Convergence of the k-Means Minimization Problem using Γ-Convergence

Matthew Thorpe[*1], Florian Theil[1], Adam M. Johansen[1], and Neil Cade[2]

[1]University of Warwick, Coventry, CV4 7AL, United Kingdom
[2]Selex-ES, Luton, LU1 3PG, United Kingdom

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.

1 Introduction

The k-means algorithm [19] 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 [10, 15–17, 25] are the only papers known to the authors. 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 [11, 20], 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 paper 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. For example in smoothing problems data are points in $\mathbb{R}^d$ and cluster centers are continuous functions into $\mathbb{R}^d$. The k-means method addresses this problem by clustering data points by associating each with one of k curves. For $k = 1$ the problem reduces to smoothing and coincides with the limiting problem studied in [13].

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 $j = 1, 2, \ldots, k$. 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 $X = \mathbb{R}^d$, where typically we choose $d(x, y) = |x - y|^2 = \sum_{i=1}^d (x_i - y_i)^2$. 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, $a_1, a_2, \ldots, a_k$, $\bigwedge_{j=1}^k a_j := \min\{a_1, \ldots, a_k\}$. 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, in a sense which will be made precise later, that $\xi_i \stackrel{\text{iid}}{\sim} P$ for some suitable probability, P, and define

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

In Section 3 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 μ^n minimizes f_n and μ^∞ 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 $\mu = (\mu_1, \mu_2, \ldots, \mu_k) \in Y^k$ 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 where 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 4 to justifying our approach. Particularly we motivate treating $\lambda_n = \lambda$ as a constant independent of n. We are able to repeat the results from Section 4; namely that the minimum and a subsequence of minimizers still converge.

Early results assumed $Y = X$ were Euclidean spaces and addressed the convergence of minimizers [14, 21] showing that the minimizers converge to the appropriate limit. 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 [4]. In this paper 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 [22] also assuming a unique minimizer to the limiting energy. Other rates of convergence have been shown in [2, 3, 7, 18].

Still assuming $Y = X$ the convergence of the minimization problem in a reflexive and separable Banach space has been proved in [17] and a similar result in metric spaces in [16]. In [15] the existence of a weakly converging subsequence was inferred using the results of [17].

In the following section we introduce the notation and preliminary material used in this paper.

We then in Section 3 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 the energies defined without regularization. Theorem 3.4 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 μ^{n_m} of minimizers of f_{n_m} which weakly converge to some μ^∞ which minimizes f_∞.

This result is generalized in Section 4 to an arbitrary X and Y. The analogous result to Theorem 3.4 is Theorem 4.6. We first motivate the problem and in particular our choice of scaling in the regularization in Section 4.1 before proceeding to the results in Section 4.2. The conditions on the cost function d and regularization term r are non-trivial and so we show an application to the smoothing-data association problem in Section 4.3.

To demonstrate the generality of the results in this paper a range of applications are considered in Section 5. Example 1 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 $\mathbb{R}^2$.

2 Preliminaries

In this section we introduce some notation and background theory which will be used in Sections 3 and 4 to establish our convergence results. In these sections we show the existence of optimal cluster centers using the direct method. By imposing conditions such that our energies are weakly lower semi-continuous we can deduce the existence of minimizers. Further conditions ensure the minimizers are uniformly bounded. The Γ-convergence framework (eg [5]) allows us to establish the convergence of the minimum and also the convergence of minimizers.

We have the following definition of Γ-convergence with respect to weak convergence.

Definition 2.1 (Γ-convergence). *A sequence $f_n : A \to \mathbb{R}$ is said to Γ-converge on the domain A to $f_\infty : A \to \mathbb{R}$ with respect to weak convergence on A, and we write $f_\infty = \Gamma\text{-}\lim_n f_n$, if for all $x \in A$ we have*

(i) (liminf inequality) for every sequence (x_n) weakly converging to x

$$f_\infty(x) \le \liminf_n f_n(x_n);$$

(ii) (recovery sequence) there exists a sequence (x_n) weakly converging to x such that

$$f_\infty(x) \ge \limsup_n f_n(x_n).$$

When it exists the Γ-limit is always weakly lower semi-continuous, and thus admits minimizers. An important property of Γ-convergence is that it implies the convergence of minimizers. In particular, we will make extensive use of the following well-known result.

Theorem 2.1 (Convergence of Minimizers). *Let $f_n : A \to \mathbb{R}$ be a sequence of functionals on a Banach space $(A, \|\cdot\|_A)$ and assume that there exists $N > 0$ and a weakly compact subset $K \subset A$ with*

$$\inf_A f_n = \inf_K f_n \quad \forall n > N.$$

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

$$\min_A f_\infty = \liminf_n \inf_A f_n.$$

Furthermore if each f_n is weakly lower semi-continuous then for each f_n there exists a minimizer $x_n \in K$ and a subsequence x_{n_m} weakly converging to $x \in K$ that minimizes f_∞.

A proof of the theorem can be found in [5].

The problems which we address involve random observations. 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. 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.

3 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.

Suppressing the $\{\xi_i\}$ dependence of f_n we define

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

$$f_\infty : X^k \to \mathbb{R}, \quad f_\infty(\mu) = P g_\mu = \int_X \bigwedge_{j=1}^{k} d(x, \mu_j) P(\mathrm{d}x) \tag{2}$$

where

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

and P is a probability measure on $(X, \mathcal{X})$ with empirical measure P_n associated with $\xi_1, \ldots, \xi_n$ and defined weakly via

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

for any $\mathcal{X}$-measurable function $h : X \to \mathbb{R}$.

We assume ξ_i are iid according to P with $P = \mathbb{P} \circ \xi_i^{-1}$ which ensures $P_n \rightharpoonup^* P$. Generalization to the non-iid case is possible provided $P_n \rightharpoonup^* P$.

We wish to show

$$\hat{\theta}_n \to \theta \quad \text{almost surely as } n \to \infty \tag{3}$$

where

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

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

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

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

To prove that (3) is satisfied we first establish in Theorem 3.2 that $f_\infty \overset{\text{a.s.}}{=} \Gamma\text{-}\lim_n f_n$ on X^k. Then we show in Proposition 3.3 that the minimizers of $\inf_\mu f_n(\mu)$ lie in a bounded set $K \subset X^k$ independent of n. We may then apply Theorem 2.1 to infer (3) and the existence of a weakly converging subsequence of minimizers.

The key assumptions on d and P are given in Assumptions 1. 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. As such we should not expect to be able to bound our minimizers. We see that this choice of d violates the first assumption. The second assumption implies that the functionals f_n are weakly lower semi-continuous. The third allows us to view $d(\xi_i, y)$ as a collection of random variables. The fourth assumption implies that we have at least k open balls with positive probability and therefore we are not overfitting clusters to data. The fifth assumption is essentially a moment condition which is used to apply the strong law of large numbers to $P_n \bigwedge_{j=1}^{k} d(\cdot, \mu_j)$.

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

1. *There exists continuous, strictly increasing functions $m, M : [0, \infty) \to [0, \infty)$ such that*

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

   with $m(r) \to \infty$ as $r \to \infty$ and $M(0) = 0$.

2. *For each $x \in X$ we have that $d(x, \cdot)$ is weakly lower semi-continuous.*

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

4. *There exist k unique 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 $B(\mu_l, \delta) := \{x \in X : \|\mu_l - x\|_X < \delta\}$.

5. *For every $\nu \in X$ we have that $P d(\cdot, \nu) < \infty$.*

4

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

Remark 3.1. *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 trivially get the first assumption. For $p = 1$ the second assumption follows from the Hahn-Banach theorem which implies that if $z_n \rightharpoonup z$ then $\liminf_{n\to\infty} \|z_n\|_X \geq \|z\|_X$. Putting this inequality to the pth power implies that $d(x, \cdot)$ is weakly lower semi-continuous for all $p > 0$. The third assumption holds by the Borel measurability of metrics on complete separable metric spaces. □

Before proceeding to the main results of this section, we demonstrate that the lower semi-continuity of d is required by considering an example in which d is not weakly lower semi-continuous. Let $k = 1$ and $X = \mathbb{R}$. We assume that $\xi_i \overset{\text{iid}}{\sim} P = \text{Uniform}(-1, 1)$ for $i = 1, 2, \ldots, n$ and define the cost function $d : \mathbb{R}^2 \to [0, \infty)$ by

$$d(x, y) = \begin{cases} |x - y| + 1 & \text{if } y \geq 0 \\ |x - y| & \text{if } y < 0. \end{cases}$$

The corresponding energy $f_n : \mathbb{R} \to [0, \infty)$ for n data points is

$$f_n(\mu) = \frac{1}{n} \sum_{i=1}^n d(\xi_i, \mu).$$

However the Γ-limit is

$$f_\infty(\mu) = \int_X \tilde{d}(x, \mu) P(\mathrm{d}x) = \begin{cases} \frac{3 + \mu^2}{2} & \text{if } \mu > 0 \\ \frac{1 + \mu^2}{2} & \text{if } \mu \leq 0 \end{cases}$$

$$\neq \int_X d(x, \mu) P(\mathrm{d}x)$$

where

$$\tilde{d}(x, y) = \begin{cases} |x - y| + 1 & \text{if } y > 0 \\ |x - y| & \text{if } y \leq 0 \end{cases}$$

is the lower semi-continuous envelope of d. Note that the limit one might naïvely expect

$$\int_X d(x, \mu) P(\mathrm{d}x) = \begin{cases} \frac{3 + \mu^2}{2} & \text{if } \mu \geq 0 \\ \frac{1 + \mu^2}{2} & \text{if } \mu < 0 \end{cases}$$

does not have a minimizer. However as the example suggests it is possible to drop the weak lower semi-continuity condition by modifying the Γ-limit by taking the lower semi-continuous envelope of d. We have opted against this generalization for three reasons. The first is it is notationally cleaner to have d in the Γ limit which in the form given is intuitively strongly linked to the strong law of large numbers. The second reason is for practical implementation we are more interested in situations where the optimal centers exist for finite n. Without weak lower semi-continuity we are not guaranteed this. The third reason is that the applications we have in mind use the underlying norm as the cost function which is weak lower semi-continuous.

Theorem 3.2. *Let $(X, \|\cdot\|_X)$ be a reflexive and separable Banach space with Borel σ-algebra, $\mathcal{X}$; let $\{\xi_i\}_{i=1}^n$ be a sequence of independent X-valued random elements with common law P. Assume $d : X \times X \to [0, \infty)$ and P satisfy the conditions in Assumptions 1. Define $f_n : X^k \to \mathbb{R}$ and $f_\infty : X^k \to \mathbb{R}$ by (1) and (2) respectively. Then*

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

for $\mathbb{P}$-almost every sequence of observations $\xi_1, \xi_2, \ldots$.

Proof. We need to show the liminf inequality and the existence of a recovery sequence. We start by showing the liminf inequality, allowing $\mu^n \in X^k$ to denote any sequence which converges weakly to $\mu \in X^k$:

$$0 \leq \liminf_n \left(f_n(\mu^n) - f_\infty(\mu) \right).$$

Consider

$$\liminf_n \left(f_n(\mu^n) - f_\infty(\mu) \right) = \liminf_n \left(P_n g_{\mu^n} - P g_\mu \right)$$

$$= \liminf_n \left(P_n g_{\mu^n} - P g_\mu + P_n g_\mu - P g_\mu \right)$$

$$\geq \liminf_n \left(P_n g_{\mu^n} - P_n g_\mu \right) + \liminf_n \left(P_n g_\mu - P g_\mu \right). \tag{4}$$

By the strong law of large numbers and Assumption 1.5:

$$\lim_n \left(P_n g_\mu - P g_\mu \right) = 0 \quad \text{almost surely}$$

Which shows the second term on the right hand side of (4) is greater than or equal to zero for almost every sequence of observations and hence almost every sequence P_n.

To show the first term on the right hand side of (4) is also greater than or equal to zero we recall Fatou's lemma with varying measure [24, Chapter 11, Proposition 17]: if we suppose Q_n is a sequence of measures on a measurable space $(A, \mathcal{A})$ and

$$Q_n(U) \to Q(U) \quad \forall U \in \mathcal{A}$$

then for non-negative integrable functions h_n we have

$$\int_A \liminf_{n \to \infty} h_n \ \mathrm{d}Q \leq \liminf_{n \to \infty} \int_A h_n \ \mathrm{d}Q_n.$$

Letting $A = X$, $\mathcal{A} = \mathcal{X}$, $Q_n = P_n$ and $Q = P$ then since, by the strong law of large numbers, $P_n(U) \to P(U)$ almost surely for all $U \in \mathcal{X}$ we can deduce

$$\liminf_n \left(P_n g_{\mu^n} - P_n g_\mu \right) = \liminf_n \int \left(g_{\mu^n}(x) - g_\mu(x) \right) P_n(\mathrm{d}x)$$

$$\geq \int \liminf_n \left(g_{\mu^n}(x) - g_\mu(x) \right) P(\mathrm{d}x) \quad \text{a.s..}$$

Since, for each x, we have by Assumption 1.2 that $d(x, \cdot)$ is weakly lower semi-continuous,

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

By taking the minimum over j we have

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

a.s. as required. Therefore we have shown that the liminf inequality holds almost surely.

The existence of a recovery sequence is trivial. Let $\mu^n = \mu \in X^k$. Then by the strong law of large numbers

$$\limsup_n P_n g_{\mu^n} = \lim_n P_n g_\mu = P g_\mu \quad \text{almost surely.}$$

$\square$

Now we have established almost sure Γ-convergence we establish the boundedness condition in Proposition 3.3 so we can apply Theorem 2.1.

Proposition 3.3. *Assuming the conditions of Theorem 3.2, there exists $R > 0$ such that*

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

almost surely. In particular R is independent of n.

Proof. The structure of the proof is similar to [16, 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.

It is easy to show that at least one cluster center is uniformly bounded. We now show that all cluster centers are uniformly bounded. Assume that the cluster centers $\mu^n \in X^k$ are almost minimizers, i.e.

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

for some sequence $\varepsilon_n > 0$ such that

$$\lim_{n \to \infty} \varepsilon_n = 0. \tag{5}$$

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

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

6

Therefore, (5) implies that

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

By Fatou's lemma with varying measure:

$$0 \leq \liminf_{n\to\infty} \int \left(\bigwedge_{j=1}^{k} d(x, \mu_j^n) - \bigwedge_{j\neq l_n} d(x, \mu_j^n) \right) P_n(\mathrm{d}x) \quad \text{a.s.}$$

and moreover

$$\liminf_{n\to\infty} \left(f_n\left(\mu^n\right) - f_n\left((\mu_j^n)_{j\neq l_n}\right) \right) \geq 0 \quad \text{a.s.,} \tag{6}$$

where we interpret f_n accordingly. It suffices to demonstrate that

$$\liminf_{n\to\infty} \left(\inf_{\mu \in X^k} f_n(\mu) - \inf_{\mu \in X^{k-1}} f_n(\mu) \right) < 0 \tag{7}$$

for n sufficiently large almost surely. Indeed, if (7) holds, then

$$\liminf_{n\to\infty} \left(f_n\left(\mu^n\right) - f_n\left((\mu_j^n)_{j\neq l_n}\right) \right)$$

$$= \lim_{n\to\infty} \underbrace{\left(f_n\left(\mu^n\right) - \inf_{\mu \in X^k} f_n(\mu) \right)}_{\leq \varepsilon_n} + \liminf_{n\to\infty} \left(\inf_{\mu \in X^k} f_n(\mu) - f_n\left((\mu_j^n)_{j\neq l_n}\right) \right)$$

$$< 0 \quad \text{a.s. by (5) and (7),}$$

but this contradicts (6).

We now establish (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 $l \in \{1, 2, \ldots, k\}$ such that we have

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

Proceeding with this choice of l then 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 $j = 1, 2, \ldots, k-1$. Also

$$g_l(\mu) := \min_{j=1,2,\ldots,k-1} d(x, \mu_j) - d(x, \mu_l^\dagger) \geq m(\frac{\delta_1}{2} - \delta_2) - M(\delta_2). \tag{8}$$

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

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

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

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

Define the characteristic function

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

where $l(\mu)$ is the maximizer in (8). For each realization $\{\xi_i\}_{i=1}^n \subset X$ one obtains

$$\inf_{\mu \in X^{k-1}} f_n(\mu) = \inf_{\mu \in X^{k-1}} \frac{1}{n} \sum_{i=1}^{n} \bigwedge_{j=1}^{k-1} d(\xi_i, \mu_j)$$

$$\geq \inf_{\mu \in X^{k-1}} \frac{1}{n} \sum_{i=1}^{n} \left[\bigwedge_{j=1}^{k-1} d(\xi_i, \mu_j) \left(1 - \chi_\mu(\xi_i)\right) + \left(d(\xi_i, \mu_{l(\mu)}^\dagger) + \epsilon \right) \chi_\mu(\xi_i) \right]$$

$$\geq \inf_{\mu \in X^k} f_n(\mu) + \epsilon \min_{l=1,2,\ldots,k} P_n(B(\mu_l^\dagger, \delta_2)).$$

Which since $P_n(B(\mu_l^\dagger, \delta_2)) \to P(B(\mu_l^\dagger, \delta_2)) > 0$ implies (7).

$\square$

We now easily prove convergence by application of Theorem 2.1.

Theorem 3.4. *Assuming the conditions of Theorem 3.2 and Proposition 3.3 the minimization problem associated with the k-means method converges. I.e.*

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

Furthermore for any sequence of minimizers μ^n of f_n there exists a weakly converging subsequence $\mu^{n_m} \rightharpoonup \mu^\infty$ where μ^∞ minimizes f_∞.

4   The Case of General Y

In the previous section both the data ξ_i and cluster centers μ_j took values in a common space, X. We now remove this restriction and let $\xi_i : \Omega \to X$ and $\mu_j \in Y$. Typically we may want to use this framework to deal with finite dimensional data and infinite dimensional cluster centers. This 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 4.2. This type of regularization is understood as penalized likelihood estimation [12].

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 [13] in which the consistency of penalized splines is studied using, what in this paper 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 [13], 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 [6] for which optimal scaling of the regularization is known [1, 27]. 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 4.6. That is, we prove that the k-means method converges where Y is a general separable 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. The assumptions required in this subsection are non-trivial to meet so in the third subsection we show an application to a smoothing-data association problem.

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

$$f_n(\mu) = a_n \sum_{j=0}^{n-1} |\mu(t_j) - y_j|^2 + b_n \|\partial^2 \mu\|_{L^2}^2. \tag{9}$$

Data are points in space-time and the cluster center is a function $\mu : [0, 1] \to \mathbb{R}^d$. 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 4.2.

Proposition 4.1. *For data $\Psi_n = \{(t_j, y_j)\}_{j=1}^n$ with $t_j = \frac{j}{n}$ under the assumption $y_j = m(t_j) + \epsilon_j$ for ϵ_j iid noise with finite variance and $m \in L^2$. If μ is 1-periodic then $\inf_\mu f_n(\mu)$ defined by (9) stays bounded (in expectation) if $a_n = O(\frac{1}{n})$ for any choice of b_n.*

Proof. The minimizer μ^n of f_n is

$$\hat{\mu}^n = \left(1 + \frac{\gamma_n k^4}{n}\right)^{-1} \hat{y}$$

where $\gamma_n = \frac{16\pi^4 b_n}{a_n}$ and $\hat{\mu}, \hat{y}$ are the Fourier transforms of μ, y respectively, i.e. for odd n we have

$$\mu(t) = \frac{1}{n} \sum_{k=-\frac{n-1}{2}}^{\frac{n-1}{2}} \hat{\mu}_k e^{2\pi i k t} \quad \text{for } t \in [0, 1]$$

$$y_j = \frac{1}{n} \sum_{k=-\frac{n-1}{2}}^{\frac{n-1}{2}} \hat{y}_k e^{\frac{2\pi i k j}{n}} \quad \text{for } j = 0, 1, \ldots, n-1.$$

It follows that the minimum is

$$\mathbb{E}\left(f_n(\mu^n)\right) = \frac{a_n}{n} \mathbb{E}\left(\left\langle \left(1 + \frac{n}{\gamma_n k^4}\right)^{-1} \hat{y}, \hat{y} \right\rangle\right) \leq a_n \sum_{j=0}^{n-1} \mathbb{E}y_j^2 \lesssim 2a_n n \left(\|m\|_{L^2}^2 + \text{Var}(\epsilon)\right).$$

Similar expressions can be obtained for the case of even n. $\qquad\square$

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 $\gamma_n = 16\pi^4 \lambda n^{p+1}$. We immediately have f_n 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 $m(t) = 0$.

Proposition 4.2. *In addition to the assumptions of Proposition 4.1 let $a_n = \frac{1}{n}$, $b_n = \lambda n^p$, $\epsilon_j \overset{iid}{\sim} N(0, \sigma^2)$ and assume that μ^n is the minimizer of f_n.*

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*

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

If we further assume that $m(t) = 0$, then the following statements are true.

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

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

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

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

9

Proof. The first two statements follow from

$$\mathbb{E}\|\mu^n\|_{L^2}^2 \lesssim 2\left(\|m\|_{L^2}^2 + \text{Var}(\epsilon)\right)$$

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

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

Following the calculation in the proof of Prop. 4.1 it is easily shown that

$$\mathbb{E}\|\partial^2\mu^n\|_{L^2}^2 = \frac{16\pi^4\sigma^2}{n}\sum_{k=-\frac{n-1}{2}}^{\frac{n-1}{2}}\frac{k^4}{(1+16\pi^4\lambda n^p k^4)^2} =: S(n). \tag{10}$$

To show $S(n) \to \infty$ we will manipulate the Riemann sum approximation of

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

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

$$\int_{-\frac{1}{2}}^{\frac{1}{2}}\frac{x^4}{(1+16\pi^4\lambda x^4)^2}\, dx = n^{1+\frac{p}{4}}\int_{-\frac{1}{2}n^{-1-\frac{p}{4}}}^{\frac{1}{2}n^{-1-\frac{p}{4}}}\frac{n^{4+p}w^4}{(1+16\pi^4\lambda n^{4+p}w^4)^2}\, dw \quad \text{where } x = n^{1+\frac{p}{4}}w$$

$$\approx n^{\frac{5p}{4}}\sum_{k=-\left\lfloor\frac{1}{2}n^{-\frac{p}{4}}\right\rfloor}^{\left\lfloor\frac{1}{2}n^{-\frac{p}{4}}\right\rfloor}\frac{k^4}{(1+16\pi^4\lambda n^p k^4)^2} =: R(n).$$

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

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

$$\leq 16\pi^4\sigma^2 R(n) + \frac{2\pi^4\sigma^2}{n^{\frac{1}{5}}(1+\pi^4\lambda)^2}.$$

The remaining cases $p \in [-\frac{4}{5}, 0]$ are a consequence of (10) which implies that $p \to \mathbb{E}(\partial^2\mu)$ is non-increasing. $\quad\square$

By the Poincaré inequality we have shown 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 $\|\partial^2\mu^n\|_{L^2}$ is bounded (in expectation) in this case but we wish to make sure that we don't have the stronger result that $\|\partial^2\mu^n\|_{L^2} \to 0$.

Proposition 4.3. *With the assumptions of Proposition 4.1 and $a_n = \frac{1}{n}$, $b_n = \lambda n^p$ with $p \in [-\frac{4}{5}, 0]$ there exists a choice of m and a constant $M > 0$ such that if μ^n is the minimizer of f_n then*

$$\mathbb{E}\|\partial^2\mu^n\|_{L^2}^2 \geq M. \tag{11}$$

Proof. We only need to prove the proposition for $p = 0$ (the strongest regularization) and find one m such that (11) is true. Let $m(t) = 2\cos(2\pi t) = e^{2\pi i t} + e^{-2\pi i t}$. Then $\hat{m}_k = 0$ for $k \neq \pm 1$ and $\hat{m}_k = n$ for $k = \pm 1$. So,

$$\mathbb{E}\|\partial^2\mu^n\|_{L^2}^2 = \frac{16\pi^4}{n^2}\sum_{k=-\frac{n-1}{2}}^{\frac{n-1}{2}}\frac{k^4}{1+16\pi^4\lambda k^4}\mathbb{E}|\hat{y}_k|^2$$

$$\gtrsim \frac{16\pi^4}{n^2}\sum_{k=-\frac{n-1}{2}}^{\frac{n-1}{2}}\frac{k^4}{1+16\pi^4\lambda k^4}|\hat{m}_k|^2$$

$$= \frac{32\pi^4}{(1+16\pi^4\lambda)^2} > 0.$$

$\square$

We have shown that the minimizer is bounded for any $p \geq -\frac{4}{5}$ and $\|\partial^2 \mu^n\|_{L^2} \to 0$ 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 paper we consider the case $p = 0$. This has the advantage that not only $\mathbb{E}\|\partial^2 \mu^n\|_{L^2}$ but also $f_n(\mu^n)$ 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.

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 5.1 we show the k-means methodology can be used to determine whether two tracks have crossed.

4.2 Convergence For General Y

Let $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$ be Banach spaces where X and Y are reflexive and X is separable. 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 in Section 3 it is possible to generalize to the case where the ξ_i are not iid provided that there exists a measure P such that the empirical measure $P_n \rightharpoonup^* P$ for $\mathbb{P}$-almost every sequence of observations. 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:

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

$$f_n : Y^k \to \mathbb{R}, \quad f_n(\mu) = P_n g_\mu + \lambda r(\mu), \tag{12}$$

$$f_\infty : Y^k \to \mathbb{R}, \quad f_\infty(\mu) = P g_\mu + \lambda r(\mu). \tag{13}$$

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

$$\hat{\theta}_n = \inf_{\mu \in Y^k} f_n(\mu) \to \inf_{\mu \in Y^k} f_\infty(\mu) = \theta \quad \text{almost surely as } n \to \infty.$$

The key assumptions are given in Assumptions 2; they imply that f_n is weakly lower semi-continuous and coercive.

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

1. *For every $x \in X$ we have that $d(x, \cdot)$ is weakly lower semi-continuous.*

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

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

4. *r is coercive.*

5. *For all $\nu \in Y$ we have that $Pd(\cdot, \nu) < \infty$.*

We will follow the structure of Section 3. We start by showing that under certain conditions f_n Γ-converges to f_∞. We then show that the regularization term guarantees that the minimizers to f_n lie in a bounded set. An application of Theorem 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 norm $\|\mu\|_k := \max_j \|\mu_j\|_Y$.

Theorem 4.4. *Let $(X, \|\cdot\|_X)$ be a separable and reflexive Banach space and $(Y, \|\cdot\|_Y)$ a reflexive Banach space. 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 $\{\xi_i\}_{i=1}^n$ from P define P_n to be the empirical measure and $f_n : Y^k \to \mathbb{R}$ and $f_\infty : Y^k \to \mathbb{R}$ by (12) and (13) respectively and where $\lambda > 0$. Then*

$$f_\infty = \Gamma\text{-}\lim_n f_n \quad \text{almost surely.}$$

Proof. The proof is similar to the proof of Theorem 3.2 since r and $d(x, \cdot)$ are weakly lower semi-continuous by assumption. $\qquad\square$

Proposition 4.5. *Assuming the conditions of Theorem 4.4, there exists $R > 0$ such that*

$$\min_{\mu \in Y^k} f_n(\mu) = \min_{\|\mu\|_k \leq R} f_n(\mu) < \inf_{\|\mu\|_k > R} f_n(\mu) \quad \forall n \text{ sufficiently large}$$

almost surely. In particular R is independent of n.

Proof. Pick $\nu = (\nu_1, \nu_1, \ldots, \nu_1) \in Y^k$ then observe that $f_n(\nu) < \infty$ by Assumption 2.5. By coercivity of r there exists R such that if $\|\mu\|_k > R$ then $\lambda r(\mu) \geq f_n(\nu) + 1$. Hence $f_n(\mu) \geq f_n(\nu) + 1$. Any minimizer μ^n of f_n must therefore satisfy $\|\mu^n\|_k \leq R$. $\qquad\square$

The convergence results now follows by applying Theorem 4.4 and Proposition 4.5 to Theorem 2.1.

Theorem 4.6. *Assuming the conditions of Theorem 4.4 and Proposition 4.5 the k-means method converges in the following sense:*

$$\min_{\mu \in Y^k} f_\infty(\mu) = \lim_{n \to \infty} \min_{\mu \in Y^k} f_n(\mu) \quad \text{almost surely.}$$

Furthermore for any sequence of minimizers μ^n of f_n there exists a weakly converging subsequence $\mu^{n_m} \rightharpoonup \mu^\infty$ where μ^∞ minimizes f_∞.

It was not necessary to assume that cluster centers are in the same space. A trivial generalization would allow each $\mu_j \in Y^{(j)}$ with the cost and regularization terms appropriately defined; in this setting Theorem 4.6 holds.

4.3 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 problem is the problem of fitting a curve to a data set. 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 $\mathbb{R}^d$ be given by $|\cdot|$. Let $X = \mathbb{R} \times \mathbb{R}^d$ 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 $X' \subset [-N, N]^d$. We define the cluster center space to be $Y = H^2([0, T])$, the Sobolev space of functions from $[0, T]$ to $\mathbb{R}^d$. Clearly X is separable and reflexive and Y is reflexive. Since the support of P is contained in $([0, T] \times [-N, N]^d)^n$ we may redefine Y to be:

$$Y = \{\mu_j \in H^2([0, T]) : |\mu_j(t)| \leq N \, \forall t \in [0, T]\}. \tag{14}$$

The cost function $d : X \times Y \to [0, \infty)$ is defined by

$$d(\xi, \mu_j) = |y - \mu_j(t)|^2 \tag{15}$$

where $\mu_j \in Y$ and $\xi = (t, y) \in X$. We introduce a regularization term that penalizes the second derivative. This is a common choice in the smoothing literature, e.g. [23]. The regularization term $r : Y^k \to [0, \infty)$ is given by

$$r(\mu) = \sum_{j=1}^{k} \|\partial^2 \mu_j\|_{L^2}^2. \tag{16}$$

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

$$f_n(\mu) = \frac{1}{n} \sum_{i=1}^{n} \bigwedge_{j=1}^{k} d(\xi_i, \mu_j) + \lambda r(\mu) = \frac{1}{n} \sum_{i=1}^{n} \bigwedge_{j=1}^{k} |y_i - \mu_j(t_i)|^2 + \lambda \sum_{j=1}^{k} \|\partial^2 \mu_j\|_{L^2}^2. \tag{17}$$

Which we show will Γ-converge $\mathbb{P}$-almost surely to

$$f_\infty(\mu) = \int_X \bigwedge_{j=1}^{k} d(x, \mu_j) P(\mathrm{d}x) + \lambda r(\mu) = \int_X \bigwedge_{j=1}^{k} |y - \mu_j(t)|^2 P(\mathrm{d}x) + \lambda \sum_{j=1}^{k} \|\partial^2 \mu_j\|_{L^2}^2. \tag{18}$$

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 4.7. *Let $X = [0, T] \times \mathbb{R}^d$ and define Y by (14). Define $d : X \times Y \to [0, \infty)$ by (15) and $r : Y^k \to [0, \infty)$ by (16). For independent samples $\{\xi_i\}_{i=1}^n$ from P which has compact support $\tilde{X} \subset X$ define P_n to be the empirical measure and $f_n, f_\infty : Y^k \to \mathbb{R}$ by (17) and (18) respectively.*

Then (1) for any sequence of minimizers $\mu^n \in Y^k$ of $\mathbb{P}$-almost any f_n, there exists a weakly converging (in H^2) subsequence to some $\mu^\infty \in Y^k$ that minimizes f_∞ and (2) after potentially passing to a further subsequence the convergence is uniform (in C^0).

To prove the first part of Theorem 4.7 we are required to check that $d(x, \cdot)$ is weakly lower semi-continuous (Proposition 4.8) and r is weakly lower semi-continuous and coercive (Proposition 4.9). This statement is then a straightforward application of Theorem 4.6. Note that we will have shown the result of Theorem 4.4 holds: $f_\infty = \Gamma\text{-}\lim_n f_n$.

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 4.8. *Let $\tilde{X} = [0, T] \times [-N, N]^d$ and define Y by* (14). *Define $d : \tilde{X} \times Y \to [0, \infty)$ by* (15). *Then $d(x, \cdot)$ is weakly lower semi-continuous for any $x \in \tilde{X}$.*

Proof. Fix $x = (t, y) \in \tilde{X}$. We are required to show that for sequences $\mu_1^n, \mu_1 \in Y$ if $\mu_1^n \rightharpoonup \mu_1$ then

$$d(x, \mu_1) = |y - \mu_1(t)| \leq \liminf_{n \to \infty} |y - \mu_1^n(t)| = \liminf_{n \to \infty} d(x, \mu_1^n). \tag{19}$$

We start by showing that the sequence $\|\mu_1^n\|_Y$ is bounded in H^2. Each μ_1^n can be associated with $\Lambda^n \in Y^{**}$ by $\Lambda^n(\nu) = \nu(\mu_1^n)$ for $\nu \in Y^*$. As μ_1^n is weakly convergent it is weakly bounded. So,

$$\sup_{n \in \mathbb{N}} |\Lambda^n(\nu)| = \sup_{n \in \mathbb{N}} |\nu(\mu_1^n)| \leq M_\nu$$

for some $M_\nu < \infty$. By the uniform boundedness principle [8]

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

And so,

$$\sup_{n \in \mathbb{N}} \|\mu_1^n\|_Y = \sup_{n \in \mathbb{N}} \|\Lambda^n\|_{Y^{**}} < \infty.$$

Hence there exists $M > 0$ such that $\|\mu_1^n\|_Y \leq M$. Therefore

$$|\mu_1^n(r) - \mu_1^n(s)| = \left| \int_s^r \partial \mu_1^n \, \mathrm{d}t \right| \leq \int_s^r |\partial \mu_1^n| \, \mathrm{d}t = \int_0^T \mathbb{I}_{[s,r]} |\partial \mu_1^n| \, \mathrm{d}t \leq \|\mathbb{I}_{[s,r]}\|_{L^2} \|\partial \mu_1^n\|_{L^2} \leq M \sqrt{|r - s|}.$$

Since μ_1^n is uniformly bounded and equi-continuous then by the Arzelà–Ascoli theorem there exists a uniformly converging subsequence, say $\mu_1^{n_m} \to \nu$. By uniqueness of the weak limit $\mu_1 = \nu$. But this implies that

$$|y - \mu_1^n(t)| \to |y - \mu_1(t)|,$$

and thereby (19). Therefore $d(x, \cdot)$ is weakly lower semi-continuous as required. $\qquad\square$

Proposition 4.9. *Define Y by* (14) *and $r : Y^k \to [0, \infty)$ by* (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 $\partial^2 \mu_1^n \rightharpoonup \partial^2 \mu_1$ weakly in L^2. Hence it follows that r is weakly lower semi-continuous.

To show r is coercive let $\hat{r}(\mu_1) = \|\partial^2 \mu_1\|_{L^2}^2$ 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

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

for some C independent of μ_1. Therefore

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

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

Finally, the existence of a strongly convergent subsequence in Theorem 4.7 follows from the uniform boundedness (in L^∞) and the equi-continuity shown in Proposition 4.8. By the Arzelà–Ascoli Theorem there exists a uniformly convergent subsequence.

5 Examples

5.1 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

$$y_i^j = x^j(t_i^j) + \epsilon_i^j$$

where ϵ_i^j are iid noise terms.

The observations take the form $\xi_i = (t_i, y_i)$ 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)

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

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 Monte Carlo trials.

When the noise is bounded this is precisely the problem described in Section 4.2 with $d = 1$, 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 [19] 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 each step of this algorithm, this iteration converges to a local minimum, but not necessarily to a global minimum.

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

1. Initialize φ^0 arbitrarily.

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

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

   The above is approximated using cubic splines.

3. Data is repartitioned using the cluster centers μ^r

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

4. If $\varphi^{r+1} \neq \varphi^r$ then return to Step 2.

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

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

and the energy

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

where

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

Figure 1 shows the raw data and output of the k-means algorithm for one realization of the model. Monte Carlo trials for increasing numbers of data points, shown in Figure 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

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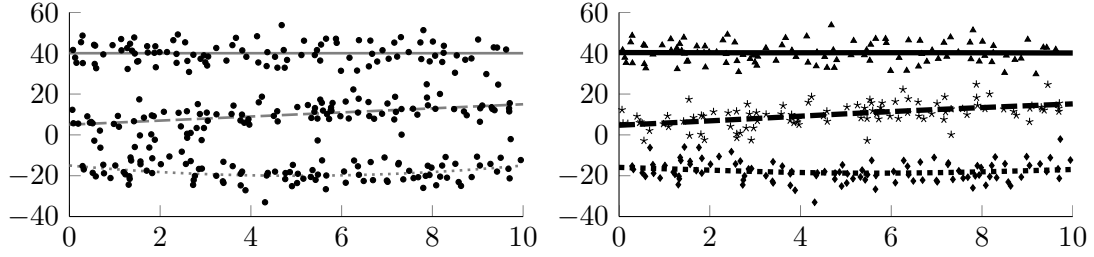
The figure on the left shows the raw data with the data generating model. That on the right shows the output of the k-means algorithm. The parameters used are: $k = 3$, $T = 10$, ϵ_i^j from a $N(0,5)$ truncated at ± 100, $\lambda = 1$, $x^1(t) = -15 - 2t + 0.2t^2$, $x^2(t) = 5 + t$ and $x^3(t) = 40$.

Figure 2: Convergence of the k-means method for 10^3 Monte Carlo trials.
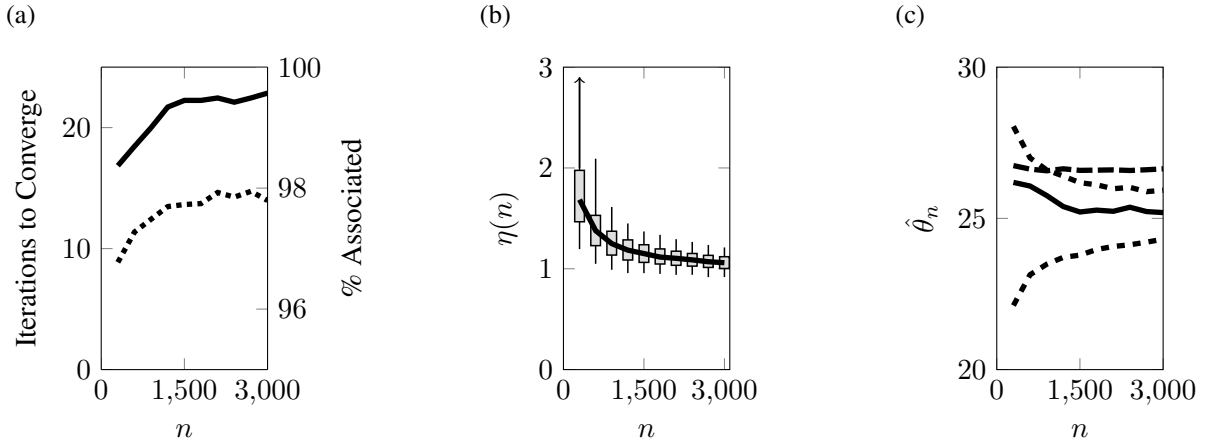
Convergence results for the parameters given in Figure 1. In (a) the dotted line corresponds to the number of iterations taken for the method to converge and the solid line corresponds to the percentage of data points correctly identified. (b) shows the median value of $\eta(n)$ (solid), interquartile 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.

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 selects not necessarily global minima of the optimization problem, we tested the algorithm on two targets that intersect paths as shown in Figure 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:
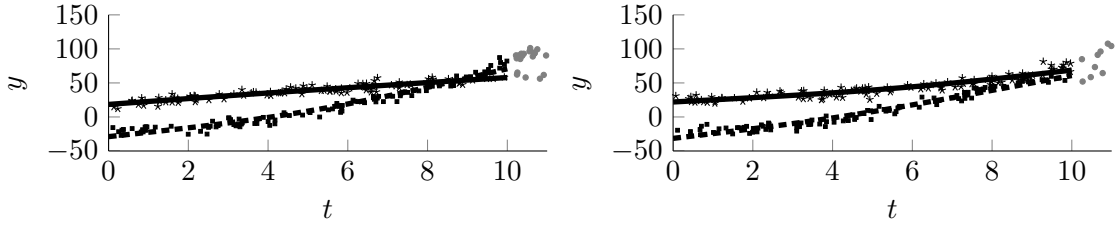
$$E_c = \frac{1}{T} f_n(\mu_c)$$

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

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

$$\Delta E = E_c - E_{nc}.$$

Figure 3: Crossing Tracks in the k-Means Method.

Typical data sets for times up to T_{max} with cluster centers, fitted up till T, exhibiting crossing and non-crossing behaviour. The parameters used are $k = 2$, $T_{min} = 9.6 \leq T \leq 11 = T_{max}$, $\epsilon_i^j \overset{iid}{\sim} N(0, 5)$, $x^1(t) = -20 + t^2$ and $x^2(t) = 20 + 4t$. There are $n = 220$ data points uniformly distributed over $[0, 11]$ with 110 observations for each track.

Figure 4: Energy Differences in Crossing Tracks in the k-Means Method.
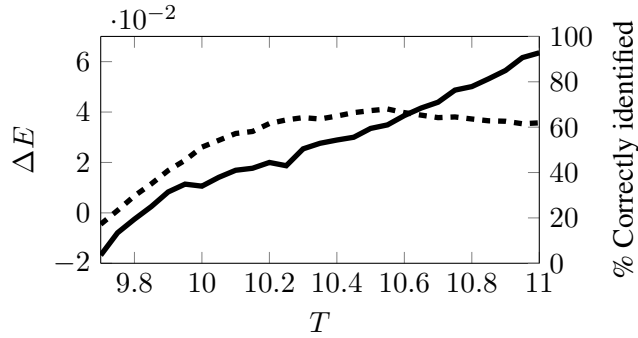
Mean results are shown for data obtained using the parameters given in Figure 3 for times, T, between T_{min} and T_{max}. The solid line shows the mean ΔE. The dashed line shows the percentage of times we correctly identified the tracks as crossing. Results were obtained by generating data and applying the k-means method until we had 10^3 estimates that crossed tracks and 10^3 estimates that did not.

The results in Figure 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 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 $T \approx 10.4$ the probability of finding the correct minima stabilizes at approximately 64%. Of course the energy itself correctly identifies the correct minimum for times greater than 9.8.

16

5.2 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 covers. 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 parametrized 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 identity 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 $j = 1, 2, \ldots k$ 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 $j = 1, 2, \ldots, k$. 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 y_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

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

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

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

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 y_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 ith 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 jth target's trajectory and time offset. The k-means minimization problem is

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

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

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

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

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

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

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

and superscript T denotes the transpose.

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

$$\mu_j^r = \underset{\mu_j}{\text{argmin}} \sum_{i \in (\varphi^{r-1})^{-1}(j)} d((t_i, a_i, p_i), \mu_j^0) \quad \text{for } 1 \leq j \leq k.$$

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

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

The method has converged when $\varphi^r = \varphi^{r-1}$ for some r. Typical simulated data and resulting trajectories are shown in Figure 5.

To illustrate the convergence result achieved above we performed a test on a set of data simulated from the same model as in Figure 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 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 6(c).

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] A. Braides. Γ-*Convergence for Beginners*. Oxford University Press, 2002.

Figure 5: Representative Data and Resulting Tracks for the Passive Tracking Example.
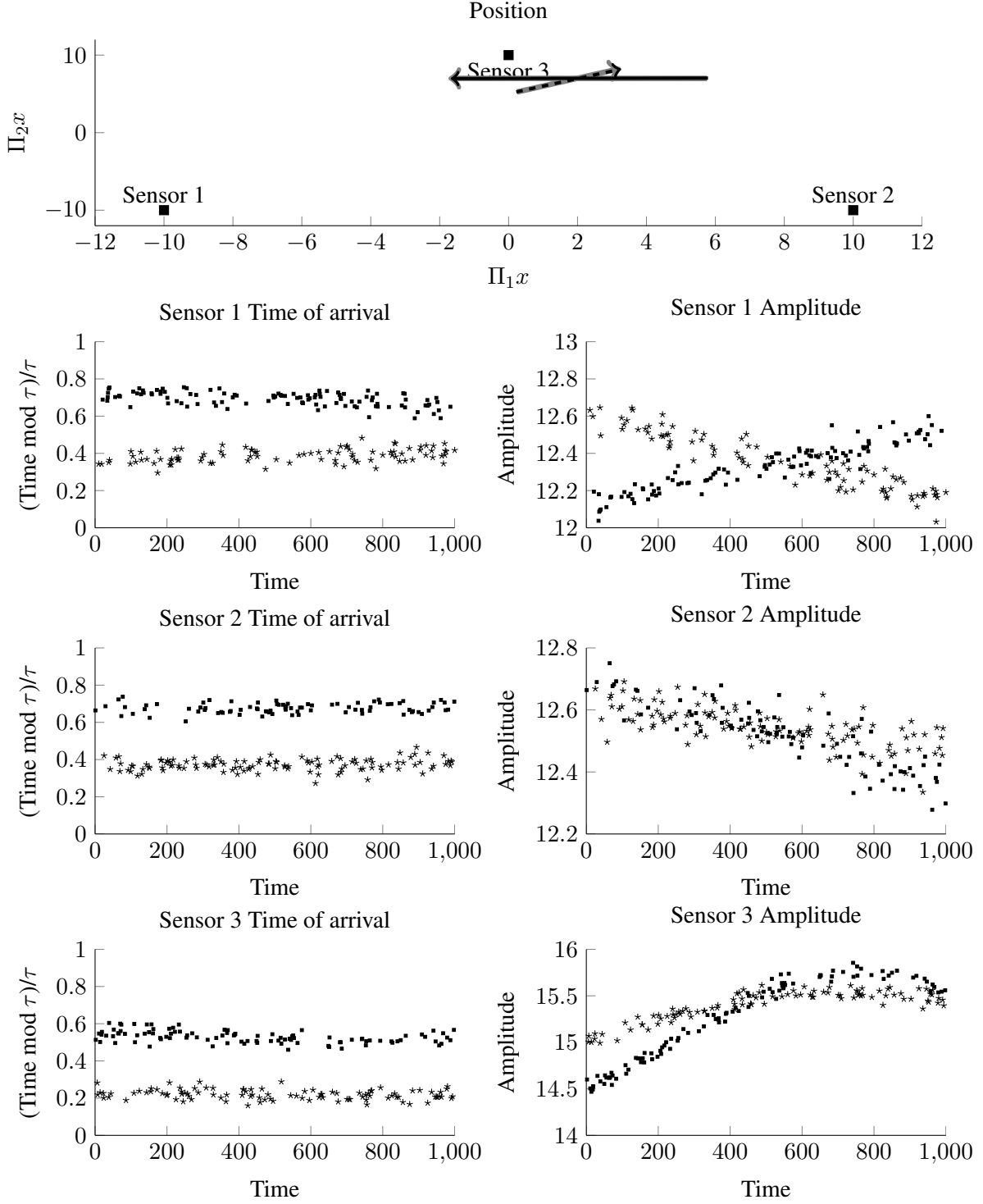
Representative data is shown for the parameters $k = 2$, $\tau = 1$, $T = 1000$, $c = 100$, $y_1 = (-10, -10)$, $y_2 = (10, -10)$, $y_3 = (0, 10)$, $\epsilon_j^p(m) \overset{\text{iid}}{\sim} N(0, 0.03^2)$, $\delta_j^p(m) \overset{\text{iid}}{\sim} N(0, 0.05^2)$, $\alpha = 10^8$, $\beta = 5$, $x_1(t) = \frac{\sqrt{2}t}{400}(1, 1) + (0, 5)$, $x_2(t) = (6, 7) - \frac{t}{125}(1, 0)$, $o_1 = 0.3$ 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.

Figure 6: Monte Carlo Convergence Results.
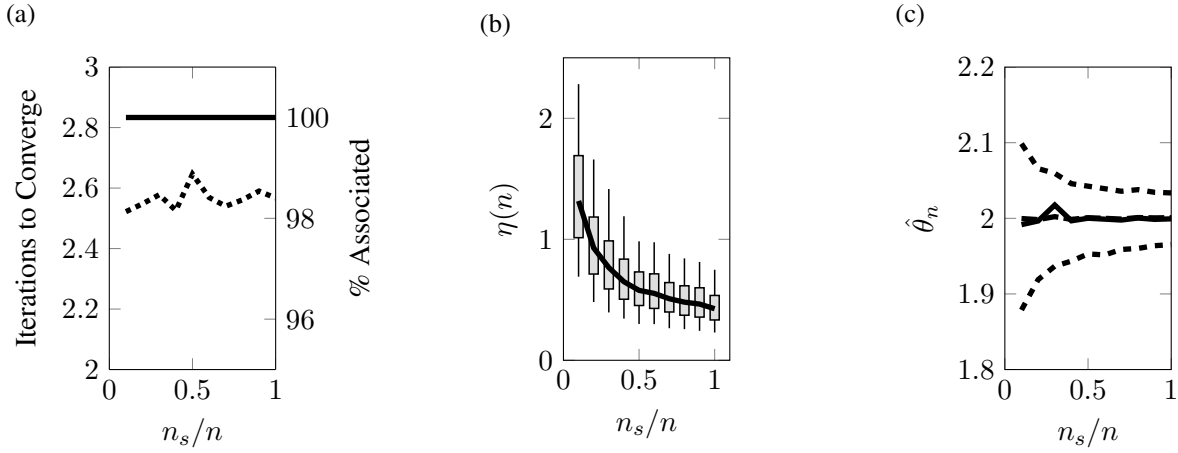
Convergence results for 10^3 Monte Carlo trials with the parameters given in Figure 5; expressed with the notation used in Figure 2. In order to increase the chance of finding a global minimum for each Monte Carlo trial out of 5 different initializations the one which had the smallest energy on terminating was recorded.

[6] L. D. Brown and M. G. Low. Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, 24(6):2384–2398, 1996.

[7] 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.

[8] J. B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer, 1990.

[9] 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.

[10] J. A. Cuesta-Albertos and R. Fraiman. Impartial trimmed k-means for functional data. *Computational Statistics & Data Analysis*, 51(10):4864–4877, 2007.

[11] 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.

[12] I. J. Good and R. A. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971.

[13] P. Hall and J. D. Opsomer. Theory for penalised spline regression. *Biometrika*, 92(1):105–118, 2005.

[14] J. Hartigan. Asymptotic distributions for clustering criteria. *The Annals of Statistics*, 6(1):117–131, 1978.

[15] T. Laloë. L_1-quantization and clustering in Banach spaces. *Mathematical Methods of Statistics*, 19(2):136–150, 2010.

[16] J. Lember. On minimizing sequences for k-centres. *Journal of Approximation Theory*, 120:20–35, 2003.

[17] T. Linder. *Principles of Nonparametric Learning*, chapter Learning-Theoretic Methods in Lossy Data Compression, pages 163–210. Springer, 2002.

[18] 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.

[19] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[20] F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1(4):502–518, 1986.

[21] D. Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135–140, 1981.

[22] D. Pollard. A central limit theorem for k-means clustering. *The Annals of Statistics*, 4(10):919–926, 1982.

[23] E. Randall. *Nonparametric regression and spline smoothing*. New York : Marcel Dekker, 1999.

[24] H. L. Royden. *Real Analysis*. Macmillan, 3rd edition, 1988.

[25] T. Tarpey and K. K. J. Kinateder. Clustering functional data. *Journal of Classification*, 2003.

[26] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[27] L. H. Zhao. Bayesian aspects of some nonparametric problems. *The Annals of Statistics*, 28(2):532–552, 2000.