Euro. Jnl of Applied Mathematics (2017), *vol.* 28, *pp.* 886–921.

doi:10.1017/S0956792517000201

A new analytical approach to consistency and overfitting in regularized empirical risk minimization

NICOLÁS GARCÍA TRILLOS[1] and RYAN MURRAY[2]

[1]*Division of Applied Mathematics, Brown University, Providence, RI, 02912, USA*
email: nicolas_garcia_trillos@brown.edu
[2]*Mathematics Department, The Pennsylvania State University, University Park, PA 16802, USA*
email: rwm22@psu.edu

(*Received 3 October 2016; revised 25 June 2017; accepted 26 June 2017; first published online 20 July 2017*)

This work considers the problem of binary classification: given training data $\mathbf{x}_1, \dots, \mathbf{x}_n$ from a certain population, together with associated labels $\mathbf{y}_1, \dots, \mathbf{y}_n \in \{0, 1\}$, determine the best label for an element $\mathbf{x}$ not among the training data. More specifically, this work considers a variant of the regularized empirical risk functional which is defined intrinsically to the observed data and does not depend on the underlying population. Tools from modern analysis are used to obtain a concise proof of asymptotic consistency as regularization parameters are taken to zero at rates related to the size of the sample. These analytical tools give a new framework for understanding overfitting and underfitting, and rigorously connect the notion of overfitting with a loss of compactness.

Key words: Overfitting, empirical risk minimization, graph total variation, discrete to continuum, classification

1991 *Mathematics Subject Classification.* 49J55, 49J45, 60D05, 68R10, 62G20

1 Introduction

The problem of classification is one of the most important problems in machine learning and statistics. In this paper, we consider the problem of binary classification: given training data $\mathbf{x}_1, \dots, \mathbf{x}_n$ from a population, together with associated labels $\mathbf{y}_1, \dots, \mathbf{y}_n \in \{0, 1\}$, determine the best label for an element $\mathbf{x}$ not among the training data. The $\mathbf{x}$ variables represent the values of certain features identifying individuals/objects in a given population; on the other hand, the $\mathbf{y}$ variables represent a group each individual belongs to. The classification problem is thus to construct, using the available training data $(\mathbf{x}_i, \mathbf{y}_i)_{i=1\dots n}$, a function, called a classifier, mapping features $\mathbf{x}$ to labels $u(\mathbf{x})$, which reflects patterns or trends exhibited in the samples. In some sense, the goal can be posed as "learning" relevant aspects of the underlying geometry of the population by observing only a finite number of samples.

Here, we follow the standard assumption that the data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_i$ are independent samples of some unknown ground-truth distribution $\boldsymbol{\nu}$. This means that $\mathbf{y}_i$ is not simply

obtained by evaluating a function at $\mathbf{x}_i$, but instead $\mathbf{y}_i$ is randomly chosen from a distribution that depends on $\mathbf{x}_i$. In other words, the labels in the training data are randomly obtained from a distribution that depends on the feature values:

$$\mathbf{y}_i \sim \mathbb{P}(\mathbf{y}_i = \cdot | \mathbf{x} = \mathbf{x}_i).$$

For our purposes, this assumption gives a robust means to account for external sources of noise and for internal uncertainty associated to an object/individual (for example, the features may not always give all of the relevant information about an individual). It is also reasonable to assume that objects with similar features have similar labels, which in this probabilistic setting means that the distribution $\mathbb{P}(\mathbf{y} = \cdot \,|\mathbf{x} = x)$ varies continuously in x.

By way of definition, a classifier is a function $u : D \to \{0, 1\}$, where we use $D \subseteq \mathbb{R}^d$ to denote the space of features for the given population. The performance, or "goodness" of any classifier is measured in terms of some risk functional. The risk functional that we consider in this paper is the average misclassifications error for data sampled from the distribution $\boldsymbol{\nu}$. More precisely, given a classifier $u : D \to \{0, 1\}$, we define its *risk* as

$$R(u) := \mathbb{E}(|u(\mathbf{x}) - \mathbf{y}|) = \int_{D \times \{0,1\}} |u(x) - y| d\boldsymbol{\nu}(x, y).$$

With respect to this risk functional, the best classifier (i.e. the one that minimizes the risk) is the *Bayes classifier*, which is the function u_B defined as

$$u_B(x) := \begin{cases} 1 & \text{if } \mathbb{P}(\mathbf{y} = 1 | \mathbf{x} = x) > 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

A central difficulty in the classification problem is that $\boldsymbol{\nu}$ is unknown, and thus we cannot compute either $R(u)$ or u_B. In fact, in some cases, the extent of D, or in other words, the support of $\boldsymbol{\nu}$, may be unknown. Given that the Bayes classifier is the best classifier, a reasonable goal is then to construct a classifier based completely on the training data, in such a way that it approximates the Bayes classifier in some asymptotic sense (as $n \to \infty$). A result of this type, namely that a family of classifiers approximates the Bayes classifier as $n \to \infty$, is known as an *asymptotic consistency result*.

One of the key difficulties in bridging the gap between the finite training sample and the unknown distribution $\boldsymbol{\nu}$ is balancing between *overfitting* and *underfitting*. When one constructs a very "complex" classifier so as to be faithful to the labels associated to the training data, it is said that the classifier overfits the data. On the other hand, when one oversimplifies the classifier by sacrificing faithfulness to the observed data, it is said that the classifier underfits the data. The so called one nearest neighbour classifier is a typical example of a classifier that overfits: for a given $x \in D$, define the label of x to be that of the point $\mathbf{x}_i$ closest to x. On the other hand, the classifier constructed by setting the label of every $x \in D$ to be the most common label among the training data, is the most extreme case of a classifier that underfits. Figure 1 shows examples of these situations. The natural question is thus: How does one construct an "ideal" classifier which neither overfits nor underfits a finite set of training data?

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

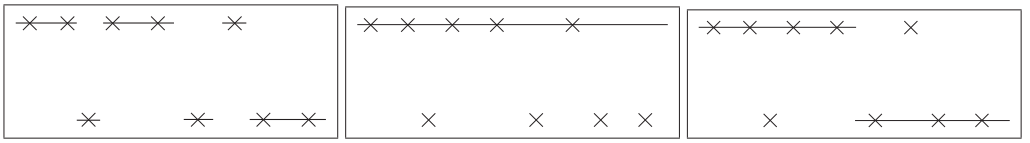

FIGURE 1. Three different classifiers for a family of data points; the x-axis represents location and the y-axis represents the labels 0 or 1. The first classifier, namely the nearest neighbour classifier $u^n_{\mathbb{1}}$, overfits the data. The second classifier picks the most common label, and underfits the data. The third classifier is the Bayes classifier.

To answer the previous question, one needs a clear mathematical notion of overfitting and underfitting. One central purpose of this paper is to give precise definitions for overfitting, underfitting and consistency as asymptotic notions ($n \to \infty$) in a concrete analytical setting introduced in Section 1.1.

Before we describe our setting, it is helpful to consider the one nearest neighbour classifier so as to get a better understanding of the problem of overfitting and the classical approaches to mitigating the same. Let $l_n : \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \to \{0, 1\}$ be the label function defined by $l_n(\mathbf{x}_i) = \mathbf{y}_i$. The one nearest neighbour classifier, $u^n_{\mathbb{1}}$, is constructed by extending the function l_n, which is only defined on the point cloud $\{\mathbf{x}_i\}_{i=1}^n$, to the whole domain D as described earlier. Since the labels $\mathbf{y}_i$ are random variables given $\mathbf{x}_i$, the function l_n may take very different values at neighbouring $\mathbf{x}_i$ and $\mathbf{x}_j$. The highly oscillatory nature of l_n means that as $n \to \infty$ the function l_n may not resemble any function u defined on the whole domain D. The function l_n will instead resemble a distribution, where at each point $x \in D$, one may have the value 1 with certain probability and the value 0 with certain probability. In the language of modern analysis, we do not have compactness in the space of measurable functions, but instead in the space of Young measures. However, each classifier $u^n_{\mathbb{1}}$ is a function that when restricted to the training data coincides with the label function l_n. In particular, it minimizes the empirical risk, which for a function $u : D \to \mathbb{R}$ is defined as

$$R_n(u) := \frac{1}{n} \sum_{i=1}^{n} |u(\mathbf{x}_i) - \mathbf{y}_i| = \frac{1}{n} \sum_{i=1}^{n} |u(\mathbf{x}_i) - l_n(\mathbf{x}_i)|.$$

Thus, if one seeks to construct a classifier via unconstrained empirical risk minimization, then even basic properties, such as being a function, may be lost in the limit. This is partly due to the limitation that the functional R_n is truly a functional defined for functions on the point cloud: $u_n : \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \to \mathbb{R}$.

Classically, the main approach for avoiding the problem of overfitting is to restrict, either explicitly or implicitly, the family of classifiers considered when trying to minimize the empirical risk R_n. After a family $\mathcal{F}$ of functions is specified, one must then prove asymptotic consistency, usually obtained by analysing the variance and the bias associated to $\mathcal{F}$. One of the first main theoretical tools developed for the purpose of analysing the variance is VC (Vapnik–Chervonenkis) theory. In VC theory, the shattering number $\mathcal{N}(\mathcal{F}, n)$ of a family of functions $\mathcal{F}$ is defined by

$$\mathcal{N}(\mathcal{F}, n) := \max_{(x_i)_{i=1\dots n}} |\mathcal{F}_{(x_i)}|,$$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

where $\mathcal{F}_{(x_i)}$ is the restriction of the functions in $\mathcal{F}$ to the set (x_i) and $|\mathcal{F}_{(x_i)}|$ is the number of distinguishable elements in $\mathcal{F}_{(x_i)}$. In essence, the shattering number gives one relevant measure of the capacity of the family of functions $\mathcal{F}$ to overfit a set of data points. One of the central results in VC theory is that if

$$\frac{\log_2 \mathcal{N}(\mathcal{F}, n)}{n} \to 0,$$

then the empirical risk R_n converges in probability uniformly (over $\mathcal{F}$) towards R. VC theory, and its many extensions, provide a powerful tool for proving asymptotic consistency. However, in many situations, estimating the shattering number of a class of functions can be a challenging combinatorial problem.

As stated, the shattering number is defined in terms of some explicit family of classifiers $\mathcal{F}$. However, it is also possible to implicitly restrict the family of classifiers by minimizing a regularized empirical risk function of the form

$$\min_{u:D\to\mathbb{R}} R_n(u) + \lambda\Omega(u),$$

where Ω is some functional measuring the complexity of the classifier u. For example, Ω may be some integral of ∇u, i.e. a TV or Sobolev norm. In this setting, λ is known as a regularization parameter, which specifies a trade-off between fidelity (R_n) and smoothness (Ω). In this context, VC theory can still be applied to the family of functions $\mathcal{F} = \{u : \Omega(u) < C\}$ if suitable combinatorial estimates are satisfied. A helpful overview of some of the classical techniques used to prove consistency is [24], and a standard reference addressing some of these topics is [21].

The classical theory outlined previously is based on classifiers that are *extrinsic* to the data, in the sense that in both cases, one considers a notion of complexity of families of functions defined on the *whole underlying domain* D. This approach is very powerful in many settings, but can be difficult to apply in practice. The extrinsic approach may also be challenging when information about D is limited and one is forced to work with families of functions defined on the whole ambient space $\mathbb{R}^d$ which may not be tailored to the geometry of D. In this paper, we take a different point of view and consider an *intrinsic* approach, namely we first seek to construct a suitable function defined on the point cloud. In particular, we focus on a *regularized empirical risk minimization* problem of the form

$$\min_{u_n} R_n(u_n) + \lambda\Omega_n(u_n), \tag{1.1}$$

where u_n is a function taking values on the point cloud $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and Ω_n is a regularizer constructed from the point cloud. This paper specifically addresses the asymptotic behaviour of minimizers of the above regularized empirical risk minimization problem when Ω_n is the *graph total variation* defined in (1.13) below. This functional depends on the construction of a proximity graph based on the point cloud and a parameter ε which specifies the connectivity of the graph.

In establishing a consistency result, we need a suitable metric for comparing functions on a point cloud, namely minimizers of (1.1), with functions defined on all of $D \subseteq \mathbb{R}^d$, namely u_B. In particular, we utilize the $TL^1(D)$ metric space introduced in [12] (see

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

(1.17) below for its definition). The $TL^1(D)$ metric space turns out to be very useful when stating our definitions of (asymptotic) overfitting, underfitting and consistency for different asymptotic regimes of λ. We show that if the regularizer is too weak (λ small with respect to ε), then the minimizers of the regularized empirical risk, despite forming a Cauchy sequence in $TL^1(D)$, do not converge to an element in the metric space $TL^1(D)$. In the completion of this metric space, the limit can be interpreted as a distribution, or Young measure, and not a function: this is an overfitting regime. If the regularizer is too strong (λ not decaying to zero), then the minimizers obtained are too regular and in the limit ($TL^1(D)$-limit), one recovers a regular function; when $\lambda \to \infty$, one recovers the most extreme case of underfitting. Finally, there is an "ideal" scaling regime where one recovers the Bayes classifier u_B in the limit: this is an asymptotic consistency result. The fact that there may be separate asymptotic regimes in the scaling parameter has been observed previously, see e.g. [1] and [15] for results of this type in the context of Bayesian inverse problems.

We also provide a simple means of constructing a classifier $u : \mathbb{R}^d \to \{0,1\}$ from the minimizer of the problem (1.1). To this end, define the Voronoi extension (or one nearest neighbour extension) of a function $u_n : \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \to \{0,1\}$ by

$$u_n^V(x) = \sum_{i=1}^{n} u_n(\mathbf{x}_i)\mathbf{1}_{V_i^n}(x), \tag{1.2}$$

where V_i^n is the set of points in D whose closest point among $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is $\mathbf{x}_i$; this set is called the Voronoi cell of the point $\mathbf{x}_i$. In simple words, the label assigned to a point $x \in D$ is the value of u_n at its closest neighbour in the set $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. The last theorem in this work proves that the Voronoi extensions of the minimizers of (1.1) indeed converge to the Bayes classifier when λ scales appropriately.

In summary, we decompose the process of constructing a classifier into two steps. The first step involves solving a discrete, convex optimization problem, namely finding a minimizer of (1.1). The second step involves extending the minimizer via the Voronoi extension. This process is *intrinsic* in the sense that it assumes no *a priori* information about the distribution, and uses only information derived from the point cloud.

There are several noteworthy features of this approach. First, the (limiting) family of classifiers attainable by this method is very broad, namely the family of BV classifiers. In other words, the structural assumptions on the limit are quite weak, giving the method significant flexibility. Second, very little information is required about the initial distribution $\boldsymbol{\nu}$. In particular, no information is needed about the support of $\boldsymbol{\nu}$, besides it being supported on some open, sufficiently regular set. The case in which $\boldsymbol{\nu}$ is supported on an embedded submanifold $\mathcal{M} \subseteq \mathbb{R}^d$ (with lower intrinsic dimension) can be addressed with similar techniques, but we will present the details elsewhere.

Our analytical framework differs from the classical learning theory approach in two main aspects. First, *regularity* of a minimizer of the functional (1.1) is enforced by the Ω_n term and an appropriate choice of the parameter λ_n. In turn, this regularity guarantees the needed compactness in the appropriate metric space so as to guarantee the asymptotic consistency and avoid overfitting. Second, we directly compare minimizers of the empirical energies with minimizers of the analogous continuum (population level)

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

energies, as opposed to studying only bounds on energy differences. Our point of view is amenable to analysis using transparent, modern tools from mathematics. These tools can be used both to prove important theoretical results, such as the consistency result of this paper, as well as to provide new insights into certain phenomena. For example, the metric that we use in this paper provide clear means for defining asymptotic notions of over and underfitting. In particular, overfitting can be seen in terms of a loss of compactness, or convergence towards a non-trivial Young measure.

Finally, we remark that while we have focussed our efforts on supervised learning, many of these ideas may be applicable in other settings. In particular, one could try to consider the semi-supervised learning problem (i.e. *graph-based semi-supervised learning*), in the setting where the labelled data is a fixed fraction of the data points. In that setting, the only difference in our set-up would be that the fidelity term in the energy (1.14) would only contain a fraction of the data points. We do not seek to address this (interesting) question in the present work.

1.1 Set-up and assumptions

To start developing the ideas presented in the introduction, we first need to be more precise about the notions and assumptions we consider in this paper.

Let $D \subseteq \mathbb{R}^d$ be a bounded, connected, open set with Lipschitz boundary. We measure the distance between two elements in D with the Euclidean distance in $\mathbb{R}^d$. We will assume that $d \geqslant 2$ throughout this work.

We let ν, the distribution of features, be given by $d\nu = \rho dx$, where $\rho : D \to \mathbb{R}$ is a continuous density function defined on D. We will assume that ρ is bounded above and below by positive constants, that is, we assume that there are constants $0 < m, M$ such that

$$m \leqslant \rho(x) \leqslant M, \quad \forall x \in D. \tag{1.3}$$

We let $\boldsymbol{\nu}$, the joint distribution of features and labels, be given by a Borel probability measure on $\mathbb{R}^d \times \mathbb{R}$ whose support is contained in $\overline{D} \times \{0, 1\}$ and whose first marginal is ν. That is, for every Borel set $A \subseteq \mathbb{R}^d$,

$$\boldsymbol{\nu}(A \times \{0, 1\}) = \nu(A \cap D) = \int_{A \cap D} \rho(x) dx.$$

For a random variable $(\mathbf{x}, \mathbf{y})$ distributed according to $\boldsymbol{\nu}$, we let $\boldsymbol{\nu}_x$ be the conditional distribution of $\mathbf{y}$ given $\mathbf{x} = x$. That is, we use the disintegration theorem to write $\boldsymbol{\nu}$ as

$$\boldsymbol{\nu}(A \times I) = \int_A \left(\int_I d\boldsymbol{\nu}_x(y) \right) d\nu(x),$$

for all Borel subsets A of D and for every interval $I \subseteq \mathbb{R}$. Expressed simply, $\boldsymbol{\nu}_x$ represents the distribution of labels of an object/individual with features $\mathbf{x} = x$.

We let $\mu : D \to \mathbb{R}$ be the conditional mean function, defined by

$$\mu(x) := \int_{\{0,1\}} y d\boldsymbol{\nu}_x(y) = \boldsymbol{\nu}_x(\{1\}) = \mathbb{P}(\mathbf{y} = 1 | \mathbf{x} = x). \tag{1.4}$$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

The Bayes classifier $u_B : D \to \mathbb{R}$ is defined by

$$u_B(x) := \begin{cases} 1, & \text{if } \mu(x) \geqslant 1/2 \\ 0, & \text{otherwise.} \end{cases} \tag{1.5}$$

It is straightforward to check that u_B is a minimizer over $L^1(v)$ of the risk functional

$$R(u) := \int_{D\times\mathbb{R}} |u(x) - y| d\boldsymbol{v}(x, y) = \int_D \left(\int_{\mathbb{R}} |u(x) - y| d\boldsymbol{v}_x(y) \right) dv(x), \tag{1.6}$$

where $L^1(v)$ is the space of real-valued functions integrable with respect to the measure v.

For ease of presentation, it will be desirable for u_B to be the unique minimizer of R. To this end, observe that on the set $\{x \in D : \mu(x) = 1/2\}$, we may modify $u(x)$ to take any value in $[0, 1]$ without increasing the value of R. Thus, for u_B to be unique, it is necessary to assume that

$$v\left(\{x \in D \ : \ \mu(x) \neq 1/2\}\right) = 1. \tag{1.7}$$

In light of (1.3), this is equivalent to the statement $\mu \neq 1/2$ Lebesgue-a.e.

The condition (1.7) is in fact sufficient for u_B to be the unique minimizer of the risk functional R over the class of $L^1(v)$-functions. Indeed, suppose that u minimizes R. It is clear that if the set where u takes values not in $[0, 1]$ has non-zero measure, then u cannot be a minimizer of R; hence, u takes values in $[0, 1]$ only. Now, given that u takes values in $[0, 1]$ only, we can write

$$\begin{aligned} R(u) &= \int_D \left(\int_{\mathbb{R}} |u(x) - y| d\boldsymbol{v}_x(y) \right) dv(x), \\ &= \int_D \left(|u(x) - 1|\mu(x) + |u(x)|(1 - \mu(x)) \right) dv(x), \\ &= \int_D \left((1 - u(x))\mu(x) + u(x)(1 - \mu(x)) \right) dv(x), \\ &= \int_D \mu(x) dv(x) + \int_D (1 - 2\mu(x))u(x) dv(x). \end{aligned} \tag{1.8}$$

Now, by the definition of u_B, for any $u(x)$ only taking values in $[0, 1]$, we have that $(1-2\mu(x))u(x) \geqslant (1-2\mu(x))u_B(x)$ for all $x \in D$. Under the assumption (1.7), this inequality can only be an equality at v a.e. x if $u = u_B$. From this, it follows that R has a unique minimizer (the Bayes classifier) if and only if the set of x with $\mu(x) = 1/2$ has v-measure zero.

In addition to assumption (1.7), which guarantees the uniqueness of minimizers for R, we also assume that $v(\{x \in D \ : \ u_B(x) = 1\}) \neq 1/2$, or in other words that the Bayes classifier has only one median. We denote by u^∞ the median of u_B, that is,

$$u^\infty := \begin{cases} 1 & \text{if } v(\{x \in D \ : \ u_B(x) = 1\}) > 1/2 \\ 0 & \text{otherwise.} \end{cases} \tag{1.9}$$

It is then straightforward to check that u^∞ is the unique minimizer of $\min_{y\in\mathbb{R}} R(y)$.

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

We additionally make some weak regularity assumptions on the functions μ and u_B. We assume that the function μ is continuous at v-a.e. $x \in D$. In particular, μ is allowed to have discontinuities as long as the set at which μ is discontinuous is v-negligible. This assumption models the continuity of the law of $\mathbf{y}$ given that $\mathbf{x} = x$, as x changes. Also, we assume that u_B is a function with finite total variation (we recall the definition of total variation in (1.16)). We notice that the assumption on the regularity of the Bayes classifier, that is the regularity of the interface between the regions where $u_B = 1$ and $u_B = 0$, is very mild. Specifically, it only requires that the interface has finite perimeter; the notion of perimeter we use is that of Caccioppoli (see [2]).

Now let us consider $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ i.i.d. samples from $\boldsymbol{v}$. These are the training data representing n objects/individuals with features $\mathbf{x}_i$ and corresponding labels $\mathbf{y}_i$. We denote by $\boldsymbol{v}_n$ the empirical measure

$$\boldsymbol{v}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{(\mathbf{x}_i, \mathbf{y}_i)},$$

and by v_n the measure

$$v_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{\mathbf{x}_i}.$$

Observe that $\boldsymbol{v}_n$ is a measure on $D \times \mathbb{R}$ and v_n a measure on D.

The labels $\mathbf{y}_i$ define a *label function* $l_n \in L^1(v_n)$, where $l_n : \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \to \{0, 1\}$ and

$$l_n(\mathbf{x}_i) := \mathbf{y}_i, \quad \forall i = 1, \ldots, n. \tag{1.10}$$

In the above and in the remainder of the paper, $L^1(v_n)$ represents the space of integrable functions with respect to the measure v_n, i.e. real-valued functions whose domain is the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

Associated to the sample $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$, we consider the *empirical risk* functional $R_n : L^1(v_n) \to \mathbb{R}$ given by

$$R_n(u_n) := \int_D |u_n(x) - l_n(x)| dv_n(x) = \frac{1}{n} \sum_{i=1}^{n} |u_n(\mathbf{x}_i) - \mathbf{y}_i|, \quad u_n \in L^1(v_n).$$

We notice that the risk functional is intrinsic to the data, as it can be defined completely in terms of the values of $(\mathbf{x}_i, \mathbf{y}_i)$ for any arbitrary function $u_n \in L^1(v_n)$. We remark that if u_n takes only values in $\{0, 1\}$, then $R_n(u_n)$ is simply the fraction of discrepancies between u_n and the labels $\mathbf{y}_i$. We also observe that using the empirical measure $\boldsymbol{v}_n$, the empirical risk functional R_n may be written as

$$R_n(u_n) = \int_{D \times \mathbb{R}} |u_n(x) - y| d\boldsymbol{v}_n(x, y).$$

When written in this form, we see that R_n resembles the true risk (1.6). The main difference between R_n and R is that the argument of R_n is a function $u_n \in L^1(v_n)$, whereas the argument of R is a function $u \in L^1(v)$.

As we stated previously, the unique minimizer of the true risk functional (1.6) is the Bayes classifier u_B defined in (1.5). On the other hand, it is evident that the function l_n

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

is the unique minimizer of the empirical risk R_n among functions $u_n \in L^1(v_n)$. Despite the resemblance between R_n and R, we cannot expect to obtain u_B as the limit of the functions l_n in any reasonable topology. As discussed in the introduction, this is due to the fact that the functions l_n are "highly oscillatory" as $n \to \infty$, and hence cannot converge to a function. To buffer the high oscillation of the functions l_n, while still being faithful to the labels $\mathbf{y}_i$, one seeks to minimize a risk functional with an extra "regularizing" term. To be more precise, we first consider a kernel $\eta : [0, \infty) \to [0, \infty)$ not identically equal to zero and satisfying the following assumptions:

(K1) η is non-increasing.

(K2) The integral $\int_0^\infty \eta(r) r^d dr$ is finite.

We note that the class of admissible kernels is broad and includes both Gaussian kernels and discontinuous kernels like one defined by η of the form $\eta = 1$ for $r \leqslant 1$ and $\eta = 0$ for $r > 1$. The assumption (K2) is equivalent to imposing that the quantity

$$\sigma_\eta := \int_{\mathbb{R}^d} \eta(|h|)|h_1|dh, \tag{1.11}$$

is finite, where h_1 is the first coordinate of the vector h. We refer to σ_η as the *surface tension* of the kernel η. Also, we will often use a slight abuse of notation and for a vector $h \in \mathbb{R}^d$ write $\eta(h)$ instead of $\eta(|h|)$.

We make an additional assumption on η, namely,

$$\eta(r) \geqslant 1, \quad \forall r \in [0, 2]. \tag{1.12}$$

This assumption is mainly for convenience, since any kernel satisfying (K1) and (K2) can be rescaled to satisfy (1.12).

Having chosen the kernel η, we choose $\varepsilon > 0$ and construct a weighted geometric graph with vertices $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$; the parameter ε defines a length scale which determines the connectivity of the point cloud. The weights of this graph are given by

$$W_{ij} := \eta_\varepsilon(\mathbf{x}_i - \mathbf{x}_j),$$

where

$$\eta_\varepsilon(z) := \frac{1}{\varepsilon^d} \eta\left(\frac{z}{\varepsilon}\right).$$

For a function $u_n \in L^1(v_n)$, namely a function whose domain is the vertices of the graph $(\{\mathbf{x}_n\}, W)$, we define the *graph total variation* by

$$GTV_{n,\varepsilon}(u_n) := \frac{1}{n^2 \varepsilon^{d+1}} \sum_{i=1}^n \sum_{j=1}^n \eta\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{\varepsilon}\right) |u_n(\mathbf{x}_i) - u_n(\mathbf{x}_j)|. \tag{1.13}$$

The graph total variation was previously used in [12, 14] in connection to approaches to clustering using balanced graph cuts.

In this work, we will analyse the *regularized empirical risk functional* given by

$$R_{n,\lambda}(u_n) := \lambda GTV_{n,\varepsilon}(u_n) + R_n(u_n), \quad u_n \in L^1(v_n). \tag{1.14}$$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

Here, $\lambda > 0$ is a parameter whose role is to emphasize or de-emphasize the effect of the regularizer $GTV_{n,\varepsilon}$. We will generally assume that λ and ε are allowed to vary as $n \to \infty$ (written λ_n and ε_n); this is natural in light of the results in [12], which require specific decay rates on ε_n.

The functional $R_{n,\lambda}$ is similar to the Rudin–Osher–Fatemi model with L^1-fidelity term used in the context of image denoising (see [5, 18]), but our setting and motivation is different from that in [5, 18], as the functional $R_{n,\lambda}$ is constructed from a random sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1\ldots n}$ of an unknown distribution $\boldsymbol{\nu}$. We remark that the L^1-fidelity term is well suited for the task of classification because it naturally generates functions valued in $\{0, 1\}$, or, in other words, sparse functions. Numerical methods designed to find an approximate minimizer of (1.14) can be found in [23]; on the other hand, an augmented Lagrangian approach to find the exact minimizer of (1.14) can be found in [8]; See also [5] and the references within.

The analogue of the functional $R_{n,\lambda}$ in the continuous setting is the functional

$$R_\lambda(u) := \lambda\sigma_\eta TV(u) + R(u), \quad u \in L^1(v), \tag{1.15}$$

where in the above, TV denotes the (weighted by ρ^2) *total variation* of the function $u \in L^1(v)$, which is defined by

$$TV(u) := \sup\left\{\int_D \operatorname{div}(\phi)u dx \ : \ \phi \in C_c^1(D : \mathbb{R}^d), \text{ and } \|\phi(x)\| \leqslant \rho^2(x), \quad \forall x \in D\right\}. \tag{1.16}$$

If the above quantity is finite, we say that $u \in L^1(v)$ is a function with bounded (weighted by ρ^2) variation. We have included the surface tension σ_η in the definition of R_λ in light of the results from [12] which state that $\sigma_\eta TV$ is the Γ-limit (we will make this precise in Theorem 2.8 below) of the functionals $GTV_{n,\varepsilon}$, when ε scales with n appropriately.

In order to state the main results of the paper, one needs a suitable metric for comparing functions in $L^1(v_n)$ with functions in $L^1(v)$. We consider the TL^1-metric space that was introduced in [12].

We denote by $\mathcal{P}(D)$ the set of Borel probability measures on D. The set $TL^1(D)$ is defined as

$$TL^1(D) := \{(\theta, f) \ : \ \theta \in \mathcal{P}(D), f \in L^1(D, \theta)\}. \tag{1.17}$$

That is, elements in $TL^1(D)$ are of the form (θ, f), where θ is a probability measure on D (in this paper, we will take v or v_n), and $f \in L^1(\theta)$, that is, f is integrable with respect to θ. This space can be seen as a formal fibre bundle over $\mathcal{P}(D)$; the fibres are the different L^1-spaces corresponding to the different Borel probability measures over D.

We endow $TL^1(D)$ with the metric

$$d_{TL^1}((\theta_1, f_1), (\theta_2, f_2)) := \inf_{\pi \in \Gamma(\theta_1, \theta_2)} \left(\iint_{D\times D} |x_1 - x_2| + |f_1(x_1) - f_2(x_2)| d\pi(x_1, x_2)\right), \tag{1.18}$$

where $\Gamma(\theta_1, \theta_2)$ represents the set of couplings, or transportation plans between θ_1 and θ_2. That is, an element $\pi \in \Gamma(\theta_1, \theta_2)$ is a Borel probability measure on $D \times D$ whose marginal on the first variable is θ_1 and whose marginal on the second variable is θ_2. In [12], it is proved that d_{TL^1} is indeed a metric.

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

Let us now discuss a characterization of TL^1-convergence of a sequence of functions $\{u_n\}_{n\in\mathbb{N}}$ with $u_n \in L^1(v_n)$ towards a function $u \in L^1(v)$; we use this characterization in the remainder. We recall that a Borel map $T_n : D \to \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is said to be a transportation map between the measures v and v_n, if for all i, $T_n^{-1}\left(\{\mathbf{x}_i\}\right)$ has v-measure equal to $1/n$ (such a map is known to exist as long as v is absolutely continuous, see e.g. [22]) . The results from [13], imply that with very high probability, i.e. probability greater than $1-n^{-\beta}$ (for β any number greater than one), there exists a transportation map T_n between v and v_n, such that

$$\|T_n - Id\|_{L^\infty(v)} \leqslant \frac{C_\beta \log(n)^{p_d}}{n^{1/d}}, \tag{1.19}$$

where p_d is a constant depending on dimension and is equal to $1/d$ for $d \geqslant 3$ and equal to $3/4$ when $d = 2$; C_β is a constant that depends on β, D and the constants from (1.3). Notice that from the Borell–Cantelli lemma and the fact that $\frac{1}{n^\beta}$ is summable, we can conclude that with probability one, we can find a sequence of transportation maps $\{T_n\}_{n\in\mathbb{N}}$, such that for all large enough n, (1.19) holds. We refer the interested reader to [13] for more background and references on the problem of finding transportation maps between some distribution and the empirical measure associated to samples drawn from it.

It is shown in [12] (see Proposition 2.2 below) that $(v_n, u_n) \xrightarrow{TL^1} (v, u)$ if and only if $u_n \circ T_n \xrightarrow{L^1(v)} u$, where T_n are the maps from (1.19) (which exist with probability one). We abuse notation a bit and simply say that $u_n \xrightarrow{TL^1} u$ in that case, understanding that $u_n \in L^1(v_n)$ and $u \in L^1(v)$.

1.2 Main results

The first main result of this paper is related to the study of the limiting behaviour of u_n^* defined by

$$u_n^* := \operatorname*{arg\,min}_{u_n \in L^1(v_n)} R_{n,\lambda_n}(u_n), \tag{1.20}$$

under different asymptotic regimes for $\{\lambda_n\}_{n\in\mathbb{N}}$.

Theorem 1.1 *Suppose that $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_n, \mathbf{y}_n), \dots$ are i.i.d. random variables distributed according to $\boldsymbol{\nu}$, where $\boldsymbol{\nu}$ satisfies the assumptions from Section 1.1. Consider a sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$ satisfying*

$$\frac{(\log(n))^{p_d}}{n^{1/d}} \ll \varepsilon_n \ll 1, \tag{1.21}$$

where $p_d = 1/d$ when $d \geqslant 3$ and $p_2 = 3/4$. Additionally, let $\{\lambda_n\}_{n\in\mathbb{N}}$ be a sequence of positive real numbers.

(1) *If $\lambda_n \ll \varepsilon_n$ as $n \to \infty$ then, with probability one, $u_n^* = l_n$ for n sufficiently large and u_n^* does not converge in the TL^1-sense towards any function $u \in L^1(v)$. In addition,*

$$\lim_{n\to\infty} R_n(u_n^*) = 0.$$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

(2) *If* $\varepsilon_n \ll \lambda_n \ll 1$ *as* $n \to \infty$ *then, with probability one,* u_n^* *converges in the* TL^1*-sense towards the Bayes classifier* u_B*. In addition,*

$$\lim_{n\to\infty} R_n(u_n^*) = R(u_B).$$

(3) *If* $\lambda_n \to \lambda \in (0,\infty)$ *as* $n \to, \infty$ *then, with probability one, every subsequence of* $\{u_n^*\}_{n\in\mathbb{N}}$ *has a further subsequence that converges to a minimizer of* R_λ *defined in* (1.15)*. In addition,*

$$\lim_{n\to\infty} R_{n,\lambda_n}(u_n^*) = \min_{u\in L^1(v)} R_\lambda(u).$$

(4) *If* $\lambda_n \to \infty$ *as* $n \to \infty$ *then, with probability one,* u_n^* *converges in the* TL^1*-sense towards the constant function* u^∞ *defined in* (1.9)*. In addition,*

$$\lim_{n\to\infty} R_n(u_n^*) = \min_{y\in\mathbb{R}} R(y).$$

Remark 1.2 *The conclusion of the theorem continues to hold even if the sequence* $\{u_n^*\}_{n\in\mathbb{N}}$ *is only assumed to be a sequence of almost minimizers of the energies* R_{n,λ_n}*. That is, we only have to assume that*

$$\lim_{n\to\infty}\left(R_{n,\lambda_n}(u_n^*) - \min_{u_n\in L^1(v_n)} R_{n,\lambda_n}(u_n)\right) = 0$$

for the conclusions of the theorem to be true.

Remark 1.3 *The assumption* (1.21) *provides a natural setting under which the geometric graph is sufficiently well-connected. This was studied in detail in [12].*

Theorem 1.1 provides a clear characterization of the asymptotic behaviour of u_n^* depending on the scaling of the parameter λ_n.

In the regime $\varepsilon_n \ll \lambda_n \ll 1$, we obtain the Bayes classifier as the limit of the functions u_n^* in the TL^1-sense. Here, we find the balance between enough regularization (so that the limit of u_n^* is a function) and enough fidelity (so that the limit of u_n^* is not just any function, but the Bayes classifier). We illustrate this regime in Figure 2. In that example, we have chosen D to be the unit square $(0,1)^2$ and the measure v was chosen to be the uniform distribution on D. The function μ determining the conditional distribution of $\mathbf{y}$ given $\mathbf{x} = x$ was chosen to take two values 0.45 and 0.55; in the upper left corner and lower right corner $\mu = 0.55$ whereas in the upper right corner and lower left corner $\mu = 0.45$. A number of samples from the resulting distribution $\boldsymbol{\nu}$ are shown in Figure 2(a). The function u_n^* was constructed using the algorithm proposed in [8]; in Figure 2(b), we present an appropriate level set of the function u_n^*.

In the regime $\lambda_n \ll \varepsilon_n$, which we will call the *overfitting regime*, the sequence of functions u_n^* minimizing R_{n,λ_n} does not converge to u_B in the TL^1 sense, and in fact, it does not converge to any function $u \in L^1(v)$. Instead, u_n^*, or in other words l_n, converges towards $\boldsymbol{\nu}$ in the completion of the $TL^1(D)$ space; see Section 2.1 for a discussion regarding the completion of $TL^1(D)$. It is important to highlight that the limit of u_n^* is not a function, but a measure (a Young measure more precisely). This type of limit is a consequence of

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

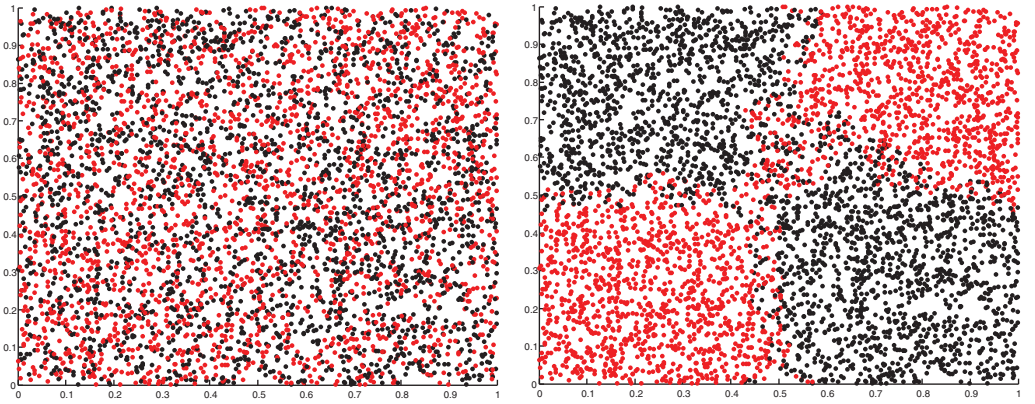

FIGURE 2. Example of consistency regime. (a) $n = 10,000$ random samples from $\boldsymbol{\nu}$ and (b) u_n^* using $\varepsilon = n^{-1/3}$ and $\lambda = n^{-1/4}$.

using a regularizer term in the functional R_{n,λ_n} that is not strong enough to control the oscillations of the label function l_n. In light of this, one could intuitively define overfitting as an asymptotic tendency towards Young measures.

When $\lambda_n \to \infty$, the functions u_n^* approach the constant function u^∞ (the median of the Bayes classifier). We may view this regime as an *underfitting regime*: the limit of the functions u_n^* is a very regular function (a constant function) that is as faithful to the labels as possible given the strong regularity constraint.

Finally, the regime $\lambda_n \to \lambda \in (0, \infty)$, interpolates between the regime in which we recover u_B and the regime in which we recover u^∞. Indeed, in this regime, we recover (up to subsequence) a function u_λ minimizing the regularized risk functional R_λ defined in (1.15). For small values of λ, u_λ should resemble the Bayes classifier, whereas for λ large u_λ should resemble u^∞. This may be viewed as a weak underfitting regime, which in the limit recovers a regularized version of the Bayes classifier.

Theorem 1.1 provides a type of consistency result for regularized empirical risk minimization as the sample size n goes to infinity. Moreover, this consistency result gives a means of characterizing the statistical notions of overfitting and underfitting through modern analytical notions (such as loss of compactness and Young measures). In this particular case, it is also possible to quantify precisely the notions of underfitting/overfitting by means of the asymptotic behaviour of the sequence $\{\lambda_n\}_{n\in\mathbb{N}}$.

However, at this stage, we have not truly addressed the classification problem. We have only given a means of constructing a suitable function u_n^* defined on the geometric graph $(\{\mathbf{x}_n\}, W)$. Thus, the natural question at this stage is how to construct a "good" classifier using u_n^*.

Given the definition of TL^1 convergence, we know that there exists a family of transportation maps T_n so that $u_n^* \circ T_n \to u_B$ in $L^1(\nu)$. However, without explicit knowledge of D and ν, it is not possible to construct the transport maps T_n. Thus, we see that while the TL^1 space and the transportation maps T_n are useful for the asymptotic analysis of the regularized empirical risk minimization problem, they *do not* immediately build a bridge between such a minimization problem and the problem of classification.

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

Fortunately, it is possible to construct a good classifier from u_n^* by simply considering its Voronoi extension. We will show that these extensions converge under slightly less general assumptions than those from Theorem 1.1 towards the Bayes classifier. This is the content of our last main result.

Theorem 1.4 *Suppose that* $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_n, \mathbf{y}_n), \ldots$ *are i.i.d. random variables distributed according to* $\boldsymbol{\nu}$*, where* $\boldsymbol{\nu}$ *satisfies the assumptions from Section 1.1. Consider a sequence* $\{\varepsilon_n\}_{n\in\mathbb{N}}$ *satisfying*

$$\frac{(\log(n))^{p_d}}{n^{1/d}} \ll \varepsilon_n \ll 1,$$

where $p_d = 1/d$ *when* $d \geqslant 3$ *and* $p_2 = 3/4$*. Additionally, let* $\{\lambda_n\}_{n\in\mathbb{N}}$ *be a sequence of positive real numbers satisfying*

$$(\log(n))^{d\cdot p_d}\varepsilon_n \ll \lambda_n \ll 1.$$

Then, with probability one,

$$u_n^{*V} \xrightarrow{L^1(\nu)} u_B, \qquad as\ n \to \infty,$$

where u_n^* *is a minimizer of* R_{n,λ_n} *and* u_n^{*V} *is the Voronoi extension (as defined in* (1.2)*) of* u_n^**.*

The bottom line is that, for $\{\lambda_n\}_{n\in\mathbb{N}}$ chosen appropriately, it is possible to construct an "intrinsic" classifier which converges towards the Bayes classifier u_B. This is constructed by first finding u_n^* using convex optimization, and then by extending using the Voronoi partition.

Remark 1.5 *In general, it is unknown whether convergence in* TL^1 *is equivalent to convergence of Voronoi extensions. The work here (e.g. the proof of Theorem 1.4) suggests that this is at least plausible under certain regularity conditions. In any case, we do not seek to address the question of the convergence of the Voronoi extensions of* u_n^* *without the hypotheses in Theorem 1.4.*

Remark 1.6 *Although in Theorem 1.4 the convergence of Voronoi extensions is only considered for a regime where one recovers the Bayes classifier, one can follow the same proof and deduce the convergence of the Voronoi extensions* u_n^{*V} *towards* u_λ *in case* $\lambda_n \to \lambda \in (0, \infty]$ *and* ε_n *satisfies the assumptions in Theorem 1.1. Intuitively, this is expected because larger* λ_n *forces* u_n^* *to be more regular. On the other hand, it is straightforward to see that Voronoi extensions in the overfitting regime will not converge towards any function in* $L^1(\nu)$ *given that for large enough n,* $u_n^*(\mathbf{x}_i) = \mathbf{y}_i$ *for all i (see Section 3.1).*

1.3 Outline

The rest of the paper is organized as follows. In Section 2, we present preliminary results that we use in the remainder of the paper. Specifically, in Section 2.1, we present some

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

relevant properties of the TL^1 space and its completion; in Section 2.2, we present the main results from [12] together with some other auxiliary results that we use in the remainder of the paper. In Section 3, we prove Theorem 1.1; we do this in three steps: in Section 3.1, we consider the overfitting regime; in Section 3.2, we consider the underfitting regime and finally in Section 3.3, we consider the intermediate regime where one obtains convergence towards the Bayes classifier. In Section 4, we establish Theorem 1.4. We conclude the paper with a discussion of our results and future work.

2 Preliminaries

2.1 The metric space TL^1

This section states some important properties of the TL^1 space.

To begin, we demonstrate that $(TL^1(D), d_{TL^1})$ is a metric space. This is accomplished by identifying the set $TL^1(D)$ with a subset of a space of probability measures over $D \times \mathbb{R}$ and by identifying the metric d_{TL^1} with the *earth mover's distance* over such space of measures.

In order to develop this idea, denote by $\mathcal{P}_1(\overline{D} \times \mathbb{R})$ the set of Borel probability measures whose support is contained in $\overline{D} \times \mathbb{R}$ and that have finite first moments, that is, $\boldsymbol{\theta} \in \mathcal{P}(\overline{D} \times \mathbb{R})$ belongs to $\mathcal{P}_1(\overline{D} \times \mathbb{R})$ if

$$\int_{D\times\mathbb{R}} (|x| + |y|) d\boldsymbol{\theta}(x, y) < \infty.$$

We define a distance d_1 between two measures $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{P}_1(\overline{D} \times \mathbb{R})$ by

$$d_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := \inf_{\pi\in\Gamma(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)} \iint_{(D\times\mathbb{R})\times(D\times\mathbb{R})} (|x_1 - x_2| + |y_1 - y_2|) d\pi(x_1, y_1, x_2, y_2).$$

The distance d_1 is a particular case of the *earth mover's distance*.

Now, given a measure $\theta \in \mathcal{P}(D)$ and a Borel map $\mathbb{T} : D \to D \times \mathbb{R}$, define the *push forward* of θ by $\mathbb{T}$ as the measure $\mathbb{T}_\sharp\theta$ in $\mathcal{P}(D \times \mathbb{R})$ defined by

$$\mathbb{T}_\sharp\theta(A \times I) = \theta\left(\mathbb{T}^{-1}(A \times I)\right), \quad \forall A \subseteq D \text{ Borel}, \quad \forall I \subseteq \mathbb{R} \text{ Borel}.$$

With the previous definitions in hand, we may now identify elements in $TL^1(D)$ with probability measures in $\mathcal{P}_1(\overline{D} \times \mathbb{R})$ using the map

$$(\theta, f) \in TL^1 \longmapsto (Id \times f)_\sharp\theta \in \mathcal{P}_1(\overline{D} \times \mathbb{R}), \tag{2.1}$$

where $Id \times f$ is the map $x \in D \mapsto (x, f(x)) \in D \times \mathbb{R}$. In other words, (θ, f) is identified with a measure supported on the graph of the function f. Notice that indeed $(Id \times f)_\sharp\theta$ has first integrable moments, due to the boundedness of the set D and the fact that $f \in L^1(D, \theta)$. Furthermore, $d_{TL^1}((\theta_1, f_1), (\theta_2, f_2)) = d_1((Id \times f_1)_\sharp\theta_1, d_1((Id \times f)_\sharp\theta))$ for any two elements $(\theta_1, f_1), (\theta_2, f_2) \in TL^1(D)$ (see [12]). That is, the map (2.1) is an isometric embedding of $TL^1(D)$ into $\mathcal{P}_1(\overline{D} \times \mathbb{R})$.

A simple example suffices to demonstrate that $(TL^1(D), d_{TL^1})$ is not a complete metric space.

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

Example 2.1 *Let $D = (0,1)$, θ be the Lebesgue measure and $f_{n+1} := \operatorname{sign}\sin(2^n \pi x)$ for $x \in (0,1)$. By constructing transport maps that swap neighbouring regions valued at ± 1, it can be shown that $d_{TL^1}((\theta, f_n),(\theta, f_{n+1})) \leqslant 1/2^n$. This implies that the sequence $\{(\theta, f_n)\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $(TL^1(D), d_{TL^1})$. However, if this was a convergent sequence, it would have to converge to an element of the form (θ, f) (see Proposition 2.2 below), but then, by Remark 2.3, it would be true that $f_n \xrightarrow{L^1(\theta)} f$. This is impossible because $\{f_n\}_{n\in\mathbb{N}}$ is not a convergent sequence in $L^1(D,\theta)$.*

The previous example illustrates the idea that highly oscillating functions (in this case, the functions f_n) do not converge to any element of $TL^1(D)$. On the other hand, since $\{(\theta, f_n)\}$ was a Cauchy sequence, it will converge in the completion of $TL^1(D)$. In fact, we can actually interpret the limit as a *Young measure* or *parameterized measure* (see [9,10,19]). Young measures are a type of generalized function, which associate each point $x \in D$ with a probability measure η_x over $\mathbb{R}$. In the example presented above, the Young measure obtained in the limit is $\eta_x = 1/2\delta_{-1} + 1/2\delta_1$. Young measures can naturally be associated with elements of $\mathcal{P}_1(\overline{D} \times \mathbb{R})$. We claim that the space $(\mathcal{P}_1(\overline{D} \times \mathbb{R}), d_1)$ is the completion of $TL^1(D)$. To see this, first note that $TL^1(D)$ can be embedded isometrically into $\mathcal{P}_1(\overline{D} \times \mathbb{R})$. Second, note that $(\mathcal{P}_1(\overline{D} \times \mathbb{R}), d_1)$ is a complete metric space (see [3]). Finally, it is shown in [12] that $TL^1(D)$ is dense in $\mathcal{P}_1(\overline{D} \times \mathbb{R})$. From the previous facts, the claim follows.

Having discussed the TL^1-space and its completion, we state a useful characterization of TL^1-convergence in terms of convergence in L^1 after composition with *transportation maps*. We recall that given two measures $\theta_1, \theta_2 \in \mathcal{P}(D)$, a Borel map $T : D \to D$ is a *transportation map* between θ_1 and θ_2, if $\theta_2 = T_\sharp \theta_1$. Notice that the condition $\theta_2 = T_\sharp \theta_1$ can be equivalently stated in terms of the change of variables formula

$$\int_D f(T(x)) d\theta_1(x) = \int_D f(z) d\theta_2(z), \tag{2.2}$$

which holds for every Borel function $f : D \to \mathbb{R}$. The following result can be found in [12].

Proposition 2.2 (Characterization of TL^1-convergence) *Let $(\theta, f) \in TL^1(D)$ and let $\{(\theta_n, f_n)\}_{n\in\mathbb{N}}$ be a sequence in $TL^1(D)$. The following statements are equivalent:*

(*i*) $(\theta_n, f_n) \xrightarrow{TL^1} (\theta, f)$ *as* $n \to \infty$.

(*ii*) $\theta_n \xrightarrow{w} \theta$ *(to be read θ_n converges weakly towards θ) and for every sequence of transportation plans $\{\pi_n\}_{n\in\mathbb{N}}$ (with $\pi_n \in \Gamma(\theta, \theta_n)$) satisfying*

$$\lim_{n\to\infty} \int |x - y| d\pi_n(x,y) = 0, \tag{2.3}$$

we have

$$\iint_{D\times D} |f(x) - f_n(y)| \, d\pi_n(x,y) \to 0, \text{ as } n \to \infty. \tag{2.4}$$

(*iii*) $\theta_n \xrightarrow{w} \theta$ *and there exists a sequence of transportation plans $\{\pi_n\}_{n\in\mathbb{N}}$ (with $\pi_n \in \Gamma(\theta, \theta_n)$) satisfying* (2.3) *for which* (2.4) *holds.*

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

Moreover, if the measure θ is absolutely continuous with respect to the Lebesgue measure, the following are equivalent to the previous statements:

(iv) $\theta_n \xrightarrow{w} \theta$ *and for every sequence of transportation maps* $\{T_n\}_{n\in\mathbb{N}}$ *(with* $T_{n\sharp}\theta = \theta_n$*) satisfying*

$$\lim_{n\to\infty} \int |T_n(x) - x| d\theta(x) = 0, \tag{2.5}$$

we have

$$\int_D |f(x) - f_n(T_n(x))| \, d\theta(x) \to 0, \text{ as } n \to \infty. \tag{2.6}$$

(v) $\theta_n \xrightarrow{w} \theta$ *and there exists a sequence of transportation maps* $\{T_n\}_{n\in\mathbb{N}}$ *(with* $T_{n\sharp}\theta = \theta_n$*) satisfying* (2.5) *for which* (2.6) *holds.*

The previous result allows us to abuse notation and talk about convergence of functions in TL^1 without having to specify the measures they are associated to. More precisely, suppose that the sequence $\{\theta_n\}_{n\in\mathbb{N}}$ in $\mathcal{P}(D)$ converges weakly to $\theta \in \mathcal{P}(D)$. We say that the sequence $\{u_n\}_{n\in\mathbb{N}}$ (with $u_n \in L^1(\theta_n)$) converges in the TL^1 sense to $u \in L^1(\theta)$, if $\{(\theta_n, u_n)\}_{n\in\mathbb{N}}$ converges to (θ, u) in the TL^1 metric space. In this case, we write $u_n \xrightarrow{TL^1} u$ as $n \to \infty$. Also, we say that the sequence $\{u_n\}_{n\in\mathbb{N}}$ (with $u_n \in L^1(\theta_n)$) is relatively compact in TL^1 if the sequence $\{(\theta_n, u_n)\}_{n\in\mathbb{N}}$ is relatively compact in TL^1. In the remainder of the paper, we use the previous proposition and observation as follows: we let $\theta_n = v_n$ (the empirical measure associated to the samples from the measure v) and let $\theta = v$; we know that with probability one, $v_n \xrightarrow{w} v$. We also know that with probability one, the maps from (1.19) exist and so for a sequence of functions $\{u_n\}_{n\in\mathbb{N}}$ with $u_n \in L^1(v_n)$, we can say $u_n \xrightarrow{TL^1} u$ for $u \in L^1(v)$ if and only if $u_n \circ T_n \xrightarrow{L^1(v)} u$. Notice that this was the characterization used right before stating Theorem 1.1.

Remark 2.3 *We finish this section by noticing that from Proposition 2.2, we can think of the convergence in TL^1 as a generalization of weak convergence of measures and of L^1 convergence of functions. That is, $\{\theta_n\}_{n\in\mathbb{N}}$ in $\mathcal{P}(D)$ converges weakly to $\theta \in \mathcal{P}(D)$ if and only if $(\theta_n, 1) \xrightarrow{TL^1} (\theta, 1)$ as $n \to \infty$; and that for fixed $\theta \in \mathcal{P}(D)$, a sequence $\{f_n\}_{n\in\mathbb{N}}$ in $L^1(\theta)$ converges in $L^1(\theta)$ to f if and only if $(\theta, f_n) \xrightarrow{TL^1} (\theta, f)$ as $n \to \infty$.*

2.2 Auxiliary properties and results

We now present the following additional properties that, as we will see, prove to be useful when establishing the main results of the paper.

Given a sequence $\{u_n\}_{n\in\mathbb{N}}$ with $u_n \in L^1(v_n)$, we say that $\{u_n\}_{n\in\mathbb{N}}$ *converges weakly* to $u \in L^1(v)$ (and denote this convergence by $u_n \rightharpoonup u$) if the sequence of functions $\{u_n \circ T_n\}_{n\in\mathbb{N}}$ converges weakly to u; the maps T_n are as in (1.19). We recall that the statement "$u_n \circ T_n$ converges weakly to u (in $L^1(v)$)", means that for every $f \in L^\infty(v)$, it is true that

$$\lim_{n\to\infty} \int_D u_n \circ T_n(x) f(x) dv(x) = \int_D u(x) f(x) dv(x).$$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

Remark 2.4 *We remark that the notion of weak convergence mentioned previously is not the same as the notion of weak convergence for measures. See [10] for more on weak convergence in $L^1(\nu)$. Although we use weak convergence for convergence of functions and convergence of measures, there should be no confusion as to what is the meaning we give to weak convergence in every specific context.*

Our first simple observation concerns the weak limit of the sequence of functions $\{l_n\}_{n\in\mathbb{N}}$.

Lemma 2.5 *With probability one, $l_n \rightharpoonup \mu$, where l_n is defined in* (1.10) *and μ is defined in* (1.4).

Proof First, recall that with probability one, the empirical measures $\boldsymbol{\nu}_n$ converge weakly to the probability measure $\boldsymbol{\nu}$ (see [4]). Second, we know that with probability one, the maps $\{T_n\}_{n\in\mathbb{N}}$ from (1.19) exist. We work on a set with probability one where both $\boldsymbol{\nu}_n \xrightarrow{w} \boldsymbol{\nu}$ and the transportation maps T_n from (1.19) exist.

Now, because $|l_n| \leqslant 1$, by the Dunford–Pettis theorem (see for example [10]), the sequence $\{l_n \circ T_n\}_{n\in\mathbb{N}}$ is weakly sequentially pre-compact, that is, every subsequence of $\{l_n \circ T_n\}$ has a further subsequence which converges weakly. Because of this, we may without the loss of generality assume that the sequence $\{l_n \circ T_n\}_{n\in\mathbb{N}}$ converges weakly to some $g \in L^1(\nu)$. Our goal is to show that $g = \mu$.

Let $f \in C_c^\infty(D)$. Then,

$$\int_D l_n \circ T_n(x) f(x) d\nu(x) = \int_D l_n \circ T_n(x)(f(x) - f(T_n(x)) d\nu(x) + \int_D l_n \circ T_n(x) f(T_n(x)) d\nu(x).$$

Observe that, again because $|l_n| \leqslant 1$,

$$\left| \int_D l_n \circ T_n(x)(f(x) - f(T_n(x)) d\nu(x) \right| \leqslant \|\nabla f\|_{L^\infty(\nu)} \cdot \int_D |x - T_n(x)| d\nu(x) \to 0, \qquad \text{as } n \to \infty.$$

Hence,

$$\int_D g(x) f(x) d\nu(x) = \lim_{n\to\infty} \int_D l_n \circ T_n(x) f(x) d\nu(x) = \lim_{n\to\infty} \int_D l_n \circ T_n(x) f(T_n(x)) d\nu(x).$$

Using the change of variables formula (2.2), and using the fact that $\boldsymbol{\nu}_n$ converges to $\boldsymbol{\nu}$ weakly, it follows that

$$\int_D g(x) f(x) d\nu(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i)\mathbf{y}_i = \int_{D\times\mathbb{R}} f(x) y d\boldsymbol{\nu}(x, y) = \int_D \mu(x) f(x) d\nu(x).$$

Since the above formula is true for every $f \in C_c^\infty(D)$, we conclude that $g = \mu$. □

We now determine the "strong" limit of the functions l_n. Indeed, we show that the functions l_n converge towards the measure $\boldsymbol{\nu}$ in the completion of $TL^1(D)$. In particular, this shows that l_n does not converge to any function $u \in L^1(\nu)$ in the TL^1-sense.

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

Lemma 2.6 *With probability one,*

$$(v_n, l_n) \xrightarrow{d_1} \boldsymbol{v} \text{ as } n \to \infty.$$

In the above, we should interpret (v_n, l_n) *as a measure in* $\mathcal{P}_1(\overline{D} \times \mathbb{R})$ *according to the identification* (2.1) *and* d_1 *is the earth mover's distance in* $\mathcal{P}_1(\overline{D} \times \mathbb{R})$.

Proof The result follows from the following simple observations. First, $(Id \times l_n)_\sharp v_n$ is nothing but $\boldsymbol{v}_n$. On the other hand, with probability one, $\boldsymbol{v}_n \xrightarrow{w} \boldsymbol{v}$. Finally, since the measures $\{\boldsymbol{v}_n\}_{n\in\mathbb{N}}$ have support contained in $\overline{D} \times [0,1]$ (a bounded subset of $\mathbb{R}^d \times \mathbb{R}$), we conclude that they have uniformly integrable first moments, and hence $\boldsymbol{v}_n \xrightarrow{w} \boldsymbol{v}$, implies that $\boldsymbol{v}_n \xrightarrow{d_1} \boldsymbol{v}$ (see Chapter 7 in [3]). □

The next observation that we will use in the remainder, concerns the continuity of the risk functionals R_n in the TL^1-sense.

Proposition 2.7 (Continuity of risk functional in the TL^1-sense) *With probability one, the following statement holds: Let* $\{u_n\}_{n\in\mathbb{N}}$ *be a sequence of* $[0,1]$*-valued functions, with* $u_n \in L^1(v_n)$. *If* $u_n \xrightarrow{TL^1} u$ *as* $n \to \infty$, *then*

$$\lim_{n\to\infty} R_n(u_n) = R(u).$$

Proof Because u_n takes values in $[0,1]$ and l_n takes values in $\{0,1\}$, we can write

$$R_n(u_n) = \int_D u_n(1-l_n)dv_n + \int_D (1-u_n)l_n dv_n = \int_D u_n dv_n + \int_D (1-2u_n)l_n dv_n.$$

Hence,

$$\lim_{n\to\infty} R_n(u_n) = \lim_{n\to\infty} \int_D u_n dv_n + \lim_{n\to\infty} \int_D (1-2u_n)l_n dv_n = \int_D u dv + \int_D (1-2u)\mu dv,$$

noticing that in the last equality we used the fact that $u_n \xrightarrow{TL^1} u$, $l_n \rightharpoonup \mu$, $|l_n| \leqslant 1$, $|u| \leqslant 1$, and Lemma 2.5. Finally, observe that the function u must take values in $[0,1]$ and thus the last expression in the above formula can be rewritten as $R(u)$. This concludes the proof. □

To finish this section, we present the main results from [12] which state that under the same assumptions on $\{\varepsilon_n\}_{n\in\mathbb{N}}$ in Theorem 1.1, the functional $\sigma_\eta TV$ is the Γ-limit of the functionals GTV_{n,ε_n} in the TL^1-sense. This result will be useful when proving Theorem 1.1 in the regime $\lambda_n \to \lambda \in (0,\infty]$.

Theorem 2.8 (Theorem 1.1, Theorem 1.2 and Corollary 1.3 in [12]) *Let the domain* D, *measure* v, *kernel* η, *sequence* $\{\varepsilon_n\}_{n\in\mathbb{N}}$, *sample points* $\{\mathbf{x}_i\}_{i\in\mathbb{N}}$, *be as in the*

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

statement of Theorem 1.1. Then, with probability one, all of the following statements hold simultaneously:

- ***Liminf inequality:*** *For every function* $u \in L^1(v)$ *and for every sequence* $\{u_n\}_{n\in\mathbb{N}}$ *with* $u_n \xrightarrow{TL^1} u$, *we have that*
$$\sigma_\eta TV(u) \leqslant \liminf_{n\to\infty} GTV_{n,\varepsilon_n}(u_n).$$
- ***Limsup Inequality:*** *For every function* $u \in L^1(v)$, *there exists a sequence* $\{u_n\}_{n\in\mathbb{N}}$ *with* $u_n \xrightarrow{TL^1} u$, *such that*
$$\limsup_{n\to\infty} GTV_{n,\varepsilon_n}(u_n) \leqslant \sigma_\eta TV(u).$$
- ***Compactness:*** *Every sequence* $\{u_n\}_{n\in\mathbb{N}}$ *satisfying*
$$\sup_{n\in\mathbb{N}} GTV_{n,\varepsilon_n}(u_n) < +\infty,$$
and
$$\sup_{n\in\mathbb{N}} \|u_n\|_{L^1(v_n)} < +\infty,$$
is pre-compact in TL^1.

Moreover, if $u \in L^1(v)$ *takes only values in* $\{0,1\}$, *then in the limsup inequality above, one may choose the functions* $u_n \in L^1(v_n)$ *to take values in* $\{0,1\}$ *as well.*

3 Proof of Theorem 1.1

3.1 Overfitting regime $\lambda_n \ll \varepsilon_n$

To prove Theorem 1.1 in the regime $\lambda_n \ll \varepsilon_n$, we use standard tools from convex analysis. The idea is simply to find the optimality conditions for u_n^*.

First, let us write $R_{n,\lambda_n}(u_n)$ as

$$\frac{\lambda_n}{\varepsilon_n n^2} J_n(u_n) + \frac{1}{n}\sum_{i=1}^n |u_n(\mathbf{x}_i) - l_n(\mathbf{x}_i)|,$$

where

$$J_n(u_n) := \sum_{i,j} \eta_{\varepsilon_n}(\mathbf{x}_i - \mathbf{x}_j)\,|u_n(\mathbf{x}_i) - u_n(\mathbf{x}_j)|\,.$$

In what follows we identify functions $f \in L^1(v_n)$ with vectors in $\mathbb{R}^n$. Namely, a function $f \in L^1(v_n)$ is identified with the vector $(f(\mathbf{x}_1),\ldots,f(\mathbf{x}_n))$. From the minimality of u_n^*, we must have

$$0 \in \frac{\lambda_n}{\varepsilon_n n^2}\partial J_n(u_n^*) + \frac{1}{n}\partial\left(\sum_{i=1}^n |u_n^*(\mathbf{x}_i) - l_n(\mathbf{x}_i)|\right),$$

where the ∂ symbol denotes sub-gradient. This inclusion implies that there exists a

$$w \in \partial\left(\sum_{i=1}^n |u_n^*(\mathbf{x}_i) - l_n(\mathbf{x}_i)|\right)$$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

such that $\frac{-n\varepsilon w}{\lambda_n} \in \partial J_n(u_n^*)$. The form of the sub-gradient of the absolute value then implies that there exists $w \in \mathbb{R}^n$ such that

$$w_i \in \begin{cases} \{1\} & \text{if } u_n^*(\mathbf{x}_i) > \mathbf{y}_i \\ [-1,1] & \text{if } u_n^*(\mathbf{x}_i) = \mathbf{y}_i \\ \{-1\} & \text{if } u_n^*(\mathbf{x}_i) < \mathbf{y}_i \end{cases} \tag{3.1}$$

for every $i = 1, \dots, n$; and such that

$$-\frac{n\varepsilon_n w}{\lambda_n} \in \partial J_n(u_n^*).$$

The Fenchel dual of J_n is defined by

$$J_n^*(f) := \sup_{g \in \mathbb{R}^n} \left\{ \sum_{i=1}^n g_i f_i - J_n(g) \right\}.$$

A straightforward consequence of this definition and the fact that $-\frac{n\varepsilon_n w}{\lambda_n} \in \partial J_n(u_n^*)$ is that (see e.g. Theorem 23.5 in [20])

$$u_n^* \in \partial J_n^* \left(-\frac{n\varepsilon_n w}{\lambda_n} \right). \tag{3.2}$$

Now, from the fact that J_n is 1-homogeneous (as can be checked easily), it follows that J_n^* has the form:

$$J_n^*(f) = \begin{cases} 0 & \text{if } f \in \mathcal{C}_n \\ \infty & \text{if } f \notin \mathcal{C}_n, \end{cases} \tag{3.3}$$

where $\mathcal{C}_n$ is a closed, convex subset of $\mathbb{R}^n$ (see e.g. Theorem 13.2 in [20]). In this case, we can give an explicit characterization of $\mathcal{C}_n$ using the following divergence operator. Given $p \in \mathbb{R}^{n^2}$, we define $\operatorname{div}(p) \in \mathbb{R}^n$ by

$$\operatorname{div}(p)_i := \sum_{j=1}^n \eta_{\varepsilon_n}(\mathbf{x}_i - \mathbf{x}_j)(p_{ji} - p_{ij}), \quad i \in \{1, \dots, n\}.$$

By reordering sums, one obtains an analogue of the divergence theorem, namely

$$\sum_{i=1}^n v_i \operatorname{div}(p)_i = \sum_{i,j=1\dots n} \eta_{\varepsilon_n}(\mathbf{x}_i - \mathbf{x}_j) p_{ij}(v_j - v_i).$$

This readily implies that

$$J_n(f) = \sup \left\{ \sum f_i r_i \,:\, r_i = \operatorname{div}(p)_i, |p_{ij}| \leqslant 1 \right\}.$$

Since $(J_n^*)^* = J_n$, we have that $J_n(f) = \sup_{r \in \mathcal{C}_n} \sum_{i=1}^n f_i r_i$, and thus we find that

$$\mathcal{C}_n = \left\{ \operatorname{div}(p) \;:\; p \in \mathbb{R}^{n^2} \text{ s.t } |p_{ij}| \leqslant 1, \ \forall i, j \right\}.$$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

From (3.2), we know in particular that $\partial J_n^*(-\frac{n\varepsilon_n w}{\lambda_n}) \neq \emptyset$. On the other hand, from (3.3), we conclude that $-\frac{n\varepsilon_n w}{\lambda_n} \in \mathcal{C}_n$. In turn, this implies that there exists $p \in \mathbb{R}^{n^2}$ with $|p_{ij}| \leqslant 1$ for all i, j and such that

$$\operatorname{div}(p) = -\frac{n\varepsilon_n w}{\lambda_n}.$$

In particular, for all $i = 1, \ldots, n$,

$$\begin{aligned} |w_i| &\leqslant \frac{2\lambda_n}{\varepsilon_n} \frac{1}{n} \sum_{j=1}^{n} \eta_{\varepsilon_n}(\mathbf{x}_i - \mathbf{x}_j) \\ &= \frac{2\lambda_n}{\varepsilon_n} \int_D \eta_{\varepsilon_n}(T_n(x) - \mathbf{x}_i)\, dv(x), \end{aligned} \tag{3.4}$$

where in the above T_n is a transportation map between v and v_n satisfying (1.19).

Let us introduce the kernel $\hat{\eta} : [0, \infty) \to \mathbb{R}$ given by

$$\hat{\eta}(r) := \begin{cases} \eta(0) & \text{if } r \in [0, 1] \\ \eta(r-1) & \text{if } r > 1. \end{cases}$$

Notice that from (1.19) and the assumptions on ε_n (i.e. (1.21)), it follows that for all large enough n, $\frac{\|Id - T_n\|_{L^\infty(v)}}{\varepsilon_n} \leqslant 1$. In particular, for all large enough n, it follows from the definition of $\hat{\eta}$ that for all $i = 1, \ldots, n$ and for all $x \in D$

$$\eta_{\varepsilon_n}(T_n(x) - \mathbf{x}_i) \leqslant \hat{\eta}_{\varepsilon_n}(x - \mathbf{x}_i).$$

Going back to (3.4), this shows that for every $i = 1, \ldots, n$

$$|w_i| \leqslant \frac{2\lambda_n}{\varepsilon_n} \int_D \hat{\eta}_{\varepsilon_n}(x - \mathbf{x}_i)\, dv(x) \leqslant \frac{2M\lambda_n}{\varepsilon_n} \int_{\mathbb{R}^d} \hat{\eta}(x) dx,$$

where we have used (1.3). Because of this, the fact that $\hat{\eta}$ is integrable (since η is integrable and $\eta(0)$ is finite), and the fact that $\frac{\lambda_n}{\varepsilon_n} \to 0$, we conclude that if n is large enough, $|w_i| < 1$ for all $i = 1, \ldots, n$.

Thus by (3.1), for n sufficiently large, we have that

$$u_n^*(\mathbf{x}_i) = \mathbf{y}_i, \quad \forall i = 1, \ldots, n.$$

In short, this means that for all large enough n, $u_n^* = l_n$. Since l_n does not converge in TL^1 to a function as $n \to \infty$ (see Lemma 2.6), we conclude that the same is true for the sequence $\{u_n^*\}_{n \in \mathbb{N}}$.

3.2 Underfitting regime: $\lambda_n \to \lambda \in (0, \infty]$

Now we establish Theorem 1.1 in the underfitting regime $\lambda_n \to \lambda \in (0, \infty]$. The main tool we have at hand to study this regime is Theorem 2.8. In particular, we will use the compactness result from Theorem 2.8.

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

First of all, notice that for every $n \in \mathbb{N}$

$$\lambda_n GTV_{n,\varepsilon_n}(u_n^*) \leqslant R_{n,\lambda_n}(u_n^*) \leqslant \inf_{y\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n |y - \mathbf{y}_i| \leqslant 1, \tag{3.5}$$

and so in particular, $GTV_{n,\varepsilon_n}(u_n^*) \leqslant \frac{1}{\lambda_n}$. Since $\lambda_n \to \lambda \in (0,\infty]$, we conclude that

$$\sup_{n\in\mathbb{N}} GTV_{n,\varepsilon_n}(u_n^*) < +\infty.$$

From the compactness statement in Theorem 2.8, we deduce that $\{u_n^*\}_{n\in\mathbb{N}}$ is pre-compact in TL^1.

Case 1 Let us assume first that $\lambda_n \to \infty$. In this case, from (3.5), we actually deduce that

$$\lim_{n\to\infty} GTV_{n,\varepsilon_n}(u_n^*) = 0. \tag{3.6}$$

Now, by the pre-compactness of $\{u_n^*\}_{n\in\mathbb{N}}$, we know that up to subsequence (that we do not relabel), $\{u_n^*\}_{n\in\mathbb{N}}$ converges in the TL^1-sense towards some $u \in L^1(v)$. From the lower semi-continuity of the graph total variation (i.e. the liminf inequality in Theorem 2.8) and from (3.6), we deduce that $TV(u) = 0$. The connectedness of the domain D implies that u is constant on D. That is, $u \equiv a$ for some $a \in \mathbb{R}$. Because, $\boldsymbol{v}_n \xrightarrow{w} \boldsymbol{v}$, we know that for every $b \in \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n |b - \mathbf{y}_i| = \int_{D\times\mathbb{R}} |b - y| d\boldsymbol{v}(x,y) = R(b).$$

On the other hand, for a given $b \in \mathbb{R}$,

$$\frac{1}{n}\sum_{i=1}^n |u_n^*(\mathbf{x}_i) - \mathbf{y}_i| \leqslant R_{n,\lambda_n}(u_n^*) \leqslant \frac{1}{n}\sum_{i=1}^n |b - \mathbf{y}_i|.$$

Additionally, from $u_n^* \xrightarrow{TL^1} a$, it is straightforward to check that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n |u_n^*(\mathbf{x}_i) - \mathbf{y}_i| = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n |a - \mathbf{y}_i| = R(a).$$

From the previous computations, we deduce that $R(a) \leqslant R(b)$ for every $b \in \mathbb{R}$. This shows that $a = u^\infty$ where u^∞ is defined in (1.9). We have just shown that for every subsequence of $\{u_n^*\}_{n\in\mathbb{N}}$, there is a further subsequence converging towards u^∞. Thus, the full sequence $\{u_n^*\}_{n\in\mathbb{N}}$ converges towards u^∞ in the TL^1-sense as we wanted to show. Finally, from Proposition 2.7, it follows that $\lim_{n\to\infty} R_n(u_n^*) = R(u^\infty) = \min_{y\in\mathbb{R}} R(y)$.

Case 2: Let us now assume that $\lambda_n \to \lambda \in (0,\infty)$. From Proposition 2.7 and from the Γ-convergence of GTV_{n,ε_n} towards $\sigma_\eta TV$ (Theorem 2.8), it is immediate that $R_{n,\lambda_n} \xrightarrow{\Gamma} R_\lambda$ as $n \to \infty$ in the TL^1-sense. Indeed, in [6], the Γ-convergence of continuous perturbations of a Γ-converging sequence is considered: in our case, we are perturbing the functionals

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

$\lambda_n GTV_{n,\varepsilon_n}$ with R_n. From the fact that $R_{n,\lambda_n} \xrightarrow{\Gamma} R_\lambda$ in the TL^1-sense and the fact that $\{u_n^*\}_{n\in\mathbb{N}}$ is pre-compact in TL^1, it follows that every subsequence of u_n^* has a further subsequence converging to a minimizer of R_λ. From the properties of Γ-convergence (see [6]), it also follows that $\lim_{n\to\infty} R_{n,\lambda_n}(u_n^*) = \min_{u\in L^1(v)} R_\lambda(u)$.

3.3 Regime $\varepsilon_n \ll \lambda_n \ll 1$

The idea of the proof of Theorem 1.1 in the regime $\varepsilon_n \ll \lambda_n \ll 1$ is as follows. We establish that if the sequence $\{u_n^*\}_{n\in\mathbb{N}}$ converges weakly to some function $u \in L^1(v)$ (recall the definition of weak convergence given at the beginning of Section 2.2), then the convergence also happens in the TL^1-sense. Then, we establish that if u_n^* converges weakly to some function $u \in L^1(v)$, and additionally

$$\lim_{n\to\infty} \int_D u_n^*(x) l_n(x) dv_n(x) = \int_D u(x)\mu(x) dv(x), \tag{3.7}$$

then u has to be equal to the Bayes classifier u_B. So in order to establish that $u_n^* \xrightarrow{TL^1} u_B$, it will be enough to show that u_n^* converges weakly to some u and that (3.7) is satisfied.

Now, since D is a bounded set in $\mathbb{R}^d$ and since all the functions $u_n^* \circ T_n$ are uniformly bounded in $L^\infty(v)$, it follows from Dunford–Pettis theorem (see for example [10]), that the sequence $\{u_n^* \circ T_n\}_{n\in\mathbb{N}}$ is weakly sequentially pre-compact, that is, every subsequence of $\{u_n^* \circ T_n\}_{n\in\mathbb{N}}$ has a further subsequence which converges weakly. Because of this, we may without the loss of generality assume that the sequence $\{u_n^*\}_{n\in\mathbb{N}}$ converges weakly to some $u \in L^1(v)$. Hence, the task is to show that (3.7) holds in the regime $\varepsilon_n \ll \lambda_n \ll 1$.

To establish (3.7), we heuristically observe that the oscillations of the functions u_n^* happen at a scale larger than ε_n, whereas the oscillations of l_n happen at a scale smaller than ε_n; the statement regarding the oscillations of the functions u_n^* is related to the fact that the energies $\lambda_n GTV_{n,\varepsilon_n}(u_n^*)$ are uniformly bounded and the fact that $\varepsilon_n \ll \lambda_n \ll 1$, on the other hand, the statement regarding the oscillations of the functions l_n is a direct consequence of concentration inequalities. Heuristically, we may think of the function u_n^* as constant on balls of radius ε_n, whereas we may view the functions l_n as rapidly oscillating on those same neighbourhoods; because of this, when integrating over such neighbourhoods, the functions l_n behave like their weak limit (i.e. the function μ, see Lemma 2.5).

There are certain connections between the ideas in the proofs here and the theory of fractional Sobolev spaces. In particular, the consistency regime $\lambda_n GTV_{n,\varepsilon_n}(u_n^*)$ has scaling similar to a fractional Sobolev seminorm. Hence, the argument that we use by approximating u_n^* with functions that are constant on a length scale ε_n is not unlike the argument used to prove the compactness of fractional Sobolev spaces, see e.g. the proof of Theorem 7.1 in [7].

With this road-map in mind, let us start making the previous statements precise.

Lemma 3.1 *With probability one, the following statement holds: Let* $\{u_n\}_{n\in\mathbb{N}}$ *be a sequence of* $[0,1]$*-valued functions, with* $u_n \in L^1(v_n)$*, and such that* $u_n \rightharpoonup u$ *for some function* $u \in L^1(v)$ *taking only the values* 0 *and* 1*. Then,* $u_n \xrightarrow{TL^1} u$ *as* $n \to \infty$.

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

Proof We may work on a set of probability one, where all the statements in Theorem 2.8 hold. Let the sequence $\{u_n\}_{n\in\mathbb{N}}$ and the function u satisfy the hypothesis in the statement of the lemma. We know that there exists a sequence $\{w_n\}_{n\in\mathbb{N}}$ with

$$w_n \xrightarrow{TL^1} u$$

and such that $w_n \in \{0,1\}$. The existence of such sequence of functions follows in particular from the last statement in Theorem 2.8. Then, from the fact that $w_n \in \{0,1\}$ and $u_n \in [0,1]$, it is straightforward to see that

$$\int_D |w_n - u_n| dv_n = \int_D u_n dv_n + \int_D (1-2u_n) w_n dv_n.$$

Using the fact that $w_n \xrightarrow{TL^1} u$ (strong convergence), $u_n \rightharpoonup u$ (weak convergence), and that u_n, w_n are uniformly bounded, we deduce that

$$\lim_{n\to\infty} \int_D |w_n - u_n| dv_n = \lim_{n\to\infty} \int_D u_n dv_n + \lim_{n\to\infty} \int_D (1-2u_n) w_n dv_n = \int_D u dv + \int_D (1-2u) u dv = 0;$$

note that in the last equality, we have used the fact that $u^2 = u$. Given that $w_n \xrightarrow{TL^1} u$, we conclude that $u_n \xrightarrow{TL^1} u$ as well. □

Lemma 3.2 *With probability one, the following statement holds: if a sequence of minimizers $\{u_n^*\}_{n\in\mathbb{N}}$ of the energies R_{n,λ_n} satisfies $u_n^* \rightharpoonup u$ for some function $u \in L^1(v)$ and in addition condition* (3.7) *holds, then $u = u_B$.*

Proof We know that with probability one, for the function u_B, there exists a sequence $\{u_n\}_{n\in\mathbb{N}}$ of $\{0,1\}$-valued functions with $u_n \in L^1(v_n)$, such that $u_n \xrightarrow{TL^1} u_B$ as $n \to \infty$ and such that $\limsup_{n\to\infty} GTV_{n,\varepsilon_n}(u_n) \leqslant \sigma_\eta TV(u_B) < +\infty$; this follows from the last statement in Theorem 2.8 and the fact that we assumed that u_B has finite total variation. From this, the fact that $\lambda_n \to 0$ and Proposition 2.7, we deduce that

$$\limsup_{n\to\infty} \lambda_n GTV_{n,\varepsilon_n}(u_n) + R_n(u_n) = R(u_B).$$

On the other hand, since u_n^* minimizes R_{n,λ_n}, we conclude that

$$\limsup_{n\to\infty} R_{n,\lambda_n}(u_n^*) \leqslant \limsup_{n\to\infty} \left(\lambda_n GTV_{n,\varepsilon_n}(u_n) + R_n(u_n)\right) = R(u_B). \tag{3.8}$$

Now, given that u_n^* minimizes R_{n,λ_n}, it is clear that u_n^* takes values in $[0,1]$ only, and thus we can write

$$R_n(u_n^*) = \int_D l_n dv_n + \int_D (1-2l_n) u_n^* dv_n.$$

From (3.7), Lemma 2.5, and the fact that $u_n^* \rightharpoonup u$, we deduce that

$$\lim_{n\to\infty} R_n(u_n^*) = \lim_{n\to\infty} \left(\int_D l_n dv_n + \int_D (1-2l_n) u_n^* dv_n \right) = \int_D \mu dv + \int_D (1-2\mu) u dv = R(u),$$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

where the last equality follows from the fact that u must take values in $[0,1]$. Since we clearly have $R_n(u_n^*) \leqslant R_{n,\lambda_n}(u_n^*)$ for every n, we deduce from the above equality and (3.8), that

$$R(u) \leqslant R(u_B).$$

The fact that u_B is the unique minimizer of R implies that $u = u_B$ as we wanted to show. □

In light of Lemmas 3.1 and 3.2, the fact that u_B takes values in $\{0,1\}$ and the discussion at the beginning of this subsection, to show that $u_n^* \xrightarrow{TL^1} u_B$, it remains to show that when $u_n^* \rightharpoonup u$ for some $u \in L^1(v)$, (3.7) holds. The remainder of the section is devoted to this purpose.

Let us consider a sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$ of positive numbers converging to zero satisfying (1.21). For every $n \in \mathbb{N}$, we consider a family of disjoint balls $B(z_1, \varepsilon_n/4), \ldots, B(z_{k_n}, \varepsilon_n/4)$ satisfying the following conditions:

(1) Every z_i belongs to D.
(2) The family of balls is maximal, in the sense that every ball $B(z, \varepsilon_n/4)$ with $z \in D$, intersects at least one of the balls $B(z_i, \varepsilon_n/4)$.

We let $S_n := \{z_1, \ldots, z_{k_n}\}$. By the maximality property of the family of balls $\{B(z, \varepsilon_n/4)\}_{z\in S_n}$, we see that $\{B(z, \varepsilon_n/2)\}_{z\in S_n}$ covers D. Moreover, we claim that there is a constant $C > 0$ such that

$$|S_n| \leqslant \frac{C}{\varepsilon_n^d}. \tag{3.9}$$

To see this, we may use the regularity assumption on the boundary of D as follows. From the fact that D is an open and bounded set with Lipschitz boundary, it follows (see [16], Theorem 1.2.2.2) that there exists a cone $\mathcal{C} \subseteq \mathbb{R}^d$ with non-empty interior and vertex at the origin, a family of rotations $\{R_x\}_{x\in D}$ and a number $1 > \zeta > 0$ such that for every $x \in D$,

$$x + R_x(\mathcal{C} \cap B(0,\zeta)) \subseteq D.$$

Thus,

$$v(B(x, \varepsilon_n/4)) = \int_{B(x,\varepsilon_n/4)\cap D} \rho(x)dx \geqslant \int_{x+\frac{\varepsilon_n}{4}(R_x(\mathcal{C}\cap B(0,\zeta)))} \rho(x)dx \geqslant \frac{m|\mathcal{C} \cap B(0,\zeta)|}{4^d}\varepsilon_n^d,$$

where $|\mathcal{C} \cap B(0,\zeta)|$ denotes the volume of $\mathcal{C} \cap B(0,\zeta)$. The bottom line is that there exists a constant $c > 0$ such that for every $x \in D$, we have

$$v(B(x, \varepsilon_n/4)) \geqslant c\varepsilon_n^d. \tag{3.10}$$

The inequality in (3.9) follows now immediately from

$$c|S_n| \cdot \varepsilon_n^d \leqslant \sum_{z\in S_n} v(B(z, \varepsilon_n/4)) = v(\cup_{z\in S_n} B(z, \varepsilon_n/4)) \leqslant v(D) = 1.$$

Let $\{\psi_z\}_{z\in S_n}$ be a smooth partition of unity subordinated to the open covering $\{B(z, \varepsilon_n)\}_{z\in S_n}$, meaning that each ψ_z is positive, smooth, has support only on $B(z, \varepsilon_n)$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

and $\sum_{z\in S_n}\psi_z(x)=1$ for all $x\in D$. We remark that the functions ψ_z can be chosen to satisfy

$$\|\nabla\psi_z\|_{L^\infty(\mathbb{R}^d)}\leqslant\frac{C}{\varepsilon_n}, \tag{3.11}$$

where $C>0$ is a constant independent of n or $z\in S_n$ (see e.g. the construction in Theorem C.21 in [17]).

The following lemma is an important first step in proving (3.7). The proof uses concentration inequalities to control oscillations on a small length scale.

Lemma 3.3 *Let $(\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2),\dots,(\mathbf{x}_n,\mathbf{y}_n),\dots$ be i.i.d. samples from $\boldsymbol{\nu}$. Assume that $\{\varepsilon_n\}_{n\in\mathbb{N}}$ is a sequence of positive numbers satisfying*

$$\frac{(\log(n))^{p_d}}{n^{1/d}}\ll\varepsilon_n\ll 1.$$

Then, with probability one

$$\lim_{n\to\infty}\sum_{z\in S_n}\left|\frac{1}{n}\sum_{i=1}^{n}(\mu(\mathbf{x}_i)-\mathbf{y}_i)\cdot\psi_z(\mathbf{x}_i)\right|=0.$$

Proof Fix $\beta>2$. Let $z\in S_n$ and let $N_z:=\#\{i\in\{1,\dots,n\}\ :\ \mathbf{x}_i\in B(z,\varepsilon_n)\}$. In the event where the transportation map T_n from (1.19) exists (this event occurs with probability at least $1-1/n^\beta$), we have that

$$\|T_n-Id\|_{L^\infty(\nu)}\leqslant C_\beta\frac{\log(n)^{p_d}}{n^{1/d}}\ll\varepsilon_n,$$

and from this, it follows that

$$\bigcup_{\mathbf{x}_i\in B(z,\varepsilon_n)}T_n^{-1}(\{\mathbf{x}_i\})\subseteq B(z,2\varepsilon_n).$$

We conclude that with probability at least $1-1/n^\beta$,

$$\frac{N_z}{n}\leqslant\nu(B(z,2\varepsilon_n))\leqslant MC_d\varepsilon_n^d, \tag{3.12}$$

where M is as in (1.3) and C_d is a constant only depending on dimension.

On the other hand, conditioned on $\mathbf{x}_i=x_i$ for $i=1,\dots,n$, the variables $\{\mathbf{y}_i\cdot\psi_z(\mathbf{x}_i)\}_{i=1,\dots,n}$ are conditionally independent and have conditional distribution:

$$\mathbf{y}_i\psi_z(\mathbf{x}_i)=\begin{cases}\psi_z(\mathbf{x}_i) & \text{with prob. } \mu(\mathbf{x}_i)\\ 0 & \text{with prob. } 1-\mu(\mathbf{x}_i).\end{cases}$$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

Hence by Hoeffding's inequality, for every $t > 0$, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\mu(\mathbf{x}_i)-\mathbf{y}_i)\cdot\psi_z(\mathbf{x}_i)\right| > t \mid \mathbf{x}_i = x_i,\ \forall i\in\{1,\dots,n\}\right) \leqslant 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(\psi_z(x_i))^2}\right) \leqslant 2\exp\left(-\frac{2n^2t^2}{N_z}\right), \quad (3.13)$$

where the second inequality follows from the fact that ψ_z is always less than 1.

From (3.12) and (3.13) (taking $t = \sqrt{\frac{\beta M C_d \log(n)\varepsilon_n^d}{2n}}$), we deduce that with probability at least $1 - 2/n^\beta$, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\mu(\mathbf{x}_i)-\mathbf{y}_i)\cdot\psi_z(\mathbf{x}_i)\right| \leqslant \sqrt{\frac{MC_d\beta\log(n)\varepsilon_n^d}{2n}}. \quad (3.14)$$

In the previous estimate, we used $z \in S_n$ fixed. Now, using a union bound (where the index set is S_n), we deduce from (3.9) that with probability at least $1 - \frac{2C}{n^\beta\varepsilon_n^d}$, (3.14) holds for every $z \in S_n$.

Therefore, with probability at least $1 - \frac{2C}{n^\beta\varepsilon_n^d}$,

$$\sum_{z\in S_n}\left|\frac{1}{n}\sum_{i=1}^{n}(\mu(\mathbf{x}_i)-\mathbf{y}_i)\cdot\psi_z(x_i)\right| \leqslant \frac{C}{\varepsilon_n^d}\cdot\sqrt{\frac{\beta\log(n)\varepsilon_n^d}{2n}} = C\sqrt{\frac{\beta\log(n)}{2n\varepsilon_n^d}},$$

where we have absorbed C_d, M and β into the constant C.

Since $\frac{1}{n^\beta\varepsilon_n^d}$ is summable (notice that $\frac{1}{n^\beta\varepsilon_n^d} \ll \frac{1}{n^{\beta-1}}$), we can use the Borel–Cantelli lemma to conclude that with probability one,

$$\lim_{n\to\infty}\sum_{z\in S_n}\left|\frac{1}{n}\sum_{i=1}^{n}(\mu(\mathbf{x}_i)-\mathbf{y}_i)\cdot\psi_z(\mathbf{x}_i)\right| = 0,$$

which is what we wanted to show. □

With all the previous lemmas at hand, we are now ready to complete the proof of Theorem 1.1.

Proof of Theorem 1.1, part (2) Following the arguments at the start of the section, we may safely assume that $u_n^* \rightharpoonup u$ for some $u \in L^1(v)$. Lemmas 3.1 and 3.2 then imply that if (3.7) holds, then $u_n^* \xrightarrow{TL^1} u_B$, which is the desired result. Hence, the remainder of the proof aims to show (3.7).

First of all, observe that

$$\sup_{n\in\mathbb{N}}\lambda_n GTV_{n,\varepsilon_n}(u_n^*) \leqslant 1, \quad (3.15)$$

which follows from the fact that for every $n \in \mathbb{N}$,

$$\lambda_n GTV_{n,\varepsilon_n}(u_n^*) \leqslant R_{n,\lambda_n}(u_n^*) \leqslant R_{n,\lambda_n}(1) = R_n(1) \leqslant 1.$$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

Consider $u_n^T := u_n^* \circ T_n$, where T_n is the transportation map from (1.19). Likewise, define $l_n^T(x) := l_n \circ T_n(x)$. Observe that for almost every $x, w \in D$, we have

$$\frac{|T_n(x) - T_n(w)|}{\varepsilon_n} \leqslant \frac{|x - w|}{\varepsilon_n} + \frac{2\|Id - T_n\|_{L^\infty(\nu)}}{\varepsilon_n}.$$

Now, given (1.19) and (1.21), we conclude that for all large enough n and for almost every $x, w \in D$, we have

$$\hat{\eta}\left(\frac{x - w}{\varepsilon_n}\right) \leqslant \eta\left(\frac{T_n(x) - T_n(w)}{\varepsilon_n}\right),$$

where $\hat{\eta}$ is defined as

$$\hat{\eta}(r) := \eta(r + 1) \text{ for } r \geqslant 0. \tag{3.16}$$

In particular, from (3.15), we deduce that

$$\sup_{n \in \mathbb{N}} \frac{\lambda_n}{\varepsilon_n} \int_{D \times D} \hat{\eta}_{\varepsilon_n}(x - w)|u_n^T(x) - u_n^T(w)|d\nu(x)d\nu(w) < \infty, \tag{3.17}$$

where we have used the change of variables (2.2) to write integrals with respect to ν_n as integrals with respect to ν.

Using again the change of variables (2.2), we can restate our original goal to be

$$\lim_{n \to \infty} \int_D u_n^T l_n^T d\nu = \int_D u\mu d\nu. \tag{3.18}$$

We show (3.18) in several steps.

First, for $z \in S_n$, we consider the average

$$\overline{u_n^T}(z) := \frac{1}{\nu\left(B(z, \varepsilon_n)\right)} \int_{B(z,\varepsilon_n)} u_n^T(w)d\nu(w).$$

Then, we notice that

$$\begin{aligned}
&\left| \int_D u_n^T(x) l_n^T(x) d\nu(x) - \sum_{z \in S_n} \overline{u_n^T}(z) \int_{B(z,\varepsilon_n)} l_n^T(x)\psi_z(x)d\nu(x) \right| \\
&= \left| \sum_{z \in S_n} \int_D u_n^T(x) l_n^T(x)\psi_z(x)d\nu(x) - \sum_{z \in S_n} \int_{B(z,\varepsilon_n)} \overline{u_n^T}(z) l_n^T(x)\psi_z(x)d\nu(x) \right| \\
&= \left| \sum_{z \in S_n} \int_{B(z,\varepsilon_n)} u_n^T(x) l_n^T(x)\psi_z(x)d\nu(x) - \sum_{z \in S_n} \int_{B(z,\varepsilon_n)} \overline{u_n^T}(z) l_n^T(x)\psi_z(x)d\nu(x) \right| \\
&\leqslant \sum_{z \in S_n} \frac{1}{\nu\left(B(z, \varepsilon_n)\right)} \int_{B(z,\varepsilon_n)} \int_{B(z,\varepsilon_n)} |u_n^T(x) - u_n^T(w)| l_n^T(x)\psi_z(x)d\nu(w)d\nu(x) \\
&\leqslant \frac{C}{\varepsilon_n^d} \int_D \int_{B(x,\varepsilon_n)} |u_n^T(x) - u_n^T(w)|d\nu(w)d\nu(x) \\
&\leqslant C \int_D \int_{B(x,\varepsilon_n)} \hat{\eta}_{\varepsilon_n}(x - w)|u_n^T(x) - u_n^T(w)|d\nu(w)d\nu(x),
\end{aligned}$$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

where in the first equality, we have used the fact that the functions $\{\psi_z\}_{z\in S_n}$ form a partition of unity; in the second equality, we have used the fact that ψ_z is supported in $B(z,\varepsilon_n)$; we have also used the fact that $|l_n^T|$ and ψ_z are bounded above by one and the fact that $v(B(z,\varepsilon_n)) \geqslant c\varepsilon_n^d$ (see (3.10)); the last inequality follows from the assumption (1.12) and the definition of $\hat{\eta}$ in (3.16).

From (3.17) and the fact that $\frac{\varepsilon_n}{\lambda_n} \to 0$ (by assumption), we deduce that

$$\lim_{n\to\infty} \left| \int_D u_n^T(x) l_n^T(x) dv(x) - \sum_{z\in S_n} \overline{u_n^T}(z) \int_{B(z,\varepsilon_n)} l_n^T(x)\psi_z(x)dv(x) \right| = 0. \tag{3.19}$$

In a similar fashion, we obtain

$$\begin{aligned} &\left| \int_D u_n^T(x)\mu(x)dv(x) - \sum_{z\in S_n} \overline{u_n^T}(z) \int_{B(z,\varepsilon_n)} \mu(x)\psi_z(x)dv(x) \right| \\ &\quad \leqslant C \int_D \int_{B(x,\varepsilon_n)} \hat{\eta}_{\varepsilon_n}(x-y)|u_n^T(x) - u_n^T(w)|dv(w)dv(x), \end{aligned} \tag{3.20}$$

and thus

$$\lim_{n\to\infty} \left| \int_D u_n^T(x)\mu(x)dv(x) - \sum_{z\in S_n} \overline{u_n^T}(z) \int_{B(z,\varepsilon_n)} \mu(x)\psi_z(x)dv(x) \right| = 0. \tag{3.21}$$

On the other hand, because $\mu \in L^\infty(D)$, the weak convergence of u_n^T towards u implies that

$$\lim_{n\to\infty} \left| \int_D u_n^T(x)\mu(x)dv(x) - \int_D u(x)\mu(x)dv(x) \right| = 0. \tag{3.22}$$

From (3.19), (3.21), (3.22) and the triangle inequality, it follows that in order to show (3.18), it is enough to show that

$$\lim_{n\to\infty} \left| \sum_{z\in S_n} \overline{u_n^T}(z) \int_{B(z,\varepsilon_n)} \mu(x)\psi_z(x)dv(x) - \sum_{z\in S_n} \overline{u_n^T}(z) \int_{B(z,\varepsilon_n)} l_n^T(x)\psi_z(x)dv(x) \right| = 0.$$

However, notice that

$$\begin{aligned} &\left| \sum_{z\in S_n} \overline{u_n^T}(z) \int_{B(z,\varepsilon_n)} \mu(x)\psi_z(x)dv(x) - \sum_{z\in S_n} \overline{u_n^T}(z) \int_{B(z,\varepsilon_n)} \mu(T_n(x))\psi_z(x)dv(x) \right| \\ &\quad \leqslant \sum_{z\in S_n} \int_{B(z,\varepsilon_n)} |\mu(x) - \mu(T_n(x))|\psi_z(x)dv(x) \\ &\quad = \sum_{z\in S_n} \int_D |\mu(x) - \mu(T_n(x))|\psi_z(x)dv(x) \\ &\quad = \int_D |\mu(x) - \mu(T_n(x))|dv(x), \end{aligned}$$

and this last term goes to zero as $n \to \infty$; this follows from the fact that μ is continuous

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

at v-a.e. $x \in D$ (as it was assumed in Section 1.1) and so $\lim_{n\to\infty} \mu(T_n(x)) = \mu(x)$ for v-a.e. $x \in D$, and by the dominated convergence theorem. Thus, to show (3.18), it is enough to show that

$$\lim_{n\to\infty} \mathcal{I}_n = 0,$$

where $\mathcal{I}_n$ is given by

$$\mathcal{I}_n := \left| \sum_{z\in S_n} \overline{u_n^T}(z) \int_{B(z,\varepsilon_n)} \mu(T_n(x))\psi_z(x)dv(x) - \sum_{z\in S_n} \overline{u_n^T}(z) \int_{B(z,\varepsilon_n)} l_n^T(x)\psi_z(x)dv(x) \right|.$$

Now, for fixed $z \in S_n$,

$$\begin{aligned}\int_{B(z,\varepsilon_n)} (\mu(T_n(x)) - l_n(T_n(x)))\,\psi_z(x)dv(x) &= \int_D (\mu(T_n(x)) - l_n(T_n(x)))\,\psi_z(x)dv(x)\\ &= \int_D (\mu(T_n(x)) - l_n(T_n(x)))\,(\psi_z(x) - \psi_z(T_n(x)))dv(x)\\ &\quad + \int_D (\mu(T_n(x)) - l_n(T_n(x)))\,\psi_z(T_n(x))dv(x). \quad (3.23)\end{aligned}$$

Observe that

$$\begin{aligned}&\left| \int_D (\mu(T_n(x)) - l_n(T_n(x)))\,(\psi_z(x) - \psi_z(T_n(x)))dv(x) \right|\\ &\quad = \left| \int_{B(z,2\varepsilon_n)} (\mu(T_n(x)) - l_n(T_n(x)))\,(\psi_z(x) - \psi_z(T_n(x)))dv(x) \right|\\ &\quad \leqslant v(B(z,2\varepsilon_n)) \cdot \sup_{x\in B(z,2\varepsilon_n)} |\psi_z(x) - \psi_z(T_n(x))|\\ &\quad \leqslant Cv(B(z,2\varepsilon_n)) \frac{\|Id - T_n\|_{L^\infty(v)}}{\varepsilon_n},\end{aligned}$$

where the first equality comes from the fact that $\|Id - T_n\|_{L^\infty(v)} < \varepsilon_n$ and the last inequality follows from (3.11). The previous computations imply that

$$\begin{aligned}\mathcal{I}_n &\leqslant \frac{C\|Id - T_n\|_{L^\infty(v)}}{\varepsilon_n} \cdot \sum_{z\in S_n} \overline{u_n^T}(z) v(B(z,2\varepsilon_n))\\ &\quad + \sum_{z\in S_n} \overline{u_n^T}(z) \left| \int_D (\mu(T_n(x)) - l_n(T_n(x)))\,\psi_z(T_n(x))dv(x) \right|\\ &\leqslant \frac{C\|Id - T_n\|_{L^\infty(v)}}{\varepsilon_n} \cdot \sum_{z\in S_n} v(B(z,2\varepsilon_n)) + \sum_{z\in S_n} \left| \frac{1}{n} \sum_{i=1}^n (\mu(\mathbf{x}_i) - \mathbf{y}_i) \cdot \psi_z(\mathbf{x}_i) \right|\\ &\leqslant \frac{C\|Id - T_n\|_{L^\infty(v)}}{\varepsilon_n} + \sum_{z\in S_n} \left| \frac{1}{n} \sum_{i=1}^n (\mu(\mathbf{x}_i) - \mathbf{y}_i) \cdot \psi_z(\mathbf{x}_i) \right|, \quad (3.24)\end{aligned}$$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

where in the above we have used the change of variables formula (2.2) to write

$$\int_D (\mu(T_n(x)) - l_n(T_n(x)))\, \psi_z(T_n(x)) dv(x) = \frac{1}{n} \sum_{i=1}^{n} (\mu(\mathbf{x}_i) - \mathbf{y}_i)\, \psi_z(\mathbf{x}_i),$$

we have also used the fact that $\overline{u_n^T}(z)$ is less than one for every $z \in S_n$, and the fact that $\sum_{z \in S_n} v(B(z, 2\varepsilon_n))$ is bounded by (3.9) and (3.12) (here we are letting C change from line to line).

The first term in the last line of (3.24) converges to zero as $n \to \infty$ (this follows from (1.19) and (1.21)); on the other hand, Lemma 3.3 shows that the second term also converges to zero. Hence, $\lim_{n\to\infty} \mathcal{I}_n = 0$ and this finishes the proof. □

4 Proof of Theorem 1.4

We now move to the proof of Theorem 1.4. We impose the additional constraint:

$$(\log(n))^{d \cdot p_d} \varepsilon_n \ll \lambda_n \ll 1.$$

Let us again denote by u_n^T the function $u_n^T := u_n^* \circ T_n$, where $\{T_n\}_{n\in\mathbb{N}}$ is the sequence of transportation maps from (1.19). Up to this point, we have established that when ε_n satisfies (1.21) and λ_n satisfies $\varepsilon_n \ll \lambda_n \ll 1$, then with probability one, the functions u_n^* converge in the TL^1 sense towards the Bayes classifier u_B; by the very definition of TL^1 convergence, this is equivalent to saying that u_n^T converges in the $L^1(v)$ sense towards u_B. Now we would like to say that the same convergence result holds for the sequence of functions $\{u_n^V\}_{n\in\mathbb{N}}$, where u_n^V is the Voronoi extension (as defined in (1.2)) of the function u_n^*. Some of the ideas that follow are also present in [11], a paper that is in preparation.

Let us consider $\tilde{\varepsilon}_n := \varepsilon_n - 2\|T_n - Id\|_{L^\infty(v)}$. From the assumptions on ε_n and from (1.19), it is clear that for large enough n, $\tilde{\varepsilon}_n > 0$, so without the loss of generality we assume this holds for all n.

Now,

$$\begin{aligned}
&\int_D |u_n^T(x) - u_n^V(x)| dv(x) \\
&\quad = \int_D \left(\frac{1}{v(B(x, \tilde{\varepsilon}_n))} \int_{B(x,\tilde{\varepsilon}_n)} |u_n^T(x) - u_n^V(x)| dv(w) \right) dv(x) \\
&\quad = \int_D \left(\frac{1}{v(B(x, \tilde{\varepsilon}_n))} \int_{B(x,\tilde{\varepsilon}_n)} |u_n^T(x) - u_n^T(w) + u_n^T(w) - u_n^V(x)| dv(w) \right) dv(x) \\
&\quad \leqslant \int_D \left(\frac{1}{v(B(x, \tilde{\varepsilon}_n))} \int_{B(x,\tilde{\varepsilon}_n)} |u_n^T(x) - u_n^T(w)| dv(w) \right) dv(x) \\
&\qquad + \int_D \left(\frac{1}{v(B(x, \tilde{\varepsilon}_n))} \int_{B(x,\tilde{\varepsilon}_n)} |u_n^T(w) - u_n^V(x)| dv(w) \right) dv(x)
\end{aligned}$$

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

$$\begin{aligned}
&\leqslant \frac{C}{\tilde{\varepsilon}_n^d}\int_D\int_{B(x,\tilde{\varepsilon}_n)} |u_n^T(x)-u_n^T(w)|dv(w)dv(x)\\
&\quad + \frac{C}{\tilde{\varepsilon}_n^d}\int_D\int_{B(x,\tilde{\varepsilon}_n)} |u_n^T(w)-u_n^V(x)|dv(w)dv(x)\\
&=: C(\mathcal{I}_n^1+\mathcal{I}_n^2),
\end{aligned}$$

where the last inequality follows from (3.10). We will show now that $\int_D |u_n^T(x)-u_n^V(x)|dv(x)$ converges to zero as $n\to\infty$ by showing that each of the terms $\mathcal{I}_n^1, \mathcal{I}_n^2$ converges to zero as $n\to\infty$. Since $u_n^T \xrightarrow{L^1(v)} u_B$ as $n\to\infty$, this will establish that $u_n^V \xrightarrow{L^1(v)} u_B$ as $n\to\infty$.

Let us first show that $\mathcal{I}_n^1\to 0$ as $n\to\infty$. Notice that for almost every $x,w\in D$, it is true that if $|T_n(x)-T_n(w)|>\varepsilon_n$, then $|x-w|>\tilde{\varepsilon}_n$. In particular, we see that for almost every $x,w\in D$

$$\frac{1}{\tilde{\varepsilon}_n^d}\mathbb{1}_{|x-w|\leqslant\tilde{\varepsilon}_n} \leqslant \frac{1}{\tilde{\varepsilon}_n^d}\mathbb{1}_{|T_n(x)-T_n(w)|\leqslant\varepsilon_n} \leqslant \left(\frac{\varepsilon_n}{\tilde{\varepsilon}_n}\right)^d \eta_{\varepsilon_n}(T_n(x)-T_n(w)),$$

where the last inequality follows using (1.12). Then, it follows that

$$\mathcal{I}_n^1 \leqslant \left(\frac{\varepsilon_n}{\tilde{\varepsilon}_n}\right)^d \int_D\int_D \eta_{\varepsilon_n}(T_n(x)-T_n(w))|u_n^T(x)-u_n^T(w)|dv(w)dv(x).$$

From the previous inequality and the change of variables formula (2.2), we deduce that

$$\mathcal{I}_n^1 \leqslant \frac{\varepsilon_n}{\lambda_n}\left(\frac{\varepsilon_n}{\tilde{\varepsilon}_n}\right)^d \lambda_n GTV_{n,\varepsilon_n}(u_n^*).$$

From (3.15), the fact that $\frac{\varepsilon_n}{\lambda_n}\to 0$ and $\frac{\varepsilon_n}{\tilde{\varepsilon}_n}\to 1$, it follows that $\mathcal{I}_n^1\to 0$ as $n\to\infty$.

Now let us estimate the term $\mathcal{I}_n^2$. Let us denote by $U_1^n,\dots,U_n^n$ the partition of D induced by T_n, that is,

$$U_i^n := T_n^{-1}(\mathbf{x}_i).$$

Also, let us denote by $V_1^n,\dots,V_n^n$ the Voronoi partition of D associated to the points $\mathbf{x}_1,\dots,\mathbf{x}_n$, that is,

$$V_i^n := \left\{x\in D \,:\, |x-\mathbf{x}_i| = \min_{j=1,\dots,n}|x-\mathbf{x}_j|\right\}.$$

Observe that if $x\in U_i^n$ and $w\in V_j^n$, then

$$\begin{aligned}
|\mathbf{x}_i-\mathbf{x}_j| &\leqslant |\mathbf{x}_i-x|+|x-w|+|w-\mathbf{x}_j|\\
&= |T_n(x)-x|+|x-w|+|w-\mathbf{x}_j|\\
&\leqslant |T_n(x)-x|+|x-w|+|w-T_n(w)|\\
&\leqslant |x-w|+2\|T_n-Id\|_{L^\infty(v)},
\end{aligned}$$

where the second inequality follows from the fact that the closest point to w among the

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is $\mathbf{x}_j$. In particular, we see that for $x \in U_i^n$ and $w \in V_j^n$,

$$\frac{1}{\tilde{\varepsilon}_n^d} \mathbb{1}_{|x-w| \leqslant \tilde{\varepsilon}_n} \leqslant \frac{1}{\tilde{\varepsilon}_n^d} \mathbb{1}_{|\mathbf{x}_i - \mathbf{x}_j| \leqslant \varepsilon_n} \leqslant \left(\frac{\varepsilon_n}{\tilde{\varepsilon}_n}\right)^d \eta_{\varepsilon_n}(\mathbf{x}_i - \mathbf{x}_j).$$

From the previous observation, we see that

$$\begin{aligned}
\mathcal{I}_n^2 &= \frac{1}{\tilde{\varepsilon}_n^d} \sum_{i,j} \int_{U_i^n} \int_{V_j^n} \mathbb{1}_{|x-w| \leqslant \tilde{\varepsilon}_n} \cdot |u_n^*(\mathbf{x}_i) - u_n^*(\mathbf{x}_j)| dv(w) dv(x) \\
&\leqslant \left(\frac{\varepsilon_n}{\tilde{\varepsilon}_n}\right)^d \sum_{i,j} \int_{U_i^n} \int_{V_j^n} \eta_{\varepsilon_n}(\mathbf{x}_i - \mathbf{x}_j) |u_n^*(\mathbf{x}_i) - u_n^*(\mathbf{x}_j)| dv(w) dv(x) \\
&= \left(\frac{\varepsilon_n}{\tilde{\varepsilon}_n}\right)^d \sum_{i,j} \eta_{\varepsilon_n}(\mathbf{x}_i - \mathbf{x}_j) |u_n^*(\mathbf{x}_i) - u_n^*(\mathbf{x}_j)| v(V_j^n) v(U_i^n) \\
&= \left(\frac{\varepsilon_n}{\tilde{\varepsilon}_n}\right)^d \frac{1}{n} \sum_{i,j} \eta_{\varepsilon_n}(\mathbf{x}_i - \mathbf{x}_j) |u_n^*(\mathbf{x}_i) - u_n^*(\mathbf{x}_j)| v(V_j^n) \\
&\leqslant \left(\frac{\varepsilon_n}{\tilde{\varepsilon}_n}\right)^d \cdot \left(\max_{j=1,\ldots,n} n \cdot v(V_j^n)\right) \cdot \frac{1}{n^2} \sum_{i,j} \eta_{\varepsilon_n}(\mathbf{x}_i - \mathbf{x}_j) |u_n^*(\mathbf{x}_i) - u_n^*(\mathbf{x}_j)| \\
&= \left(\frac{\varepsilon_n}{\tilde{\varepsilon}_n}\right)^d \cdot \left(\max_{j=1,\ldots,n} n \cdot v(V_j^n)\right) \frac{\varepsilon_n}{\lambda_n} \lambda_n G T V_{n,\varepsilon_n}(u_n^*),
\end{aligned} \tag{4.1}$$

where the third equality follows from the fact that $v(U_i^n) = \frac{1}{n}$ for every $i = 1, \ldots, n$. Now, for an arbitrary $j = 1, \ldots, n$, notice that if $w \in V_j^n$, then $|w - \mathbf{x}_j| \leqslant |w - T_n(w)| \leqslant C \frac{(\log(n))^{p_d}}{n^{1/d}}$ which follows from (1.19). Thus, V_j^n is contained in a ball with radius $C \frac{(\log(n))^{p_d}}{n^{1/d}}$ and so $v(V_j^n) \leqslant C \frac{(\log(n))^{d \cdot p_d}}{n}$ for some constant C that depends on dimension and the constant M from (1.3). Therefore,

$$\mathcal{I}_n^2 \leqslant \left(\frac{\varepsilon_n}{\tilde{\varepsilon}_n}\right)^d \cdot \left(\frac{C \varepsilon_n (\log(n))^{d \cdot p_d}}{\lambda_n}\right) \cdot \lambda_n G T V_{n,\varepsilon_n}(u_n^*) \to 0, \quad \text{as } n \to \infty,$$

given the assumptions on ε_n, λ_n. This concludes the proof.

5 Conclusions and future work

Our work establishes the consistency of the empirical risk minimization problem (1.1) by showing that with the right choice of scaling for λ_n, the minimizer u_n^* converges towards the Bayes classifier in the TL^1-sense. Although the function u_n^* is only defined on the cloud $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, one may extend the function u_n^* in a simple way to the whole ambient space so as to obtain a classifier that in the limit converges towards the desired Bayes classifier. We remark that we do not use the notion of VC dimension explicitly in our analysis given that we do not consider classes of functions defined on the ambient space as feasible elements in the empirical risk minimization problem. Instead, we work directly with the graph and its natural space of functions; in our analysis, we exploit the level of

https://doi.org/10.1017/S0956792517000201 Published online by Cambridge University Press

regularity of minimizers of R_{n,λ_n} (enforced by the graph total variation) and we use the TL^1 distance to compare the solutions of the discrete problem with the Bayes classifier.

We suspect a close connection between *regularity* of a solution of a discrete problem like the one considered in this paper and the VC dimension of a certain implicit family of functions. A natural setting in which to investigate notions of regularity (along with their connection to VC theory) would be in the linear setting in which one attempts to minimize an energy of the form

$$E_{n,\lambda_n}(u_n) := \frac{\lambda_n}{n^2\varepsilon^{d+2}} \sum_{i=1}^{n}\sum_{j=1}^{n} \eta\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{\varepsilon}\right)(u_n(\mathbf{x}_i) - u_n(\mathbf{x}_j))^2 + \frac{1}{n}\sum_{i=1}^{n}(u_n(\mathbf{x}_i) - \mathbf{y}_i)^2, \quad u_n \in L^2(v_n),$$

with the goal of approximating the Bayes regressor $u(x) := \mathbb{E}(\mathbf{y}|\mathbf{x} = x)$, where the variable $\mathbf{y}$ follows a law of the form

$$\mathbf{y} \sim \mathbb{P}(\mathbf{y} \in dy|\mathbf{x} = x).$$

The minimizer of the energy E_{n,λ_n} can be found by solving a linear system of equations involving the graph Laplacian associated to the graph $(\{\mathbf{x}_i\}, W)$, which can be interpreted as an elliptic PDE on the graph. Appropriate analogues of techniques from elliptic theory, such as Schauder estimates, and convex analysis, might then be powerful tools for analysis. We anticipate that these tools will permit a finer analysis of the problem, including detailed estimates on rates of convergence. The development of these tools, as well as their application, is the subject of current investigation.

Finally, we notice that the setting that we have considered in this paper is that in which the support of the measure v is an open domain $D \subseteq \mathbb{R}^d$. It is natural to consider the case in which the support of v is actually a sub-manifold $\mathcal{M}$ embedded in $\mathbb{R}^d$. We believe that the consistency results presented in this paper can be extended to the sub-manifold setting in a relative straightforward way. In the interest of clarity, we defer the details to a future work. In the linear problem described above, we anticipate that the desired rates of convergence will depend only on geometric quantities of $\mathcal{M}$ and not on the ambient space $\mathbb{R}^d$.

Acknowledgements

The authors are grateful to the anonymous reviewers and to the editors for their many suggestions to improve this paper. The authors are also thankful to Moritz Gerlach, Matthias Hein, Dejan Slepčev and Kavita Ramanan for enlightening discussions.

References

[1] Agapiou, S., Larsson, S. & Stuart, A. M. (2013) Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stoch. Process. Appl.* **123**, 3828–3860.

[2] Ambrosio, L., Fusco, N. & Pallara, D. (2000) *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Mathematical Monographs, The Clarendon Press Oxford University Press, New York.

[3] Ambrosio, L., Gigli, N. & Savaré, G. (2008) *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*, Lectures in Mathematics, Birkhäuser, Basel.

[4] Billingsley, P. (2012) *Probability and Measure*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ.
[5] Chambolle, A., Caselles, V., Cremers, D., Novaga, M. & Pock, T. (2010) An introduction to total variation for image analysis. In: Massimo Fornasier (editor), *Theoretical Foundations and Numerical Methods for Sparse Recovery*, Radon Ser. Comput. Appl. Math., vol. 9, Walter de Gruyter, Berlin, pp. 263–340.
[6] Dal Maso, G. (1993) *An Introduction to Γ-Convergence*, Springer, Birkhäuser Boston.
[7] Di Nezza, E., Palatucci, G. & Valdinoci, E. (2012) Hitchhiker's guide to the fractional Sobolev spaces. *Bull. Sci. Math.* **136**, 521–573.
[8] Esser, E. (2009) *Applications of Lagrangian Based Alternating Direction Methods and Connections to Split Bregman*, CAM Report 09-31, UCLA.
[9] Evans, L. C. (1990) *Weak Convergence Methods for Nonlinear Partial Differential Equations*, vol. 74, American Mathematical Soc, Providence, RI.
[10] Fonseca, I. & Leoni, G. (2007) *Modern Methods in the Calculus of Variations: L^p Spaces*, Springer Monographs in Mathematics, Springer, New York.
[11] García Trillos, N., Gerlach, M., Hein, M. & Slepčev, D. (2017) Spectral convergence of emprical graph laplacians. In preparation.
[12] García Trillos, N. & Slepčev, D. (2016) Continuum limit of total variation on point clouds. *Arch. Ration. Mech. Anal.* **220**(1), 193–241.
[13] Garcia Trillos, N. & Slepcev, D. (2015) On the rate of convergence of empirical measures in ∞-transportation distance. *Canad. J. Math.* **67**, 1358–1383.
[14] García Trillos, N., Slepčev, D., von Brecht, J., Laurent, T. & Bresson, X. (2016) Consistency of Cheeger and ratio graph cuts. to appear in *J. Mach. Learning Res.* **17**, Paper No. 181, 46 pages.
[15] Ghosal, S., Ghosh, J. K. & van der Vaart, A. W. (2000) Convergence rates of posterior distributions. *Ann. Statist.* **28**, 500–531.
[16] Grisvard, P. (1985) *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics, vol. 24, Pitman (Advanced Publishing Program), Boston, MA.
[17] Leoni, G. (2009) *A First Course in Sobolev Spaces*, Graduate Studies in Mathematics, vol. 24, American Mathematical Society, Providence, RI.
[18] Nikolova, M. (2004) A variational approach to remove outliers and impulse noise. *J. Math. Imaging Vision* **20**, 99–120.
[19] Pedregal, P. (1997) *Parametrized Measures and Variational Principles*, Progress in Nonlinear Differential Equations and their Applications, vol. 30, Birkhäuser Verlag, Basel.
[20] Rockafellar, R. T. (1970) *Convex Analysis*, Princeton Mathematical Series, vol. 28, Princeton University Press, Princeton, NJ.
[21] Vapnik, V. N. (1998) *Statistical Learning Theory*, vol. 1, John Wiley & Sons, Inc., New York.
[22] Villani, C. (2003) *Topics in Optimal Transportation*, Graduate Studies in Mathematics, vol. 58, American Mathematical Society, Providence, RI.
[23] Vogel, C. R. & Oman, M. E. (1996) Iterative methods for total variation denoising. *SIAM J. Sci. Comput.* **17**, 227–238.
[24] von Luxburg, U. & Schölkopf, B. (2011) Statistical learning theory: Models, concepts, and results. In: Dov M. Gabbay, Stephan Hartmann and John Woods (editors), *Handbook of the History of Logic, Vol. 10: Inductive Logic*Elsevier, North Holland, pp. 651–706.