\right] = \mathbb{E}\left[\|G_{n,\lambda_n}^{-1}U_n\mu^\dagger\|^2\right] + \mathbb{E}\left[\|G_{n,\lambda_n}^{-1}\nu^{(n)}\|^2\right]$$

we conclude the proof. □

5.4 Application to the Special Spline Model

Consider the application to the special spline case, $L_i\mu = \mu(t_i)$. We let

$$\mathcal{H} = H^m := \left\{g : [0,1] \to \mathbb{R} \text{ s.t } \nabla^i g \text{ abs. cts. for } i = 1, 2, \dots, m-1 \text{ and } \nabla^m g \in L^2\right\}.$$

For $m \geq 1$, $\mathcal{H}$ is a reproducing kernel Hilbert space and therefore L_i as defined are linear and bounded operators on $\mathcal{H}$. See [25, 171] for more details on reproducing kernel Hilbert spaces. This section discusses the following points.

1. The decomposition $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $\mathcal{H}_0$ is finite dimensional.

2. The function η_t corresponding to $(\eta_t, \mu) = L_t \mu = \mu(t)$.

The other assumptions needed to apply Theorem 5.3.1 are Assumption 4.3 and Assumption 4.6. Assumption 4.3 is

$$\mu(t) = \mu(r) \quad \text{for all polynomial } \mu \text{ of degree at most } m-1 \text{ then } t = r$$

which clearly holds. Assumption 4.6 becomes

$$\int_0^1 |\nu(t)|^2 \phi_T(\mathrm{d}t) = 0 \Leftrightarrow \nu = 0$$

which, for example, is true if $\phi_T(\mathrm{d}t) = \hat{\phi}_T(t)\ \mathrm{d}t$ and $\hat{\phi}_T(t) > 0$ for all $t \in [0,1]$.

1. The decomposition $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$. For $\mu \in \mathcal{H}$ by Taylor expanding μ from 0 we can write:

$$\mu(t) = \sum_{i=0}^{m-1} \frac{\nabla^i \mu(0)}{i!} t^i + R(t)$$

where $\nabla^i R(0) = 0$ for all $i = 0, 1, \ldots, m-1$. Hence $R \in \mathcal{H}_1$ where

$$\mathcal{H}_1 = \left\{ g \in H^m : \nabla^i g(0) = 0 \text{ for all } i = 0, 1, \ldots, m-1 \right\}.$$

A Poincaré inequality holds on this space so $\|\mu\|_1^2 = \int_0^1 |\nabla^m \mu(t)|^2\ \mathrm{d}t$ is a norm on $\mathcal{H}_1$.

We define $\mathcal{H}_0$ to be the span of the functions ζ_i defined by

$$\zeta_i(t) = \frac{t^i}{i!} \qquad \text{for } i = 0, 1, \ldots, m-1.$$

The space is coupled with the inner product

$$(\mu, \nu)_0 = \sum_{i=0}^{m-1} \nabla^i \mu(0) \nabla^i \nu(0).$$

The space $\mathcal{H}_0$ has $\dim(\mathcal{H}_0) = m$.

2. The functions η_t. In the above R is given by

$$R(t) = \int_0^1 \frac{(t-u)_+^{m-1}}{(m-1)!} \nabla^m \mu(u)\ \mathrm{d}u = \int_0^1 G(t,u) \nabla^m \mu(u)\ \mathrm{d}u$$

where $(u)_+ = \max\{0, u\}$ and

$$G(t,u) = \frac{(t-u)_+^{m-1}}{(m-1)!}$$

is the Green's function for $\nabla^m \mu = \nu$ and boundary conditions $\nabla^j \mu(0) = 0$ for all $0 \le j \le m-1$.

We claim that $\eta_t \in H^m$ satisfying $(\eta_t, \mu) = \mu(t)$ are given by

$$\eta_t(r) = \sum_{i=0}^{m-1} \zeta_i(t)\zeta_i(r) + \int_0^1 G(t,u)G(r,u)\ \mathrm{d}u =: \eta_t^0(r) + \eta_t^1(r).$$

Furthermore $\eta_t^0 \in \mathcal{H}_0$ and $\eta_t^1 \in \mathcal{H}_1$ for all $t \in [0,1]$. The proof follows directly from calculating

$$(\eta_t, \mu) = \sum_{i=0}^{m-1} \nabla^i \eta_t(0) \nabla^i \mu(0) + \int_0^1 \nabla^m \eta_t(u) \nabla^m \mu(u)\ \mathrm{d}u$$

and noticing

$$\nabla^i \eta_t(r) = \sum_{j=1}^{m-1} \zeta_j(t) \left[\nabla^i \zeta_j(r)\right]_{r=0} = \zeta_i(t) \quad \text{for } i < m$$

$$\nabla^m \eta_t(r) = \nabla_r^m \int_0^1 G(t,u)G(r,u)\ \mathrm{d}u = G(t,r).$$

One can easily show that $\|\eta_t\| \le 1$ for all $t \in [0,1]$.

Continuity of η_t follows easily. As each polynomial is Lipschitz continuous on the interval $[0,1]$, there exists a constant C_i (depending on the order of the polynomial i) such that $|\zeta_i(t) - \zeta_i(s)| \le C_i |t-s|$. Now for the integral term let $m \ge 2$ and $s \ge t$ then:

$$\begin{aligned}
&\left| \int_0^1 \left(G(s,u) - G(t,u)\right) G(r,u)\ \mathrm{d}u \right| \\
&\quad = \left| \int_0^1 \left(\mathbb{I}_{s>u} \frac{(s-u)^{m-1}}{(m-1)!} - \mathbb{I}_{t>u} \frac{(t-u)^{m-1}}{(m-1)!} \right) G(r,u)\ \mathrm{d}u \right| \\
&\quad \le \int_t^s \frac{(s-u)^{m-1}}{(m-1)!} G(r,u)\ \mathrm{d}u + \frac{1}{(m-2)!} \int_0^t |s-t|\, g(r,u)\ \mathrm{d}u \\
&\quad \le \frac{m|s-t|}{[(m-1)!]^2}.
\end{aligned}$$

The case $m = 1$ is similar. It follows that $\|L_s - L_t\|_{\mathcal{H}^*} = \|\eta_s - \eta_t\| \le C|s-t|$ for some $C < \infty$ and hence L_t is continuous.

Chapter 6

Asymptotic Analysis of the Ginzburg-Landau Functional on Point Clouds

Abstract

The Ginzburg-Landau functional is a phase transition model which is suitable for clustering or classification type problems. We study the asymptotics of a sequence of Ginzburg-Landau functionals with anisotropic interaction potentials on point clouds Ψ_n where n denotes the number data points. In particular we show the limiting problem, in the sense of Γ-convergence, is related to the total variation norm restricted to functions taking binary values; which can be understood as a surface energy. We generalize the result known for isotropic interaction potentials to the anisotropic case and add a result concerning the rate of convergence.

6.1 Introduction

6.1.1 Finite Dimensional Modeling

Graphical models are used across a very broad spectrum of problems from social science type problems, such as identifying communities [51, 65, 130, 166, 175], to image segmentation [20, 84], to cell biology [33], to modeling the world wide web [24, 31, 33, 62] and many more. We use an anisotropic model which, for example, is suitable for cosmological models [83, 101], modeling outbreaks of disease [99] and image recognition [178]. With this type of problem there is often no data generating model available. For example in [84] the authors use a graphical model to identify features in a picture. In problems of this type there is no physical model. For this reason graphical models are a very popular choice of modeling technique. The types of problems that motivate the graphical modeling methodology can be seen more generally as clustering or classification problems.

The problem is given data $\Psi_n = \{\xi_i\}_{i=1}^n \subset X$ where $X \subset \mathbb{R}^d$ find $\mu : \Psi_n \to \mathbb{R}$ that labels each data point. The labeling is constructed so that $\mu(\xi_i) = 0$ means that ξ_i is associated

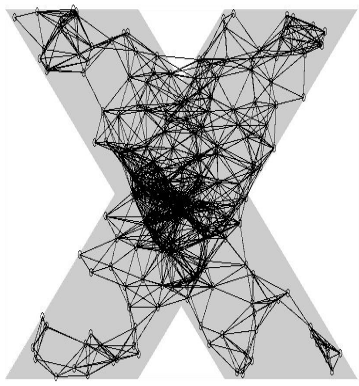


Figure 6.1: An example graph. For the classifier estimates see Figure 6.2.

with the cluster labeled 0 and $\mu(\xi_i) = 1$ means that ξ_i is associated with the cluster labeled 1. For a finite number of observations we allow a soft classification however the scaling is chosen such that in the data rich limit classifiers are binary valued. The motivation for our approach is to validate approximating the hard classification problem by a soft classification problem. The soft classification problem is in general numerically easier [68] and therefore more appealing to the practitioner. However one also wants to be precise in regards to which class a data point belongs. Minimizers of the Ginzburg-Landau functional are used as a classification tool [163] in order to allow for phase transitions which allow a soft classification approach whilst also penalizing states that are not close to a hard classification. A consequence of our proofs is an insight into the ratio of data points that receive a hard classification, i.e. the asymptotic behavior of $\{\xi_i : \mu(\xi_i) \in \{0,1\}\}$.

Another important application for this work is in designing classifiers. By not assuming that the model is isotropic we allow greater flexibility which allows one to choose some features as more important than others. The next subsection contains a simple example which shows how the design choice can affect the classification. In particular one can use the methodology to map infinite dimensional data onto a finite dimensional space in order to classify the infinite dimensional data set.

Assessing the validity of such an approach is of high importance. This is especially true as one cannot intuitively link the model to the data generating process. When one can make such a connection then the model can be heuristically motivated. Without such connection one needs to do more in order to justify the approach.

The primary results of this chapter concern showing that $\mu^{(n)}$ converges to a minimizer of a limiting model. We also give some preliminary results into characterizing the rate of convergence in a simplified example. We believe these results will hold under more generality than stated here and it is the objective of ongoing work to extend them.

Our approach is motivated by [5, 69, 163]. Classifiers are constructed as the solution of a variational problem which is common in statistical problems, e.g. maximum likelihood and maximum-a-posterior problems. In particular minimizers of the Ginzburg-Landau funtional, a phase transition model popular in material science and image segmentation, are used as classifiers. Classifiers $\mu^{(n)} : \Psi_n \to \mathbb{R}$ are constructed as follows. Let $V : \mathbb{R} \to [0, \infty)$ be a potential such that states taking the value 0 or 1 is favored. For example $V(t) = t^2(t-1)^2$. A graph is constructed by taking the vertices as the set Ψ_n and weighting edges

$$W_{ij} = \eta_\epsilon(\xi_i - \xi_j)$$

for $\eta_\epsilon : \mathbb{R}^d \to [0, \infty)$ and we say that there is an edge between ξ_i and ξ_j if $W_{ij} > 0$, for example see Figure 6.1. Assume that $\eta_\epsilon : \mathbb{R}^d \to [0, \infty)$ is of the form

$$\eta_\epsilon(x) = \frac{1}{\epsilon^d} \eta \left(x/\epsilon \right) \tag{6.1}$$

with scaling parameter $\epsilon \in (0, \infty)$ and $\eta : \mathbb{R}^d \to [0, \infty)$. For example one (isotropic) choice of η is $\eta(x) = 1$ if $|x| < 1$ and $\eta(x) = 0$ if $|x| \geq 1$. For a function μ on Ψ_n the graph energy $\mathcal{E}_n(\mu^{(n)}) \in [0, \infty]$ is defined by

$$\mathcal{E}_n(\mu) = \frac{1}{\epsilon} \frac{1}{n} \sum_{i=1}^n V(\mu(\xi_i)) + \frac{1}{\epsilon} \frac{1}{n^2} \sum_{i,j} W_{ij} |\mu(\xi_i) - \mu(\xi_j)|. \tag{6.2}$$

Our classifier to the clustering problem is then given as the minimizer of (6.2).

This is similar to the approach taken in [69] where they consider pairwise interactions only. In particular one can define the graph total variation by

$$GTV_n(\mu) := \frac{1}{\epsilon} \frac{1}{n^2} \sum_{i,j} W_{ij} |\mu(\xi_i) - \mu(\xi_j)|. \tag{6.3}$$

In the special case that $\mu(\xi_i) \in \{0, 1\}$ this reduces to the graph cut of Ψ_n, i.e. if $\mu^{-1}(0) = A_0$ and $\mu^{-1}(1) = A_1$ then

$$GTV_n(\mu) = \frac{1}{\epsilon} \frac{1}{n^2} \sum_{\xi_i \in A_0} \sum_{\xi_j \in A_1} W_{ij}.$$

We wish to allow for soft clustering however the total variation term is not enough to be able to do this informatively. The clustering approach is made more robust by including a first order term which penalizes associating a data point to more than one cluster. See, for example, Figure 6.2 for a comparison. It is not trivial that the convergence results in [69] will survive adding a penalty term.

Finding minimizers of $\mathcal{E}_n$ is also an important problem but is not addressed in this thesis. We instead refer to [29, 30] for numerical methods.

6.1.2 Example: Classification Dependence on the Choice of η

Through a toy problem we demonstrate how the interaction potential can be used to pick out features of the practitioners choice. Data points are functions $f_i : [0, 1] \to \mathbb{R}$ generated from four classes. For a fixed α the interaction potential $\eta : L^2([0, 1]) \to [0, \infty)$ is defined by

$$\eta(f) = \begin{cases} 1 & \text{if} \left| (\|f\|_{L^2}, \|\nabla f\|_{L^2}) \right|_\alpha \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

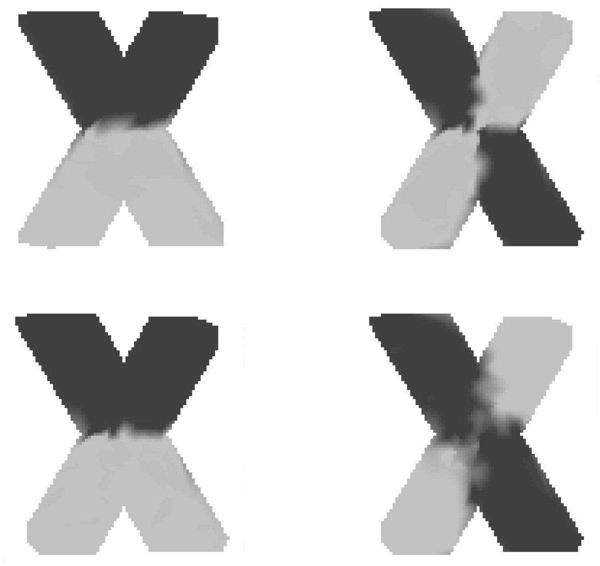


Figure 6.2: The top row shows the minimizers of $\mathcal{E}_n$ and the bottom row shows the minimizers of GTV_n for the graph given in Figure 6.1 conditioned on the node closest to each corner taking either 0 or 1. The left column is conditioned to have 0 in the bottom corners and 1 in the top corners. The right column has 0 in the bottom left and top right corners and 1 in the top left and bottom right corners. There is very little difference between the outputs on the left but on the right the GTV_n term fails to pick out the singularity at the center.

where $|\cdot|_\alpha$ is a transformed Euclidean norm on $\mathbb{R}^2$:

$$|x|_\alpha^2 = x_1^2(1-\alpha) + x_2^2\alpha.$$

For $\alpha \approx 1$ the potential favors vertical interactions whilst for $\alpha \approx 0$ the potential favors horizontal interactions. As α varies from $\frac{1}{8}$ to $\frac{7}{8}$ the energy $\mathcal{E}_n$ changes behavior by assigning a lower energy to a vertical partitioning of the data (compared with a horizontal partitioning). More precisely, let

$$\mu^{(1)}(f_i) = \begin{cases} 1 & \text{if } \|\nabla f_i\|_{L^2} \leq c_1 \\ 0 & \text{otherwise} \end{cases}$$

$$\mu^{(2)}(f_i) = \begin{cases} 1 & \text{if } \|f_i\|_{L^2} \leq c_2 \\ 0 & \text{otherwise.} \end{cases}$$

Then define

$$\Delta\mathcal{E}_n = \mathcal{E}_n(\mu^{(1)}) - \mathcal{E}_n(\mu^{(2)}).$$

The results are given in Figure 6.3. The results show that functions with similar L^2 norms are clustered together for smaller α and as α increases the classification favors functions whose derivatives have similar L^2 norm.

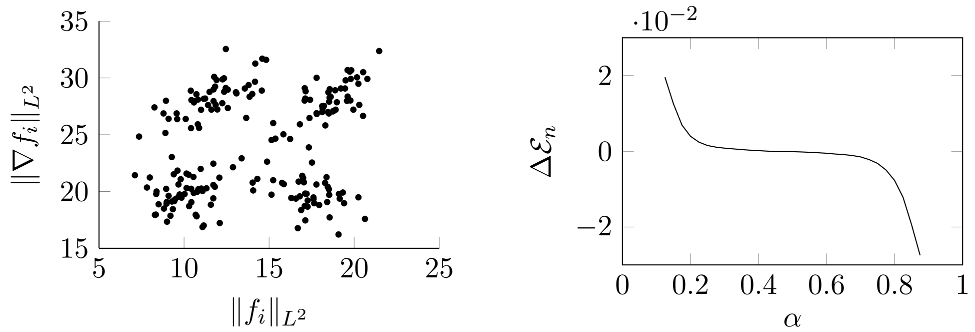


Figure 6.3: Random functions are projected onto $\mathbb{R}^2$ by $f_i \mapsto (\|f_i\|_{L^2}, \|\nabla f_i\|_{L^2})$. The distribution of the functions is chosen so that each random function falls in one of four classes which can be seen in the figure on the left. The interactions are parameterized by a potential α which favors horizontal partitions for $\alpha \approx 0$ and vertical partitions for $\alpha \approx 1$ as shown by the figure on the right.

6.1.3 The Limiting Model

Rather surprisingly the problem of soft classifications for finite data sets and hard classification in the limit has received relatively little attention in the literature. However it is well known that for finite data one can recover the k-means algorithm (hard classification) from the expectation-maximization algorithm (soft classification) in the zero-variance limit for the Gaussian mixture model and the Dirichlet process mixture model [92, 106].

The results of this chapter concern the asymptotics of the minimum and minimizers of $\mathcal{E}_n$, where $\epsilon = \epsilon_n$ and $\epsilon_n \to 0$ as $n \to \infty$. The advantages of scaling ϵ_n to zero are two-fold. The first is that the matrix $W = (W_{ij})_{ij}$ is sparse and therefore the minimization is numerically easier. The second is to improve resolution of the boundary. One can think of soft classification as estimating the probability that a data point belongs to a certain class and the hard classification problem as estimating the boundaries where one class is more likely than all others. By scaling $\epsilon_n \to 0$ it will be shown that the limiting minimization problem is a hard classification. For example, Figure 6.4 shows (for a fixed number of data points) improved resolution in the boundary between clusters as $\epsilon \to 0$. See also [68].

Define $\mathcal{E}_\infty : L^1(X) \to [0, \infty]$ by

$$\mathcal{E}_\infty(\mu) = \begin{cases} \int_{\partial\{\mu=1\}} \sigma(n(x))\rho^2(x)\, \mathrm{d}\mathcal{H}^{d-1}(x) & \text{if } \mu \in L^1(X; \{0,1\}) \\ \infty & \text{otherwise} \end{cases} \tag{6.4}$$

where $n(x)$ is the outward unit normal for the set $\partial\{\mu = 1\}$, $\mathcal{H}^{d-1}$ is the $d-1$ dimensional Hausdorff measure and

$$\sigma(\nu) = \int_{\mathbb{R}^d} \eta(x)|x \cdot \nu|\, \mathrm{d}x. \tag{6.5}$$

One can also define $\mathcal{E}_\infty$ by

$$\mathcal{E}_\infty(\mu) = \begin{cases} TV(\mu; \rho, \eta) & \text{if } \mu \in L^1(X; \{0,1\}) \\ \infty & \text{otherwise} \end{cases}$$

where TV is defined by (2.1-2.3). It will be shown, in the sense of Γ-convergence, that $\mathcal{E}_\infty$

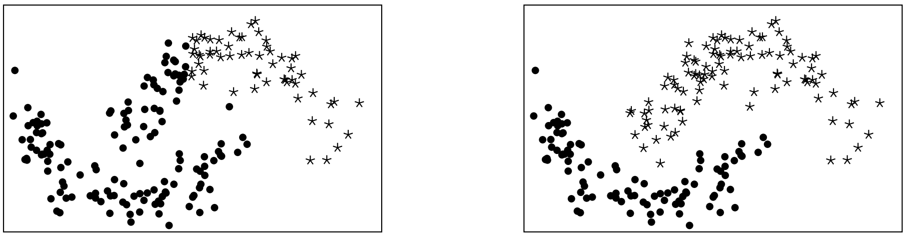


Figure 6.4: Both figures were classified using the Ginzburg-Landau functional. The one on the left used a larger ϵ than the one on the right. The figure shows that the smaller value of ϵ gives a much better resolution in the boundary.

is the limiting problem and any sequence such that $\sup_{n\in\mathbb{N}} \mathcal{E}_n(\mu^{(n)}) < \infty$ is precompact. In particular this allows one to apply the results of this chapter to infer the consistency of the constrained minimization problem (see Section 6.2.1 and Chapter 7).

Note that since each $\mu^{(n)}$ is defined on a different space (the domain of each $\mu^{(n)}$ is Ψ_n) that it is not straightforward what is meant by convergence of $\mu^{(n)} \to \mu^{(\infty)}$. We use the results of Section 2.5 to define a map $T_n : X \to \Psi_n$ so that one can compare $\mu^{(n)}$ to $\mu^{(\infty)}$. Approximately we can say $\mu^{(n)} \to \mu^{(\infty)}$ in TL^1 if $\mu^{(n)} \circ T_n \to \mu^{(\infty)}$ in L^1.

We also include preliminary results towards characterizing the rate of convergence by considering a simple example when $\mu = \mathbb{I}_E$ for a polyhedral set $E \subset X$ and looking at the convergence in mean square:

$$\mathbb{E}\left|\mathcal{E}_n(\mu) - \mathcal{E}_\infty(\mu)\right|^2 = \mathbb{E}\left|GTV_n(\mu) - TV(\mu;\rho,\eta)\right|^2.$$

It is shown that the above convergence is dominated by a bias due to approximating edges of the set E and is of order $O(\epsilon_n^2)$. The second largest term corresponds to the convergence along faces of E and obeys a $\frac{\kappa}{n}$ decay for a given constant κ. A further overview of these results is given in Section 6.2.2 and the proofs in Section 6.5.

The outline of the chapter is as follows. In Section 6.2 we state the main result, Theorem 6.2.2 (the convergence of the unconstrained minimization problem). We also include an overview of the preliminary rate of convergence results to be found in Section 6.5. In Section 6.3 the proof of the first part of Theorem 6.2.2 (the compactness result) is given. And in Section 6.4 the proof is completed with the Γ-convergence result. Finally in Section 6.5 we make the preliminary calculation regarding the rate of convergence of "$\mathcal{E}_n \to \mathcal{E}_\infty$".

6.2 Main Results and Assumptions

Throughout this chapter $X \subset \mathbb{R}^d$ where $d \geq 2$ is open, connected and bounded with Lipschitz boundary.

The data points ξ_i are assumed to be independent and identically distributed (iid) from a probability measure P supported on X which has density (with respect to the Lebesgue measure) ρ. It is assumed that ρ is continuous on X and bounded above and below by strictly positive constants, i.e. there exists constants $0 < c \leq C < \infty$ such that $c \leq \rho(x) \leq C$ for every $x \in X$, and $\rho(x) = 0$ for every $x \in \mathbb{R}^d \setminus X$.

To reduce computational expense and increase boundary resolution one should scale the graph edge weights to zero as quickly as possible. Our proofs require the following lower bound on ϵ_n:

$$\lim_{n\to\infty} \frac{n\epsilon_n^d}{\log n} = \infty \qquad \text{if } d \geq 3 \tag{6.6}$$

$$\lim_{n\to\infty} \frac{n\epsilon_n^2}{(\log n)^{\frac{3}{2}}} = \infty \qquad \text{if } d = 2. \tag{6.7}$$

The lower bound is required to ensure the graph with vertices Ψ_n and edges weighted by W_{ij} is (with probability one) connected [121, Theorem 13.2]. Furthermore by application of Theorem 2.5.1 there exists an optimal transport map $T_n : X \to \Psi_n$ that one can use to define the convergence of $\mu^{(n)} \to \mu^{(\infty)}$ where $\mu^{(n)} \in L^1(\Psi_n)$ and $\mu^{(\infty)} \in L^1(X)$ and

$$\frac{\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n} \to 0.$$

The assumptions on V and η are given in the following definition.

Definition 6.2.1. *We say that the pair (V, η) where $V : \mathbb{R} \to [0,\infty)$ and $\eta : \mathbb{R}^d \to [0,\infty)$ are admissible if*

1. $\int_{\mathbb{R}^d} \eta(y)|y| \, \mathrm{d}y < \infty$,
2. *η satisfies the following strengthened Lipschitz condition: there exists L such that for all $x, y \in \mathbb{R}^d$:*
$$|\eta(x) - \eta(y)| \leq L \max\{\eta(x), \eta(y)\} |x - y|,$$
3. $\eta(0) > 0$,
4. *there exists two decreasing functions $\eta^-, \eta^+ : [0,\infty) \to [0,\infty)$ and constants $\gamma_1, \gamma_2 > 0$ such that for all $x \in \mathbb{R}^d$*
$$\eta^-(|x|) \leq \eta(x) \leq \eta^+(|x|) \qquad \textit{and} \qquad \eta^+(|x|) \leq \gamma_1 \eta^-(|x|/\gamma_2),$$
5. *$V(y) = 0$ if and only if $y \in \{0, 1\}$,*
6. *V is continuous,*
7. *there exists $r > 0$ and $\tau > 0$ such that if $|t| \geq r$ then $V(t) \geq \tau|t|$.*

Remark 6.2.1. *A sufficient condition for condition 2 in Definition 6.2.1 is that $\eta \in C^1(\mathbb{R}^d)$ satisfies the following:*

1. *there exists L_1 such that for all $x, y \in \mathbb{R}^d$ and $t \in [0, 1]$*
$$|\nabla\eta(tx + (1-t)y)| \leq L_1 t |\nabla\eta(x)| + L_1(1-t) |\nabla\eta(y)|,$$
2. *there exists L_2 such that $|\nabla\eta(x)| \leq L_2\eta(x)$ for all $x \in \mathbb{R}^d$.*

When the above holds then so does condition 2 with $L = L_1 L_2$.

Note that the integrability condition on η implies that $\sigma(\nu) < \infty$ for all $\nu \in \mathbb{R}^d$. We now state the main result.

Theorem 6.2.2. *Let $X \subset \mathbb{R}^d$ for $d \geq 2$ be a bounded, open, connected domain with Lipschitz boundary and P a probability distribution with support on X and density ρ. Assume that ρ is continuous and bounded above and below by strictly positive constants on X. The data is distributed $\xi_i \overset{\text{iid}}{\sim} P$ and let $\Psi_n = \{\xi_i\}_{i=1}^n$. Let $\epsilon_n \to 0$ be a sequence satisfying the bound* (6.6) *if $d \geq 3$ or* (6.7) *if $d = 2$. For any function $\mu^{(n)}$ on Ψ_n define $\mathcal{E}_n(\mu^{(n)}) \in [0, \infty]$ by* (6.2) *where the weights W_{ij} are given by $W_{ij} = \eta_{\epsilon_n}(\xi_i - \xi_j)$ and η_{ϵ_n} is given by* (6.1). *Define $\mathcal{E}_\infty : L^1(X) \to [0, \infty]$ by* (6.4-6.5). *Assume (V, η) are admissible functions. Then, with probability one, the following hold*

1. *Compactness: Let $\mu^{(n)}$ be a sequence of functions on Ψ_n such that $\sup_{n\in\mathbb{N}} \mathcal{E}_n(\mu^{(n)}) < \infty$ then $\mu^{(n)}$ is relatively compact in TL^1 and each cluster point is in $BV(X; \rho, \eta) \cap L^1(X; \{0, 1\})$.*
2. *Γ-limit: we have*
$$\Gamma\text{-}\lim_{n\to\infty} \mathcal{E}_n = \mathcal{E}_\infty.$$

The compactness result is proved in Proposition 6.3.1 and the Γ-convergence result in Theorem 6.4.1. We apply the above theorem for when $\eta = \mathbb{I}_E$ in the following corollary.

Corollary 6.2.3. *Let $X \subset \mathbb{R}^d$ for $d \geq 2$ be a bounded, open, connected domain with Lipschitz boundary and P a probability distribution with support on X and density ρ. Assume that ρ is continuous and bounded above and below by strictly positive constants on X. The data is distributed $\xi_i \overset{\text{iid}}{\sim} P$ and let $\Psi_n = \{\xi_i\}_{i=1}^n$. Let $\epsilon_n \to 0$ be a sequence satisfying the bound* (6.6) *if $d \geq 3$ or* (6.7) *if $d = 2$. For any function $\mu^{(n)}$ on Ψ_n define $\mathcal{E}_n(\mu^{(n)}) \in [0, \infty]$ by* (6.2) *where the weights W_{ij} are given by $W_{ij} = \eta_{\epsilon_n}(\xi_i - \xi_j)$ and η_{ϵ_n} is given by* (6.1) *where $\eta = \mathbb{I}_E$ for an open, bounded set E with Lipschitz boundary and $0 \in E$. Define $\mathcal{E}_\infty : L^1(X) \to [0, \infty]$ by* (6.4-6.5). *Assume V satisfies condition 5 to 7 in Definition 6.2.1. Then, with probability one, the following hold*

1. *Compactness: Let $\mu^{(n)}$ be a sequence of functions on Ψ_n such that $\sup_{n\in\mathbb{N}} \mathcal{E}_n(\mu^{(n)}) < \infty$ then $\mu^{(n)}$ is relatively compact in TL^1 and each cluster point is in $BV(X; \rho, \eta) \cap L^1(X; \{0, 1\})$.*
2. *Γ-limit: we have*
$$\Gamma\text{-}\lim_{n\to\infty} \mathcal{E}_n = \mathcal{E}_\infty.$$

Proof. The compactness property holds analogously to Proposition 6.3.1 in Section 6.3. For the Γ-convergence define

$$\eta^{(\delta)}(x) = \begin{cases} \alpha & \text{if } x \in E \\ \alpha \exp\left(-\frac{\operatorname{dist}(x, \partial E)}{\delta}\right) & \text{otherwise.} \end{cases}$$

Note that $\eta^{(\delta)} \geq \eta$ and $\eta^{(\delta)}$ satisfies condition 2 in Definition 6.2.1 which in light of Remark 6.2.1 is not difficult to check. Verifying conditions 1 and 3 for $\eta^{(\delta)}$ is trivial and one can show that for all δ

$$\eta^-(t) = \begin{cases} \alpha & \text{if } x \in B(0,r) \\ \alpha \exp\left(-\frac{\text{dist}(x,\partial B(0,r))}{\delta}\right) & \text{otherwise} \end{cases}$$

$$\eta^+(t) = \begin{cases} \alpha & \text{if } x \in B(0,R) \\ \alpha \exp\left(-\frac{\text{dist}(x,\partial B(0,R))}{\delta}\right) & \text{otherwise} \end{cases}$$

satisfies

$$\eta^-(|x|) \leq \eta^{(\delta)}(x) \leq \eta^+(|x|) \qquad \text{and} \qquad \eta^+(|x|) \leq \eta^-\left(\frac{r|x|}{R}\right),$$

where r and R are chosen so that $B(0,r) \subseteq E \subseteq B(0,R)$. Hence $\eta^{(\delta)}$ is an admissible function.

Also observe that $\eta^{(\delta)} \geq \eta$. The liminf inequality then follows from

$$\begin{aligned} \liminf_{n\to\infty} GTV_n(\mu^{(n)};\eta) &\leq \liminf_{n\to\infty} GTV_n(\mu^{(n)};\eta^{(\delta)}) \quad \text{for any } \delta \\ &\leq TV(\mu;\rho,\eta^{(\delta)}) \quad \text{by Lemma 6.4.3} \\ &\to TV(\mu;\rho,\eta) \quad \text{as } \delta \to 0 \text{ by the monotone convergence theorem.} \end{aligned}$$

The recovery sequence is similar by redefining the approximation

$$\eta^{(\delta)}(x) = \begin{cases} 0 & \text{if } x \in \mathbb{R}^d \setminus E \\ \alpha \exp\left(-\delta \text{dist}(x,\partial E)\right) & \text{otherwise,} \end{cases}$$

applying Lemma 6.4.4 and the monotone convergence theorem as $\delta \to 0$. □

6.2.1 Comments on the Main Result

The classical Ginzburg-Landau functional:

$$F_\epsilon(\mu) = \frac{1}{\epsilon}\int_X V(\mu(x))\,\mathrm{d}x + \frac{1}{\epsilon}\int_{X^2} \eta_\epsilon(x-y)\,|\mu(x)-\mu(y)|^2\,\mathrm{d}x\,\mathrm{d}y$$

has been well studied and its convergence to a total variation functional

$$F_0(\mu) = \int_{\{\mu=1\}} \sigma_\eta(n(x))\,\mathrm{d}\mathcal{H}^{d-1}(x)$$

known for some time [5,114]. More recent results have studied this functional on a (deterministic) regular graph. In [163] the authors show the Γ-convergence and compactness of two variants of the Ginzburg-Landau functional where $\{\xi_i\}_{i=1}^n \subset \mathbb{R}^2$ form a 4-regular graph. Let us exploit the structure of the graph by writing data as $\{\xi_{i,j}\}_{i,j=1}^n$ where $\xi_{i,j}$, $\xi_{i,j+1}$ are neighbors, as are

$\xi_{i,j}$ and $\xi_{j+1,i}$. The two variants of the Ginzburg-Landau functional considered in [163] are

$$h_{n,\epsilon}(\mu) = \frac{1}{\epsilon}\sum_{i,j=1}^{n} V(\mu(\xi_{i,j})) + \frac{1}{n}\sum_{i,j=1}^{n}\left(|\mu(\xi_{i+1,j}) - \mu(\xi_{i,j})|^2 + |\mu(\xi_{i,j+1}) - \mu(\xi_{i,j})|^2\right)$$
$$k_{n,\epsilon}(\mu) = \frac{1}{\epsilon n^2}\sum_{i,j=1}^{n} V(\mu(\xi_{i,j})) + \epsilon\sum_{i,j=1}^{n}\left(|\mu(\xi_{i+1,j}) - \mu(\xi_{i,j})|^2 + |\mu(\xi_{i,j+1}) - \mu(\xi_{i,j})|^2\right).$$

The first functional $h_{n,\epsilon}$ Γ-converges as $\epsilon \to 0$ (for a fixed n to a total variation function $h_{n,0}$ in a discrete setting defined by

$$h_{n,0}(\mu) = \begin{cases} \frac{1}{n}\sum_{i,j=1}^{n}\left(|\mu(\xi_{i+1,j}) - \mu(\xi_{i,j})|^2 + |\mu(\xi_{i,j+1}) - \mu(\xi_{i,j})|^2\right) & \text{if } \mu \in L^1(\Psi_n;\{0,1\}) \\ \infty & \text{otherwise.} \end{cases}$$

As $\epsilon \to 0$ and $n \to \infty$ sequentially or for $\epsilon = n^{-\alpha}$ for α within some range then $h_{n,\epsilon}$ Γ-converges to an anisotropic total variation in a continuous setting

$$\Gamma\text{-}\lim_{\substack{n\to\infty \\ \epsilon\to 0}} h_{n,\epsilon} = \int_{\mathbb{T}^2} \left|\frac{\partial \mu}{\partial x}\right| + \left|\frac{\partial \mu}{\partial y}\right|$$

and $k_{n,\epsilon}$ Γ-converges to an isotropic total variation

$$\Gamma\text{-}\lim_{\substack{n\to\infty \\ \epsilon\to 0}} k_{n,\epsilon} = \int_{\mathbb{T}^2} |\nabla \mu|$$

upto renormalization. Also discussed in [163] is the application to the constrained minimization problem

Convergence of the Graph Total Variation. As a consequence of our proofs a very similar result holds for GTV_n. In particular it is shown:

1. Compactness: Let $\mu^{(n)}$ be any sequence of functions on Ψ_n such that $\mu^{(n)}$ is bounded in TL^1 and $\sup_{n\in\mathbb{N}} GTV_n(\mu^{(n)}) < \infty$ then $\mu^{(n)}$ is relatively compact in TL^1 and each cluster point is in $BV(X;\rho,\eta) \cap L^1(X;\{0,1\})$.

2. Γ-limit: we have
$$\Gamma\text{-}\lim_{n\to\infty} GTV_n = TV(\cdot;\rho,\eta).$$

The same result can be found in [69] for the isotropic case. A key contribution of [69], and the related paper [70], is to identify the scale at which interactions between nodes is important and define a notion of convergence suitable for comparing functions on different domains.

Convergence of minimizers. The results of the Theorem 6.2.2 can be understood as implying the convergence of minimizers in the following sense. For a sequence of closed sets $\Theta_n \subseteq L^1(\Psi_n)$ and $\Theta \subseteq L^1(X)$ which we assume respect the Γ-convergence, that is if $\zeta \in \Theta$ then

$\zeta^{(n)} := \zeta\big|_{\Psi_n} \in \Theta_n$, there exists $\zeta \in \Theta$ such that $\mathcal{E}_\infty(\zeta) < \infty$ and $\zeta^{(n)} \to \zeta$ implies $\zeta \in \Theta$ if $\zeta^{(n)} \in \Theta_n$. Then (with probability one):

1. $\lim_{n\to\infty} \inf_{\Theta_n} \mathcal{E}_n = \min_\Theta \mathcal{E}_\infty$, and
2. if $\mu^{(n)} \in L^1(\Psi_n)$ are a sequence of (almost-)minimizers of $\mathcal{E}_n$ then this sequence is precompact in TL^1 and furthermore any cluster point minimizes $\mathcal{E}_\infty$.

The proof is a simple consequence of Theorem 6.2.2 and Theorem 2.2.1.

Alternatively one could let $g_n : L^1(\Psi_n) \to [0,\infty)$ be a sequence that continuously converges to $g : L^1(X) \to [0,\infty)$, i.e. $g_n(\zeta^{(n)}) \to g(\zeta)$ whenever $\zeta^{(n)} \to \zeta$ in L^1, and then since the Γ-convergence is stable under continuous perturbations the results of this chapter imply (with probability one) that

1. $\lim_{n\to\infty} \inf_{L^1(\Psi_n)} (\mathcal{E}_n + g_n) = \min_{L^1(X)} (\mathcal{E}_\infty + g)$, and
2. if $\mu^{(n)} \in L^1(\Psi_n)$ are a sequence of (almost-)minimizers of $\mathcal{E}_n + g_n$ then this sequence is precompact in TL^1 and furthermore any cluster point minimizes $\mathcal{E}_\infty + g$.

For example one could use this in order to fit data, e.g.

$$g_n(\mu;\zeta) = \lambda \int_X |\mu(T_n(x)) - \zeta(x)| \, \mathrm{d}x$$

where ζ is a known function (data) and T_n is a sequence of stagnating transport maps. In this case $g(\mu) = \lambda \int_X |\mu(x) - \zeta(x)| \, \mathrm{d}x$.

Choice of scaling. The natural choice of scaling in $\mathcal{E}_n$ between the two terms is not a-priori obvious. One could write

$$\mathcal{E}_n(\mu) = \frac{1}{\gamma_n} \sum_{i=1}^n V(\mu(\xi_i)) + \frac{1}{\epsilon_n} \frac{1}{n^2} \sum_{i,j} W_{ij} |\mu(\xi_i) - \mu(\xi_j)|.$$

The proof of Theorem 6.3.1 requires $\frac{\gamma_n}{\epsilon_n} = O(1)$. One can show that Theorem 6.2.2 holds for $\gamma_n = O(\epsilon_n)$. For simplicity it is assumed that $\gamma_n = \epsilon_n$.

Extension to L^p spaces. One does not have to use L^1 type distances for finite data. Define

$$\mathcal{E}_n(\mu) = \frac{1}{\epsilon_n} \sum_{i=1}^n V(\mu(\xi_i)) + \frac{1}{\epsilon_n^p} \frac{1}{n^2} \sum_{i,j} W_{ij} \left|\mu(\xi_i) - \mu(\xi_j)\right|^p.$$

Then the results of Theorem 6.2.2 hold for the same limiting energy $\mathcal{E}_\infty$ assuming

$$\int_{\mathbb{R}^d} \eta(y)|y|^{2p} \, \mathrm{d}y < \infty.$$

In particular the convergence of almost minimizers is still in TL^1.

Size of the phase transition. Fix $\mu = \mathbb{I}_E \in BV(X;\rho,\eta)$ and define a sequence of functions $\mu^{(n)} = \mu|_{\Psi_n}$. Then one has that the rate of convergence of $\mu^{(n)}$ to μ in L^p is of order $\|T_n - \mathrm{Id}\|_{L^\infty(X)}^{\frac{1}{p}}$. In particular there is a phase transition around ∂E of width $\|T_n - \mathrm{Id}\|_{L^\infty(X)}^{\frac{1}{p}}$.

6.2.2 Preliminary Results on the Rate of Convergence

We include some preliminary results concerning the rate of convergence for $\inf \mathcal{E}_n \to \min \mathcal{E}_\infty$. The problem is simplified by looking at the convergence $GTV_n(\mu) \to TV(\mu;\rho,\eta)$ for $\mu = \mathbb{I}_E$ where E is a polyhedral set. To characterize the rate of convergence we look for convergence in mean square. It is first shown that

$$\mathbb{E}GTV_n(\mu) \to TV(\mu;\rho,\eta)$$

as $n \to \infty$. Even though (by Lemma 6.4.4) one has $GTV_n(\mu) \to TV(\mu;\rho,\eta)$ almost surely the above convergence does not immediately follow. For example one needs to apply the dominated convergence theorem in order to show almost sure convergence implies convergence in mean. We state and prove the result in Theorem 6.5.1 in Section 6.5

After this result the statement and proof of convergence in mean square is given. In particular, in Theorem 6.5.2 it is shown that

$$\mathbb{E}\left|GTV_n(\mu) - TV(\mu;\rho,\eta)\right|^2 = O(\epsilon_n^2) + \frac{\kappa}{n} + O\left(\frac{1}{n^2\epsilon_n^{d+1}}\right) \tag{6.8}$$

for some constant κ given by

$$\kappa = 2|\partial E| \int_{B(0,1)} \int_{B(0,1)} \min\{|z_d|,|y_d|\}\,\mathrm{d}z\,\mathrm{d}y - 4TV(\mu;\rho,\eta)^2.$$

The first term on the RHS of (6.8) corresponds to approximating μ along edges of E. It is the error in the edges causes a bias in the estimate. For example if one considers the function $\mu = \mathbb{I}_{H\cap X}$ where H is any half plane then μ is a polyhedral function with no edges in X and then one can show $\mathbb{E}GTV_n(\mu) = TV(\mu;\rho,\eta)$ and it follows from our proofs that $\mathbb{E}\left|GTV_n(\mu) - TV(\mu;\rho,\eta)\right|^2 = \frac{\kappa}{n} + O\left(\frac{1}{n^2\epsilon_n^{d+1}}\right)$.

These results are preliminary and are leading towards characterizing the rate of convergence of the minima and minimum. Let $\Theta_n \subset L^1(\Psi_n)$ and $\Theta \subset L^1(X)$ be subsets so that the minimization $\inf_{\mu\in\Theta_n} \mathcal{E}_n$ and $\min_{\mu\in\Theta}\mathcal{E}_\infty$ is non-trivial. In future works we aim to find the convergence of $\inf_{\Theta_n}\mathcal{E}_n \to \min_\Theta \mathcal{E}_\infty$ and $\mu^{(n)} \to \mu^{(\infty)}$ where $\{\mu^{(n)}\}_{n=1}^\infty$ is a sequence of almost minimizers of $\mathcal{E}_n$ over Θ_n and $\mu^{(\infty)}$ is a minimizer of $\mathcal{E}_\infty$ in Θ.

6.3 The Compactness Property

In this section we prove the first part of Theorem 6.2.2 and establish that sequences bounded in $\mathcal{E}_n$ are precompact in TL^1 with cluster points in $L^1(X;\{0,1\})$. Our proofs compare $\mathcal{E}_n$ to its continuous analogue $\mathcal{C}_\epsilon : L^1(X) \to [0,\infty]$ defined by

$$\mathcal{C}_\epsilon(\mu) = \frac{1}{\epsilon}\int_X V(\mu(x))\rho(x)\,\mathrm{d}x + \frac{1}{\epsilon}\int_{X^2} \eta_\epsilon(y-x)\,|\mu(y)-\mu(x)|\,\rho(x)\rho(y)\,\mathrm{d}y\,\mathrm{d}x. \quad (6.9)$$

The transport map T_n between the measures P_n and P is used to compare a function $\mu^{(n)} : \Psi_n \to \mathbb{R}$ to its continuous version $\tilde{\mu}^{(n)} : X \to \mathbb{R}$, i.e. $\tilde{\mu}^{(n)} = \mu^{(n)} \circ T_n$. One then uses standard results to conclude the compactness of $\tilde{\mu}^{(n)}$ in L^1 and show that this implies compactness of $\mu^{(n)}$ in TL^1.

Proposition 6.3.1. *Under the same conditions as Theorem 6.2.2. If $\mu^{(n)} \in L^1(\Psi_n)$ is a sequence with*

$$\sup_{n\in\mathbb{N}} \mathcal{E}_n(\mu^{(n)}) < \infty$$

then, with probability one, there exists a subsequence $\mu^{(n_m)}$ and $\mu \in L^1(X;\{0,1\})$ such that $\mu^{(n_m)} \to \mu$ in TL^1.

Proof. Recall the following preliminary compactness result. If $\{\mu^{(n)}\}_{n=1}^\infty$ is a sequence in $L^1(X)$ such that

$$\sup_{n\in\mathbb{N}} \mathcal{C}_{\epsilon_n}(\mu^{(n)}) < \infty, \quad (6.10)$$

where $\mathcal{C}_{\epsilon_n} : L^1(X) \to [0,\infty]$ is defined by (6.9), then there exists a subsequence $\mu^{(n_m)}$ and $\mu \in L^1(X;\{0,1\})$ such that $\mu^{(n_m)} \to \mu$ in L^1. A proof can be found, for example, in [5].

For clarity we will denote the dependence of η on $\mathcal{E}_n$ by $\mathcal{E}_n(\cdot;\eta)$. Since η is continuous at 0 and $\eta(0) > 0$ there exists $b > 0$ and $a > 0$ such that $\eta(x) \geq a$ for all $|x| < b$. Define $\tilde{\eta}$ by $\tilde{\eta}(x) = a$ for $|x| < b$ and $\tilde{\eta}(x) = 0$ otherwise. As $\tilde{\eta} \leq \eta$ then $\mathcal{E}_n(\mu^{(n)};\eta) \geq \mathcal{E}_n(\mu^{(n)};\tilde{\eta})$.

Let T_n be such that $T_{n\#}P = P_n$ and the conclusions of Theorem 2.5.1 hold. We want to show $\{\mu^{(n)} \circ T_n\}_{n=1}^\infty$ satisfies

$$\sup_{n\in\mathbb{N}} \mathcal{C}_{\tilde{\epsilon}_n}(\mu^{(n)} \circ T_n;\tilde{\eta}) < \infty \quad (6.11)$$

for a sequence $\tilde{\epsilon}_n > 0$ with $\tilde{\epsilon}_n \to 0$ and $\frac{\epsilon_n}{\tilde{\epsilon}_n} \to 1$ that will be chosen shortly. If so then by (6.10) there exists a subsequence $\mu^{(n_m)} \circ T_{n_m}$ and $\mu \in L^1(X;\{0,1\})$ such that $\mu^{(n_m)} \circ T_{n_m} \to \mu$ in L^1 and therefore $\mu^{(n_m)} \to \mu$ in TL^1. To show (6.11) we write

$$\begin{aligned}\mathcal{C}_{\tilde{\epsilon}_n}(\mu^{(n)} \circ T_n;\tilde{\eta}) = {} & \frac{1}{\tilde{\epsilon}_n}\int_X V(\mu^{(n)}(T_n(x))\rho(x)\,\mathrm{d}x \\ & + \frac{1}{\tilde{\epsilon}_n}\int_{X^2} \tilde{\eta}_{\tilde{\epsilon}_n}(y-x)\left|\mu^{(n)}(T_n(x)) - \mu^{(n)}(T_n(y))\right|\rho(x)\rho(y)\,\mathrm{d}y\,\mathrm{d}x.\end{aligned}$$

The first term is uniformly bounded, since by (2.9)

$$\frac{1}{\tilde{\epsilon}_n}\int_X V(\mu^{(n)}(T_n(x))\rho(x)\,\mathrm{d}x = \frac{\epsilon_n}{\tilde{\epsilon}_n}\frac{1}{\epsilon_n}\sum_{i=1}^n V(\mu^{(n)}(\xi_i)).$$

Assume that $\left|\frac{y-x}{\tilde{\epsilon}_n}\right| < b$ then

$$\begin{aligned}|T_n(x) - T_n(y)| &\leq |T_n(x) - x| + |x - y| + |y - T_n(y)| \\ &\leq 2\|\mathrm{Id} - T_n\|_{L^\infty(X)} + |x - y| \\ &\leq 2\|\mathrm{Id} - T_n\|_{L^\infty(X)} + b\tilde{\epsilon}_n.\end{aligned}$$

Choose $\tilde{\epsilon}_n$ satisfying

$$2\|T_n - \mathrm{Id}\|_{L^\infty(X)} + b\tilde{\epsilon}_n = b\epsilon_n,$$

i.e. $\tilde{\epsilon}_n = \epsilon_n - \frac{2\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{b}$. By the decay assumption on ϵ_n (6.6-6.7) for n sufficiently large (with probability one) $\tilde{\epsilon}_n > 0$, $\tilde{\epsilon}_n \to 0$ and $\frac{\tilde{\epsilon}_n}{\epsilon_n} \to 1$. Also

$$\tilde{\eta}\left(\frac{x-y}{\tilde{\epsilon}_n}\right) = a \Rightarrow \tilde{\eta}\left(\frac{T_n(x) - T_n(y)}{\epsilon_n}\right) = a.$$

Therefore

$$\tilde{\eta}_{\tilde{\epsilon}_n}(y - x) = \frac{1}{\tilde{\epsilon}_n^d}\tilde{\eta}\left(\frac{x-y}{\tilde{\epsilon}_n}\right) \leq \frac{1}{\tilde{\epsilon}_n^d}\tilde{\eta}\left(\frac{T_n(x) - T_n(y)}{\epsilon_n}\right) = \frac{\epsilon_n^d}{\tilde{\epsilon}_n^d}\tilde{\eta}_{\epsilon_n}\left(T_n(x) - T_n(y)\right).$$

So,

$$\begin{aligned}&\frac{1}{\tilde{\epsilon}_n}\int_{X^2} \tilde{\eta}_{\tilde{\epsilon}_n}(y - x)\left|\mu^{(n)}(T_n(x)) - \mu^{(n)}(T_n(y))\right|\rho(x)\rho(y)\,\mathrm{d}y\,\mathrm{d}x \\ &\quad\leq \frac{\epsilon_n^d}{\tilde{\epsilon}_n^{d+1}}\int_{X^2} \tilde{\eta}_{\epsilon_n}\left(T_n(x) - T_n(y)\right)\left|\mu^{(n)}(T_n(x)) - \mu^{(n)}(T_n(y))\right|\rho(x)\rho(y)\,\mathrm{d}y\,\mathrm{d}x \\ &\quad= \frac{\epsilon_n^d}{\tilde{\epsilon}_n^{d+1}}\sum_{i,j}\tilde{\eta}_{\epsilon_n}\left(\xi_i - \xi_j\right)\left|\mu^{(n)}(\xi_i) - \mu^{(n)}(\xi_j)\right| \quad \text{by (2.9)} \\ &\quad= \frac{\epsilon_n^{d+1}}{\tilde{\epsilon}_n^{d+1}}\mathcal{E}_n(\mu^{(n)};\tilde{\eta}) \\ &\quad\leq \frac{\epsilon_n^{d+1}}{\tilde{\epsilon}_n^{d+1}}\mathcal{E}_n(\mu^{(n)};\eta).\end{aligned}$$

It follows that the second term is also uniformly bounded in n. □

6.4 Γ-Convergence

The main result of this section is Theorem 6.4.1, which states that $\mathcal{E}_n$ Γ-converges to $\mathcal{E}_\infty$ for almost every sequence $\xi_1, \xi_2, \dots$ which we state now. The proof is a consequence of Lemmas 6.4.3 and 6.4.4. The proofs follow closely from [69] who in turn based their methodology on [5] and [129]. In particular in [69] the authors show the Γ-convergence of the second term.

Theorem 6.4.1. *Under the same conditions as Theorem 6.2.2*

$$\mathcal{E}_\infty = \Gamma\text{-}\lim_{n\to\infty}\mathcal{E}_n$$

in the TL^1 sense and with probability one.

We start with an elementary observation on η.

Proposition 6.4.2. *Let $\eta : \mathbb{R}^d \to [0,\infty)$ satisfy*

$$|\eta(x) - \eta(y)| \leq L \max\{|\eta(x)|, |\eta(y)|\}|x-y|$$

for any $x, y \in \mathbb{R}^d$. If $w, z \in \mathbb{R}^d$, $T_n : \mathbb{R}^d \to \mathbb{R}^d$ and $\epsilon > 0$ then

$$\frac{1}{1 + \frac{2L\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon}} \eta\left(\frac{w-z}{\epsilon}\right) \leq \eta\left(\frac{T_n(w) - T_n(z)}{\epsilon}\right) \leq \frac{1}{1 - \frac{2L\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon}} \eta\left(\frac{z-z}{\epsilon}\right).$$

Proof. A short calculation gives

$$\begin{aligned}
\eta\left(\frac{T_n(w) - T_n(z)}{\epsilon}\right) &\geq \eta\left(\frac{w-z}{\epsilon}\right) \\
&\quad - L \max\left\{\left|\eta\left(\frac{w-z}{\epsilon}\right)\right|, \left|\eta\left(\frac{T_n(w)-T_n(z)}{\epsilon}\right)\right|\right\} \frac{2\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon} \\
&\geq \eta\left(\frac{w-z}{\epsilon}\right) - \frac{2L\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon} \eta\left(\frac{w-z}{\epsilon}\right) \\
&\quad + \frac{4L^2\|T_n - \mathrm{Id}\|^2_{L^\infty(X)}}{\epsilon^2} \max\left\{\left|\eta\left(\frac{w-z}{\epsilon}\right)\right|, \left|\eta\left(\frac{T_n(w)-T_n(z)}{\epsilon}\right)\right|\right\}.
\end{aligned}$$

By induction we have

$$\begin{aligned}
&\eta\left(\frac{T_n(w) - T_n(z)}{\epsilon}\right) \\
&\qquad \geq \eta\left(\frac{w-z}{\epsilon}\right)\left(1 - \frac{2L\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon} + \frac{4L^2\|T_n - \mathrm{Id}\|^2_{L^\infty(X)}}{\epsilon^2} - \dots\right) \\
&\qquad = \frac{1}{1 + \frac{2L\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon}} \eta\left(\frac{z-z}{\epsilon}\right).
\end{aligned}$$

Similarly for the other inequality. □

We now proceed to the lim inf inequality.

Lemma 6.4.3 (The lim inf inequality)**.** *Under the same conditions as Theorem 6.2.2 if $\mu \in L^1(X)$ and $\mu^{(n)} \to \mu$ in TL^1 then*

$$\mathcal{E}_\infty(\mu) \leq \liminf_{n\to\infty} \mathcal{E}_n(\mu^{(n)})$$

with probability one.

Proof. Let $\mu^{(n)} \in L^1(\Psi_n), \mu \in L^1(X)$ with $\mu^{(n)} \to \mu$ in TL^1. Let $\nu^{(n)} = \mu^{(n)} \circ T_n \in L^1(X)$ where $T_n : X \to \Psi_n$ is as in Theorem 2.5.1 (with probability one) so $\nu^{(n)} \to \mu$ in $L^1(X)$. Without loss of generality we assume that

$$\liminf_{n\to\infty} \mathcal{E}_n(\mu^{(n)}) < \infty$$

else there is nothing to prove. By Theorem 6.3.1 $\mu \in L^1(X; \{0,1\})$ hence the proof is complete if

$$\liminf_{n\to\infty} GTV_n(\mu^{(n)}) \geq TV(\mu; \rho, \eta) \tag{6.12}$$

where GTV_n is defined by (6.3).

We now show (6.12) in three steps by starting with stronger assumptions than needed and progressively relaxing the conditions.

Step 1. Assume η is has compact support and ρ is Lipschitz continuous.

Step 2. Remove the compact support condition on η whilst ρ is still assumed Lipschitz.

Step 3. Assume η is any admissible function and ρ is continuous.

Step 1. Let η have compact support in $B(0, M)$ and let X' be a compact subset of X. Then

$$\begin{aligned}
&GTV_n(\mu^{(n)}) \\
&= \frac{1}{\epsilon_n}\frac{1}{n^2}\sum_{i,j} W_{ij}\left|\mu^{(n)}(\xi_i) - \mu^{(n)}(\xi_j)\right| \\
&= \frac{1}{\epsilon_n}\int_{X^2} \eta_{\epsilon_n}(T_n(x) - T_n(y))\left|\nu^{(n)}(x) - \nu^{(n)}(y)\right| \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \quad \text{using (2.9)} \\
&\geq \frac{1}{\epsilon_n}\int_{X'}\int_{X} \eta_{\epsilon_n}(T_n(x) - T_n(y))\left|\nu^{(n)}(x) - \nu^{(n)}(y)\right| \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{\epsilon_n}\int_{X'}\int_{y+\epsilon_n z\in X} \eta\left(\frac{T_n(y+\epsilon_n z) - T_n(y)}{\epsilon_n}\right)\left|\nu^{(n)}(y+\epsilon_n z) - \nu^{(n)}(y)\right| \rho^2(y)\,\mathrm{d}z\,\mathrm{d}y + a_n
\end{aligned}$$

where

$$\begin{aligned}
a_n = \frac{1}{\epsilon_n}\int_{X'}\int_{y+\epsilon_n z\in X} &\eta\left(\frac{T_n(y+\epsilon_n z) - T_n(y)}{\epsilon_n}\right)\left|\nu^{(n)}(y+\epsilon_n z) - \nu^{(n)}(y)\right| \\
&\times \rho(y)\left(\rho(y+\epsilon_n z) - \rho(y)\right)\,\mathrm{d}z\,\mathrm{d}y.
\end{aligned}$$

The following shows $a_n = O(\epsilon_n)$:

$$\begin{aligned}
|a_n| &\leq M\mathrm{Lip}(\rho)\int_X\int_{y+\epsilon_n z\in X} \eta\left(\frac{T_n(y+\epsilon_n z) - T_n(y)}{\epsilon_n}\right) \\
&\qquad\times\left|\nu^{(n)}(y+\epsilon_n z) - \nu^{(n)}(y)\right|\rho(y)\,\mathrm{d}z\,\mathrm{d}y \\
&\leq \frac{M\mathrm{Lip}(\rho)}{\inf_{x\in X}\rho(x)}\epsilon_n GTV_n(\mu^{(n)}).
\end{aligned}$$

Returning to GTV_n and applying Proposition 6.4.2

$$\begin{aligned}
&GTV_n(\mu^{(n)}) \\
&\geq \frac{1}{1+\frac{2L\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n}}\frac{1}{\epsilon_n}\int_{X'}\int_{y+\epsilon_n z\in X}\eta(z)\left|\nu^{(n)}(y+\epsilon_n z) - \nu^{(n)}(y)\right|\rho^2(y)\,\mathrm{d}z\,\mathrm{d}y + o(1) \\
&= \frac{1}{1+\frac{2L\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n}}\int_{\mathbb{R}^d}\eta(z)f_n(z)\,\mathrm{d}z - b_n + o(1)
\end{aligned}$$

where $f_n(z) = \frac{1}{\tilde{\epsilon}_n}\int_{X'} \left|\nu^{(n)}(y+\tilde{\epsilon}_n z) - \nu^{(n)}(y)\right| \rho^2(y)\,\mathrm{d}y$ and

$$b_n = \frac{1}{1+\frac{2L\|T_n-\mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n}} \int_{X'}\int_{y+\tilde{\epsilon}_n\hat{z}\notin X} \eta(z)\left|\nu^{(n)}(y)\right|\rho^2(y)\,\mathrm{d}\hat{z}\,\mathrm{d}y.$$

Since X' and X^c are both closed and disjoint then $\tau := \mathrm{dist}(X', X^c) > 0$ and therefore if $y \in X'$ and $y + \tilde{\epsilon}_n\hat{z} \notin X$ then $\tilde{\epsilon}_n|\hat{z}| \geq \tau$. Since η has compact support then for n sufficiently large $\eta(\hat{z}) = 0$ for all $|\hat{z}| \geq \tau/\epsilon$. Therefore $b_n = 0$ for n sufficiently large.

By applying Fatou's lemma, Proposition 2.4.1 and $\frac{1}{1+\frac{2L\|T_n-\mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n}} \to 1$ one has

$$\liminf_{n\to\infty} GTV_n(\mu^{(n)}) \geq \int_{\mathbb{R}^d} \eta(z)TV_z(\mu;\rho,X')\,\mathrm{d}z = TV(\mu;\rho,\eta,X')$$

Now take $X' \to X$ and therefore $TV(\mu;\rho,\eta,X') \to TV(\mu;\rho,\eta,X) := TV(\mu;\rho,\eta)$ by the monotone convergence theorem.

Step 2. Let η be an admissible function and define

$$\eta^{(l)}(x) = \max\left\{\eta(x) - \frac{1}{l}, 0\right\}.$$

Note that $\eta^{(l)}$ is an increasing sequence of functions satisfing the same strengthened Lipschitz condition as η and converging (Lebesgue) almost everywhere to η as $l \to \infty$. We claim that $\eta^{(l)}$ has compact support so we may apply step 1. Since $\eta(x) \leq \eta^+(|x|)$ and $\eta^+(|x|)$ is decreasing there exists M such that $\eta(x) \leq \eta^+(|x|) \leq \frac{1}{l}$ for all $|x| \geq M$. Hence $\eta^{(l)}(x) = 0$ for $|x| \geq M$.

Now let $GTV_n(\cdot;\eta)$ be the graph total variation using the interaction potential η. By step 1:

$$TV(\mu;\rho,\eta^{(l)}) \leq \liminf_{n\to\infty} GTV_n(\mu^{(n)};\eta^{(l)}) \leq \liminf_{n\to\infty} GTV_n(\mu^{(n)};\eta).$$

By the monotone convergence theorem

$$\sigma_l(\nu) = \int_{\mathbb{R}^d} \eta^{(l)}(x)|x\cdot\nu|\,\mathrm{d}x \to \int_{\mathbb{R}^d} \eta(x)|x\cdot\nu|\,\mathrm{d}x = \sigma(\nu).$$

And therefore in light of Theorem 2.4.2 and the monotone convergence theorem:

$$TV(\mu;\rho,\eta^{(l)}) \to TV(\mu;\rho,\eta) \quad \text{as } l \to \infty$$

since σ_l converges monotonically.

Step 3. Denote the dependence of ρ on GTV_n by $GTV_n(\cdot;\rho)$. Assume $\rho : X \to [0,\infty)$ is continuous and let $\rho_k : \mathbb{R}^d \to [0,\infty)$ be defined by

$$\rho_k(x) = \begin{cases} \inf_{y\in X}\left(\rho(y) + k|x-y|\right) & \text{if } x \in X \\ 0 & \text{otherwise.} \end{cases}$$

Clearly $\rho_k(x) \le \rho(x)$ for all $x \in X$. Then it follows that

$$\rho_k(x) \ge \inf \left\{ \rho(y) : y \in B\left(x, \frac{\rho(x)}{k}\right) \right\}.$$

As ρ is continuous on X then $\rho_k(x) \to \rho(x)$ for each $x \in X$. It is also clear that $\rho_k(x) \ge \inf_{x\in X} \rho(x) > 0$. Furthermore, for $x, z \in X$

$$\begin{aligned}
\rho_k(x) - \rho_k(z) &= \inf_{y_1\in X} \sup_{y_2\in X} \rho(y_1) - \rho(y_2) + k\left(|x-y_1| - |z-y_2|\right) \\
&\le \sup_{y_2\in X} k\left(|x-y_2| - |z-y_2|\right) \\
&\le k|x-z|
\end{aligned}$$

so ρ_k is Lipschitz in X. By step 2

$$\liminf_{n\to\infty} GTV_n(\mu^{(n)};\rho) \ge \liminf_{n\to\infty} GTV_n(\mu^{(n)};\rho_k) \ge TV(\mu;\rho_k,\eta).$$

By Theorem 2.4.2

$$TV(\mu;\rho_k,\eta) = \int_X \rho_k^2(x)\sigma(x)\ \hat{\lambda}(\mathrm{d}x).$$

And therefore by the monotone convergence theorem one has

$$\lim_{k\to\infty} TV(\mu;\rho_k,\eta) = \int_X \rho(x)^2\sigma(x)\ \hat{\lambda}(\mathrm{d}x) = TV(\mu;\rho,\eta)$$

which completes the proof. □

For $\mu \notin L^1(X;\{0,1\})$ the recovery sequence is trivial as $\mathcal{E}_\infty(\mu) = \infty$. For $\mu \in L^1(X;\{0,1\})$ we divide the proof into two steps. First assume that μ is a polyhedral function (defined below) and we show the existence of a recovery sequence. Then we extend the result for any function of bounded variation in $L^1(X;\{0,1\})$. Recall that $\mathcal{H}^k$ is the k-dimensional Hausdorff measure.

Definition 6.4.1. *A (d-dimensional) polyhedral set in $\mathbb{R}^d$ is an open set F whose boundary is a Lipschitz manifold contained in the union of finitely many affine hyperplanes. We say $\mu \in BV(X;\{0,1\})$ is a polyhedral function if there exists a polyhedral set F such that ∂F is transversal to ∂X (i.e. $\mathcal{H}^{d-1}(\partial F \cup \partial X) = 0$) and $\mu(x) = 1$ for $x \in X \cap F$, $\mu(x) = 0$ for $x \in X \setminus F$.*

Lemma 6.4.4 (The existence of a recovery sequence for Theorem 6.4.1)**.** *Under the same conditions as Theorem 6.2.2 for any $\mu \in L^1(X)$ there exists a sequence $\mu^{(n)} \to \mu$ in TL^1 such that*

$$\mathcal{E}_\infty(\mu) \ge \limsup_{n\to\infty} \mathcal{E}_n(\mu^{(n)}) \tag{6.13}$$

with probability one.

Proof. Without loss of generality assume $\mu \in BV(X;\{0,1\})$. By the following argument it is enough to prove the lemma for polyhedral functions. Suppose the lemma holds for polyhedral functions and let $\mu \in BV(X;\rho,\eta) \cap L^1(X;\{0,1\})$. There exists a sequence of smooth sets F_n such that $\mu^{(n)} := \mathbb{I}_{F_n} \to \mu$ in L^1 and $TV(\mu^{(n)};\rho,\eta) \to TV(\mu;\rho,\eta)$, for example see [111, Section 9.1.3]. And therefore by approximating each $\mu^{(n)}$ by a polyhedral function (in L^1 and TV) there exists a sequence of polyhedral functions $\zeta^{(n)}$ such that

$$\zeta^{(n)} \to \mu \text{ in } L^1(X) \qquad \text{and} \qquad TV(\zeta^{(n)};\rho,\eta) \to TV(\mu;\rho,\eta).$$

By assumption that the lemma holds for polyhedral functions and a diagonalization argument we can then conclude the existence of a sequence $\nu^{(n)}$ such that $\limsup_{n\to\infty} GTV_n(\nu^{(n)}) \leq TV(\mu;\rho,\eta)$ as required.

Therefore assume $\mu \in BV(X;\{0,1\})$ is a polyhedral function corresponding to the polyhedral set F, i.e. $\mu = \mathbb{I}_F$. Let $\mu^{(n)}$ be the restriction of μ to Ψ_n. Define T_n as in Theorem 2.5.1 and use it to create a partition of X, $X = \cup_{i=1}^n T_n^{-1}(\xi_i)$. If $x, y \in T_n^{-1}(\xi_i)$ then

$$|x-y| \leq |x - T_n(x)| + |T_n(x) - T_n(y)| + |T_n(y) - y| \leq 2\|\mathrm{Id} - T_n\|_{L^\infty(X)}.$$

Let $x \in T_n^{-1}(\xi_i)$ and assume $\mathrm{dist}(\partial F, x) > 2\|\mathrm{Id} - T_n\|_{L^\infty}$. If $y \in T_n^{-1}(\xi_i)$ then $\mu(y) = \mu(x)$ (since $T_n^{-1}(\xi_i) \subset B(x, 2\|\mathrm{Id} - T_n\|_{L^\infty(X)})$ and $B(x, 2\|\mathrm{Id} - T_n\|_{L^\infty(X)}) \cap \partial F = \emptyset$). Therefore

$$\int_{T_n^{-1}(\xi_i)} \left| \mu^{(n)}(T_n(y)) - \mu(y) \right| \, P(\mathrm{d}y) = 0.$$

In particular

$$\int_X \left| \mu^{(n)}(T_n(y)) - \mu(y) \right| \, P(\mathrm{d}y) = \int_{X_n} \left| \mu^{(n)}(T_n(y)) - \mu(y) \right| \, P(\mathrm{d}y) \leq \|\rho\|_{L^\infty(X)} \mathrm{Vol}(X_n)$$

where

$$X_n = \left\{ y \in X \; : \; \mathrm{dist}(\partial F, y) \leq 2\|\mathrm{Id} - T_n\|_{L^\infty(X)} \right\}. \tag{6.14}$$

Clearly $\mathrm{Vol}(X_n) = O(\|\mathrm{Id} - T_n\|_{L^\infty(X)}) = o(1)$ and therefore $\mu^{(n)} \to \mu$ in TL^1. Define $\nu^{(n)} = \mu^{(n)} \circ T_n$ then since $\nu^{(n)}, \mu \in L^1(X;\{0,1\})$ we have that (6.13) is equivalent to

$$TV(\mu;\rho,\eta) \geq \limsup_{n\to\infty} GTV_n(\mu^{(n)}).$$

We complete the proof in three steps.

Step 1. Assume that η is has compact support and ρ is Lipschitz continuous.

Step 2. Now let η be any admissible function with ρ still Lipschitz continuous.

Step 3. Finally let η be any admissible functions and ρ any continuous function satisfying the criteria in the lemma.

Step 1. By application of Proposition 6.4.2

$$\begin{aligned}
&GTV_n(\mu^{(n)}) \\
&= \frac{1}{\epsilon_n}\int_{X^2} \eta_{\epsilon_n}(T_n(x)-T_n(y)) \left|\nu^{(n)}(x)-\nu^{(n)}(y)\right| \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
&\qquad \leq \frac{1}{1-\frac{2L\|T_n-\mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n}} \frac{1}{\epsilon_n^{d+1}} \int_{X^2} \eta\left(\frac{x-y}{\epsilon_n}\right) \left|\nu^{(n)}(x)-\nu^{(n)}(y)\right| \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
&\qquad = \frac{1}{1-\frac{2L\|T_n-\mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n}} CTV_n(\nu^{(n)})
\end{aligned}$$

with CTV_n defined below. Let us approximate μ by a sequence $\zeta^{(n)} \in C^\infty(\mathbb{R}^d)\cap BV(X)$ such that $\zeta^{(n)} \to \mu$ in $L^1(X)$ and $TV(\zeta^{(n)};\rho,\eta) \to TV(\mu;\rho,\eta)$. Without loss of generality assume that $\zeta^{(n)}(x)=0$ for all $x\in\mathbb{R}^d\setminus X$ and $\|\mu-\zeta^{(n)}\|_{L^1(X)} = o(\epsilon_n)$. Then

$$\begin{aligned}
CTV_n(\zeta^{(n)}) &:= \frac{1}{\epsilon_n}\int_{X^2} \eta_{\epsilon_n}(x-y)\left|\zeta^{(n)}(x)-\zeta^{(n)}(y)\right|\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{\epsilon_n}\int_{X^2} \eta_{\epsilon_n}(x-y)\left|\int_0^1 \nabla\zeta^{(n)}(y+\xi(x-y))\cdot(x-y)\,\mathrm{d}\xi\right|\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
&\leq \frac{1}{\epsilon_n}\int_X\int_0^1\int_X \eta_{\epsilon_n}(x-y)\left|\nabla\zeta^{(n)}(y+\xi(x-y))\cdot(x-y)\right|\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}\xi\,\mathrm{d}y \\
&= \int_X\int_0^1\int_{\mathcal{Z}_{nh\xi}} \eta(z)\left|\nabla\zeta^{(n)}(h)\cdot z\right|\rho(h+(1-\xi)\epsilon_n z)\rho(h-\epsilon_n\xi z)\,\mathrm{d}z\,\mathrm{d}\xi\,\mathrm{d}h \\
&\leq TV(\zeta^{(n)};\rho,\eta)+c_n
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{Z}_{nh\xi} &= \left\{z\in\mathbb{R}^d : h+(1-\xi)\epsilon_n z\in X \text{ and } h-\epsilon_n\xi z\in X\right\} \\
c_n &= \int_X\int_0^1\int_{\mathcal{Z}_{nh\xi}} \eta(z)\left|\nabla\zeta^{(n)}(h)\cdot z\right|\left(\rho(h+(1-\xi)\epsilon_n z)\rho(h-\epsilon_n\xi z)-\rho^2(h)\right)\,\mathrm{d}z\,\mathrm{d}\xi\,\mathrm{d}h.
\end{aligned}$$

If

$$\lim_{n\to\infty} c_n = 0 \tag{6.15}$$

$$\text{and} \quad \lim_{n\to\infty}\left|CTV_n(\nu^{(n)})-CTV_n(\zeta^{(n)})\right| = 0 \tag{6.16}$$

then

$$\limsup_{n\to\infty} CTV_n(\nu^{(n)}) = \limsup_{n\to\infty} CTV_n(\zeta^{(n)}) \leq \limsup_{n\to\infty} TV(\zeta^{(n)};\rho,\eta) = TV(\mu;\rho,\eta).$$

We now show (6.15). It is an easy exercise to show

$$\begin{aligned}
\left|\rho^2(h)-\rho(h+(1-\xi)\epsilon_n z)\rho(h+\epsilon_n\xi z)\right| &\leq \|\rho\|_{L^\infty(X)}\mathrm{Lip}(\rho)\left(|(1-\xi)\epsilon_n z|+|\epsilon_n\xi z|\right) \\
&\leq 2\epsilon_n\|\rho\|_{L^\infty(X)}\mathrm{Lip}(\rho)M
\end{aligned}$$

where $\mathrm{spt}(\eta) \subseteq B(0,M)$. Then

$$\begin{aligned} |c_n| &\leq 2\|\rho\|_{L^\infty(X)} \mathrm{Lip}(\rho) \epsilon_n M \int_X \int_{\mathbb{R}^d} \eta(z) \left| \nabla \zeta^{(n)}(h) \cdot z \right| \, \mathrm{d}z \, \mathrm{d}h \\ &\leq \frac{2\|\rho\|_{L^\infty(X)} \mathrm{Lip}(\rho) \epsilon_n M}{\inf_{x \in X} \rho^2(x)} \int_X \int_{\mathbb{R}^d} \eta(z) \left| \nabla \zeta^{(n)}(h) \cdot z \right| \rho^2(h) \, \mathrm{d}z \, \mathrm{d}h \\ &= \frac{2\|\rho\|_{L^\infty(X)} \mathrm{Lip}(\rho) \epsilon_n M}{\inf_{x \in X} \rho^2(x)} TV(\zeta^{(n)}; \rho, \eta). \end{aligned}$$

To complete step 1 we show (6.16). This follows by:

$$\begin{aligned} &\left| CTV_n(\nu^{(n)}) - CTV_n(\zeta^{(n)}) \right| \\ &\leq \frac{\|\rho\|^2_{L^\infty(X)}}{\epsilon_n} \int_{X^2} \eta_{\epsilon_n}(y - x) \left(\left| \nu^{(n)}(x) - \zeta^{(n)}(x) \right| + \left| \nu^{(n)}(y) - \zeta^{(n)}(y) \right| \right) \, \mathrm{d}x \, \mathrm{d}y \\ &\leq \frac{2\|\eta\|_{L^\infty(\mathbb{R}^d)} \mathrm{Vol}(X) \|\rho\|^2_{L^\infty(X)}}{\epsilon_n} \int_X \left| \nu^{(n)}(x) - \zeta^{(n)}(x) \right| \, \mathrm{d}y \\ &\leq \frac{2\|\eta\|_{L^\infty(\mathbb{R}^d)} \mathrm{Vol}(X) \|\rho\|^2_{L^\infty(X)}}{\epsilon_n} \left(\|\nu^{(n)} - \mu\|_{L^1(X)} + \|\mu - \zeta^{(n)}\|_{L^1(X)} \right) \\ &\to 0 \end{aligned}$$

where the last line follows as $\|\mu - \zeta^{(n)}\|_{L^1(X)} = o(\epsilon_n)$ and $\|\nu^{(n)} - \mu\|_{L^1(X)} = O(\|T_n - \mathrm{Id}\|_{L^\infty(X)}) = o(\epsilon_n)$.

Step 2. Let η be an admissible function and define $\eta^{(i)}$ by

$$\eta^{(i)}(x) = \begin{cases} \eta(x) & \text{if } x \in B(0,i) \\ \max\left\{ \eta(y) - \frac{|x-y|}{L}, 0 \right\} & \text{if } x \in B(0, i + \frac{1}{L}) \setminus B(0,i) \\ 0 & \text{otherwise} \end{cases}$$

for $y = \mathrm{argmin}_{z \in B(0,i)} |x - z|$ where $\eta^{(i)}$ has been constructed so that it satisfies $\eta^{(i)} \leq \eta$, $\eta^{(i)} = \eta$ on $B(0,i)$, $\eta^{(i)}$ satisfies the same strengthened Lipschitz condition as η and η has compact support.

By step 1

$$TV(\mu; \rho, \eta) \geq TV(\mu; \rho, \eta^{(i)}) \geq \limsup_{n \to \infty} GTV_n(\mu^{(n)}; \eta^{(i)}).$$

We can write

$$GTV_n(\mu^{(n)}; \eta) \leq GTV_n(\mu^{(n)}; \eta^{(i)}) + d_{n,i}$$

where

$$\begin{aligned} d_{n,i} &= \frac{1}{\epsilon_n} \int_{X_{n,i}} \eta_{\epsilon_n}(T_n(x) - T_n(y)) \left| \nu^{(n)}(x) - \nu^{(n)}(y) \right| \rho(x) \rho(y) \, \mathrm{d}x \, \mathrm{d}y, \\ X_{n,i} &= \left\{ (x,y) \in X^2 : \left| \frac{T_n(x) - T_n(y)}{\epsilon_n} \right| \geq i \right\} \end{aligned}$$

and $\nu^{(n)} = \mu^{(n)} \circ T_n = \mu \circ T_n$. We will show $\limsup_{i\to\infty} \limsup_{n\to\infty} d_{n,i} \geq 0$. Assume n is sufficiently large so that $\frac{2\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n} \leq \frac{1}{2} \leq \frac{i}{2}$. Then for $(x,y) \in X_{n,i}$

$$\left|\frac{x-y}{\epsilon_n}\right| \leq \left|\frac{T_n(x)-T_n(y)}{\epsilon_n}\right| + \frac{2\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n} \leq 2\left|\frac{T_n(x)-T_n(y)}{\epsilon_n}\right|.$$

In particular

$$\begin{aligned}\eta\left(\frac{T_n(x)-T_n(y)}{\epsilon_n}\right) &\leq \eta^+\left(\left|\frac{T_n(x)-T_n(y)}{\epsilon_n}\right|\right) \leq \eta^+\left(\left|\frac{x-y}{2\epsilon_n}\right|\right) \\ &\leq \gamma_1 \eta^-\left(\left|\frac{x-y}{2\gamma_2\epsilon_n}\right|\right) \leq \gamma_1 \eta\left(\frac{x-y}{2\gamma_2\epsilon_n}\right).\end{aligned}$$

Similarly if $(x,y) \in X_{n,i}$ then

$$\left|\frac{x-y}{\epsilon_n}\right| \geq \left|\frac{T_n(x)-T_n(y)}{\epsilon_n}\right| - \frac{2\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n} \geq \frac{i}{2}.$$

Hence

$$X_{n,i} \subseteq \hat{X}_{n,i} := \left\{(x,y) \in X^2 : \left|\frac{x-y}{\epsilon_n}\right| \geq \frac{i}{2}\right\}.$$

Therefore

$$\begin{aligned}d_{n,i} &\leq \frac{\gamma_1}{\epsilon_n^{d+1}} \int_{\hat{X}_{n,i}} \eta\left(\frac{x-y}{2\gamma_2\epsilon_n}\right) \left|\nu^{(n)}(x) - \nu^{(n)}(y)\right| \rho(x)\rho(y) \; \mathrm{d}x \; \mathrm{d}y \\ &\leq \frac{\gamma_1}{\epsilon_n^{d+1}} \int_{\hat{X}_{n,i}} \eta\left(\frac{x-y}{2\gamma_2\epsilon_n}\right) |\mu(x) - \mu(y)| \rho(x)\rho(y) \; \mathrm{d}x \; \mathrm{d}y \\ &\qquad + \frac{\gamma_1\|\rho\|^2_{L^\infty(X)}}{\epsilon_n^{d+1}} \int_{\hat{X}_{n,i}} \eta\left(\frac{x-y}{2\gamma_2\epsilon_n}\right) \left(\left|\nu^{(n)}(x) - \nu^{(n)}(y)\right| - |\mu(x) - \mu(y)|\right) \; \mathrm{d}x \; \mathrm{d}y.\end{aligned}$$

Let

$$\begin{aligned}e_{n,i} &= \frac{\gamma_1}{\epsilon_n^{d+1}} \int_{\hat{X}_{n,i}} \eta\left(\frac{x-y}{2\gamma_2\epsilon_n}\right) |\mu(x) - \mu(y)| \rho(x)\rho(y) \; \mathrm{d}x \; \mathrm{d}y \\ f_{n,i} &= \frac{\gamma_1\|\rho\|^2_{L^\infty(X)}}{\epsilon_n^{d+1}} \int_{\hat{X}_{n,i}} \eta\left(\frac{x-y}{2\gamma_2\epsilon_n}\right) \left(\left|\nu^{(n)}(x) - \nu^{(n)}(y)\right| - |\mu(x) - \mu(y)|\right) \; \mathrm{d}x \; \mathrm{d}y.\end{aligned}$$

By changing coordinates one has

$$f_{n,i} \leq \frac{2(2\gamma_2)^d\gamma_1\|\rho\|^2_{L^\infty(X)}}{\epsilon_n} \int_{\mathbb{R}^d} \eta(z) \; \mathrm{d}z \int_{X_n} \left|\nu^{(n)}(x) - \mu(x)\right| \; \mathrm{d}x$$

where X_n is defined by (6.14) and since $\int_{X_n} \left|\nu^{(n)}(x) - \mu(x)\right| \; \mathrm{d}x = O(\mathrm{Vol}(X_n)) = O(\|T_n - \mathrm{Id}\|_{L^\infty(X)}) = o(\epsilon_n)$ then $\limsup_{n\to\infty} f_{n,i} = 0$.

For $e_{n,i}$ we have

$$\begin{aligned} e_{n,i} &\leq \frac{(2\gamma_2)^d\gamma_1\|\rho\|_{L^\infty(X)}}{\epsilon_n \inf_{x\in X}\rho(x)} \int_{|z|\geq \gamma_2 i}\int_X \eta(z)\left|\mu(y+2\gamma_2\epsilon_n z)-\mu(y)\right|\rho^2(y)\,\mathrm{d}y\,\mathrm{d}z \\ &\leq \frac{(2\gamma_2)^{d+1}\gamma_1\|\rho\|_{L^\infty(X)}}{\inf_{x\in X}\rho^2(x)} \int_{|z|\geq\gamma_2 i} \eta(z) TV_z(\mu;\rho)\,\mathrm{d}z \end{aligned}$$

where the above follows from

$$\int_X \left|\mu(x+\epsilon z)-\mu(x)\right|\rho^2(x)\,\mathrm{d}x \leq \epsilon TV_z(\mu;\rho) \quad \text{for Lebesque-a.e. } z\in X$$

and TV_z is defined by (2.4). The proof is analogous to the well known result: if $\mu \in \hat{BV}(\mathbb{R}^d)$ then $\int_{\mathbb{R}^d} |\mu(x+h)-\mu(x)|\,\mathrm{d}x \leq |h|\hat{TV}(\mu)$, see for example [97, Lemma 13.33]. Hence

$$\limsup_{i\to\infty}\limsup_{n\to\infty} e_{n,i} = 0$$

which completes step 2.

Step 3. Let $GTV(\cdot;\rho)$ be the graph total variation defined using ρ. Let ρ be continuous but not necessarily Lipschitz and define $\rho_k : \mathbb{R}^d \to [0,\infty)$ by

$$\rho_k(x) = \begin{cases} \sup_{y\in X}\rho(y) - k|x-y| & \text{if } x\in X \\ 0 & \text{otherwise.} \end{cases}$$

Similarly to Lemma 6.4.3 step 3 we can check that ρ_k is bounded above and below by positive constants, Lipschitz continuous on X and converges pointwise to ρ from above. We have

$$\limsup_{n\to\infty} GTV_n(\mu^{(n)};\rho) \leq \limsup_{n\to\infty} GTV_n(\mu^{(n)};\rho_k) \leq TV(\mu;\rho_k,\eta).$$

By the monotone convergence theorem and Theorem 2.4.2 we have

$$\lim_{k\to\infty} TV(\mu;\rho_k,\eta) = TV(\mu;\rho,\eta)$$

which completes the proof. □

6.5 Preliminary Results for the Rate of Convergence

In this section we fix $\mu = \mathbb{I}_E$ where E is a polyhedral set and look at the convergence

$$\mathbb{E}\left|\mathcal{E}_n(\mu)-\mathcal{E}_\infty(\mu)\right|^2 = \mathbb{E}\left|GTV_n(\mu) - TV(\mu;\rho,\eta)\right|^2 \to 0.$$

For convenience (and to reduce bias) let us redefine the normalization on GTV_n so that

$$GTV_n(\mu) = \frac{1}{\epsilon_n}\frac{1}{n(n-1)}\sum_{i,j}\eta_{\epsilon_n}(\xi_i-\xi_j)\left|\mu(\xi_i)-\mu(\xi_j)\right|. \tag{6.17}$$

For simplicity we make the following simplifications. Assume $X = (0,1)^d$ where $d \geq 2$ and that $\rho \equiv 1$ on X. We use an isotropic interaction potential $\eta = \mathbb{I}_{B(0,1)}$. These assumptions allow us to greatly simplify the calculations without losing the important features of the problem. We start with the convergence of the expectation.

Theorem 6.5.1. *Let $X = (0,1)^d$ with $d \geq 2$, $\rho \equiv 1$ and ϵ_n be any sequence converging to zero. The data is distributed $\xi_i \overset{\text{iid}}{\sim} \rho$ and let $\Psi_n = \{\xi_i\}_{i=1}^n$. Define $GTV_n : L^1(\Psi_n) \to [0,\infty]$ by* (6.17) *where the weights are given by $W_{ij} = \eta_{\epsilon_n}(\xi_i - \xi_j)$ and $\eta_{\epsilon_n}(x) = \frac{1}{\epsilon_n^d}\mathbb{I}_{|x|\leq\epsilon_n}$. Define $TV(\cdot;\rho,\eta) : L^1(X) \to [0,\infty]$ by* (2.1-2.3)*. Let $\mu = \mathbb{I}_E$ be a polyhedral function. Then*

$$|\mathbb{E}GTV_n(\mu) - TV(\mu;\rho,\eta)| = O(\epsilon_n).$$

Note that we do not need a lower bound on the decay of ϵ_n. By taking the expectation we are immediately in the continuous setting and therefore lose all the graphical structure. Our proof shows that along faces of E and sufficiently far from edges in some sense the expected graph total variation is equal to the total variation. To be more precise let $Y \subset X$ be a strip centered around one of the faces of E with non-zero width. Then one can show for the graph total variation and total variation of $\mu = \mathbb{I}_E$ on Y that $GTV_n(\mu;Y) = TV(\mu;\rho,\eta,Y)$. The discrepancy between $\mathbb{E}GTV_n(\mu)$ and $TV(\mu;\rho,\eta)$ is a consequence of having to approximate corners. In our proof we approximate the corners at a cost of $O(\epsilon_n)$. In particular $GTV_n(\mu)$ is a biased estimator of $TV(\mu;\rho,\eta)$.

Proof of Theorem 6.5.1. Let $\partial E = \sum_{i=1}^N \partial E_i$. We first calculate $TV(\mu;\rho,\eta)$,

$$TV(\mu;\rho,\eta) = \int_{\partial\{\mu=1\}} \sigma(n(x)) \, \mathrm{d}\mathcal{H}^{d-1}(x) = \sum_{i=1}^N |\partial E_i| \, \sigma(n_i)$$

where n_i is the outward unit normal for side ∂E_i and we use $|\cdot|$ to denote the $\mathcal{H}^{d-1}$ measure. Observe

$$\sigma(n_i) = \int_{B(0,1)} |x \cdot n_i| \, \mathrm{d}x = \int_{B(0,1)} |x_d| \, \mathrm{d}x =: \sigma.$$

So $TV(\mu;\rho,\eta) = \sigma|\partial E|$.

Now consider $\mathbb{E}GTV_n(\mu)$,

$$\begin{aligned}\mathbb{E}GTV_n(\mu) &= \frac{1}{\epsilon_n}\int_{(0,1)^d}\int_{(0,1)^d} \eta_{\epsilon_n}(x-y)\,|\mu(x)-\mu(y)| \, \mathrm{d}y \, \mathrm{d}x \\ &= \frac{2}{\epsilon_n^{d+1}} \sum_{i=1}^N \int_{S_i^{(n)}\cap E}\int_{S_i^{(n)}\cap E^c} \mathbb{I}_{|x-y|\leq\epsilon_n} \, \mathrm{d}y \, \mathrm{d}x + O(\epsilon_n)\end{aligned}$$

where $S_i^{(n)}$ is the strip of width $2\epsilon_n$ centered around ∂E_i. Consider $S_i^{(n)}$ then after rotation we may assume

$$\frac{2}{\epsilon_n^{d+1}}\int_{S_i^{(n)}\cap E}\int_{S_i^{(n)}\cap E^c}\mathbb{I}_{|x-y|\leq\epsilon_n}\;\mathrm{d}y\;\mathrm{d}x$$
$$=\frac{2}{\epsilon_n^{d+1}}\int_{\otimes_{j=1}^{d-1}[\epsilon_n,L_j-\epsilon_n]}\int_{-\epsilon_n}^{0}\int_{\otimes_{j=1}^{d-1}[0,L_j]}\int_0^{\epsilon_n}\mathbb{I}_{|x-y|\leq\epsilon_n}\;\mathrm{d}y_d\;\mathrm{d}y_{1:d-1}\mathrm{d}x_d\;\mathrm{d}x_{1:d-1}+O(\epsilon_n)$$

where $\partial E_i=\otimes_{j=1}^{d-1}[0,L_j]\times\{0\}$. Fix $x_{1:d-1}\in\otimes_{j=1}^{d-1}[\epsilon_n,L_j-\epsilon_n]$ and $y_d\in[0,\epsilon_n]$ then elementary geometry reveals that the integral $\int_{-\epsilon_n}^{0}\int_{\otimes_{j=1}^{d-1}[0,L_j]}\mathbb{I}_{|x-y|\leq\epsilon_n}\;\mathrm{d}y_{1:d-1}\mathrm{d}x_d$ is the volume of a segment of the d-dimensional ball. More precisely

$$\int_{-\epsilon_n}^{0}\int_{\otimes_{j=1}^{d-1}[0,L_j]}\mathbb{I}_{|x-y|\leq\epsilon_n}\;\mathrm{d}y_{1:d-1}\mathrm{d}x_d=\int_{B(0,\epsilon_n)}\mathbb{I}_{z_d\geq y_d}\;\mathrm{d}z.$$

Integrating the above over y_d and exchanging the order of integration one has

$$\begin{aligned}\int_{-\epsilon_n}^{0}\int_{\otimes_{j=1}^{d-1}[0,L_j]}\int_0^{\epsilon_n}\mathbb{I}_{|x-y|\leq\epsilon_n}\;\mathrm{d}y_d\;\mathrm{d}y_{1:d-1}\mathrm{d}x_d&=\int_{B(0,\epsilon_n)}\int_0^{\epsilon_n}\mathbb{I}_{z_d\geq y_d}\;\mathrm{d}y_d\;\mathrm{d}z\\&=\int_{B(0,\epsilon_n)}z_d\mathbb{I}_{z_d\geq 0}\;\mathrm{d}z\\&=\frac{\epsilon_n^{d+1}}{2}\int_{B(0,1)}|z_d|\;\mathrm{d}z\\&=\frac{\epsilon_n^{d+1}\sigma}{2}.\end{aligned}$$

Therefore

$$\frac{2}{\epsilon_n^{d+1}}\int_{S_i^{(n)}\cap E}\int_{S_i^{(n)}\cap E^c}\mathbb{I}_{|x-y|\leq\epsilon_n}\;\mathrm{d}y\;\mathrm{d}x=\left(\prod_{j=1}^{d-1}(L_j-2\epsilon_n)\right)\sigma=|\partial E_i|\sigma+O(\epsilon_n).$$

And in particular

$$\mathbb{E}GTV_n(\mu)=TV(\mu;\rho,\eta)+O(\epsilon_n)$$

which completes the proof. □

The above theorem can be developed by looking at higher order expansions. The theorem below gives the two leading terms in $\mathbb{E}\,|GTV_n(\mu)-TV(\mu;\rho,\eta)|^2$. As previously discussed the approximation of the corners leads to an error of $O(\epsilon_n)$. The multiplicative constant of this approximation will depend on the angles between faces of E and therefore will be difficult and not particularly interesting to characterize. The next dominant term gives the convergence (in mean square) of $GTV_n(\mu)$ to $TV(\mu;\rho,\eta)$ along face of E and is therefore of more interest. For this reason we keep track of the constants in order to better understand the convergence of $GTV_n(\mu)$ to $TV(\mu;\rho,\eta)$.

Theorem 6.5.2. *Under the same conditions as Theorem 6.5.1*

$$\mathbb{E}\left|GTV_n(\mu) - TV(\mu;\rho,\eta)\right|^2 = \frac{(n-2)(n-3)}{n(n-1)}\alpha_n^2$$
$$+ \frac{4}{n(n-1)}\left((n-2)|\partial E|V - \left(n-\frac{3}{2}\right)TV(\mu;\rho,\eta)^2\right) + O\left(\frac{1}{n^2\epsilon_n^{d+1}}\right)$$

where $\alpha_n = \mathbb{E}GTV_n(\mu) - TV(\mu;\rho,\eta) = O(\epsilon_n)$ *is the error in the edge terms and*

$$V = \frac{1}{2}\int_{B(0,1)}\int_{B(0,1)} \min\{|z_d|,|y_d|\}\;\mathrm{d}z\;\mathrm{d}y.$$

Proof. We can write

$$\begin{aligned}\mathbb{E}\left|GTV_n(\mu) - TV(\mu;\rho,\eta)\right|^2 &= \mathbb{E}GTV_n(\mu)^2 + TV(\mu;\rho,\eta)^2 - 2TV(\mu;\rho,\eta)\mathbb{E}GTV_n(\mu)\\ &= \mathbb{E}GTV_n(\mu)^2 - TV(\mu;\rho,\eta)^2 - 2TV(\mu;\rho,\eta)\alpha_n.\end{aligned}$$

Let $X_{ij} = \frac{1}{\epsilon_n}\eta_{\epsilon_n}(\xi_i-\xi_j)\left|\mu(\xi_i)-\mu(\xi_j)\right|$ then

$$GTV_n(\mu) = \frac{1}{n(n-1)}\sum_{i,j}X_{ij} \quad \text{and} \quad GTV_n(\mu)^2 = \frac{1}{n^2(n-1)^2}\sum_{i,j,k,l}X_{ij}X_{kl}.$$

Let i,j,k,l be distinct, then $GTV_n(\mu)^2$ has the following contributions:

1. $2n(n-1)$ terms consisting of X_{ij}^2,
2. $4n(n-1)(n-2)$ terms consisting of $X_{ij}X_{ik}$ and
3. $n(n-1)(n-2)(n-3)$ terms consisting of $X_{ij}X_{kl}$.

By independence we have (3): $\mathbb{E}X_{ij}X_{kl} = TV(\mu;\rho,\eta)^2 + 2\alpha_n TV(\mu;\rho,\eta) + \alpha_n^2$. For (1):

$$\begin{aligned}\mathbb{E}X_{ij}^2 &= \frac{1}{\epsilon_n^{2d+2}}\int_{(0,1)^d}\int_{(0,1)^d}\mathbb{I}_{|x-y|\le\epsilon_n}\left|\mu(x)-\mu(y)\right|\;\mathrm{d}x\;\mathrm{d}y\\ &= \frac{1}{\epsilon_n^{d+1}}\mathbb{E}GTV_n(\mu)\\ &= \frac{1}{\epsilon_n^{d+1}}TV(\mu;\rho,\eta) + O\left(\frac{1}{\epsilon_n^d}\right) \quad \text{by Theorem 6.5.1.}\end{aligned}$$

Now consider (2). We have

$$\begin{aligned}\mathbb{E}X_{ij}X_{ik} &= \frac{1}{\epsilon_n^{2d+2}}\int_{(0,1)^d}\int_{(0,1)^d}\int_{(0,1)^d}\mathbb{I}_{|x-y|\le\epsilon_n}\mathbb{I}_{|x-w|\le\epsilon_n}\\ &\qquad\times\left|\mu(x)-\mu(y)\right|\left|\mu(x)-\mu(w)\right|\;\mathrm{d}x\;\mathrm{d}y\;\mathrm{d}w\\ &= \frac{2}{\epsilon_n^{2d+2}}\sum_{l=1}^N\int_{S_l^{(n)}\cap E^c}\int_{S_l^{(n)}\cap E^c}\int_{S_l^{(n)}\cap E}\mathbb{I}_{|x-y|\le\epsilon_n}\mathbb{I}_{|x-w|\le\epsilon_n}\;\mathrm{d}x\;\mathrm{d}y\;\mathrm{d}w + O(\epsilon_n)\\ &= \frac{2}{\epsilon_n^{2d+2}}\sum_{l=1}^N\int_{S_l^{(n)}\cap E}\left(\int_{S_l^{(n)}\cap E^c}\mathbb{I}_{|x-y|\le\epsilon_n}\;\mathrm{d}y\right)^2\;\mathrm{d}x + O(\epsilon_n)\end{aligned}$$

where $S_l^{(n)}$ is as in the proof of Theorem 6.5.1. Considering one $S_l^{(n)}$, after rotating,

$$
\begin{aligned}
&\frac{2}{\epsilon_n^{2d+2}} \int_{S_l^{(n)} \cap E} \left(\int_{S_l^{(n)} \cap E^c} \mathbb{I}_{|x-y| \leq \epsilon_n} \, \mathrm{d}y \right)^2 \mathrm{d}x \\
&\quad = \frac{2}{\epsilon_n^{2d+2}} \int_{\otimes_{r=1}^{d-1} [\epsilon_n, L-\epsilon_n]} \int_{-\epsilon_n}^{0} \left(\int_{B(0,\epsilon_n)} \mathbb{I}_{z_d \geq -x_d} \, \mathrm{d}z \right)^2 \mathrm{d}x_d \, \mathrm{d}x_{1:d-1} + O(\epsilon_n) \\
&\quad = 2 \left(\prod_{r=1}^{d-1} (L_r - 2\epsilon_n) \right) \int_{-1}^{0} \left(\int_{B(0,1)} \mathbb{I}_{z_d \geq -x_d} \, \mathrm{d}z \right)^2 \mathrm{d}x_d + O(\epsilon_n) \\
&\quad = 2 \left(\prod_{r=1}^{d-1} (L_r - 2\epsilon_n) \right) \int_{-1}^{0} \int_{B(0,1)} \int_{B(0,1)} \mathbb{I}_{\min\{z_d, y_d\} \geq -x_d} \, \mathrm{d}z \, \mathrm{d}y \, \mathrm{d}x_d + O(\epsilon_n) \\
&\quad = 2 \left(\prod_{r=1}^{d-1} (L_r - 2\epsilon_n) \right) \int_{B(0,1)} \int_{B(0,1)} \mathbb{I}_{\min\{z_d, y_d\} \geq 0} \min\{z_d, y_d\} \, \mathrm{d}z \, \mathrm{d}y + O(\epsilon_n) \\
&\quad = \frac{1}{2} \left(\prod_{r=1}^{d-1} (L_r - 2\epsilon_n) \right) \int_{B(0,1)} \int_{B(0,1)} \min\{|z_d|, |y_d|\} \, \mathrm{d}z \, \mathrm{d}y \\
&\quad = |\partial E_l| V + O(\epsilon_n)
\end{aligned}
$$

where the second line follows from noticing that the inner integral is the volume of a segment formed by the intersection of a d-dimensional ball centered at the origin with the half plane $\{z_d \geq -x_d\}$. Collecting terms we have

$$
\begin{aligned}
&\mathbb{E} \left| GTV_n(\mu) - TV(\mu; \rho, \eta) \right|^2 \\
&\quad = \frac{2n(n-1)}{n^2(n-1)^2} \frac{1}{\epsilon_n^{d+1}} TV(\mu; \rho, \eta) + O\left(\frac{1}{n^2 \epsilon_n^{d+1}} \right) + \frac{4n(n-1)(n-2)}{n^2(n-1)^2} |\partial E| V \\
&\qquad + O\left(\frac{\epsilon_n}{n} \right) + \frac{n(n-1)(n-2)(n-3)}{n^2(n-1)^2} TV(\mu; \rho, \eta)^2 \\
&\qquad + 2\alpha_n \frac{n(n-1)(n-2)(n-3)}{n^2(n-1)^2} TV(\mu; \rho, \eta) + \alpha_n^2 \frac{n(n-1)(n-2)(n-3)}{n^2(n-1)^2} \\
&\qquad - TV(\mu; \rho, \eta)^2 - 2\alpha_n TV(\mu; \rho, \eta) \\
&\quad = \alpha_n^2 \frac{(n-2)(n-3)}{n(n-1)} + \frac{6-4n}{n(n-1)} TV(\mu; \rho, \eta)^2 + \frac{4(n-2)}{n(n-1)} |\partial E| V \\
&\qquad + 2\alpha_n TV(\mu; \rho, \eta) \frac{6-4n}{n(n-1)} + O\left(\frac{1}{n^2 \epsilon_n^{d+1}} \right)
\end{aligned}
$$

which completes the proof. □

Chapter 7

The Constrained Ginzburg-Landau Functional on Point Clouds

Abstract

We use minimizers of the Ginzburg-Landau functional as a classification tool for data sets Ψ_n where n denotes the number of data points. To obtain non-trivial minimizers it is necessary to impose conditions. In this chapter we study the asymptotics of the minimization problem with a mass constraint. We show that, with probability one, both the minimum and minimizers converge in the data rich limit to the minimum and minimizers of a limiting functional which is closely related to a total variation norm restricted to functions taking binary values and respecting the constraints. We produce similar results for the graph total variation minimization problem. Another approach we consider is to add a data term to the Ginzburg-Landau functional (respectively the graph total variation functional). We give the corresponding convergence results.

7.1 Introduction

In the previous chapter the asymptotics of the unconstrained Ginzburg-Landau functional in a discrete setting were studied (see also [69, 163]). To make these results useful one should impose constraints in order to obtain non-trivial minimizers. A common choice is to impose a mass conservation constraint. Given data $\Psi_n = \{\xi_i\}_{i=1}^n \subset X$ where $X \subset \mathbb{R}^d$ we construct our classifier $\mu : \Psi_n \to \mathbb{R}$ as the minimizer to $\mathcal{E}_n$ (defined in the previous chapter). We assume there are two classes and interpret $\mu(\xi_i) = 0$ as data point ξ_i belongs to class 0 and $\mu(\xi_i) = 1$ as data point ξ_i belongs to class 1. We do not (for finite data sets) impose a hard classification. This means we allow $\mu(\xi_i) \notin \{0, 1\}$.

For mass conservation we impose the condition that $\frac{1}{n}\sum_{i=1}^n \mu(\xi_i) = m$ where $m \in [0, 1]$ is a fixed constant. The main results of this chapter are to show that as $n \to \infty$ the minimizers and minimum of the constrained problem converge, with probability one, to minimizers and minimum of a limiting functional $\mathcal{E}_\infty$. One can also define the graph total variation (GTV) on point clouds and ask the same questions.

Recall that we construct the Ginzburg-Landau functional on the point cloud $\Psi_n = \{\xi_i\}_{i=1}^n$ as follows. We use an anisotropic interaction potential between nodes ξ_i by weighting edges

$$W_{ij} = \eta_\epsilon(\xi_i - \xi_j)$$

where $\eta_\epsilon : \mathbb{R}^d \to [0,\infty)$ and say there is an edge between ξ_i and ξ_j if $W_{ij} > 0$. We assume that η_ϵ is of the form

$$\eta_\epsilon(x) = \frac{1}{\epsilon^d}\eta(x/\epsilon) \tag{7.1}$$

with scaling parameter ϵ and $\eta : \mathbb{R}^d \to [0,\infty)$. One motivational example is $\eta(x) = 1$ for $|x| < 1$ and $\eta(x) = 0$ for $|x| \geq 1$. We scale $\epsilon = \epsilon_n$ to zero as $n \to \infty$ in order to reduce the computational expense and increase the boundary resolution. We define the GTV by

$$GTV_n(\mu) := \frac{1}{\epsilon_n}\frac{1}{n^2}\sum_{i,j} W_{ij}\left|\mu(\xi_i) - \mu(\xi_j)\right|. \tag{7.2}$$

Now we let $V : \mathbb{R} \to [0,\infty)$ be a potential such that states taking the value 0 or 1 are favored, e.g. $V(t) = t^2(t-1)^2$, and we define the Ginzburg-Landau functional by

$$\mathcal{E}_n(\mu) := \frac{1}{\epsilon_n}\frac{1}{n}\sum_{i=1}^n V(\mu(\xi_i)) + \frac{1}{\epsilon_n}\frac{1}{n^2}\sum_{i,j} W_{ij}\left|\mu(\xi_i) - \mu(\xi_j)\right|. \tag{7.3}$$

As $n \to \infty$ our estimators converge in the TL^1 sense (see Section 2.5) to functions on X. We will show that the limiting functional (in the sense of Γ-convergence) is given by $\mathcal{E}_\infty$ which is defined by

$$\mathcal{E}_\infty(\mu) = \begin{cases} \int_{\partial\{\mu=1\}} \sigma(n(x))\rho^2(x)\,\mathrm{d}\mathcal{H}^{d-1}(x) & \text{if } \mu \in L^1(X;\{0,1\}) \\ \infty & \text{otherwise} \end{cases} \tag{7.4}$$

where $n(x)$ is the outward unit normal for the set $\partial\{\mu = 1\}$, $\mathcal{H}^{d-1}$ is the $d-1$ dimensional Hausdorff measure and

$$\sigma(\nu) = \int_{\mathbb{R}^d} \eta(x)|x\cdot\nu|\,\mathrm{d}x. \tag{7.5}$$

We let Θ_n, Θ be the set of functions on Ψ_n, X respectively that satisfy the mass constraint. Then the results of this chapter show that for any sequence $\mu^{(n)}$ of almost minimizers of $\mathcal{E}_n$ in Θ_n with probability one

$$\lim_{n\to\infty}\inf_{\Theta_n}\mathcal{E}_n = \min_{\Theta}\mathcal{E}_\infty$$
$$\mu^{(n)} \to \mu^{(\infty)}$$

for some $\mu^{(\infty)}$ that minimizes $\mathcal{E}_\infty$ in Θ. We define $TV(\cdot;\rho,\eta)$ to be a weighted total variation distance (defined in Section 2.4). If $\mu^{(n)}$ are a sequence of almost minimizers of GTV_n in Θ_n

then, with probability one,

$$\lim_{n\to\infty}\inf_{\Theta_n} GTV_n = \min_{\Theta} TV(\cdot;\rho,\eta)$$
$$\mu^{(n)} \to \mu^{(\infty)}$$

where $\mu^{(\infty)}$ minimizes TV in Θ. As is common we use the Γ-convergence framework. Since Γ-convergence is stable under continuous perturbations one could also impose any constraint so that the minimization problem can be written in the form

$$\text{minimize: } \mathcal{E}_n(\mu) + g(\mu)$$

where g is continuous and $g \geq 0$. More generally if $g_n : L^1(\Psi_n) \to [0,\infty]$ is continuously convergent to $g : L^1(X) \to [0,\infty]$, see [50], then one can also imply the convergence of minimizers of $\mathcal{E}_n + g_n$. For example this could be used to fit data, i.e. $g_n(\mu) = g_n(\mu, \zeta^{(n)})$ where $\zeta^{(n)}$ is known data, see Section 7.4.

Asymptotic results for the Ginzburg-Landau functional in a continuous setting are well known, see for example [5,114], which also discuss the constrained optimization problem. More recent results use a square lattice in $\mathbb{R}^2$ and isotropic interaction potentials [163] (who also discuss the constrained optimization problem). The unconstrained GTV_n problem for random graphs when the interaction potential is isotropic was studied in [69]. The unconstrained $\mathcal{E}_n$ problem for random graphs and anisotropic potentials was studied in the previous chapter. In particular we can use the results of the previous chapter and [69] to infer the convergence of the constrained problem.

The chapter is organized as follows. Section 7.2 reviews the unconstrained minimization problems. Then in Section 7.3 we give the convergence results for the $\mathcal{E}_n$ and GTV_n optimization problems with the mass constraint. The following section gives the convergence results for $\mathcal{E}_n$ and GTV_n with the addition of a data term. We briefly discuss the generalization to more than two classes in Section 7.5.

7.2 Convergence of the Unconstrained Optimization Problem

We start by recapping the assumptions for both the unconstrained and constrained optimization problems given in the previous chapter.

Definition 7.2.1. *We say that the pair (V,η) where $V : \mathbb{R} \to [0,\infty)$ and $\eta : \mathbb{R}^d \to [0,\infty)$ are admissible if V satisfies conditions 6 to 8 and η satisfies either conditions 1 to 4 or condition 5, where*

1. $\int_{\mathbb{R}^d} \eta(y)|y| \, \mathrm{d}y < \infty$,

2. *η satisfies the following strengthened Lipschitz condition: there exists L such that for all $x, y \in \mathbb{R}^d$ we have*

$$|\eta(x) - \eta(y)| \leq L \max\{|\eta(x)|, |\eta(y)|\} |x - y|,$$

3. $\eta(0) > 0$,

4. *there exists two decreasing functions* $\eta^-, \eta^+ : [0,\infty) \to [0,\infty)$ *and constants* $\gamma_1, \gamma_2 > 0$ *such that for all* $x \in \mathbb{R}^d$

$$\eta^-(|x|) \leq \eta(x) \leq \eta^+(|x|) \qquad \textit{and} \qquad \eta^+(|x|) \leq \gamma_1 \eta^-(|x|/\gamma_2),$$

5. $\eta = \mathbb{I}_F$ *where* F *is open, bounded, with Lipschitz boundary and* $0 \in F$,

6. $V(y) = 0$ *if and only if* $y \in \{0,1\}$,

7. V *is continuous,*

8. *there exists* $r > 0$ *and* $\tau > 0$ *such that for all* $|t| \geq r$ *we have* $V(t) \geq \tau |t|$.

We will assume the following rate of decay in ϵ_n,

$$\lim_{n\to\infty} \frac{(\log n)^{\frac{3}{4}}}{\epsilon_n n^{\frac{1}{2}}} = 0 \quad \text{if } d = 2, \tag{7.6}$$

$$\lim_{n\to\infty} \frac{(\log n)^{\frac{1}{d}}}{\epsilon_n n^{\frac{1}{d}}} = 0 \quad \text{if } d \geq 3. \tag{7.7}$$

The scaling in ϵ_n implies that the geometric random graph defined by connecting all nodes ξ_i that are within ϵ_n is, with probability one, eventually connected [121]. The bound is constructed so that

$$\frac{\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n} \to 0 \quad \text{as } n \to \infty$$

where T_n is as in Theorem 2.5.1.

Recall the following convergence result for $\mathcal{E}_n$ in the unconstrained case (given in the previous chapter).

Theorem 7.2.1. *Let* $X \subset \mathbb{R}^d$ *for* $d \geq 2$ *be a bounded, open, connected domain with Lipschitz boundary and* P *a probability distribution with support on* X *and density* ρ. *We assume that* ρ *is continuous and bounded above and below by strictly positive constants on* X. *The data is distributed* $\xi_i \overset{\text{iid}}{\sim} P$ *and we let* $\Psi_n = \{\xi_i\}_{i=1}^n$. *Let* $\epsilon_n \to 0$ *be a sequence satisfying the bound* (7.6) *if* $d = 2$ *or* (7.7) *if* $d \geq 3$. *For any function* $\mu^{(n)}$ *on* Ψ_n *define* $\mathcal{E}_n : L^1(\Psi_n) \to [0,\infty]$ *by* (7.3) *where the weights* W_{ij} *are given by* $W_{ij} = \eta_{\epsilon_n}(\xi_i - \xi_j)$ *and* η_{ϵ_n} *is given by* (7.1). *Define* $\mathcal{E}_\infty : L^1(X) \to [0,\infty]$ *by (7.4-7.5). Assume* (V,η) *are admissible functions. Then, with probability one, the following hold*

1. *Compactness: Let* $\mu^{(n)}$ *be a sequence of functions on* Ψ_n *such that* $\sup_{n\in\mathbb{N}} \mathcal{E}_n(\mu^{(n)}) < \infty$ *then* $\mu^{(n)}$ *is relatively compact in* TL^1 *and each cluster point is in* $BV(X;\rho,\eta) \cap L^1(X;\{0,1\})$.

2. Γ*-limit: we have*

$$\Gamma\text{-}\lim_{n\to\infty} \mathcal{E}_n = \mathcal{E}_\infty.$$

Similarly we have the following convergence result for the GTV_n unconstrained optimization problem. The proof is a consequence of results in the previous chapter and can also be found in [69] for isotropic interaction potentials η.

Theorem 7.2.2. *Let $X \subset \mathbb{R}^d$ for $d \geq 2$ be a bounded, open, connected domain with Lipschitz boundary and P a probability distribution with support on X and density ρ. We assume that ρ is continuous and bounded above and below by strictly positive constants on X. The data is distributed $\xi_i \overset{\text{iid}}{\sim} P$ and we let $\Psi_n = \{\xi_i\}_{i=1}^n$. Let $\epsilon_n \to 0$ be a sequence satisfying the bound* (7.6) *if $d = 2$ or* (7.7) *if $d \geq 3$. For any function $\mu^{(n)}$ on Ψ_n define $GTV_n : L^1(\Psi_n) \to [0,\infty]$ by* (7.2) *where the weights W_{ij} are given by $W_{ij} = \eta_{\epsilon_n}(\xi_i - \xi_j)$ and η_{ϵ_n} is given by* (7.1)*. Define $TV : L^1(X) \to [0,\infty]$ by* (2.1-2.3)*. Assume η is an admissible functions. Then, with probability one, the following hold*

1. *Compactness: Let $\mu^{(n)}$ be a sequence of functions on Ψ_n that are uniformly bounded in TL^1 and $\sup_{n\in\mathbb{N}} GTV_n(\mu^{(n)}) < \infty$. Then $\mu^{(n)}$ is relatively compact in TL^1.*

2. *Γ-limit: we have*
$$\Gamma\text{-}\lim_{n\to\infty} GTV_n = TV(\cdot;\rho,\eta).$$

Remark 7.2.3. *Note that we require that $\mu^{(n)}$ be bounded in GTV_n and L^1 in the compactness property in Theorem 7.2.2. In the proof it is shown that boundedness in GTV_n implies boundedness in TV hence if any sequence is bounded in GTV_n and in L^1 then, by the Rellich-Kondrachov theorem, the sequence is precompact in L^1. When considering $\mathcal{E}_n$ the growth condition on first term implies boundedness in L^1. As a consequence, when we consider the constrained GTV_n in Section 7.3.2 we will also have to show that minimizers are bounded in L^1.*

Remark 7.2.4. *One can construct the recovery sequence for $\mu \in L^1(X)$ by choosing $\mu^{(n)} \in L^1(\Psi_n)$ to be the restriction of μ onto Ψ_n. Therefore it is a simple consequence that if one chooses sets*

$$\Theta_n = \left\{\mu \in L^1(\Psi_n) : \mu(\xi_i) = 0 \text{ for } \xi_i \in E_0 \text{ and } \mu(\xi_i) = 1 \text{ for } \xi_i \in E_1\right\} \tag{7.8}$$
$$\Theta = \left\{\mu \in L^1(X) : \mu(x) = 0 \text{ for } x \in E_0 \text{ and } \mu(x) = 1 \text{ for } x \in E_1\right\} \tag{7.9}$$

for any open sets E_0, E_1 then as long as there exists $\zeta \in L^1(X)$ with $\mathcal{E}_\infty(\zeta) < \infty$ then $(\mathcal{E}_n, \mathcal{E}_\infty, \Theta_n, \Theta)$ is compatible with respect to Γ-convergence. The closure of Θ is immediate once one notices that it is weakly closed. Hence the minimum and minimizers of $\mathcal{E}_n$ in Θ_n converge to the minimum and minimizers of $\mathcal{E}_\infty$ in Θ. Similarly for $(GTV_n, TV, \Theta_n, \Theta)$ if the minimizers are uniformly bounded in L^1, see Remark 7.3.6.

7.3 Convergence of the Mass Constrained Optimization Problem

7.3.1 The Ginzburg-Landau Functional

Let $m \in [0,1]$ and define

$$\Theta_n = \left\{ \nu \in L^1(\Psi_n) : \frac{1}{n} \sum_{i=1}^n \nu(\xi_i) = m \right\} \tag{7.10}$$

$$\Theta = \left\{ \nu \in L^1(X) : \int_X \nu(x)\rho(x) \, \mathrm{d}x = m \right\}. \tag{7.11}$$

We start by proving that $(\Theta_n, \Theta, \mathcal{E}_n, \mathcal{E}_\infty)$ are compatible with respect to Γ-convergence before concluding the convergence of the constrained optimization problem.

Lemma 7.3.1. *Let $m \in [0,1]$ and define Θ_n and Θ by* (7.10) *and* (7.11) *respectively. Then under the same conditions as Theorem 7.2.1 $(\mathcal{E}_n, \mathcal{E}_\infty, \Theta_n, \Theta)$ are compatible with respect to Γ-convergence in the sense of Definition 2.2.2.*

Proof. Clearly Θ is closed and there exists $\zeta \in \Theta$ such that $\mathcal{E}_\infty(\zeta) < \infty$. Let $\zeta^{(n)} \to \zeta$ with $\zeta^{(n)} \in \Theta_n$. Then since $\zeta^{(n)}$ also converges weakly to ζ and $\int_X \cdot \rho(x) \, \mathrm{d}x$ is a bounded linear operator we have $\zeta \in \Theta$.

Let $\mu \in \Theta$ and we may assume that $\mu \in L^1(X; \{0,1\})$. Define $\mu^{(n)}$ to be the restriction of μ to $\Psi_n \setminus \{\xi_i\}_{i=2}^n$ and $\mu^{(n)}(\xi_1) = \mu(\xi_1) + \delta_n$ where $\delta_n = m - \frac{1}{n}\sum_{i=1}^n \mu(\xi_i)$. Therefore $\mu^{(n)} \in \Theta_n$. Now we can write

$$\int_X \left| \mu^{(n)}(T_n(x)) - \mu(x) \right| \, P(\mathrm{d}x) \leq \int_X |\mu(T_n(x)) - \mu(x)| \, P(\mathrm{d}x) + \int_{T_n^{-1}(\xi_1)} |\delta_n| \, P(\mathrm{d}x)$$

where T_n is as in Theorem 2.5.1. By Lemma 6.4.4 the first term converges to zero. And since $|\delta_n| \leq \max\{m, 1-m\}$ then we can bound the second term:

$$\int_{T_n^{-1}(\xi_1)} |\delta_n| \, P(\mathrm{d}x) \leq \frac{\max\{m, 1-m\}}{n} \to 0 \quad \text{as } n \to \infty.$$

Therefore $\mu^{(n)} \to \mu$ in TL^1.

The rest of the proof then follows from 6.4.4 once we show, as $n \to \infty$,

$$\frac{1}{\epsilon_n} \frac{1}{n} \sum_{i=1}^n V(\mu^{(n)}(\xi_i) \to 0 \tag{7.12}$$

$$\frac{1}{\epsilon_n} \frac{1}{n^2} \sum_{i=2}^n (W_{i1} + W_{1i}) \, |\mu(\xi_i) - \delta_n| \to 0. \tag{7.13}$$

For (7.12) using the continuity assumption on V and boundedness of δ_n we have that $V(\mu(\xi_1) + \delta_n)$ is bounded. Hence

$$\frac{1}{\epsilon_n} \frac{1}{n} \sum_{i=1}^n V(\mu^{(n)}(\xi_i)) = \frac{1}{\epsilon_n} \frac{1}{n} V(\mu(\xi_1) + \delta_n) \to 0$$

since $\frac{1}{\epsilon_n n} \to 0$. We show (7.13) by controlling η in the ball $B(\xi_1, \epsilon_n)$ and on $X \setminus B(\xi_1, \epsilon_n)$ as follows. We assume n is sufficiently large so that $\frac{\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n} \leq 1$ and we define

$$X_n = \left\{ x \in X \,:\, \frac{|T_n(x) - \xi_1|}{\epsilon_n} \geq 1 \right\}.$$

Then for $x \in X_n$ we have

$$\frac{|x - \xi_1|}{\epsilon_n} \leq \frac{|T_n(x) - \xi_1|}{\epsilon_n} + \frac{\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n} \leq \frac{2|T_n(x) - \xi_1|}{\epsilon_n}.$$

And hence for $x \in X_n$

$$\begin{aligned} \eta\left(\frac{T_n(x) - \xi_1}{\epsilon_n}\right) &\leq \eta^+\left(\left|\frac{T_n(x) - \xi_1}{\epsilon_n}\right|\right) \\ &\leq \eta^+\left(\left|\frac{x - \xi_1}{2\epsilon_n}\right|\right) \\ &\leq \gamma_1 \eta^-\left(\left|\frac{x - \xi_1}{2\gamma_2\epsilon_n}\right|\right) \\ &\leq \gamma_1 \eta\left(\frac{x - \xi_1}{2\gamma_2\epsilon_n}\right). \end{aligned}$$

We let $Y_n = X \setminus X_n$ then we note that

$$Y_n \subseteq \hat{Y}_n := \left\{ x \in X \,:\, \frac{|x - \xi_1|}{\epsilon_n} \leq 2 \right\}.$$

By application of (2.9) we may write

$$\begin{aligned} &\frac{1}{n}\sum_{i=2}^n W_{i1} \left|\mu(\xi_i) - \delta_n\right| \\ &\quad = \int_X \eta_{\epsilon_n}(T_n(x) - \xi_1) \left|\mu(\xi_i) - \delta_n\right| \rho(x) \,\mathrm{d}x \\ &\quad \leq \|\rho\|_{L^\infty(X)} \max\{m, 2-m\} \left(\int_{X_n} \eta_{\epsilon_n}(T_n(x) - \xi_1) \,\mathrm{d}x + \int_{\hat{Y}_n} \eta_{\epsilon_n}(T_n(x) - \xi_1) \,\mathrm{d}x \right) \\ &\quad \leq \|\rho\|_{L^\infty(X)} \max\{m, 2-m\} \left(\frac{\gamma_1}{\epsilon_n^d} \int_{X_n} \eta\left(\frac{x - \xi_1}{2\gamma_2\epsilon_n}\right) \,\mathrm{d}x + \frac{\eta(0)}{\epsilon_n^d} \mathrm{Vol}(B(\xi_1, 2\epsilon_n)) \right) \\ &\quad \leq \|\rho\|_{L^\infty(X)} \max\{m, 2-m\} \left(2^d \gamma_1 \gamma_2^d \int_{\mathbb{R}^d} \eta(x) \,\mathrm{d}x + \frac{\eta(0)}{\epsilon_n^d} \mathrm{Vol}(B(\xi_1, 2\epsilon_n)) \right). \end{aligned}$$

Which shows (7.13). □

We may now conclude the convergence of the $\mathcal{E}_n$ mass constrained optimization problem. The proof is a simple consequence of Corollary 2.2.3 and Theorem 7.2.1.

Corollary 7.3.2. *Under the same conditions as Theorem 7.2.1 and Lemma 7.3.1 with probability one we have the following.*

1. $\inf_{\Theta_n} \mathcal{E}_n \to \min_{\Theta} \mathcal{E}_\infty$ *as* $n \to \infty$

2. *If $\mu^{(n)}$ is a sequence of almost minimizers of $\mathcal{E}_n$ in Θ_n then the sequence is precompact. Furthermore, any cluster point of $\mu^{(n)}$ is a minimizer of $\mathcal{E}_\infty$ in Θ.*

7.3.2 The Graph Total Variation Functional

We now give the analogous result for the graph total variation term.

Lemma 7.3.3. *Let $m \in \mathbb{R}$ and define Θ_n and Θ by* (7.10) *and* (7.11) *respectively. Then under the same conditions as Theorem 7.2.2 $(GTV_n, TV, \Theta_n, \Theta)$ are compatible with respect to Γ-convergence in the sense of Definition 2.2.2.*

Proof. Clearly Θ is closed, there exists $\zeta \in \Theta$ such that $TV(\zeta; \rho, \eta) < \infty$ and any convergent sequence in Θ_n has a limit in Θ.

Let $\mu \in \Theta$ and T_n be as in Theorem 2.5.1. One could prove the lemma in the same way as Lemma 7.3.1, i.e. by defining $\mu^{(n)}$ to be the restriction of μ onto Ψ_n and adding δ_n to one coordinate in order to preserve the mass constraint. This is not the strategy we use here. Instead we construct our recovery sequence by defining $\mu^{(n)} \in L^1(\Psi_n)$ by

$$\mu^{(n)}(\xi_i) = n \int_X \mathbb{I}_{T_n^{-1}(\xi_i)}(x)\mu(x)\rho(x)\, \mathrm{d}x.$$

I.e. $\mu^{(n)}$ is the average of μ over each partition given by $\cup_{i=1}^n T_n(\xi_i) = X$. In the authors opinion this second proof is more aesthetically pleasing than adapting the proof of Lemma 7.3.1.

Since we can infer the existence of a sequence $\nu^{(m)} \in \Theta \cap C_c^\infty(\mathbb{R}^d)$ such that $\nu^{(m)} \to \mu$ in $L^1(X)$ and $TV(\nu^{(m)}; \rho, \eta) \to TV(\mu; \rho, \eta)$ then by a diagonalization argument it is enough to prove the statement for μ that are Lipschitz.

We therefore assume μ is Lipschitz and first show that $\mu^{(n)} \in \Theta_n$ and $\mu^{(n)} \to \mu$ in TL^1. We have that

$$\frac{1}{n}\sum_{i=1}^n \mu^{(n)}(\xi_i) = \sum_{i=1}^n \int_{T_n^{-1}(\xi_i)} \mu(x)\rho(x)\, \mathrm{d}x = \int_X \mu(x)\rho(x)\, \mathrm{d}x = m$$

since $T_n^{-1}(\xi_i)$ is a partition of X. Therefore $\mu^{(n)} \in \Theta_n$. We also have

$$\begin{aligned}
\|\mu - \mu^{(n)} \circ T_n\|_{L^1(X;\rho)} &= \sum_{i=1}^n \int_{T_n^{-1}(\xi_i)} \left|\mu(x) - \mu^{(n)}(T_n(x))\right| \rho(x)\, \mathrm{d}x \\
&= \sum_{i=1}^n \int_{T_n^{-1}(\xi_i)} \left|\mu(x) - n\int_{T_n^{-1}(\xi_i)} \mu(y)\rho(y)\, \mathrm{d}y\right| \rho(x)\, \mathrm{d}x \\
&\leq n\sum_{i=1}^n \int_{(T_n^{-1}(\xi_i))^2} |\mu(x) - \mu(y)|\, \rho(y)\rho(x)\, \mathrm{d}y\, \mathrm{d}x \\
&\leq n\mathrm{Lip}(\mu)\sum_{i=1}^n \int_{(T_n^{-1}(\xi_i))^2} |x - y|\, \rho(y)\rho(x)\, \mathrm{d}y\, \mathrm{d}x.
\end{aligned}$$

Now for $x, y \in T_n^{-1}(\xi_i)$ we have $T_n(x) = T_n(y) = \xi_i$ and therefore

$$|x - y| \leq |x - T_n(x)| + |T_n(x) - T_n(y)| + |T_n(y) - y| \leq 2\|T_n - \mathrm{Id}\|_{L^\infty(X)}.$$

Hence

$$\|\mu - \mu^{(n)} \circ T_n\|_{L^1(X;\rho)} \leq 2\mathrm{Lip}(\mu)\|T_n - \mathrm{Id}\|_{L^\infty(X)}$$

since $\int_{T_n^{-1}(\xi_i)} \rho(x) \, \mathrm{d}x = \frac{1}{n}$. As $\|T_n - \mathrm{Id}\|_{L^\infty(X)} \to 0$ we have $\mu^{(n)} \to \mu$ in TL^1 as required. To complete the lemma we need to show

$$TV(\mu; \rho, \eta) \geq \limsup_{n\to\infty} GTV_n(\mu^{(n)})$$

which follows from 6.4.4 (see also [69, Theorem 4.1]) with the following minor modification to step 2 (numbered as in 6.4.4). We can show $\limsup_{i\to\infty} \limsup_{n\to\infty} d_{n,i} \leq 0$ from

$$\begin{aligned}
d_{n,i} &\leq \frac{\gamma_1 \|\rho\|^2_{L^\infty(X)}}{\epsilon_n^{d+1}} \int_{\hat{X}_{n,i}} \eta\left(\frac{x-y}{2\gamma_2\epsilon_n}\right) \left|\nu^{(n)}(x) - \nu^{(n)}(y)\right| \, \mathrm{d}x \, \mathrm{d}y \\
&\leq \frac{\gamma_1 \|\rho\|^2_{L^\infty(X)} \mathrm{Lip}(\mu)}{\epsilon_n^{d+1}} \int_{\hat{X}_{n,i}} \eta\left(\frac{x-y}{2\gamma_2\epsilon_n}\right) \left(2\|T_n - \mathrm{Id}\|_{L^\infty(X)} + |x-y|\right) \, \mathrm{d}x \, \mathrm{d}y \\
&= \frac{\gamma_1 2^d \gamma_2^d \|\rho\|^2_{L^\infty(X)} \mathrm{Lip}(\mu) \mathrm{Vol}(X)}{\epsilon_n} \int_{4\gamma_2|z|\geq i} \eta(z) \left(2\|T_n - \mathrm{Id}\|_{L^\infty(X)} + 2\gamma_2\epsilon_n|z|\right) \, \mathrm{d}z \\
&= \gamma_1 2^{d+1} \gamma_2^d \|\rho\|^2_{L^\infty(X)} \mathrm{Lip}(\mu) \mathrm{Vol}(X) \frac{\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\epsilon_n} \int_{4\gamma_2|z|\geq i} \eta(z) \, \mathrm{d}z \\
&\quad + \gamma_1 2^{d+1} \gamma_2^{d+1} \|\rho\|^2_{L^\infty(X)} \mathrm{Lip}(\mu) \mathrm{Vol}(X) \int_{4\gamma_2|z|\geq i} \eta(z)|z| \, \mathrm{d}z.
\end{aligned}$$

All other details remain unchanged. □

We can now prove the convergence of the GTV_n mass constrained optimization problem. The proof is an application of Corollary 2.2.3 and Lemma 7.3.3 once we have shown that almost minimizers are uniformly bounded.

Corollary 7.3.4. *Under the same conditions as Theorem 7.2.2 and Lemma 7.3.3 with probability one we have the following.*

1. $\inf_{\Theta_n} GTV_n \to \min_\Theta TV(\cdot; \rho, \eta)$ *as* $n \to \infty$
2. *If* $\mu^{(n)}$ *is a sequence of almost minimizers of* GTV_n *in* Θ_n *then the sequence is precompact. Furthermore, any cluster point of* $\mu^{(n)}$ *is a minimizer of* $TV(\cdot; \rho, \eta)$ *in* Θ.

Proof. We have to show that almost minimizers are bounded in L^1. Let $\mu^{(n)} \in L^1(\Psi_n)$ satisfy

$$GTV_n(\mu^{(n)}) \leq \inf_{\Theta_n} GTV_n + \delta_n \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^n \mu^{(n)}(\xi_i) = m$$

for some sequence $\delta_n \to 0$, i.e. $\mu^{(n)}$ is a sequence of almost minimizers of GTV_n in Θ_n. Since, by Lemma 7.3.3, $(GTV_n, TV, \Theta_n, \Theta)$ are compatible with respect to Γ-convergence then there exists $\zeta \in \Theta$ and a sequence $\zeta^{(n)} \in \Theta_n$ such that $\limsup_{n\to\infty} GTV_n(\zeta^{(n)}) \leq TV(\zeta; \rho, \eta) =: \kappa < \infty$. Therefore we may assume that $GTV_n(\mu^{(n)}) \leq \kappa + 1$ for all n.

We want to show there exists M such that $\frac{1}{n}\sum_{i=1}^n |\mu^{(n)}(\xi_i)| \leq M$ for all n. Suppose not. Then for all M there exists a subsequence n_m such that $\frac{1}{n_m}\sum_{i=1}^{n_m} |\mu^{(n_m)}(\xi_i)| > M$. By relabeling the data points we may assume that $|\mu^{(n_m)}(\xi_1)| \geq M$ for all m. Without loss of generality suppose $\mu^{(n_m)}(\xi_1) \geq M$.

Since η is continuous at 0 and $\eta(0) > 0$ there exists $r > 0$ and $\alpha > 0$ such that $\eta(x) > \alpha \mathbb{I}_{B(0,r)}(x)$ for all $x \in \mathbb{R}^d$. Also note, by [121, Theorem 13.2], the graph defined by connecting all edges with distance less than $r\epsilon_n$ is connected with probability one. Hence there exists i with $\mathbb{I}_{B(0,r)}\left(\frac{\xi_1 - \xi_i}{\epsilon}\right) > 0$. In particular this implies

$$\left|\mu^{(n_m)}(\xi_1) - \mu^{(n_m)}(\xi_i)\right| \leq \frac{GTV_{n_m}(\mu^{(n_m)})}{W_{1i}} \leq \frac{\epsilon_n^d(\kappa+1)}{\alpha}.$$

By [58, Theorem 8] for any k there exists a constant C (independent of k and n) such that the number of edges connecting ξ_1 to ξ_k is less than $\frac{C}{\epsilon_n r}$. Hence for any k we have

$$\left|\mu^{(n_m)}(\xi_1) - \mu^{(n_m)}(\xi_k)\right| \leq \frac{\epsilon_n^{d-1} C(\kappa+1)}{\alpha r}.$$

In particular this implies that $\mu^{(n_m)}(\xi_k) \geq M - \frac{\epsilon_n^{d-1}C(\kappa+1)}{\alpha r}$. Hence by choosing M sufficiently large we have that $\mu^{(n_m)}(\xi_k) \geq m$ and therefore $\mu^{(n_m)} \notin \Theta_{n_m}$. A contradiction. □

Remark 7.3.5. *The bound on the number of the minimal number of edges between two nodes in [58] was proved when X is the unit ball and data points are uniformly iid. It is immediately clear that these results will generalize to any connected and bounded domain X and for any probability density ρ bounded above and below by strictly positive constants.*

Remark 7.3.6. *From the proof of Corollary 7.3.4 one can also see that if Θ_n and Θ were defined by* (7.8) *and* (7.9) *respectively then almost minimizers are uniformly bounded in L^1. Hence Corollary 7.3.4 also holds for constraints of this form.*

7.4 Convergence with Data

In this section we are interested in using data to obtain non-trivial minimizers to both the Ginzburg-Landau functional and the graph total variation functional. We use a data term of the form:

$$g_n(\mu^{(n)}, \zeta^{(n)}) = \frac{1}{n}\sum_{i=1}^{n}\left|\mu^{(n)}(\xi_i) - \zeta^{(n)}(\xi_i)\right| \tag{7.14}$$

where $\zeta^{(n)} \in L^1(\Psi_n)$ is the data and we assume that $\zeta^{(n)} \to \zeta$ in TL^1. Both these results are (almost) immediate once we have shown that $g_n(\cdot, \zeta^{(n)})$ converges continuously to $g(\cdot, \zeta)$.

Lemma 7.4.1. *Define g_n by* (7.14) *and $g : L^1(X) \to [0,\infty)$ by*

$$g(\mu, \zeta) = \int_X |\mu(x) - \zeta(x)|\,\rho(x)\,\mathrm{d}x.$$

Assume $\zeta^{(n)} \to \zeta$ in TL^1 then g_n converges continuously to g.

Proof. Assume $\mu^{(n)} \to \mu$ in TL^1 then

$$\begin{aligned}\left|g_n(\mu^{(n)}, \zeta^{(n)}) - g(\mu, \zeta)\right| &= \left|\int_X \left(\left|\mu^{(n)}(T_n(x)) - \zeta^{(n)}(T_n(x))\right| - |\mu(x) - \zeta(x)|\right) \rho(x)\, \mathrm{d}x\right| \\ &\leq \int_X \left(\left|\mu^{(n)}(T_n(x)) - \mu(x)\right| + \left|\zeta^{(n)}(T_n(x)) - \zeta(x)\right|\right) \rho(x)\, \mathrm{d}x \\ &\to 0.\end{aligned}$$

Which proves the lemma. □

Corollary 7.4.2. *Under the same conditions as Theorem 7.2.1 and Lemma 7.4.1 with probability one we have the following.*

1. $\inf_{L^1(\Psi_n)} \mathcal{E}_n + g_n(\cdot, \zeta^{(n)}) \to \min_{L^1(X)} \mathcal{E}_\infty + g(\cdot, \zeta)$ *as* $n \to \infty$
2. *If* $\mu^{(n)}$ *is a sequence of almost minimizers of* $\mathcal{E}_n + g_n(\cdot, \zeta^{(n)})$ *in* $L^1(\Psi_n)$ *then the sequence is precompact. Furthermore, any cluster point of* $\mu^{(n)}$ *is a minimizer of* $\mathcal{E}_\infty + g(\cdot, \zeta)$ *in* $L^1(X)$.

Remark 7.4.3. *One should note that the* TL^1 *notion of convergence is random. In particular the stagnating transport map between* P_n *and* P *is random. If we consider the sequence* $\zeta^{(n)}$ *to be the restriction of* ζ *onto* Ψ_n *then it is not immediate that* $\zeta^{(n)} \to \zeta$ *in* TL^1*. However if one assumes that* ζ *is Lipschitz then with probability one we have* $\zeta^{(n)} \to \zeta$ *which follows from the almost sure weak convergence of the empirical measure (see also the proof of Lemma 7.3.3).*

The proof of the above is immediate from Proposition 2.2.4, Lemma 7.4.1 and Theorems 2.2.1 and 7.2.1. We have the corresponding result for the graph total variation functional.

Corollary 7.4.4. *Under the same conditions as Theorem 7.2.2 and Lemma 7.4.1 with probability one we have the following.*

1. $\inf_{L^1(\Psi_n)} GTV_n + g_n(\cdot, \zeta^{(n)}) \to \min_{L^1(X)} TV(\cdot; \rho, \eta) + g(\cdot, \zeta)$ *as* $n \to \infty$
2. *If* $\mu^{(n)}$ *is a sequence of almost minimizers of* $GTV_n + g_n(\cdot, \zeta^{(n)})$ *in* $L^1(\Psi_n)$ *then the sequence is precompact. Furthermore, any cluster point of* $\mu^{(n)}$ *is a minimizer of*

   $$TV(\cdot; \rho, \eta) + g(\cdot, \zeta)$$

   in $L^1(X)$.

Proof. From Proposition 2.2.4, Lemma 7.4.1 and Theorem 7.2.2 we have that

$$\Gamma\text{-}\lim_{n\to\infty} \left(GTV_n + g_n(\cdot, \zeta^{(n)})\right) = TV(\cdot; \rho, \eta) + g(\cdot, \zeta).$$

We are left to show the compactness property. This is a simple case of showing that if

$$GTV_n(\mu^{(n)}) + g_n(\mu^{(n)}, \zeta^{(n)}) \leq M \quad \text{for all } n$$

then $\sup_{n\in\mathbb{N}} \|\mu^{(n)} \circ T_n\|_{L^1(X)} < \infty$. But this is trivial since

$$\frac{1}{n}\sum_{i=1}^n \left|\mu^{(n)}(\xi_i)\right| \leq M + \frac{1}{n}\sum_{i=1}^n \left|\zeta^{(n)}(\xi_i)\right|$$

and the RHS converges. □

7.5 Multiple Classes

Let there be k classes at points $\{p_1, \ldots, p_k\}$ where $p_i \in \mathbb{R}^m$ are affinely independent and therefore $k \leq m+1$. We then consider $\mu^{(n)} \in L^1(\Psi_n; \mathbb{R}^m)$ and $\mu \in L^1(X; \mathbb{R}^m)$ where we write $\mu = (\mu_1, \ldots, \mu_m)$. We assume $V : \mathbb{R}^m \to [0,\infty)$ has zeros at $\{p_1, \ldots, p_k\}$ only and by interpreting $|\cdot|$ as the Euclidean norm on $\mathbb{R}^m$ then $\mathcal{E}_n : L^1(\Psi_n; \mathbb{R}^m) \to [0,\infty]$ and $GTV_n : L^1(\Psi_n; \mathbb{R}^m) \to [0,\infty]$ remain unchanged. The limiting potential becomes

$$\mathcal{E}_\infty = \begin{cases} TV(\mu;\rho,\eta) & \text{if } \mu \in L^1(X; \{p_i\}_{i=1}^k) \\ \infty & \text{otherwise} \end{cases}$$

where TV is defined by

$$TV(\mu;\rho,\eta) := \sup\left\{ \int_X \sum_{i=1}^m \mu_i(x) \mathrm{div}(\phi_i(x)) \,\mathrm{d}x \,:\, \phi_i \in C_c^\infty(X;\mathbb{R}^d), \right.$$
$$\left. \sup_{x\in X} \sigma^*\left(-\frac{\phi_i(x)}{\rho^2(x)}\right) < \infty \right\}.$$

With these minor modifications all the results stated in this chapter (and the previous chapter) hold. One should note that $\mathcal{E}_\infty$ written in the form (7.4) is not well defined as the outward surface normal in more than one dimension is not unique.

Chapter 8

Closing Remarks

In this thesis we have looked at two approaches to problems based in statistical inference which include a data association component. These were the k-means approach (Chapter 3 and Chapter 4) and a graphical approach (Chapter 6 and Chapter 7). We devote the final part of this thesis to suggesting problems for future work.

8.1 Further Problems in the k-Means Method

In the smoothing data association problem we used a constant regularizationterm, i.e. cluster centers were, by definition, minimizers $\mu^{(n)}$ of f_n defined by

$$f_n(\mu) = \frac{1}{n} \sum_{i=1}^{n} \bigwedge_{j=1}^{k} |\mu_j(t_i) - y_i|^2 + \lambda_n \|\nabla^s \mu\|_{L^2}^2$$

where $\lambda_n = \lambda$ is constant. In the spline problem ($k = 1$) in Chapter 5 we showed that one could take $\lambda_n \asymp \frac{1}{\sqrt{n}}$ in order for weak convergence of minimizers and the minimum of f_n. It seems reasonable that this will also hold for $k > 1$ however the proofs in the $k = 1$ case relied upon an explicit formula for the minimizer. In particular for $k = 1$ we could write

$$\mu^{(n)} = G_{n,\lambda_n}^{-1} \nu^{(n)}$$

for a linear operator G_{n,λ_n}^{-1} and some $\nu^{(n)}$ depending on the data. We were able to study the properties of G_{n,λ_n}^{-1} which lead to the bound

$$\left\| G_{n,\lambda_n}^{-1} \nu^{(n)} \right\| = O\left(1 + \frac{1}{n\lambda_n^2}\right).$$

For $k > 1$ the minimization problem is no longer quadratic and therefore one cannot expect the existence of G_{n,λ_n}^{-1} with such ‘nice’ properties. However in some sense the problem is still locally quadratic so one would expect similar results to the $k = 1$ case.

In the general case where $\mu_j \in Y$ for some Banach space Y and data is in another Banach space X then one could also ask whether it is possible to take $\lambda_n \to 0$. It is not clear what we would expect in this case. When one looks at weak convergence one is taking a finite dimensional projection of the random variable and therefore the $\frac{1}{\sqrt{n}}$ drops out naturally as a 'central limit theorem' type result. However, it also seems reasonable that the rate will depend on the interplay between the regularization $r(\mu)$ and the data term $P_n \bigwedge_{j=1}^k d(\xi_i, \mu_j)$. When $d(\xi_i, \mu_j) = |L_i\mu_j - y_i|^2$ (for $\xi_i = (L_i, y_i) \in (Y^*, \mathbb{R})$) then this becomes the general spline coupled with data association problem and we expect $\lambda_n \asymp \frac{1}{\sqrt{n}}$. Whether this is true in greater generality is not clear and an interesting problem for future works.

8.2 Further Problems in the Graphical Approach

An interesting extension of Chapter 6 and Chapter 7 would be the application to infinite dimensional data. In order to further develop the problem we consider a simple example. We assume P is a centered measure on $L^2([a, b])$ with continuous covariance operator $K : [a, b] \times [a, b] \to \mathbb{R}$. We define the operator $T_K : L^2([a, b]) \to L^2([a, b])$ by

$$T_K(f) = \int_a^b K(s, \cdot)f(s) \, \mathrm{d}s.$$

By the Karhunen-Loéve theorem there exists an orthonormal basis of eigenfunctions of T_K which we will call ϕ_i with corresponding eigenvalues λ_i^2. Any $\xi \sim P$ can be written

$$\xi(t) = \sum_{k=1}^{\infty} \hat{\xi}(k)\phi_k(t)$$

where the convergence is in L^2 and uniform in t and the random variables $\hat{\xi}(k)$ are given by

$$\hat{\xi}(k) = (\xi, \phi_k) := \int_a^b \xi(t)\phi_k(t) \, \mathrm{d}t.$$

Furthermore $\hat{\xi}(k)$ satisfy

$$\mathbb{E}(\hat{\xi}(k)) = 0 \qquad \text{and} \qquad \mathbb{E}(\hat{\xi}(j)\hat{\xi}(k)) = \delta_{jk}\lambda_k^2.$$

We assume that $\frac{\hat{\xi}(k)}{\lambda_k}$ are distributed with density ρ_k on $\mathbb{R}$.

For data points $x, y \in L^2([a, b])$ we define the interaction potential $\eta_\epsilon : L^2([a, b]) \times L^2([a, b]) \to [0, \infty)$ by

$$\eta_\epsilon(x, y) = \sum_{k=1}^{\infty} \Psi\left(\frac{\alpha(k)(\hat{x}(k) - \hat{y}(k))}{\epsilon}\right). \tag{8.1}$$

The function $\Psi : \mathbb{R} \to \mathbb{R}$ is used to compare coefficients and for example

$$\Psi(t) = \begin{cases} 1 & \text{if } |t| < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The weights α act as a filter and we will show that by taking $\alpha(k) \to \infty$ as $k \to \infty$ sufficiently quickly will ensure that only finitely many terms of the sum in (8.1) are positive which in particular implies that the sum is finite. For a data set $\{\xi_i\}_{i=1}^n$ we define the graph total variation by:

$$GTV_n(\mu) = \frac{1}{\epsilon_n^2} \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \eta_{\epsilon_n}(\xi_i, \xi_j) \, |\mu(i) - \mu(j)|$$

for binary functions $\mu : \{1, \ldots, n\} \to \{0, 1\}$ and for some constant $p > 0$.

This problem arises naturally in image classification. For example the problem of classifying images using a rotationally invariant distance function [119,141,186] could be approached using the formulation we express here. The objective is to classify images $I_1, \ldots, I_n$ via a distance d_{RID} of the form

$$d_{RID}(I_i, I_j) = \min_{O \in SO(d)} \|I_i - O \circ I_j\|$$

where $SO(d)$ is the set of rotations on $\mathbb{R}^d$. By using a radial basis ϕ_i one should be able to use η_ϵ as an alternative to d_{RID}. This problem has applications in cryo-Electron Microscopy which concerns determining 3D macromolecular structures from noisy images at random orientations. This is a very active area of research and indeed the 2003 and 2009 Nobel prizes in Chemistry were awarded for determining the structure of various molecules.

For GTV_n to converge as $n \to \infty$ it is necessary (but not sufficient) that the scalar F_n defined by

$$F_n = \frac{1}{\epsilon_n} \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \eta_{\epsilon_n}(\xi_i, \xi_j) \tag{8.2}$$

also converges as $n \to \infty$. One would expect that if F_n is bounded then so is GTV_n (for simplicity we will consider only F_n here). First we show that η_ϵ is bounded for each ϵ and therefore F_n is finite for each n (Lemma 8.2.1). Next we show that F_n can be bounded uniformly in n. For simplicity we bound in expectation and more precisely we show

$$\sup_{n \in \mathbb{N}} \mathbb{E} F_n < \infty.$$

Taking expectations has the advantage of putting the problem into the continuous setting which greatly simplifies the proofs. The disadvantages are that we do not see the graphical structure and in particular we gain no intuition in what the natural scaling of $\epsilon_n \to 0$ should be.

Lemma 8.2.1. *Let P be a centered measure on $L^2([a,b])$ with continuous covariance operator $K : [a,b]^2 \to \mathbb{R}$. Let $\{(\lambda_k^2, \phi_k)\}_{k=1}^\infty$ be the Karhunen-Loéve basis of eigenfunctions where the Karhunen-Loéve coefficients are distributed $\frac{(x,\phi_k)}{\lambda_k} \sim \rho_k$ for a density ρ_k. Assume $\lambda_k \asymp k^r$ with $r < 0$ and let $\alpha(k) \asymp k^q$ with $q + r > 1$ and there exists $C < \infty$ such that $\sup_{k \in \mathbb{N}} \|\rho_k\|_{L^\infty} \leq C$. Then for $x, y \sim P$ independently there almost surely exists $K < \infty$ such that $\alpha(k) \, |(x - y, \phi_k)| \geq \epsilon$ for all $k \geq K$.*

Proof. By our assumptions we can write

$$\mathbb{P}\left(|(x-y,\phi_k)| \le \frac{\epsilon}{\alpha(k)}\right) = \int_{\mathbb{R}} \int_{s-\frac{\epsilon}{\alpha(k)\lambda_k}}^{s+\frac{\epsilon}{\alpha(k)\lambda_k}} \rho_k(t)\, \mathrm{d}t\, \rho_k(s)\, \mathrm{d}s$$
$$\le \frac{2C\epsilon}{\alpha(k)\lambda_k} \int_{\mathbb{R}} \rho_k(s)\, \mathrm{d}s$$
$$\le \frac{2C\epsilon}{\alpha(k)\lambda_k}.$$

Therefore

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\left|\frac{\alpha(k)(x-y,\phi_k)}{\epsilon}\right| \le 1\right) \le \sum_{k=1}^{\infty} \frac{2C\epsilon}{\alpha(k)\lambda_k} \lesssim \sum_{k=1}^{\infty} \frac{1}{k^{q+r}}$$

where the above summation is finite for $q+r>1$. By the Borel-Cantelli lemma the event

$$\{|\alpha(k)(x-y,\phi_k)| \le \epsilon\}$$

almost surely occurs finitely many times. □

The above lemma shows that $\eta_\epsilon(x,y)$ is finite for almost every $x, y \overset{\text{iid}}{\sim} P$. We now show that F_n is bounded in expectation.

Lemma 8.2.2. *Under the same conditions as Lemma 8.2.1 where data is distributed $\xi_i \overset{\text{iid}}{\sim} P$ we define F_n by* (8.2) *with η_ϵ by* (8.1) *and $\Psi(t) = \mathbb{I}_{|t|<1}$. Then F_n is bounded in expectation, i.e. there exists a constant $M < \infty$ such that*

$$\sup_{n\in\mathbb{N}} \mathbb{E} F_n \le M.$$

Proof. One has

$$\mathbb{E} F_n = \frac{1}{\epsilon_n} \sum_{k=1}^{\infty} \mathbb{E}\Psi\left(\frac{\alpha(k)(\hat{x}(k)-\hat{y}(k))}{\epsilon_n}\right) = \frac{1}{\epsilon_n} \sum_{k=1}^{\infty} \mathbb{P}\left(\alpha(k)\,|(x-y,\phi_k)| \le \epsilon_n\right).$$

By the calculation in the proof of Lemma 8.2.1 we have

$$\mathbb{E} F_n \lesssim \sum_{k=1}^{\infty} \frac{1}{k^{q+r}}.$$

For $q+r>1$ the above converges. □

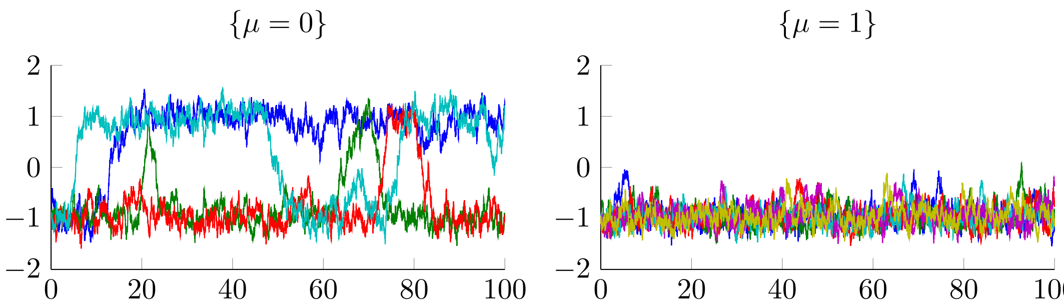
To test the methodology we perform the following numerical experiment. Let ξ_i be independent samples from the following stochastic differential equation on $[0,T]$,

$$\mathrm{d}\xi = -\sigma(\xi)\, \mathrm{d}t + \rho\, \mathrm{d}W, \qquad \xi(0) = -1 \tag{8.3}$$

where W is a Brownian motion, $\rho > 0$ a fixed constant and $\sigma(\xi) = \xi^3 - \xi$. Realizations of (8.3) have the behavior that $\xi(t)$ is close to ± 1. In particular we choose constants so that approximately half of the realizations have a jump from -1 to 1. We define a classifier μ of

$\{\xi_i\}_{i=1}^n$ by minimizing GTV_n over binary functions (conditioned on $\sum_{i=1}^n \mu(i) = m$ for some $m \in \mathbb{N}$). In Figure 8.1 we see that classifiers are able to correctly identify which paths have a jump.

Figure 8.1: Infinite dimensional classifiers



Minimizers of GTV_n partition the data as shown above. The figure on the left contains all the data points that have at least one jump. The figure on the right contains all the data points with no jumps.

An important point which we have so far not touched upon is the Γ-limit. As motivation we discuss ratio and Cheeger graph cuts for which, to a limited extent, have been considered in infinite dimensional settings and are closely related to the graph total variation. The ratio and Cheeger graph cuts for a data set $\{\xi_i\}_{i=1}^n \subset X$ (with graph weights W_{ij}) are minimizers of $\mathcal{E}_n$ defined by

$$\mathcal{E}_n(F) := \frac{\mathrm{Cut}_n(F)}{\mathrm{Bal}_n(F)}$$

over sets $F \subset X$ and where $\mathrm{Cut}_n(F)$ is the graph cut of F defined by

$$\mathrm{Cut}_n(F) = \sum_{\xi_i \in F} \sum_{\xi_j \in F^c} W_{ij}$$

and $\mathrm{Bal}_n(F)$ is a defined by either:

$$\mathrm{Bal}_n(F) = 2|F||F^c| \quad \text{for ratio cuts}$$
$$\mathrm{Bal}_n(F) = \min\{|F|, |F^c|\} \quad \text{for Cheeger cuts}$$

which, with an abuse of notation, we let $|F| = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\xi_i \in F}$. The results of [71] imply that when $X \subset \mathbb{R}^d$ then minimizers of $\mathcal{E}_n$ converge to a minimizer of $\mathcal{E}_\infty$, the ratio or Cheeger cut on X, defined by

$$\mathcal{E}_\infty(F) := \frac{\mathrm{Cut}_P(F)}{\mathrm{Bal}_P(F)}$$

where

$$\mathrm{Cut}_P(F) = \int_{\partial F} \rho^2(x) \, \mathrm{d}\mathcal{H}^{d-1}(x)$$
$$\mathrm{Bal}_P(F) = P(F)P(F^c) \quad \text{for ratio cuts}$$
$$\mathrm{Bal}_P(F) = \min\{P(F), P(F^c)\} \quad \text{for Cheeger cuts,}$$

$\xi_i \overset{\text{iid}}{\sim} P$ and ρ is the density of P. When X is infinite dimensional (and more precisely a Gauss space) one has that

$$\text{Cut}_P(F) = TV(\mathbb{I}_F; P)$$

where $TV(\mu; P)$ is the total variation defined with respect to the measure P, see [36].

There already exists some results in the literature towards understanding $\mathcal{E}_\infty$. We call a set a Cheeger set if it is a minimizer of

$$\hat{\mathcal{E}}_\infty(F) := \frac{\text{Cut}_P(F)}{P(F)}.$$

One can see that this is very closely related to the Cheeger and ratio cuts. When X is finite dimensional the existence and uniqueness of a minimizer of $\hat{\mathcal{E}}_\infty$ (under suitable conditions) has been proven in, for example, [6]. This result was successfully extended to infinite dimensions when X is a subset of the Wiener space [36]. It is most likely a straightforward generalization to show that these results on $\hat{\mathcal{E}}_\infty$ carry through to $\mathcal{E}_\infty$.

The results of the finite dimensional case suggest a candidate Γ-limit for the infinite dimensional case, that is

$$\mathcal{E}_\infty(F) = \frac{TV(\mathbb{I}_F; P)}{\text{Bal}_P(F, F)}.$$

This would also suggest that a candidate Γ-limit for GTV_n is $TV(\cdot; P)$.

Bibliography

[1] E. F. Abaya and G. L. Wise. Convergence of vector quantizers with applications to optimal quantization. *SIAM Journal on Applied Mathematics*, 44(1):183–189, 1984.

[2] R. A. Adams. *Sobolev Spaces*. Pure and applied mathematics; a series of monographs and textbooks; v. 65. Academic Press, Inc. (London) Ltd., 1975.

[3] M. Aerts, G. Claeskens, and M. P. Wand. Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference*, 103(1-2):455–470, 2002.

[4] S. Agapiou, S. Larsson, and A. M. Stuart. Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Processes and their Applications*, 123(10):3828–3860, 2013.

[5] G. Alberti and G. Bellettini. A nonlocal anisotropic model for phase transitions: Asymptotic behaviour of rescaled energies. *European Journal of Applied Mathematics*, 1998.

[6] F. Alter and V. Caselles. Uniqueness of the Cheeger set of a convex body. *Nonlinear Analysis: Theory, Methods and Applications*, 70(1):32–44, 2009.

[7] L. Ambrosio, M. Miranda Jr., S. Maniglia, and D. Pallara. BV functions in abstract Wiener spaces. *Journal of Functional Analysis*, 258(3):785–813, 2010.

[8] L. Ambrosio and A. Pratelli. Existence and stability results in the L^1 theory of optimal transportation. In *Optimal Transportation and Applications*, volume 1813 of *Lecture Notes in Mathematics*, pages 123–160. Springer Berlin Heidelberg, 2003.

[9] T. Amemiya. *Advanced Econometrics*. Havard University Press, 1985.

[10] A. Antos. Improved minimax bounds on the test and training distortion of empirically designed vector quantizers. *Information Theory, IEEE Transactions on*, 51(11):4022–4032, 2005.

[11] A. Antos, L. Gyorfi, and A. Gyorgy. Individual convergence rates in empirical vector quantizer design. *Information Theory, IEEE Transactions on*, 51(11):4013–4022, 2005.

[12] E. Arias-Castro. Clustering based on pairwise distances when the data is of mixed dimensions. *Information Theory, IEEE Transactions on*, 57(3):1692–1706, 2011.

[13] E. Arias-Castro, G. Chen, and G. Lerman. Spectral clustering based on local linear approximations. *Electronic Journal of Statistics*, 5:1537–1587, 2011.

[14] E. Arias-Castro, G. Lerman, and T. Zhang. Spectral clustering based on local PCA. *arXiv preprint arXiv:1301.2007*, 2013.

[15] H. Attouch, G. Buttazzo, and G. Michaille. *Variational Analysis in Sobolev and BV Spaces: Applications to PDE's and Optimization*. MPS-SIAM Series on Optimization, 2006.

[16] A. Baldi. Weighted BV functions. *Houston Journal of Mathematics*, 27(3), 2001.

[17] J. D. Barrow, S. P. Bhavsar, and D. H. Sonoda. Minimal spanning trees, filaments and galaxy clustering. *Monthly Notices of the Royal Astronomical Society*, 216:17–35, 1985.

[18] P. L. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *Information Theory, IEEE Transactions on*, 44(5):1802–1813, 1998.

[19] S. Ben-David, D. Pál, and H. U. Simon. Stability of k-means clustering. In *Proceedings of the Twentieth Annual Conference on Computational Learning*, pages 20–34, 2007.

[20] A. L. Bertozzi and A. Flenner. Diffuse interface models on graphs for classification of high dimensional data. *Multiscale Modeling & Simulation*, 10(3):1090–1118, 2012.

[21] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *Information Theory, IEEE Transactions on*, 54(2):781–790, 2008.

[22] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250, 2001.

[23] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM Journal on Numerical Analysis*, 45(6):2610–2636, 2007.

[24] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.

[25] V. I. Bogachev. *Gaussian Measures*. the American Mathematical Society, 1998.

[26] L. Bottou and Y. Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems 7*, pages 585–592, 1995.

[27] A. Braides. *Γ-Convergence for Beginners*. Oxford University Press, 2002.

[28] A. Braides. *Local Minimization, Variational Evolution and Γ-Convergence*. Springer International Publishing, 2014.

[29] X. Bresson, T. Laurent, D. Uminsky, and J. H. von Brecht. Convergence and energy landscape for Cheeger cut clustering. In *Advances in Neural Information Processing Systems 25*, pages 1385–1393. Curran Associates, Inc., 2012.

[30] X. Bresson, T. Laurent, D. Uminsky, and J. H. von Brecht. An adaptive total variation algorithm for computing the balanced cut of a graph. *arXiv preprint arXiv:1302.2717*, 2013.

[31] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.

[32] L. D. Brown and M. G. Low. Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, 24(6):2384–2398, 1996.

[33] G. Caldarelli. *Scale Free Networks: Complex Webs in Nature and Technology*. Oxford University Press, 2007.

[34] G. Canas, T. Poggio, and L. Rosasco. Learning manifolds with K-means and K-flats. In P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2474–2482. MIT Press, 2012.

[35] R. J. Carroll, A. C. M. Van Rooij, and F. H. Ruymgaart. Theoretical aspects of ill-posed problems in statistics. *Acta Applicandae Mathematica*, 24(2):113–140, 1991.

[36] V. Caselles, M. Miranda Jr., and M. Novaga. Total variation and Cheeger sets in Gauss space. *Journal of Functional Analysis*, 259(6):1491–1516, 2010.

[37] G. Celeux, D. Chauveau, and J. Diebolt. On stochastic versions of the EM algorithm, 1995.

[38] G. Celeux and J. Diebolt. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2(1):73–82, 1985.

[39] 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.

[40] T. Champion, L. De Pascale, and P. Juutinen. The ∞-Wasserstein distance: Local solutions and existence of optimal transport maps. *SIAM Journal on Mathematical Analysis*, 40(1):1–20, 2008.

[41] M.-T. Chiang and B. Mirkin. Experiments for the number of clusters in k-means. In J. Neves, M. Santos, and J. Machado, editors, *Progress in Artificial Intelligence*, volume 4874 of *Lecture Notes in Computer Science*, pages 395–405. Springer Berlin Heidelberg, 2007.

[42] P. A. Chou. The distortion of vector quantizers trained on n vectors decreases to the optimum as $O_p(1/n)$. In *Information Theory, 1994. Proceedings., 1994 IEEE International Symposium on*, page 457, 1994.

[43] G. Claeskens, T. Krivobokova, and J. D. Opsomer. Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3):529–544, 2009.

[44] J. B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer, 1990.

[45] D. D. Cox. Asymptotics for M-type smoothing splines. *The Annals of Statistics*, 11(2):530–551, 1983.

[46] D. D. Cox. Approximation of method of regularization estimators. *The Annals of Statistics*, 16(2):694–712, 1988.

[47] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403, 1979.

[48] J. A. Cuesta and C. Matran. The strong law of large numbers for k-means and best possible nets of Banach valued random variables. *Probability Theory and Related Fields*, 78(4):523–534, 1988.

[49] J. A. Cuesta-Albertos and R. Fraiman. Impartial trimmed k-means for functional data. *Computational Statistics & Data Analysis*, 51(10):4864–4877, 2007.

[50] G. Dal Maso. *An Introduction to* Γ*-Convergence*. Springer, 1993.

[51] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(9), 2005.

[52] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss. Map estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Problems*, 29(9):095017, 2013.

[53] C. De Boor. *A Practical Guide to Splines*. Springer–Verlag New York Inc., 1978.

[54] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[55] P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1):1–26, 1986.

[56] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

[57] P. H. C. Eilers and B. D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121, 1996.

[58] R. B. Ellis, J. L. Martin, and C. Yan. Random geometric graph diameter in the unit ball. *Algorithmica*, 2007.

[59] L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2010.

[60] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 1992.

[61] L. Fahrmeir and T. Kneib. *Bayesian Smoothing and Regression for Longitudinal, Spatial and Even History Data*. Oxford University Press Inc., New York, 2011.

[62] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pages 251–262, 1999.

[63] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and Ulthurusamy, editors. *Advances in Knowledge Discovery and Data Mining*. AAAI/The MIT Press, 1996.

[64] E. A. Feinberg, P. O. Kasyanov, and N. V. Zadoianchuk. Fatou's lemma for weakly converging probabilities. *Theory of Probability & Its Applications*, 58(4):683–689, 2014.

[65] S. Fortunato. Community detection in graphs. *Physics Reports*, 2010.

[66] D. A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403, 1963.

[67] W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

[68] C. Garcia-Cardona, A. Flenner, and A. G. Percus. Multiclass semi-supervised learning on graphs using Ginzburg-Landau functional minimization. In *Pattern Recognition Applications and Methods*, pages 119–135. Springer, 2015.

[69] N. García Trillos and D. Slepčev. Continuum limit of total variation on point clouds. *arXiv preprint arXiv:1403.6355*, 2014.

[70] N. García Trillos and D. Slepčev. On the rate of convergence of empirical measures in ∞-transportation distance. *arXiv preprint arXiv:1407.1157*, 2014.

[71] N. García Trillos, D. Slepčev, J. H. von Brecht, T. Laurent, and X. Bresson. Consistency of Cheeger and ratio graph cuts. *arXiv preprint arXiv:1411.6590*, 2014.

[72] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistic*, 28(2):500–531, 2000.

[73] A. Gkiokas, A. I. Cristea, and M. Thorpe. Self-reinforced meta learning for belief generation. In *Research and Development in Intelligent Systems XXXI*, pages 185–190. Springer International Publishing, 2014.

[74] A. Goldenshluger and S. V. Pereverzev. Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations. *Probability Theory and Related Fields*, 118(2):169–186, 2000.

[75] I. J. Good and R. A. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971.

[76] S. Graf and H. Luschgy. *Foundations of Quantization For Probability Distributions*. Springer, 2000.

[77] S. Graf and H. Luschgy. Rates of convergence for the empirical quantization error. *The Annals of Probability*, 30(2):874–897, 2002.

[78] S. Graf, H. Luschgy, and G. Pagès. The local quantization behavior of absolutely continuous probabilities. *The Annals of Probability*, 40(4):1795–1828, 2012.

[79] P. Hall and J. D. Opsomer. Theory for penalised spline regression. *Biometrika*, 92(1):105–118, 2005.

[80] J. Hartigan. Asymptotic distributions for clustering criteria. *The Annals of Statistics*, 6(1):117–131, 1978.

[81] T. J. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2008.

[82] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

[83] J. F. Hennawi and J. X. Prochaska. Quasars probing quasars. II. The anisotropic clustering of optically thick absorbers around quasars. *The Astrophysical Journal*, 655(2), 2007.

[84] H. Hu, Y. van Gennip, B. Hunter, A. L. Bertozzi, and M. A. Porter. Multislice modularity optimization in community detection and image segmentation. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pages 934–936, 2012.

[85] C. M. Hurvich, J. S. Simonoff, and C.-L. Tsai. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(2):271–293, 1998.

[86] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[87] H. Jylhä. The L^∞ optimal transport: infinite cyclical monotonicity and the existence of optimal transport maps. *Calculus of Variations and Partial Differential Equations*, 52(1-2):303–326, 2015.

[88] G. Kauermann, T. Krivobokova, and L. Fahrmeir. Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):487–503, 2009.

[89] G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

[90] G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

[91] S. C. Kou and B. Efron. Smoothers and the C_p, generalized maximum likelihood, and extended exponential criteria: A geometric approach. *Journal of the American Statistical Association*, 97(459):766–782, 2002.

[92] B. Kulis and M. I. Jordan. Revisiting k-means: New algorithms via Bayesian nonparametrics. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 513–520. ACM, 2012.

[93] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9:112–147, 1998.

[94] M.-J. Lai and L. Wang. Bivariate penalized splines for regression. *Statistica Sinica*, 23:1399–1417, 2013.

[95] T. Laloë. L_1-quantization and clustering in Banach spaces. *Mathematical Methods of Statistics*, 19(2):136–150, 2010.

[96] J. Lember. On minimizing sequences for k-centres. *Journal of Approximation Theory*, 120:20–35, 2003.

[97] G. Leoni. *A First Course in Sobolev Spaces*, volume 105. American Mathematical Society, 2009.

[98] K.-C. Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15(3):958–975, 1987.

[99] X.-Z. Li, J.-F. Wang, W.-Z. Yang, Z.-J. Li, and S.-J. Lai. A spatial scan statistic for nonisotropic two-level risk cluster. *Statistics in medicine*, 31(2):177–187, 2012.

[100] Y. Li and D. Ruppert. On the asymptotics of penalized splines. *Biometrika*, 95(2):415–436, 2008.

[101] E. V. Linder, M. Oh, T. Okumura, C. G. Sabiu, and Y.-S. Song. Cosmological constraints from the anisotropic clustering analysis using BOSS DR9. *Phys. Rev. D*, 89, 2014.

[102] T. Linder. On the training distortion of vector quantizers. *Information Theory, IEEE Transactions on*, 46(4):46, 2000.

[103] T. Linder. *Principles of Nonparametric Learning*, chapter Learning-Theoretic Methods in Lossy Data Compression, pages 163–210. Springer, 2002.

[104] T. Linder, G. Lugosi, and K. Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *Information Theory, IEEE Transactions on*, 40(6):1728–1740, 1994.

[105] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[106] D. J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, 2002.

[107] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[108] B. A. Mair and F. H. Ruymgaart. Statistical inverse estimation in Hilbert scales. *SIAM Journal on Applied Mathematics*, 56(5):1424–1444, 1996.

[109] C. L. Mallows. Some comments on C_P. *Technometrics*, 15(4):661–675, 1973.

[110] M. Mampaey and J. Vreeken. Summarizing categorical data by clustering attributes. *Data Mining and Knowledge Discovery*, 26(1):130–173, 2013.

[111] V. Maz'ya. *Sobolev Spaces: With Applications to Elliptic Partial Differential Equations*. Springer, 2010.

[112] R. E. Megginson. *An Introduction to Banach Space Theory*. Springer–Verlag New York Inc, 1998.

[113] G. Milligan and M. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[114] L. Modica. The gradient theory of phase transitions and the minimal interface criterion. *Archive for Rational Mechanics and Analysis*, 98(2), 1987.

[115] J. R. Munkres. *Topology*. Prentice Hall, Incorporated, 2000.

[116] M. Newman. The physics of networks. *Physics Today*, 61(11):33–38, 2008.

[117] D. W. Nychka and D. D. Cox. Convergence rates for regularized solutions of integral equations from discrete noisy data. *The Annals of Statistics*, 17(2):556–572, 1989.

[118] F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1(4):502–518, 1986.

[119] P. A. Penczek, J. Zhu, and J. Frank. A common-lines based method for determining orientations for $N > 3$ particle projections simultaneously. *Ultramicroscopy*, 63(3-4):205–218, 1996.

[120] X. Peng, C. Zhou, D. M. Hepburn, M. D. Judd, and W. H. Siew. Application of k-means method to pattern recognition in on-line cable partial discharge monitoring. *Dielectrics and Electrical Insulation, IEEE Transactions on*, 20(3):754–76, 2013.

[121] M. Penrose. *Random Geometric Graphs*. Oxford University Press, 2003.

[122] S. Pigeon and S. Coulombe. Efficient clustering-based algorithm for predicting file size and structural similarity of transcoded jpeg images. In *Multimedia (ISM), 2011 IEEE International Symposium on*, pages 137–142, 2011.

[123] Y. Pokern, A. M. Stuart, and J. H. van Zanten. Posterior consistency via precision operators for Bayesian nonparametric drift estimation in SDEs. *Stochastic Processes and their Applications*, 123(2):603–628, 2013.

[124] D. Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135–140, 1981.

[125] D. Pollard. A central limit theorem for empirical processes. *Journal of the Australian Mathematical Society (Series A)*, 33(2):235–248, 1982.

[126] D. Pollard. A central limit theorem for k-means clustering. *The Annals of Statistics*, 4(10):919–926, 1982.

[127] D. Pollard. Quantization and the method of k-means. *Information Theory, IEEE Transactions on*, 28(2):199–205, 1982.

[128] D. Pollard. *Convergence of Stochastic Processes*. Springer–Verlag New York Inc., 1984.

[129] A. C. Ponce. A new approach to Sobolev spaces and connections to Γ-convergence. *Calculus of Variations and Partial Differential Equations*, 19(3):229–255, 2004.

[130] M. A Porter, J.-P. Onnela, and P. J Mucha. Communities in networks. *Notices of the AMS*, 2009.

[131] D. L. Ragozin. Error bounds for derivative estimates based on spline smoothing of exact or noisy data. *Journal of Approximation Theory*, 37(4):335–355, 1983.

[132] E. Randall. *Nonparametric regression and spline smoothing*. New York : Marcel Dekker, 1999.

[133] C. H. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–183, 1967.

[134] J. Rousseau. Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *The Annals of Statistics*, 38(1):146–180, 2010.

[135] H. L. Royden. *Real Analysis*. Macmillan, 3rd edition, 1988.

[136] Y. Sakamoto, M. Ishiguro, and G. Kitagawa. *Akaike Information Criterion Statistics*. KTK Scientific Publishers (KTK), 1986.

[137] S. Z. Selim and M. A. Ismail. k-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(1):81–87, 1984.

[138] J. Shen and X. Wang. Estimation of monotone functions via P-splines: A constrained dynamical optimization approach. *SIAM Journal on Control and Optimization*, 49(2):646–671, 2011.

[139] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714, 2001.

[140] A. N. Shiryaev. *Probability (2nd Ed.)*. Springer-Verlag New York, Inc., 1995.

[141] A. Singer, Z. Zhao, Y. Shkolnisky, and R. Hadani. Viewing angle classification of cryo-electron microscopy images using eigenvectors. *SIAM Journal on Imaging Sciences*, 4(2):723–759, 2011.

[142] M. Soltanolkotabi, E. Elhamifar, and E. J. Candès. Robust subspace clustering. *The Annals of Statistics*, 42(2):669–699, 2014.

[143] P. L. Speckman. Spline smoothing and optimal rates of convergence in nonparametric regression models. *The Annals of Statistics*, 13(3):970–983, 1985.

[144] P. L. Speckman and D. Sun. Asymptotic properties of smoothing parameter selection in spline smoothing. Technical report, Department of Statistics, University of Missouri, 2001.

[145] D. Steinley. k-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1):1–34, 2006.

[146] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

[147] S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[148] T. Tarpey and K. K. J. Kinateder. Clustering functional data. *Journal of Classification*, 2003.

[149] M. Thorpe and N. Cade. Asking the right questions. `http://www.smithinst.co.uk/blog/asking-right-questions/`, 2015.

[150] M. Thorpe and A. M. Johansen. Rate of convergence for a smoothing spline with data association model. *arXiv preprint arXiv:1506.02676*, 2015.

[151] M. Thorpe and A. M. Johansen. Weak convergence of general smoothing splines. *arXiv preprint arXiv:1506.08450*, 2015.

[152] M. Thorpe and F. Theil. Asmyptotic analysis of the Ginzburg-Landau functional on point clouds. *In Preparation*, 2015.

[153] M. Thorpe and F. Theil. The constrained Ginzburg-Landau functional on point clouds. *In Preparation*, 2015.

[154] M. Thorpe, F. Theil, A. Johansen, and N. Cade. Convergence of the k-means minimization problem using Γ-convergence. *arXiv preprint arXiv:1501.01320v2*, 2015.

[155] P. Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105(1):249–252, 2000.

[156] F. I. Utreras. Optimal smoothing of noisy data using spline functions. *SIAM Journal on Scientific and Statistical Computing*, 2(3):349–362, 1981.

[157] F. I. Utreras. Natural spline functions, their associated eigenvalue problem. *Numerische Mathematik*, 42(1):107–117, 1983.

[158] F. I. Utreras. Smoothing noisy data under monotonicity constraints existence, characterization and convergence rates. *Numerische Mathematik*, 47(4):611–625, 1985.

[159] R. Valdarnini. Detection of non-random patterns in cosmological gravitational clustering. *Astronomy and Astrophysics*, 366(2):376–386, 2001.

[160] A. W. van der Vaart. *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, 2000.

[161] A. W. van der Vaart and J. H. van Zanten. *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, volume 3, chapter Reproducing kernel Hilbert spaces of Gaussian priors, pages 200–222. Institute of Mathematical Statistics, 2008.

[162] A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

[163] Y. van Gennip and A. L. Bertozzi. Γ-convergence of graph Ginzburg-Landau functionals. *Advances in Differential Equations*, 17(11-12):1115–1180, 2012.

[164] Y. van Gennip, N. Guillen, B. Osting, and A. L. Bertozzi. Mean curvature, threshold dynamics, and phase field theory on finite graphs. *Milan Journal of Mathematics*, 82(1):3–65, 2014.

[165] Y. van Gennip, H. Hu, B. Hunter, and M. A. Porter. Geosocial graph-based community detection. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pages 754–758, 2012.

[166] Y. van Gennip, B. Hunter, R. Ahn, P. Elliott, K. Luh, M. Halvorson, S. Reid, M. Valasik, J. Wo, G. Tita, A. L. Bertozzi, and P. Brantingham. Community detection using spectral clustering on sparse geosocial data. *SIAM Journal on Applied Mathematics*, 73(1):67–83, 2013.

[167] V. S. Varadarajan. On the convergence of sample probability distributions. *Sankhya*, 19(1-2):23–26, 1958.

[168] C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. American Mathematical Society, cop., 2003.

[169] G. Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(3):364–372, 1978.

[170] G. Wahba. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, 13(4):1378–1402, 1985.

[171] G. Wahba. *Spline models for observational data*. Society for Industrial and Applied Mathematics (SIAM), 1990.

[172] G. Wahba and S. Wold. A completely automatic french curve: Fitting spline functions by cross-validation. *Communications in Statistics*, 4(1), 1975.

[173] M. P. Wand. On the optimal amount of smoothing in penalised spline regression. *Biometrika*, 86(4):936–940, 1999.

[174] X. Wang, J. Shen, and D. Ruppert. On the asymptotics of penalized spline smoothing. *Electronic Journal of Statistics*, 5:1–17, 2011.

[175] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. Cambridge University Press, 1994.

[176] W. E. Wecker and C. F. Ansley. The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association*, 78(381):81–89, 1983.

[177] M. A. Wong. Asymptotic properties of univariate sample k-means clusters. *Journal of Classification*, 1(1):225–270, 1984.

[178] K. J. Worsley, M. Andermann, T. Koulis, D. MacDonald, and A. C. Evans. Detecting changes in nonisotropic images. *Human brain mapping*, 8(2-3):98–101, 1999.

[179] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1083.

[180] L. Xiao, Y. Li, T. V. Apanasovich, and D. Ruppert. Local asymptotics of P-splines. *arXiv preprint arXiv:1201.0708*, 2012.

[181] X. Yang, Y. Wang, D. Wu, and A. Ma. K-means based clustering on mobile usage for social network analysis purpose. In *Advanced Information Management and Service (IMS), 2010 6th International Conference on*, pages 223–228, 2010.

[182] T. Yoshida and K. Naito. Asymptotics for penalized additive B-spline regression. *Journal of the Japan Statistical Society*, 42(1):81–107, 2012.

[183] T. Yoshida and K. Naito. Asymptotics for penalised splines in generalised additive models. *Journal of Nonparametric Statistics*, 2014.

[184] K. Yosida. *Functional Analysis*. Springer-Verlag, 6th edition, 1980.

[185] L. H. Zhao. Bayesian aspects of some nonparametric problems. *The Annals of Statistics*, 28(2):532–552, 2000.

[186] Z. Zhao and A. Singer. Rotationally invariant image representation for viewing direction classification in cryo-EM. *Journal of structural biology*, 186(1):153–166, 2014.