LOCAL VERSIONS OF SUM-OF-NORMS CLUSTERING

ALEXANDER DUNLAP AND JEAN-CHRISTOPHE MOURRAT

Abstract. Sum-of-norms clustering is a convex optimization problem whose solution can be used for the clustering of multivariate data. We propose and study a localized version of this method, and show in particular that it can separate arbitrarily close balls in the stochastic ball model. More precisely, we prove a quantitative bound on the error incurred in the clustering of disjoint connected sets. Our bound is expressed in terms of the number of datapoints and the localization length of the functional.

1. Introduction

1.1. **Context and informal description of main result.** Let $x_1, \dots, x_N \in \mathbf{R}^d$ ($d \in \mathbf{N}$) be a collection of points, which we think of as a dataset. We consider the clustering problem, which is to find a partition of $\{x_1, \dots, x_N\}$ that collects close-together points into the same element of the partition. The problem of *K-means clustering* is to identify a global minimizer of the functional

$$(y_1, \dots, y_N) \mapsto \frac{1}{N} \sum_{n=1}^{N} |y_n - x_n|^2, \tag{1.1}$$

over all $(y_1, \dots, y_N) \in (\mathbf{R}^d)^N$ such that the cardinality of the set $\{y_1, \dots, y_N\}$ is at most K. This minimization problem is known to be NP-hard in general, even when restricted to $K = 2$ [2] or $d = 2$ [21]. Practitioners typically resort to iterative search algorithms such as Lloyd's algorithm and its refinements [20, 28], which at least identify local minimizers of (1.1). However, these methods are known to perform poorly in some cases, as will be discussed further below.

In this paper, we focus our attention on the "sum-of-norms clustering" method (also known as "convex clustering shrinkage" or "Clusterpath") introduced in [25, 16, 19]. This method can be thought of as a convex relaxation of the K-means problem. It considers the minimizer of the convex functional

$$(y_1, \dots, y_N) \mapsto \frac{1}{N} \sum_{n=1}^{N} |y_n - x_n|^2 + \frac{\lambda}{N^2} \sum_{m,n=1}^{N} w(|x_m - x_n|)|y_m - y_n| \tag{1.2}$$

over $(y_1, \dots, y_N) \in (\mathbf{R}^d)^N$, for some nonincreasing "weight function" w. (Typical choices include constant and exponential weights.) Here $|\cdot|$ denotes the Euclidean norm on $\mathbf{R}^d$. The point y_n is thought of as a "representative point" of the cluster to which x_n belongs, and thus x_n and x_m are declared to be members of the same cluster if $y_n = y_m$. The first term of (1.2) is designed to keep the representative point of a cluster close to the points in that cluster (and so encouraging having many clusters), while the second term (called the "fusion term") is designed to encourage points to merge into fewer clusters, at least if they are close together according to the weight function. The parameter λ controls the relative strength of these two effects.

The present work investigates an asymptotic regime of sum-of-norms clustering as the number of datapoints becomes very large and the weight w is simultaneously scaled in a

careful way. In order to do so, it is useful to specify a more explicit model for the dataset. We assume that the datapoints $x_1, \dots, x_N$ are independent and identically distributed. Their common law μ, a probability measure on $\mathbf{R}^d$, is supported on the union of disjoint closed sets $\overline{U}_1, \dots, \overline{U}_L$. These sets are not known to the practitioner. We would like x_i and x_j to be in the same cluster if and only if they lie in the same set $\overline{U}_\ell$ for some $\ell \in \{1, \dots, L\}$, and so we seek a clustering algorithm that can guarantee this in the limit as $N \to \infty$.

The weight function we choose is $w(r) := \gamma^{d+1} e^{-\gamma r}$, where $\gamma > 0$ is a parameter that can be tuned with the number of datapoints N. Roughly speaking, our main result states that, under modest assumptions, if we choose λ above a critical threshold not depending on N, and also choose $\gamma \simeq N^{3/(4d)}$, then in some mean-square sense, each point $x_n \in \overline{U}_\ell$ will be associated with a "representative point" y_n that is at distance of about $N^{-1/(8d)}$ from the centroid of the set $\overline{U}_\ell$ as $N \to \infty$. In particular, the clustering of the dataset is successful in the mean-square sense. The technical assumptions we need are that each set $\overline{U}_\ell$ is "effectively" star-shaped (see Definition 1.1 below), that the measure μ has a density with respect to the Lebesgue measure, and that this density is Lipschitz and bounded away from zero on its support. As an illustration, we can take μ to be the uniform measure on the union of the sets depicted in Fig. 1.2 below. The condition that the clusters be effectively star-shaped is a nontrivial geometric restriction, although it does not seem to be fundamental. See Remark 4.2 below for a weaker but more complicated sufficient condition, and further discussion.

Our result applies in particular to the case in which μ is the uniform measure on the union of disjoint balls. One of the strengths of our result is that these balls, or more generally the sets $\overline{U}_1, \dots, \overline{U}_L$, can be chosen arbitrarily close to one another, as long as they do not touch. (However, we expect that the required number of datapoints N will grow as the balls are brought closer to each other.) Another important feature is that we allow for sets $\overline{U}_1, \dots, \overline{U}_L$ that may be non-convex, as long as they are effectively star-shaped. Moreover, our result covers situations in which the convex hulls of the clusters intersect.

The unweighted version of the sum-of-norms clustering method, i.e. the case $w \equiv 1$, does not share any of these features. Indeed, the unweighted method fails to recover the clusters of datapoints sampled independently from two disjoint balls if the balls are too close together, as we showed in [13]. Moreover, the unweighted algorithm must output clusters that are contained in disjoint balls (see [23, Theorem 3] or [13, Proposition 1.8]), and in particular, it cannot separate two clusters unless their convex hulls are disjoint.

Popular alternative clustering methods such as Lloyd's algorithm and its refinements [20, 28] are also known to have important limitations. In [4, Appendix E], the authors exhibit explicit examples of configurations of disjoint balls $\overline{U}_1, \dots, \overline{U}_L$ of equal radius such that if the measure μ is the uniform probability measure on the union of these balls, then the probability that Lloyd's algorithm successfully clusters the dataset is at most $(1 - \frac{2}{9})^{L/3}$. They also construct similar examples for which a refined method called "kmeans++" also fails to successfully cluster the dataset with a probability that can be made arbitrarily close to 1.

Other convex relaxations of the K-means problem have been explored, but we are not aware of theoretical guarantees that would cover the case in which two clusters can be taken arbitrarily close to one another. Possibly the simplest way to ask the question is to consider the "stochastic ball model" [22]: we assume that the datapoints are sampled independently according to the uniform measure on the union of two disjoint balls of unit radius. In this setting, the method explored in [4] is guaranteed to recover the clusters provided that the distance between the two ball centers is above $2\sqrt{2}(1 + d^{-1/2})$. (See also [12] for the related problem of K-medians clustering.) Another convex relaxation of K-means clustering is

explored in [11]: for the stochastic ball model, that method successfully clusters the dataset provided that the distance between the centers of the balls is above $1 + \sqrt{3}$.

Several earlier works have explored the theoretical properties of sum-of-norms clustering. The unweighted method ($w \equiv 1$) was shown to separate cube-shaped clusters provided that they are sufficiently far away in [29]; for the case of two cubes of side-length 2 and equal number of datapoints falling in each cube, the criterion requires that the minimal distance between two points in each cube be at least $6\sqrt{d}$. More general conditions are derived in [24] (see in particular part 2 of Theorem 1) that imply the successful recovery of the clusters for the stochastic ball model if the distance between the ball centers is larger than 4. These results were refined and extended to the case of arbitrary weights in [26]. The problem of separating mixtures of Gaussian random variables has been considered in [27, 24, 18], and algorithmic aspects were explored in [25, 16, 9, 10, 17]. Several works have stressed the apparent advantages of using non-constant weights in sum-of-norms clustering [16, 9, 10, 23].

1.2. **Precise statement and proof strategy.** Following our previous work [13], for the purposes of mathematical analysis we consider the somewhat more general problem of clustering of measures. For a Borel measure μ on $\mathbf{R}^d$ of compact support, we abbreviate $L^2(\mu) \coloneqq L^2(\mathbf{R}^d, \mu; \mathbf{R})$ and $(L^2(\mu))^d \simeq L^2(\mathbf{R}^d, \mu; \mathbf{R}^d)$ to denote the Lebesgue spaces of μ-square-integrable functions from $\mathbf{R}^d$ to $\mathbf{R}$ and $\mathbf{R}^d$ to $\mathbf{R}^d$ respectively. (We recall that these spaces identify functions that only disagree on a set of μ-measure zero.) We define the functional $J_{\mu,\lambda,\gamma} : (L^2(\mu))^d \to \mathbf{R}$ by

$$J_{\mu,\lambda,\gamma}(u) \coloneqq \int |u(x) - x|^2 \,\mathrm{d}\mu(x) + \lambda\gamma^{d+1} \iint \mathrm{e}^{-\gamma|x-y|} |u(x) - u(y)| \,\mathrm{d}\mu(x) \,\mathrm{d}\mu(y). \tag{1.3}$$

We note that (1.2) with $w(r) = \gamma^{d+1}\mathrm{e}^{-\gamma r}$ is obtained from (1.3) by setting $\mu = \frac{1}{N}\sum_{n=1}^N \delta_{x_n}$. The map $x \mapsto u(x)$ is then the analogue of the map $x_n \mapsto y_n$ from points to cluster representative points. We denote by $u_{\mu,\lambda,\gamma}$ the minimizer of $J_{\mu,\lambda,\gamma}$, which exists and is unique because $J_{\mu,\lambda,\gamma}$ is coercive, uniformly convex, and continuous on $(L^2(\mu))^d$. (See (2.2) below.) For every Borel set U such that $\mu(U) > 0$, we let

$$\mathrm{cent}_\mu(U) \coloneqq \frac{1}{\mu(U)} \int_U x \,\mathrm{d}\mu(x)$$

be the μ-centroid of U. We also write $a \vee b \coloneqq \max(a, b)$, and define

$$d' \coloneqq \begin{cases} \infty & \text{if } d = 1, \\ \frac{4}{3} & \text{if } d = 2, \\ d & \text{if } d \geqslant 3. \end{cases} \tag{1.4}$$

Our main result considers a measure μ with support comprising a finite union of connected components, each with sufficiently regular boundary and satisfying a quantitative version of a "star-shaped" property. We also assume that μ is bounded below on its support, and is sufficiently regular on its support. We draw N datapoints independently from μ and run our clustering algorithm on these datapoints. If γ is chosen appropriately large depending on N, and λ is fixed sufficiently large independent of N, then our clustering algorithm will recover the connected components of $\operatorname{supp}\mu$. Before stating our main result, we introduce the technical condition we need on the components of $\operatorname{supp}\mu$.

Definition 1.1. For U a subset of $\mathbf{R}^d$ and $\varepsilon > 0$, let U_ε be the ε-enlargement of U, namely

$$U_\varepsilon \coloneqq \{x \in \mathbf{R}^d \mid \operatorname{dist}(x, U) \leqslant \varepsilon\}.$$

FIGURE 1.1. A set that is star-shaped but not effectively star-shaped.

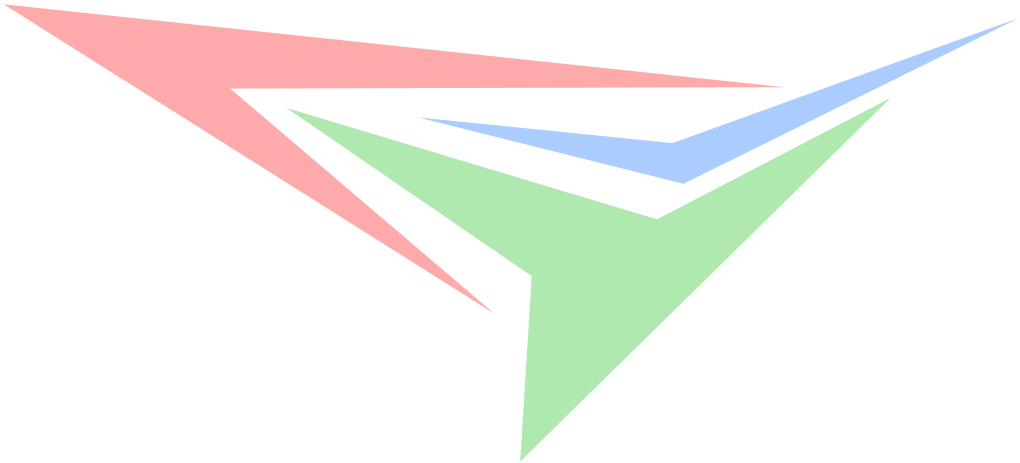

FIGURE 1.2. A set of three open sets U_1, U_2, U_3 satisfying the hypotheses of Theorem 1.2.

We say that a domain U is *effectively star-shaped* if there exists $x_* \in U$ and a constant $C_* < \infty$ such that for every $\varepsilon > 0$ sufficiently small, the image of U_ε under the mapping $x \mapsto x_* + (1 - C_*\varepsilon)(x - x_*)$ is contained in U.

For example, any convex open set is effectively star-shaped (in which case x_* can be chosen arbitrarily in the interior). Any effectively star-shaped set is star-shaped. An example of a set that is star-shaped but not effectively star-shaped is illustrated in Fig. 1.1. Now we can state our main theorem.

Theorem 1.2. *Let μ be a probability measure on $\mathbf{R}^d$ such that $\operatorname{supp}\mu = \bigcup_{\ell=1}^{L} \overline{U_\ell}$, where $U_1, \dots, U_L$ are bounded, effectively star-shaped open sets with Lipschitz boundaries, such that their closures $\overline{U_1}, \dots, \overline{U_L}$ are pairwise disjoint. Assume that μ admits a density with respect to the Lebesgue measure, and that this density is Lipschitz and bounded away from zero on $\operatorname{supp}\mu$. Then there exist $\lambda_{\mathrm{c}}, C < \infty$ such that for every $\lambda \geqslant \lambda_{\mathrm{c}}$, the following holds. Let $(X_n)_{n\in\mathbf{N}}$ be a sequence of independent random variables with law μ, $N \geqslant 1$ be an integer, $\mu_N := \frac{1}{N}\sum_{n=1}^{N} \delta_{X_n}$ be the empirical measure of the datapoints, and*

$$A_N^{(\ell)} := \{n \in \{1, \dots, N\} \mid X_n \in U_\ell\}, \qquad \ell \in \{1, \dots, L\}$$

be the set of indices of datapoints in U_ℓ. For every $\gamma \geqslant 1$, the mean-square error between the clustering algorithm and the centroids of the clusters is bounded as follows:

$$\mathbf{E}\left[\frac{1}{N}\sum_{\ell=1}^{L}\sum_{n\in A_N^{(\ell)}}|u_{\mu_N,\lambda,\gamma}(X_n)-\mathrm{cent}_\mu(U_\ell)|^2\right] \leqslant C\left(\gamma N^{-1/(d\vee 2)}(\log N)^{1/d'}+(1+\lambda)\gamma^{-1/3}\right). \tag{1.5}$$

For $d \geqslant 2$, optimizing the right-hand side of (1.5) suggests the optimal choice $\gamma \simeq N^{3/(4d)}$, in which case the mean-square error is at most of the order of $N^{-1/(4d)}$, up to logarithmic corrections. We do not know if the estimate in (1.5) is sharp. If technical issues that arise near the boundary of the domains could be avoided, then we believe that we could replace the term $\gamma^{-1/3}$ in (1.5) by $\gamma^{-1/2}$; this in turn would suggest choosing $\gamma \simeq N^{2/(3d)}$, up to a logarithmic correction.

A similar result to Theorem 1.2 can be obtained if the weight $r \mapsto \mathrm{e}^{-\gamma r}$ is replaced by a truncated version $r \mapsto \mathrm{e}^{-\gamma r}\mathbf{1}_{r\leqslant\omega}$ for an appropriate choice of ω; see Proposition 6.1 below. This result essentially says that we can choose $\omega \simeq \gamma^{-1}$, up to a logarithmic correction, without modifying the optimizer substantially. In the discrete setting, this reduces the number of pairs of points that need to be included in the sum that is the double integral in (1.3), and thus may lead to improvements in computational efficiency. (See [9] regarding efficient computational algorithms for sum-of-norms clustering, and in particular regarding the effect of the sparsity of the weights on the computational complexity.) For instance, under the assumptions of Theorem 1.2 and with the choice of $\omega \simeq \gamma^{-1} \simeq N^{-3/(4d)}$, a typical point only interacts with about $N^{1/4}$ points in its vicinity. Depending on the relative costs of computation versus the procurement of new datapoints, efficiency considerations may lead to a different choice of γ than what would be suggested by the optimal accuracy considerations discussed in the previous paragraph. We do not further pursue the question of computational efficiency in the present paper.

While we did not keep track of this explicitly, one can check from the proof that the critical value $\lambda_c < \infty$ identified in Theorem 1.2 does not change as the sets $\overline{U}_1, \dots, \overline{U}_L$ are individually translated or rotated, provided that they remain pairwise disjoint. In particular, this constant does not depend on the minimal distance between the different data clusters. As a careful examination of the arguments below shows, one can also choose the constant $C < \infty$ in Theorem 1.2 to be invariant under individual translations and rotations of the sets $\overline{U}_1, \dots, \overline{U}_L$ that do not make them intersect each other, provided that we also require γ to be sufficiently large. Roughly speaking, we would then require γ^{-1} to be larger than the minimal distance separating any pair of clusters, that is,

$$\gamma^{-1} \gtrsim \Delta := \min_{1\leqslant\ell\neq\ell'\leqslant L}\mathrm{dist}(U_\ell, U_{\ell'}).$$

The precise condition is displayed in (7.1) below. In particular, for $d \geqslant 2$, our approach would yield non-trivial information provided that the number of datapoints N is much larger than Δ^{-d}.

An important step in the proof of Theorem 1.2, which is also of independent interest, concerns the behavior of the functional $J_{u,\lambda,\gamma}$ as γ is taken to infinity. The factor γ^{d+1} in (1.3) was chosen so that $J_{u,\lambda,\gamma}$ would converge to a limiting functional as $\gamma \to \infty$, under appropriate conditions on μ. Let U be a bounded open subset of $\mathbf{R}^d$ and suppose that $\operatorname{supp}\mu = \overline{U}$. Suppose furthermore that μ is absolutely continuous with respect to the Lebesgue measure on $\overline{U}$, with density $\rho \in \mathcal{C}(\overline{U})$ bounded away from zero on $\overline{U}$. We denote by $\mathrm{BV}(U)$ the space of functions of bounded variation on U. (Some elementary properties

of the space $\mathrm{BV}(U)$ are recalled in Section 2 below; see also [3].) If $u \in (L^2(U) \cap \mathrm{BV}(U))^d$, then we can define

$$J_{\mu,\lambda,\infty}(u) := \int |u(x) - x|^2 \, \mathrm{d}\mu(x) + c\lambda \int \rho(x)^2 \, \mathrm{d}|Du|(x), \tag{1.6}$$

where

$$c := \int_{\mathbf{R}^d} \mathrm{e}^{-|y|} |y \cdot \mathrm{e}_1| \, \mathrm{d}y. \tag{1.7}$$

We will see in Proposition 2.1 below that $J_{\mu,\lambda,\infty}$ admits a unique minimizer $u_{\mu,\lambda,\infty} \in (L^2(U) \cap \mathrm{BV}(U))^d$. In Theorem 4.1, we will then show in a quantitative sense that, if U is sufficiently regular and the density ρ is Lipschitz, then $u_{\mu,\lambda,\gamma}$ converges to $u_{\mu,\lambda,\infty}$ as γ tends to infinity. The essential strategy here is to compare the functionals $J_{\mu,\lambda,\infty}$ and $J_{\mu,\lambda,\gamma}$ and use their uniform convexity. An important technical complication is that $J_{\mu,\lambda,\infty}(u)$ is only defined for functions u of bounded variation on $\overline{U}$ while the minimizer of $J_{\mu,\lambda,\gamma}$ may not (a priori) be of bounded variation. Therefore, to compare the functionals, we must first smooth their argument u in a way that respects derivatives. Convolution by a smooth function works, but we first must dilate u slightly since it is only defined on $\overline{U}$, not all of $\mathbf{R}^d$. Moreover, this modification of the optimizer for $J_{\mu,\lambda,\gamma}$ needs to be performed in such a way that the functional does not increase too much. It is this constraint that leads us to the requirement that the domains be effectively star-shaped (or that the more general condition in Remark 4.2 holds).

The utility of the gradient functional (1.6) in the proof of Theorem 1.2 is apparent in Proposition 5.1 below. This proposition states that when λ is large enough, the minimizer of the gradient functional recovers the centroids of the connected components of the support of the measure μ. The critical λ is identified in terms of the L^∞ norm of the solution to a PDE arising from the first-order conditions for the minimizer. We expect that further information about the behavior of the limiting functional could be obtained by further studying this PDE.

As mentioned, the gradient clustering functional (1.6) only makes sense for smooth measures. In order to show the convergence of the minimizers of the weighted clustering functionals (1.3) on empirical distributions, we need to relate the minimizers of the finite-γ problem for empirical distributions to the minimizers of the finite-γ problem for smooth distributions. We do this by proving a stability result with respect to the ∞-Wasserstein metric $\mathcal{W}_\infty$, which is Proposition 3.1 below. This works in combination with a quantitative Glivenko–Cantelli-type result for the ∞-Wasserstein metric proved in [15], and recalled in Proposition 7.1 below. However, since the latter result only holds for connected domains, we also need to truncate the exponential weight in (1.3), which is done in Section 6.

1.3. **Outline of the paper.** In Section 2 we establish some basic properties of $J_{\mu,\lambda,\gamma}$ and $J_{\mu,\lambda,\infty}$. In Section 3 we prove a stability result for $u_{\widetilde{\mu},\lambda,\gamma}$ as $\widetilde{\mu} \to \mu$ in the ∞-Wasserstein distance. In Section 4 we prove the convergence result for $u_{\mu,\lambda,\gamma}$ as $\gamma \to \infty$. In Section 5 we show that the limiting functional $u_{\mu,\lambda,\infty}$ recovers the centroids of the connected components of $\operatorname{supp}\mu$ as long as λ is large enough. In Section 6 we prove a stability result when the exponential weight is truncated. In Section 7 we put everything together to prove Theorem 1.2.

Acknowledgments. AD was partially supported by the NSF Mathematical Sciences Postdoctoral Fellowship program under grant no. DMS-2002118. JCM was partially supported by NSF grant DMS-1954357.

2. Basic properties of the functionals

As mentioned above, for a bounded open set $U \subseteq \mathbf{R}^d$, we denote by $\mathrm{BV}(U)$ the space of functions of bounded variation on U. This is the set of all functions $u \in L^1(U)$ whose derivatives are Radon measures. For $u \in \mathrm{BV}(U)$, we denote by Du the gradient of u, which is thus a vector-valued Radon measure, and we denote by $|Du|$ its total variation. In particular, for every open set $V \subseteq U$, we have by [3, Proposition 1.47] that

$$|Du|(V) = \sup_{\phi} \int_V \phi \cdot \mathrm{d}Du = \sup_{\phi} \sum_{i=1}^{d} \int_V \phi_i \,\mathrm{d}D_i u, \tag{2.1}$$

where the supremum is over all $\phi \in (\mathcal{C}_{\mathrm{c}}(V))^d$ (the space of $\mathbf{R}^d$-valued continuous functions supported on compact subsets of V) such that $\|\phi\|_{L^\infty(V)} \leqslant 1$, with the understanding that

$$\|\phi\|_{L^\infty(V)} = \| \, |\phi| \, \|_{L^\infty(V)} = \operatorname*{ess\,sup}_{x \in V} \left(\sum_{i=1}^{d} \phi_i^2(x) \right)^{\frac{1}{2}}.$$

When $u \in (\mathrm{BV}(U))^d$, the gradient Du is a Radon measure taking values in the space of $d \times d$ matrices. Identifying such a matrix with an element of $\mathbf{R}^{d^2}$, we can still define the total variation measure $|Du|$ as above. (Thus, if Du is in fact an $\mathbf{R}^{d\times d}$-valued function, then $|Du|(x)$ is the Frobenius norm of the matrix $Du(x)$.) We refer to [3] for a thorough exposition of the properties of BV functions.

In the remainder of this section, we collect some basic properties of the functionals $J_{\mu,\lambda,\gamma}$. It is straightforward to see that, for any $\gamma \in (0,\infty)$, the functional $J_{\mu,\lambda,\gamma}$ is uniformly convex on $(L^2(\mu))^d$. Indeed, for every $u, v \in (L^2(\mu))^d$, we have

$$\frac{1}{2}\left(J_{\mu,\lambda,\gamma}(u+v) + J_{\mu,\lambda,\gamma}(u-v)\right) - J_{\mu,\lambda,\gamma}(u) \geqslant \int v^2 \,\mathrm{d}\mu. \tag{2.2}$$

Since the functional is also coercive, the existence and uniqueness of the minimizer $u_{\mu,\lambda,\gamma}$ follow. The next proposition covers the case when $\gamma = \infty$.

Proposition 2.1. *Let U be a bounded open subset of $\mathbf{R}^d$ and suppose that $\operatorname{supp}\mu = \overline{U}$. Suppose furthermore that μ is absolutely continuous with respect to the Lebesgue measure on $\overline{U}$ with a density $\rho \in \mathcal{C}(\overline{U})$ that is bounded away from zero on $\overline{U}$. Then for any $\lambda > 0$, the functional $J_{\mu,\lambda,\infty}$ admits a unique minimizer $u_{\mu,\lambda,\infty} \in (L^2(U) \cap \mathrm{BV}(U))^d$.*

Proof. We start by observing that the convexity property (2.2) is still valid for $\gamma = \infty$, for every $u, v \in (L^2(U) \cap \mathrm{BV}(U))^d$. Let $(u_k)_k$ be a sequence of functions in $(L^2(U) \cap \mathrm{BV}(U))^d$ such that

$$\lim_{k\to\infty} J_{\mu,\lambda,\infty}(u_k) = \inf_{u \in (L^2(U)\cap \mathrm{BV}(U))^d} J_{\mu,\lambda,\infty}(u). \tag{2.3}$$

Since ρ is bounded away from zero, the functional $J_{\mu,\lambda,\infty}$ is coercive on $(L^2(U) \cap \mathrm{BV}(U))^d$. By the Banach–Alaoglu theorem and [3, Theorem 3.23] (the latter saying that sets S of functions in $\mathrm{BV}(U)$ for which $\sup_{u\in S} \int_U |u| \,\mathrm{d}x + |Du|(U) < \infty$ are weakly-$*$ precompact), by passing to a subsequence we can assume that there is a $u \in (L^2(U) \cap \mathrm{BV}(U))^d$ such that $u_k \to u$ weakly in $(L^2(U))^d$ and weakly-$*$ in $(\mathrm{BV}(U))^d$. From the weak convergence in $(L^2(U))^d$ we see that

$$\int |u(x) - x|^2 \,\mathrm{d}\mu(x) \leqslant \liminf_{k\to\infty} \int |u_k(x) - x|^2 \,\mathrm{d}\mu(x).$$

From the weak-$*$ convergence in $(\mathrm{BV}(U))^d$ we see that

$$\begin{aligned}\int_U \rho(x)^2 \,\mathrm{d}|Du|(x) &= \sup_{\phi} \int_U \rho(x)^2 \phi(x) \cdot \mathrm{d}Du(x) \\ &\leqslant \liminf_{k\to\infty} \sup_{\phi} \int_U \rho(x)^2 \phi(x) \cdot \mathrm{d}Du_k(x) \\ &= \liminf_{k\to\infty} \int_U \rho(x)^2 \,\mathrm{d}|Du_k|(x),\end{aligned}$$

where the supremum is over all $\phi \in (\mathcal{C}_c(U))^{d^2}$ such that $\|\phi\|_{L^\infty(U)} \leqslant 1$. The last two displays and (2.3) imply that $J_{\mu,\lambda,\infty}(u) = \inf J_{\mu,\lambda,\infty}$, so we can take $u_{\mu,\lambda,\infty} = u$. The uniqueness of $u_{\mu,\lambda,\infty}$ follows from the uniform convexity (2.2). □

A direct consequence of the convexity property (2.2) is that, for every $\gamma \in (0,\infty)$ and $u \in (L^2(\mu))^d$, we have

$$\begin{aligned}\int |u - u_{\mu,\lambda,\gamma}|^2 \,\mathrm{d}\mu &\leqslant 2\left(J_{\mu,\lambda,\gamma}(u) + J_{\mu,\lambda,\gamma}(u_{\mu,\lambda,\gamma})\right) - 4J_{\mu,\lambda,\gamma}\left(\frac{u_{\mu,\lambda,\gamma}+u}{2}\right) \\ &\leqslant 2\left(J_{\mu,\lambda,\gamma}(u) - \inf J_{\mu,\lambda,\gamma}\right). \end{aligned} \tag{2.4}$$

Under the assumptions of Proposition 2.1, the inequalities in (2.4) remain valid with $\gamma = \infty$, provided that we also impose that $u \in (L^2(U) \cap \mathrm{BV}(U))^d$. Another important fact will be that, for every $\gamma \in (0,\infty]$,

$$0 \leqslant \inf J_{\mu,\lambda,\gamma} \leqslant J_{\mu,\lambda,\gamma}(\mathrm{cent}_\mu(\mathbf{R}^d)) = \int |x - \mathrm{cent}_\mu(\mathbf{R}^d)|^2 \,\mathrm{d}\mu(x), \tag{2.5}$$

where we note that the right-hand side is the variance of a random variable distributed according to μ, and in particular is independent of λ and γ.

3. Stability with respect to ∞-Wasserstein perturbations of the measure

Throughout the paper, for any two measures μ and ν on $\mathbf{R}^d$, we let $\mathcal{W}_\infty(\mu,\nu)$ be the ∞-Wasserstein distance between μ and ν, namely

$$\mathcal{W}_\infty(\mu,\nu) = \inf_{\pi} \operatorname*{ess\,sup}_{(x,y)\sim\pi} |x-y|,$$

where the infimum is taken over all couplings π of μ and ν. It is classical to verify that this infimum is achieved (see e.g. [8, Proposition 2.1]). We call any π achieving this infimum an *∞-optimal transport plan* from μ to ν. In this section we prove that, for finite γ, the minimizer $u_{\mu,\lambda,\gamma}$ is stable under ∞-Wasserstein perturbations of μ.

Proposition 3.1. *There is a universal constant C such that the following holds. Let $\gamma, \lambda, M \in (0,\infty)$ and let $\mu, \widetilde{\mu}$ be two probability measures on $\mathbf{R}^d$ with supports contained in a common Euclidean ball B of diameter M. There exists an ∞-optimal transport plan π from μ to $\widetilde{\mu}$ such that*

$$\int |u_{\mu,\lambda,\gamma}(x) - u_{\widetilde{\mu},\lambda,\gamma}(\widetilde{x})|^2 \,\mathrm{d}\pi(x,\widetilde{x}) \leqslant C(M+1)^2(\gamma+1)\mathcal{W}_\infty(\mu,\widetilde{\mu}). \tag{3.1}$$

Proof. Throughout the proof, λ and γ will remain fixed, so we write $J_\mu = J_{\mu,\lambda,\gamma}$ and $u_\mu = u_{\mu,\lambda,\gamma}$. (Nonetheless, we emphasize that the constant C in the statement of the theorem does *not* depend on λ or γ.) Let π be an ∞-optimal transport plan from μ to $\widetilde{\mu}$. We write the disintegration

$$\mathrm{d}\pi(x,\widetilde{x}) = \mathrm{d}\nu(\widetilde{x} \mid x)\,\mathrm{d}\mu(x)$$

and define

$$\overline{u}(x) \coloneqq \int u_{\widetilde{\mu}}(\widetilde{x})\,\mathrm{d}\nu(\widetilde{x} \mid x).$$

We have

$$\begin{aligned}\inf J_{\widetilde{\mu}} &= \int |u_{\widetilde{\mu}}(\widetilde{x}) - \widetilde{x}|^2\,\mathrm{d}\widetilde{\mu}(\widetilde{x}) + \lambda\gamma^{d+1} \iint \mathrm{e}^{-\gamma|\widetilde{x}-\widetilde{y}|}|u_{\widetilde{\mu}}(\widetilde{x}) - u_{\widetilde{\mu}}(\widetilde{y})|\,\mathrm{d}\widetilde{\mu}(\widetilde{x})\,\mathrm{d}\widetilde{\mu}(\widetilde{y}) \\ &= \iint |u_{\widetilde{\mu}}(\widetilde{x}) - \widetilde{x}|^2\,\mathrm{d}\nu(\widetilde{x} \mid x)\,\mathrm{d}\mu(x) \\ &\qquad + \lambda\gamma^{d+1} \iiiint \mathrm{e}^{-\gamma|\widetilde{x}-\widetilde{y}|}|u_{\widetilde{\mu}}(\widetilde{x}) - u_{\widetilde{\mu}}(\widetilde{y})|\,\mathrm{d}\nu(\widetilde{x} \mid x)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(\widetilde{y} \mid y)\,\mathrm{d}\mu(y). \qquad (3.2)\end{aligned}$$

For the first term on the right side of (3.2), we write

$$\begin{aligned}|u_{\widetilde{\mu}}(\widetilde{x}) - \widetilde{x}|^2 &= |u_{\widetilde{\mu}}(\widetilde{x}) - x|^2 - |x - \widetilde{x}|^2 + 2(u_{\widetilde{\mu}}(\widetilde{x}) - \widetilde{x}) \cdot (x - \widetilde{x}) \\ &\geqslant |u_{\widetilde{\mu}}(\widetilde{x}) - x|^2 - 3M|x - \widetilde{x}|. \qquad (3.3)\end{aligned}$$

For the second term on the right side of (3.2), we note that, for μ-a.e. x, y, on the support of $\nu(\widetilde{x} \mid x) \otimes \nu(\widetilde{y} \mid y)$ we have, writing $W \coloneqq \mathcal{W}_\infty(\mu, \widetilde{\mu})$,

$$|\widetilde{y} - \widetilde{x}| \leqslant 2W + |y - x|,$$

so

$$\mathrm{e}^{-\gamma|\widetilde{x}-\widetilde{y}|} \geqslant \mathrm{e}^{-2\gamma W}\mathrm{e}^{-\gamma|y-x|}.$$

Thus we can write

$$\begin{aligned}&\iiiint \mathrm{e}^{-\gamma|\widetilde{x}-\widetilde{y}|}|u_{\widetilde{\mu}}(\widetilde{x}) - u_{\widetilde{\mu}}(\widetilde{y})|\,\mathrm{d}\nu(\widetilde{x} \mid x)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(\widetilde{y} \mid y)\,\mathrm{d}\mu(y) \\ &\quad\geqslant \mathrm{e}^{-2\gamma W} \iint \mathrm{e}^{-\gamma|x-y|} \left(\iint |u_{\widetilde{\mu}}(\widetilde{x}) - u_{\widetilde{\mu}}(\widetilde{y})|\,\mathrm{d}\nu(\widetilde{x} \mid x)\,\mathrm{d}\nu(\widetilde{y} \mid y) \right) \mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \\ &\quad\geqslant \mathrm{e}^{-2\gamma W} \iint \mathrm{e}^{-\gamma|x-y|}|\overline{u}(x) - \overline{u}(y)|\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y), \qquad (3.4)\end{aligned}$$

where we used Jensen's inequality in the last step. Substituting (3.3) and (3.4) into (3.2), we obtain

$$\begin{aligned}\inf J_{\widetilde{\mu}} &\geqslant \iint |u_{\widetilde{\mu}}(\widetilde{x}) - x|^2\,\mathrm{d}\nu(\widetilde{x} \mid x)\,\mathrm{d}\mu(x) - 3M \iint |x - \widetilde{x}|\,\mathrm{d}\pi(x, \widetilde{x}) \\ &\qquad + \lambda\gamma^{d+1}\mathrm{e}^{-2\gamma W} \iint \mathrm{e}^{-\gamma|x-y|}|\overline{u}(x) - \overline{u}(y)|\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \\ &\geqslant \int |\overline{u}(x) - x|^2\,\mathrm{d}\mu(x) + \lambda\gamma^{d+1}\mathrm{e}^{-2\gamma W} \iint \mathrm{e}^{-\gamma|x-y|}|\overline{u}(x) - \overline{u}(y)|\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) - 3MW \\ &\geqslant \mathrm{e}^{-2\gamma W} J_\mu(\overline{u}) - 3MW,\end{aligned}$$

where in the second step we again used Jensen's inequality. Therefore, we have

$$\inf J_\mu \leqslant J_\mu(\overline{u}) \leqslant \mathrm{e}^{2\gamma W}\left(\inf J_{\widetilde{\mu}} + 3MW\right) \leqslant \inf J_{\widetilde{\mu}} + 3M\mathrm{e}^{2\gamma W}W + \left(\mathrm{e}^{2\gamma W} - 1\right) M^2, \quad (3.5)$$

with the last inequality by (2.5). By symmetry, this implies that

$$\left|\inf J_{\widetilde{\mu}} - \inf J_\mu\right| \leqslant 3M\mathrm{e}^{2\gamma W}W + (\mathrm{e}^{2\gamma W} - 1)M^2. \quad (3.6)$$

Now we have, using the second and third inequalities of (3.5), as well as (2.4) and (3.6), that

$$\begin{aligned}\int |\overline{u} - u_\mu|^2\,\mathrm{d}\mu &\leqslant 2\left(J_\mu(\overline{u}) - \inf J_\mu\right) \leqslant 2\left(\inf J_{\widetilde{\mu}} - \inf J_\mu\right) + 6M\mathrm{e}^{2\gamma W}W + 2\left(\mathrm{e}^{2\gamma W} - 1\right) M^2 \\ &\leqslant 12M\mathrm{e}^{2\gamma W}W + 4(\mathrm{e}^{2\gamma W} - 1)M^2 \leqslant (M+1)^2 Q((\gamma+1)\mathcal{W}_\infty(\mu, \widetilde{\mu})), \quad (3.7)\end{aligned}$$

where we have defined $Q(t) := 12\mathrm{e}^{2t}t + 4(\mathrm{e}^{2t} - 1)$.

The remainder of the proof is very similar to the second half of the proof of [13, Proposition 5.3]. For each $\varepsilon > 0$, let μ_ε be a measure on the ball B, absolutely continuous with respect to the Lebesgue measure, and such that

$$\mathcal{W}_\infty(\mu, \mu_\varepsilon) \leqslant \varepsilon. \tag{3.8}$$

Since μ_ε is absolutely continuous with respect to the Lebesgue measure, by [8, Theorems 5.5 and 3.2] there are maps T_ε and $\widetilde{T}_\varepsilon$ from $\operatorname{supp}\mu_\varepsilon$ to $\operatorname{supp}\mu$ and $\operatorname{supp}\widetilde{\mu}$, respectively, such that $(\mathrm{id}\times T_\varepsilon)_*(\mu_\varepsilon)$ is an ∞-optimal transport plan between μ_ε and μ and similarly $(\mathrm{id}\times\widetilde{T}_\varepsilon)_*(\mu_\varepsilon)$ is an ∞-optimal transport plan between μ_ε and $\widetilde{\mu}$. We have

$$\begin{aligned}&\int |u_\mu(T_\varepsilon(x)) - u_{\widetilde{\mu}}(\widetilde{T}_\varepsilon(x))|^2 \,\mathrm{d}\mu_\varepsilon(x)\\ &\quad\leqslant 2\int |u_\mu(T_\varepsilon(x)) - u_{\mu_\varepsilon}(x)|^2 \,\mathrm{d}\mu_\varepsilon(x) + 2\int |u_{\mu_\varepsilon}(x) - u_{\widetilde{\mu}}(\widetilde{T}_\varepsilon(x))|^2 \,\mathrm{d}\mu_\varepsilon(x).\end{aligned} \tag{3.9}$$

For the first term on the right side, we use (3.7) above with $\mu \leftarrow \mu_\varepsilon$ and $\widetilde{\mu} \leftarrow \mu$ (so that $\overline{u} \leftarrow u_\mu \circ T_\varepsilon$):

$$\int |u_\mu(T_\varepsilon(x)) - u_{\mu_\varepsilon}(x)|^2 \,\mathrm{d}\mu_\varepsilon(x) \leqslant (M+1)^2 Q((\gamma+1)\varepsilon).$$

For the second term on the right side, we use (3.7) above with $\mu \leftarrow \mu_\varepsilon$ and $\widetilde{\mu} \leftarrow \widetilde{\mu}$ (so that $\overline{u} \leftarrow u_{\widetilde{\mu}} \circ \widetilde{T}_\varepsilon$):

$$\int |u_{\mu_\varepsilon}(x) - u_{\widetilde{\mu}}(\widetilde{T}_\varepsilon(x))|^2 \,\mathrm{d}\mu_\varepsilon(x) \leqslant (M+1)^2 Q((\gamma+1)\mathcal{W}_\infty(\mu_\varepsilon, \widetilde{\mu})).$$

Using the last two displays in (3.9), we get

$$\begin{aligned}&\int |u_\mu(T_\varepsilon(x)) - u_{\widetilde{\mu}}(\widetilde{T}_\varepsilon(x))|^2 \,\mathrm{d}\mu_\varepsilon(x)\\ &\quad\leqslant 2(M+1)^2 Q((\gamma+1)\varepsilon) + 2(M+1)^2 Q((\gamma+1)\mathcal{W}_\infty(\mu_\varepsilon, \widetilde{\mu})).\end{aligned} \tag{3.10}$$

We can find a sequence $\varepsilon_k \downarrow 0$ and a coupling π of μ and $\widetilde{\mu}$ such that $(T_{\varepsilon_k}, \widetilde{T}_{\varepsilon_k})_*\mu_{\varepsilon_k} \to \pi$ as $k \to \infty$. Taking $\varepsilon = \varepsilon_k$ in (3.10), and then taking the limit as $k \to \infty$, we get

$$\int |u_{\mu,\lambda,\gamma}(x) - u_{\widetilde{\mu},\lambda,\gamma}(\widetilde{x})|^2 \,\mathrm{d}\pi(x, \widetilde{x}) \leqslant 2(M+1)^2 Q((\gamma+1)\mathcal{W}_\infty(\mu, \widetilde{\mu})). \tag{3.11}$$

Hence, since, Q is smooth, $Q(0) = 0$, and the left side of (3.11) is also evidently bounded above by M^2, we obtain the desired inequality (3.1).

It remains to show that π is an ∞-optimal transport plan. This follows by using (3.8) to note that

$$\operatorname*{ess\,sup}_{x\sim\mu_\varepsilon} |T_\varepsilon(x) - \widetilde{T}_\varepsilon(x)| \leqslant \operatorname*{ess\,sup}_{x\sim\mu_\varepsilon} |T_\varepsilon(x) - x| + \operatorname*{ess\,sup}_{x\sim\mu_\varepsilon} |x - \widetilde{T}_\varepsilon(x)| \leqslant \varepsilon + \mathcal{W}_\infty(\mu_\varepsilon, \widetilde{\mu}),$$

and then taking limits along the subsequence $\varepsilon_k \downarrow 0$. □

4. Convergence as $\gamma \to \infty$

In this section we show that, under suitable assumptions on U and μ, the optimizer $u_{\mu,\lambda,\gamma}$ converges to $u_{\mu,\lambda,\infty}$ as $\gamma \to \infty$. In essence, we will obtain this by showing a quantitative version of the fact that the functional $J_{\mu,\lambda,\gamma}$ Γ-converges to $J_{\mu,\lambda,\infty}$ as γ tends to infinity.

Theorem 4.1. *Assume that $U = \operatorname{supp}\mu$ is effectively star-shaped and has a Lipschitz boundary, and that the measure μ has a density with respect to the Lebesgue measure that is Lipschitz on U and is bounded away from zero. Then there exists a constant $C < \infty$ such that, for every $\lambda \in (0,\infty)$, we have*

$$|\inf J_{\mu,\lambda,\infty} - \inf J_{\mu,\lambda,\gamma}| + \int |u_{\mu,\lambda,\infty} - u_{\mu,\lambda,\gamma}|^2 \,\mathrm{d}\mu \leqslant C\gamma^{-1/3}. \tag{4.1}$$

Proof. Without loss of generality, assume that the point x_* in Definition 1.1 is the origin, and that the constant C_* appearing there is 1. We denote by ρ the density of μ with respect to the Lebesgue measure. By [14, Theorem 5.4.1], we can and do extend ρ to a Lipschitz function on $\mathbf{R}^d$, which we can also prescribe to vanish outside of a bounded set. Throughout the proof, we will leave μ, λ fixed, and write $u_\gamma = u_{\mu,\lambda,\gamma}$ and $J_\gamma = J_{\mu,\lambda,\gamma}$. The constant C may depend on μ but not on γ or λ, and may change over the course of the argument. We let U_ε be the ε-enlargement of U as in Definition 1.1.

For every $\varepsilon \in (0,1)$, $\gamma \in (0,\infty]$, and $x \in U_\varepsilon$, we define

$$\widetilde{u}_{\gamma,\varepsilon}(x) \coloneqq u_\gamma((1-\varepsilon)x),$$

and for every $x \in U$, we define

$$u_{\gamma,\varepsilon}(x) \coloneqq (\widetilde{u}_{\gamma,\varepsilon} * \chi_\varepsilon)(x),$$

where $*$ denotes the convolution operator, $\chi \in \mathcal{C}_c^\infty(\mathbf{R}^d;\mathbf{R}_+)$ is a nonnegative smooth function with compact support in the unit ball satisfying

$$\int_{\mathbf{R}^d} \chi(x)\,\mathrm{d}x = 1 \qquad \text{and} \qquad \int_{\mathbf{R}^d} x\chi(x)\,\mathrm{d}x = 0, \tag{4.2}$$

and where we have set $\chi_\varepsilon \coloneqq \varepsilon^{-d}\chi(\varepsilon^{-1}\cdot)$.

Step 1. We show that, for every $\gamma \in (0,\infty)$,

$$\begin{aligned}\int_{U_\varepsilon} |\widetilde{u}_{\gamma,\varepsilon}(x) - x|^2\rho(x)\,\mathrm{d}x + \lambda\gamma^{d+1} \iint_{U_\varepsilon^2} \mathrm{e}^{-\gamma|x-y|}|\widetilde{u}_{\gamma,\varepsilon}(x) - \widetilde{u}_{\gamma,\varepsilon}(y)|\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\ \leqslant J_\gamma(u_\gamma) + C\varepsilon.\end{aligned} \tag{4.3}$$

To prove this, we bound the first term on the left side of (4.3) by

$$\begin{aligned}\int_{U_\varepsilon} |\widetilde{u}_{\gamma,\varepsilon}(x) - x|^2\rho(x)\,\mathrm{d}x &\leqslant (1-\varepsilon)^{-d}\int_U \left|u_\gamma(x) - \frac{x}{1-\varepsilon}\right|^2 \rho\left(\frac{x}{1-\varepsilon}\right)\,\mathrm{d}x \\ &\leqslant \int_U |u_\gamma(x) - x|^2\rho(x)\,\mathrm{d}x + C\varepsilon,\end{aligned}$$

where in the second inequality we used the fact that ρ is Lipschitz. For the second term on the left side of (4.3), we proceed similarly, noting that

$$\begin{aligned}&\gamma^{d+1} \iint_{U_\varepsilon^2} \mathrm{e}^{-\gamma|x-y|}|\widetilde{u}_{\gamma,\varepsilon}(x) - \widetilde{u}_{\gamma,\varepsilon}(y)|\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \\ &\qquad\leqslant \frac{\gamma^{d+1}}{(1-\varepsilon)^{2d}} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|/(1-\varepsilon)}|u_\gamma(x) - u_\gamma(y)|\rho\left(\frac{x}{1-\varepsilon}\right)\rho\left(\frac{y}{1-\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}y \\ &\qquad\leqslant \frac{\gamma^{d+1}}{(1-\varepsilon)^{2d}} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|}|u_\gamma(x) - u_\gamma(y)|\rho\left(\frac{x}{1-\varepsilon}\right)\rho\left(\frac{y}{1-\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}y \\ &\qquad\leqslant \frac{\gamma^{d+1}}{(1-\varepsilon)^{2d}} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|}|u_\gamma(x) - u_\gamma(y)|\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y + C\varepsilon.\end{aligned}$$

It is in this calculation that the star-shaped property is crucial: in the second inequality, we used that the map sending U_ε to U (i.e. the map $x \mapsto x/(1-\varepsilon)$) is contractive. We also used (2.5) and again the fact that ρ is Lipschitz. Combining the last two displays, we obtain (4.3).

Step 2. We show that, for every $\gamma \in (0,\infty)$,

$$J_\gamma(u_{\gamma,\varepsilon}) \leqslant J_\gamma(u_\gamma) + C\varepsilon. \tag{4.4}$$

Using (4.2), we can write

$$\begin{aligned}\int_U |u_{\gamma,\varepsilon}(x) - x|^2 \,\mathrm{d}\mu(x) &= \int_U \left| \int_{U_\varepsilon} (\widetilde{u}_{\gamma,\varepsilon}(y) - y)\chi_\varepsilon(x-y)\,\mathrm{d}y \right|^2 \rho(x)\,\mathrm{d}x \\ &\leqslant \int_{U_\varepsilon} |\widetilde{u}_{\gamma,\varepsilon}(y) - y|^2 \int_{\mathbf{R}^d} \chi_\varepsilon(x-y)\rho(x)\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

Since ρ is Lipschitz, the inner integral is close to $\rho(y)$, up to an error bounded by $C\varepsilon$, and we thus get that

$$\int_U |u_{\gamma,\varepsilon}(x) - x|^2 \,\mathrm{d}\mu(x) \leqslant \int_{U_\varepsilon} |\widetilde{u}_{\gamma,\varepsilon}(x) - x|^2 \rho(x)\,\mathrm{d}x + C\varepsilon. \tag{4.5}$$

We also have

$$\begin{aligned}&\gamma^{d+1} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|} |u_{\gamma,\varepsilon}(x) - u_{\gamma,\varepsilon}(y)| \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\ &\quad\leqslant \gamma^{d+1} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|} \left| \int_{\mathbf{R}^d} [\widetilde{u}_{\gamma,\varepsilon}(x-z) - \widetilde{u}_{\gamma,\varepsilon}(y-z)]\chi_\varepsilon(z)\,\mathrm{d}z \right| \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\ &\quad\leqslant \gamma^{d+1} \iint_{U^2} \int_{\mathbf{R}^d} \mathrm{e}^{-\gamma|x-y|} \left|\widetilde{u}_{\gamma,\varepsilon}(x) - \widetilde{u}_{\gamma,\varepsilon}(y)\right| \chi_\varepsilon(z)\rho(x+z)\rho(y+z)\,\mathrm{d}z\,\mathrm{d}x\,\mathrm{d}y \\ &\quad\leqslant \gamma^{d+1} \iint_{U_\varepsilon^2} \mathrm{e}^{-\gamma|x-y|} \left|\widetilde{u}_{\gamma,\varepsilon}(x) - \widetilde{u}_{\gamma,\varepsilon}(y)\right| \left(\int_{\mathbf{R}^d} \chi_\varepsilon(z)\rho(x+z)\rho(y+z)\,\mathrm{d}z \right) \mathrm{d}x\,\mathrm{d}y \\ &\quad\leqslant \gamma^{d+1} \iint_{U_\varepsilon^2} \mathrm{e}^{-\gamma|x-y|} \left|\widetilde{u}_{\gamma,\varepsilon}(x) - \widetilde{u}_{\gamma,\varepsilon}(y)\right| \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y + C\varepsilon,\end{aligned}$$

where in the last step we used (4.3), (2.5), and the fact that ρ is Lipschitz. Combining the last two displays with (4.3) yields (4.4).

Step 3. We show that, for every $\gamma \in [1,\infty)$ and $\varepsilon \in (0,1]$,

$$J_\infty(u_{\gamma,\varepsilon}) \leqslant J_\gamma(u_\gamma) + C\varepsilon + \frac{C}{\gamma\varepsilon^2}. \tag{4.6}$$

In view of (4.4), it suffices to show (4.6) with $J_\gamma(u_\gamma)$ replaced by $J_\gamma(u_{\gamma,\varepsilon})$. We start by using the fact that $\|D^2 u_{\gamma,\varepsilon}\|_{L^\infty(\mu)} \leqslant C\varepsilon^{-2}$ to write

$$\begin{aligned}&\gamma^{d+1} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|} |u_{\gamma,\varepsilon}(x) - u_{\gamma,\varepsilon}(y)| \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\ &\qquad\geqslant \gamma^{d+1} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|} |Du_{\gamma,\varepsilon}(x)\cdot(x-y)| \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\ &\qquad\quad - C\gamma^{d+1} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|} \frac{|x-y|^2}{\varepsilon^2} \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y.\end{aligned} \tag{4.7}$$

Since ρ is bounded and

$$\gamma^{d+1} \int_{\mathbf{R}^d} \mathrm{e}^{-\gamma|x-y|} |x-y|^2 \,\mathrm{d}y = \gamma^{-1} \int_{\mathbf{R}^d} \mathrm{e}^{-|y|} |y|^2 \,\mathrm{d}y, \tag{4.8}$$

we see that the second integral on the right-hand side of (4.7) is bounded by $C\gamma^{-1}\varepsilon^{-2}$. Next, we aim to compare the first integral on the right-hand side of (4.7) with the same quantity with $\rho(y)$ replaced by $\rho(x)$. Since ρ is Lipschitz and $\|Du_{\gamma,\varepsilon}\|_{L^\infty(\mu)} \leqslant C\varepsilon^{-1}$, the difference between these two quantities is bounded by

$$C\varepsilon^{-1}\gamma^{d+1} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|}|x-y|^2\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \leqslant C\gamma^{-1}\varepsilon^{-1},$$

using again (4.8) and the boundedness of ρ. To complete this step, it remains to argue that

$$\gamma^{d+1} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|}|Du_{\gamma,\varepsilon}(x)\cdot(x-y)|\rho(x)^2\,\mathrm{d}x\,\mathrm{d}y \geqslant c\int \rho(x)^2|Du_{\gamma,\varepsilon}(x)|\,\mathrm{d}x + C\gamma^{-1}\varepsilon^{-1}. \tag{4.9}$$

Recalling (1.7), we see that the first term on the right-hand side above can be rewritten as

$$\gamma^{d+1}\int_U\int_{\mathbf{R}^d} \mathrm{e}^{-\gamma|x-y|}|Du_{\gamma,\varepsilon}(x)\cdot(x-y)|\rho(x)^2\,\mathrm{d}y\,\mathrm{d}x.$$

For every $\delta>0$, we denote $U^\delta := \{x\in U : \operatorname{dist}(x,\partial U)\leqslant\delta\}$. Since $\|Du_{\gamma,\varepsilon}\|_{L^\infty(\mu)}\leqslant C\varepsilon^{-1}$, the inequality (4.9) will follow from the fact that

$$\gamma^{d+1}\int_U\int_{\mathbf{R}^d\setminus U} e^{-\gamma|x-y|}|x-y|\,\mathrm{d}y\,\mathrm{d}x \leqslant C\gamma^{-1}. \tag{4.10}$$

Since U has a Lipschitz boundary, there exists $\delta>0$ such that for every $0<\eta<\eta'<\delta$, the Lebesgue measure of $U^{\eta'}\setminus U^\eta$ is at most $C(\eta'-\eta)$. Therefore,

$$\begin{aligned}
&\gamma^{d+1}\int_U\int_{\mathbf{R}^d\setminus U} e^{-\gamma|x-y|}|x-y|\,\mathrm{d}y\,\mathrm{d}x\\
&\quad\leqslant C\gamma^{d+1}e^{-\delta\gamma} + \gamma^{d+1}\sum_{k=0}^{\lceil\delta\gamma\rceil}\int_{U^{(k+1)\gamma^{-1}}\setminus U^{k\gamma^{-1}}}\int_{\mathbf{R}^d\setminus U} e^{-\gamma|x-y|}|x-y|\,\mathrm{d}y\,\mathrm{d}x\\
&\quad\leqslant C\gamma^{d+1}e^{-\delta\gamma} + \gamma^{d+1}\sum_{k=0}^{\lceil\delta\gamma\rceil} e^{-\frac{\gamma k}{2}}\int_{U^{(k+1)\gamma^{-1}}\setminus U^{k\gamma^{-1}}}\int_{\mathbf{R}^d} e^{-\frac{\gamma|x-y|}{2}}|x-y|\,\mathrm{d}y\,\mathrm{d}x\\
&\quad\leqslant C\gamma^{d+1}e^{-\delta\gamma} + C\gamma^{-1}\sum_{k=0}^{\lceil\delta\gamma\rceil} e^{-\frac{\gamma k}{2}}\\
&\quad\leqslant C\gamma^{-1}.
\end{aligned}$$

This is (4.10). Combining these estimates with (4.4) yields (4.6).

Step 4. We show that

$$\int_{U_\varepsilon}|\widetilde{u}_{\infty,\varepsilon}(x)-x|^2\,\rho(x)\,\mathrm{d}x + c\lambda\iint_{U_\varepsilon^2}\rho(x)^2\,\mathrm{d}|D\widetilde{u}_{\infty,\varepsilon}|(x) \leqslant J_\infty(u_\infty)+C\varepsilon. \tag{4.11}$$

This follows from the fact that the the left side of (4.11) can be rewritten as

$$(1-\varepsilon)^{-d}\int_U\left|u_\infty(x)-\frac{x}{1-\varepsilon}\right|^2\rho\left(\frac{x}{1-\varepsilon}\right)\,\mathrm{d}x + \frac{c\lambda}{(1-\varepsilon)^{d+1}}\iint_{U^2}\rho\left(\frac{x}{1-\varepsilon}\right)^2\,\mathrm{d}|Du_\infty|(x),$$

and from the fact that ρ is Lipschitz.

Step 5. We show that

$$J_\infty(u_{\infty,\varepsilon}) \leqslant J_\infty(u_\infty)+C\varepsilon. \tag{4.12}$$

Arguing in the same way as for (4.5), we see that

$$\int_U|u_{\infty,\varepsilon}(x)-x|^2\,\mathrm{d}\mu(x) \leqslant \int_{U_\varepsilon}|\widetilde{u}_{\infty,\varepsilon}(x)-x|^2\rho(x)\,\mathrm{d}x + C\varepsilon. \tag{4.13}$$

For the second term, we notice that by [3, Proposition 3.2], we have

$$D(\widetilde{u}_{\infty,\varepsilon} * \chi_\varepsilon) = D\widetilde{u}_{\infty,\varepsilon} * \chi_\varepsilon,$$

and thus

$$\begin{aligned}\int_U \rho(x)^2 |D(\widetilde{u}_{\infty,\varepsilon} * \chi_\varepsilon)|(x)\,\mathrm{d}x &\leqslant \int_U \int_{U_\varepsilon} \rho(x)^2 \chi_\varepsilon(x-y)\,\mathrm{d}|D\widetilde{u}_{\infty,\varepsilon}|(y)\,\mathrm{d}x \\ &\leqslant \int_{U_\varepsilon} \rho(y)^2\,\mathrm{d}|D\widetilde{u}_{\infty,\varepsilon}|(y) + C\varepsilon,\end{aligned}$$

where we used (4.11), (2.5), and the fact that ρ is Lipschitz in the last step. Combining this with (4.13) and using (4.11) once more, we obtain (4.12).

Step 6. We show that

$$J_\gamma(u_{\infty,\varepsilon}) \leqslant J_\infty(u_\infty) + C\varepsilon + \frac{C}{\gamma\varepsilon^2}. \tag{4.14}$$

We decompose the fusion term of $J_\gamma(u_{\infty,\varepsilon})$ into

$$\begin{aligned}&\gamma^{d+1} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|} |u_{\infty,\varepsilon}(x) - u_{\infty,\varepsilon}(y)| \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\ &\quad\leqslant \gamma^{d+1} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|} |Du_{\infty,\varepsilon}(x)\cdot(x-y)| \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\ &\qquad + C\gamma^{d+1} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|} \frac{|x-y|^2}{\varepsilon^2} \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y,\end{aligned} \tag{4.15}$$

and estimate each of these integrals in turn. The second integral on the right side is the same as the second integral in (4.7), and thus is bounded by $C\gamma^{-1}\varepsilon^{-2}$. We next aim to compare the first integral on the right-hand side of (4.15) with the one where $\rho(y)$ is replaced by $\rho(x)$. Since ρ is Lipschitz, the difference between these two quantities is bounded by

$$C\gamma^{d+1} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|} |Du_{\infty,\varepsilon}(x)| |x-y|^2\,\mathrm{d}x\,\mathrm{d}y \leqslant C\gamma^{-1} \int_U |Du_{\infty,\varepsilon}(x)|\,\mathrm{d}x \leqslant C\gamma^{-1},$$

where we used (4.12) and the fact that ρ is bounded above and below in the last step. Then it remains to estimate

$$\begin{aligned}&\gamma^{d+1} \iint_{U^2} \mathrm{e}^{-\gamma|x-y|} |Du_{\infty,\varepsilon}(x)\cdot(x-y)| \rho(x)^2\,\mathrm{d}x\,\mathrm{d}y \\ &\quad\leqslant \int_{\mathbf{R}^2} \mathrm{e}^{-|y|} |y\cdot \mathrm{e}_1|\,\mathrm{d}y \int_U |Du_{\infty,\varepsilon}(x)| \rho(x)^2\,\mathrm{d}x = c\int_U |Du_{\infty,\varepsilon}(x)| \rho(x)^2\,\mathrm{d}x,\end{aligned}$$

where we recalled (1.7) in the last step. Thus we have

$$J_\gamma(u_{\infty,\varepsilon}) \leqslant J_\infty(u_{\infty,\varepsilon}) + C\gamma^{-1}\varepsilon^{-2},$$

and inequality (4.14) then follows using (4.12).

Step 7. We can now conclude the proof. We take $\varepsilon := \gamma^{-1/3}$, and using (4.6) and (4.14), we see that

$$J_\infty(u_\infty) \leqslant J_\infty(u_{\gamma,\gamma^{-1/3}}) \leqslant J_\gamma(u_\gamma) + C\gamma^{-1/3} \leqslant J_\gamma(u_{\infty,\gamma^{-1/3}}) + C\gamma^{-1/3} \leqslant J_\infty(u_\infty) + C\gamma^{-1/3}.$$

From this, we deduce that

$$|J_\infty(u_\infty) - J_\gamma(u_\gamma)| \leqslant C\gamma^{-1/3}, \tag{4.16}$$

and moreover that

$$0 \leqslant J_\infty(u_{\gamma,\gamma^{-1/3}}) - J_\infty(u_\infty) \leqslant C\gamma^{-1/3}. \tag{4.17}$$

By (2.4) and (4.17), we obtain

$$\int |u_{\gamma,\gamma^{-1/3}} - u_\infty|^2 \,\mathrm{d}\mu \leqslant C\gamma^{-1/3}. \tag{4.18}$$

Using (2.4) and (4.4), we also infer that

$$\int |u_{\gamma,\gamma^{-1/3}} - u_\gamma|^2 \,\mathrm{d}\mu \leqslant C\gamma^{-1/3}. \tag{4.19}$$

Combining (4.16), (4.18), and (4.19) yields (4.1). □

Remark 4.2. In the proof of Theorem 4.1, the assumption that U is effectively star-shaped could be replaced by the following weaker assumption: that there exist $L < \infty$ and, for every $\varepsilon > 0$ sufficiently small, a 1-Lipschitz injective map $P_\varepsilon : U_\varepsilon \to U$ with L-Lipschitz inverse. In Theorem 1.2, we could then assume that the same property holds for each of the sets $U_1, \dots, U_L$ in place of the assumption that these sets are effectively star-shaped.

5. Properties of the limiting functional

In this section we show that if λ is large enough, then the minimizer $u_{\mu,\lambda,\infty}$ of $J_{\mu,\lambda,\infty}$ recovers the connected components of $\operatorname{supp}\mu$.

Proposition 5.1. *Let μ be a probability measure on $\mathbf{R}^d$ satisfying the conditions of Theorem 1.2, so its support is the disjoint union of $\overline{U_1} \sqcup \cdots \sqcup \overline{U_L}$. There is a $\lambda_c < \infty$ such that if $\lambda \geqslant \lambda_c$, then $u_{\mu,\lambda,\infty}(x) = \operatorname{cent}_\mu(U_\ell)$ for all $x \in U_\ell$, $\ell \in \{1, \dots, L\}$.*

Proof. Let $u(x) = \operatorname{cent}_\mu(U_\ell)$ for all $x \in U_\ell$, $\ell \in \{1, \dots, L\}$. Since the gradient of u is zero on each U_ℓ, we have

$$J_{\mu,\lambda,\infty}(u) = \sum_{\ell=1}^{L} \int_{U_\ell} |u(x) - x|^2 \,\mathrm{d}\mu(x).$$

Let $U = \bigcup_{\ell=1}^L U_\ell$, $p > d$, and let $W^{1,p}(U)$ denote the usual Sobolev space with regularity 1 and integrability p. Note that $W^{1,p}(U)$ embeds continuously into $\mathcal{C}(\overline{U})$ by Morrey's inequality; see [1, Theorem 4.12]. Let $\psi \in (W^{1,p}(U))^{d\times d}$ be a weak solution to the PDE

$$2\rho(x)(u(x)_j - x_j) - c\sum_{k=1}^{d} D_k(\rho^2\psi_{jk})(x) = 0, \qquad x \in U, j = 1, \dots, d; \tag{5.1}$$

$$\psi|_{\partial U} \equiv 0. \tag{5.2}$$

We note that the problem (5.1)–(5.2) separates into dL problems, one for each j and ℓ. Each problem can be solved by [7, Theorem 2.4] (which follows the approach introduced in [5, 6]). We have, for every $v \in (L^2(U) \cap \mathrm{BV}(U))^d$,

$$\begin{aligned} J_{\mu,\lambda,\infty}(u+v) &= \int_U |u(x) + v(x) - x|^2 \,\mathrm{d}\mu(x) + c\lambda \int_U \rho(x)^2 \,\mathrm{d}|Dv|(x) \\ &= J_{\mu,\lambda,\infty}(u) + \int_U \left(2(u(x) - x)\cdot v(x) + |v(x)|^2\right) \,\mathrm{d}\mu(x) + c\lambda \int_U \rho(x)^2 \,\mathrm{d}|Dv|(x). \end{aligned}$$

A minor variant of (2.1) takes the form

$$\int_U \rho(x)^2 \,\mathrm{d}|Dv|(x) = \sup\left\{ \int_U \rho(x)^2 \phi(x)\cdot \mathrm{d}Dv(x),\ \phi \in (\mathcal{C}(\overline{U}))^{d\times d} \text{ s.t. } \|\phi\|_{L^\infty(U)} \leqslant 1 \right\}.$$

Selecting $\phi = \psi/\|\psi\|_{L^\infty(U)}$, and using the assumption that $\lambda \geqslant \|\psi\|_{L^\infty(U)}$, we obtain

$$\begin{aligned}
J_{\mu,\lambda,\infty}(u+v) &\geqslant J_{\mu,\lambda,\infty}(u) + \int \left(2(u(x)-x)\cdot v(x) + |v(x)|^2\right) \mathrm{d}\mu(x) \\
&\qquad + c\sum_{j,k=1}^{d} \int \rho(x)^2 \psi_{jk}(x) D_k v_j(x)\,\mathrm{d}x \\
&= J_{\mu,\lambda,\infty}(u) + \int \left(2(u(x)-x)\cdot v(x) + |v(x)|^2\right) \mathrm{d}\mu(x) \\
&\qquad - \sum_{j=1}^{d} \int 2\rho(x)(u(x)_j - x_j)(x) v_j(x)\,\mathrm{d}x \\
&= J_{\mu,\lambda,\infty}(u) + \int |v(x)|^2\,\mathrm{d}\mu(x) \\
&\geqslant J_{\mu,\lambda,\infty}(u),
\end{aligned}$$

where we used (5.1) for the first equality. This implies that $u_{\mu,\lambda,\infty} = u$, and hence the statement of the proposition with $\lambda_{\mathrm{c}} = \|\psi\|_{L^\infty(U)}$. □

6. Truncation

In this section we prove a stability result for when we truncate the exponential weight. For $\gamma, \omega \in (0,\infty)$, we define the truncated functional

$$\begin{aligned}
&\overline{J}_{\mu,\lambda,\gamma,\omega}(u) \\
&\quad := \int |u(x)-x|^2\,\mathrm{d}\mu(x) + \lambda\gamma^{d+1} \iint \mathrm{e}^{-\gamma|x-y|}\mathbf{1}\{|x-y|\leqslant\omega\}|u(x)-u(y)|\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y).
\end{aligned} \tag{6.1}$$

The functional $\overline{J}_{\mu,\lambda,\gamma,\omega}$ is uniformly convex and satisfies (2.2) and (2.4) in the same way as $J_{\mu,\lambda,\gamma}$. Let $\overline{u}_{\mu,\lambda,\gamma,\omega}$ be the (unique) minimizer of $\overline{J}_{\mu,\lambda,\gamma,\omega}$.

Proposition 6.1. *Let $\gamma, \lambda, \omega > 0$ and let μ be a probability measure on $\mathbf{R}^d$ with compact support. Let $M := \operatorname{diam}\operatorname{supp}\mu$. Then we have*

$$\int |\overline{u}_{\mu,\lambda,\gamma,\omega}(x) - u_{\mu,\lambda,\gamma}(x)|^2\,\mathrm{d}\mu(x) \leqslant 2M\lambda\gamma^{d+1}\mathrm{e}^{-\gamma\omega}. \tag{6.2}$$

In light of this statement, we define

$$\overline{u}_{\mu,\lambda,\gamma} := \overline{u}_{\mu,\lambda,\gamma,(d+4/3)\gamma^{-1}\log\gamma}. \tag{6.3}$$

Then (6.2) implies that

$$\int |\overline{u}_{\mu,\lambda,\gamma}(x) - u_{\mu,\lambda,\gamma}(x)|^2\,\mathrm{d}\mu(x) \leqslant 2M\lambda\gamma^{-1/3}. \tag{6.4}$$

Proof of Proposition 6.1. Subtracting (1.3) from (6.1), we obtain

$$\overline{J}_{\mu,\lambda,\gamma,\omega}(u) - J_{\mu,\lambda,\gamma}(u) = \lambda\gamma^{d+1} \iint \mathrm{e}^{-\gamma|x-y|}\mathbf{1}\{|x-y|>\omega\}|u(x)-u(y)|\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y).$$

Taking $u = u_{\mu,\lambda,\gamma}$, we get

$$\begin{aligned}
&\overline{J}_{\mu,\lambda,\gamma,\omega}(u_{\mu,\lambda,\gamma}) - \inf J_{\mu,\lambda,\gamma} \\
&\qquad = \lambda\gamma^{d+1} \iint \mathrm{e}^{-\gamma|x-y|}\mathbf{1}\{|x-y| > \omega\}|u_{\mu,\lambda,\gamma}(x) - u_{\mu,\lambda,\gamma}(y)|\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \\
&\qquad \leqslant M\lambda\gamma^{d+1}\mathrm{e}^{-\gamma\omega},
\end{aligned}$$

and similarly,

$$\begin{aligned}
&J_{\mu,\lambda,\gamma}(\overline{u}_{\mu,\lambda,\gamma,\omega}) - \inf \overline{J}_{\mu,\lambda,\gamma,\omega} \\
&\qquad = -\lambda\gamma^{d+1} \iint \mathrm{e}^{-\gamma|x-y|}\mathbf{1}\{|x-y| > \omega\}|\overline{u}_{\mu,\lambda,\gamma,\omega}(x) - \overline{u}_{\mu,\lambda,\gamma,\omega}(y)|\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \leqslant 0.
\end{aligned}$$

Therefore, using (2.4) and the last two displays we have

$$\begin{aligned}
&\int |\overline{u}_{\mu,\lambda,\gamma,\omega}(x) - u_{\mu,\lambda,\gamma}(x)|^2\,\mathrm{d}\mu(x) \\
&\quad \leqslant 2\left(J_{\mu,\lambda,\gamma}(\overline{u}_{\mu,\lambda,\gamma,\omega}) - \inf J_{\mu,\lambda,\gamma}\right) \\
&\quad \leqslant 2\left[J_{\mu,\lambda,\gamma}(\overline{u}_{\mu,\lambda,\gamma,\omega}) - \inf \overline{J}_{\mu,\lambda,\gamma,\omega}\right] + 2\left[\overline{J}_{\mu,\lambda,\gamma,\omega}(u_{\mu,\lambda,\gamma}) - \inf J_{\mu,\lambda,\gamma}\right] \\
&\quad \leqslant 2M\lambda\gamma^{d+1}\mathrm{e}^{-\gamma\omega},
\end{aligned}$$

as claimed. □

7. Proof of Theorem 1.2

In this section we prove Theorem 1.2. We first need a result from [15]. Recall the notation d' introduced in (1.4).

Proposition 7.1. *Let $U \subseteq \mathbf{R}^d$ be a bounded, connected domain with Lipschitz boundary. Let μ be a probability measure on U, absolutely continuous with respect to Lebesgue measure, with density bounded above and away from zero on U. For every $\alpha \geqslant 1$, there is a constant $C < \infty$, depending only on U, α, and μ, such that the following holds. If $(X_n)_{n\in\mathbf{N}}$ are independent random variables with law μ, then for every integer $N \geqslant 1$,*

$$\mathbf{P}\left(\mathcal{W}_\infty\left(\mu, \frac{1}{N}\sum_{n=1}^{N}\delta_{X_n}\right) \geqslant CN^{-1/(d\vee 2)}(\log N)^{1/d'}\right) \leqslant CN^{-\alpha}.$$

Proof. For $d \geqslant 2$, this is a restatement of [15, Theorem 1.1]. For $d = 1$, the result can be obtained from the classical Kolmogorov-Smirnov quantitative version of the Glivenko-Cantelli theorem. □

Now we can prove Theorem 1.2. For a measure μ on $\mathbf{R}^d$ and a Borel set U, we denote by $\mu \llcorner U$ the restriction of μ to the set U.

Proof of Theorem 1.2. Recalling (6.3), it is clear that if γ is so large that

$$(d + 4/3)\gamma^{-1}\log\gamma \leqslant \min_{1\leqslant\ell\neq\ell'\leqslant L} \mathrm{dist}(U_\ell, U_{\ell'}), \tag{7.1}$$

then

$$\overline{u}_{\mu_N\llcorner U_\ell,\lambda,\gamma}(x) = \overline{u}_{\mu_N,\lambda,\gamma}(x), \qquad \text{for all } x \in U_\ell, \tag{7.2}$$

and similarly

$$\overline{u}_{\mu\llcorner U_\ell,\lambda,\gamma}(x) = \overline{u}_{\mu,\lambda,\gamma}(x), \qquad \text{for all } x \in U_\ell. \tag{7.3}$$

Also, we have by the definitions and Proposition 5.1 that there exists λ_{c} such that for every $\lambda \geqslant \lambda_{\mathrm{c}}$,

$$u_{\mu \llcorner U_\ell,\lambda,\infty}(x) = u_{\mu,\lambda,\infty}(x) = \mathrm{cent}_\mu(U_\ell), \qquad \text{for all } x \in U_\ell. \tag{7.4}$$

By (7.4) and Theorem 4.1, we have

$$\int_{U_\ell} |\,\mathrm{cent}_\mu(U_\ell) - u_{\mu \llcorner U_\ell,\lambda,\gamma}|^2 \,\mathrm{d}\mu = \int_{U_\ell} |u_{\mu \llcorner U_\ell,\lambda,\infty} - u_{\mu \llcorner U_\ell,\lambda,\gamma}|^2 \,\mathrm{d}\mu \leqslant C\gamma^{-1/3}.$$

By (7.3) and (6.4), we have, as long as (7.1) holds,

$$\int_{U_\ell} |\overline{u}_{\mu,\lambda,\gamma} - u_{\mu \llcorner U_\ell,\lambda,\gamma}|^2 \,\mathrm{d}\mu = \int_{U_\ell} |\overline{u}_{\mu \llcorner U_\ell,\lambda,\gamma} - u_{\mu \llcorner U_\ell,\lambda,\gamma}|^2 \,\mathrm{d}\mu \leqslant 2M\lambda\gamma^{-1/3}.$$

Combining the last two displays, we see that

$$\int_{U_\ell} |\overline{u}_{\mu,\lambda,\gamma} - \mathrm{cent}_\mu(U_\ell)|^2 \,\mathrm{d}\mu \leqslant C(1+\lambda)\gamma^{-1/3}.$$

Using (6.4) again, this implies that

$$\int_{U_\ell} |u_{\mu,\lambda,\gamma} - \mathrm{cent}_\mu(U_\ell)|^2 \,\mathrm{d}\mu \leqslant C(1+\lambda)\gamma^{-1/3}. \tag{7.5}$$

On the other hand, by Proposition 7.1, we have for each ℓ that

$$\mathbf{P}\left(\mathcal{W}_\infty\left(\frac{\mu \llcorner U_\ell}{\mu(U_\ell)}, \frac{\mu_N \llcorner U_\ell}{\mu_N(U_\ell)}\right) \geqslant CN^{-1/(d\vee 2)}(\log N)^{1/d'}\right) \leqslant CN^{-100}. \tag{7.6}$$

By Proposition 3.1, for each ℓ there is an ∞-optimal transport plan $\pi_{\ell,N}$ between $\frac{\mu \llcorner U_\ell}{\mu(U_\ell)}$ and $\frac{\mu_N \llcorner U_L}{\mu_N(U_\ell)}$ such that, using also (7.2) and (7.3), we have

$$\iint_{U_\ell^2} |u_{\mu,\lambda,\gamma}(x) - u_{\mu_N,\lambda,\gamma}(\widetilde{x})|^2 \,\mathrm{d}\pi_{\ell,N}(x,\widetilde{x}) \leqslant C(\gamma+1)\mathcal{W}_\infty\left(\frac{\mu \llcorner U_\ell}{\mu(U_\ell)}, \frac{\mu_N \llcorner U_\ell}{\mu_N(U_\ell)}\right).$$

Combining this with (7.5), we see that

$$\begin{aligned}
&\frac{1}{\mu_N(U_\ell)} \int_{U_\ell} |u_{\mu_N,\lambda,\gamma} - \mathrm{cent}_\mu(U_\ell)|^2 \,\mathrm{d}\mu_N \\
&\qquad = \iint_{U_\ell^2} |u_{\mu_N,\lambda,\gamma}(\widetilde{x}) - \mathrm{cent}_\mu(U_\ell)|^2 \,\mathrm{d}\pi(x,\widetilde{x}) \\
&\qquad \leqslant C\left((\gamma+1)\mathcal{W}_\infty\left(\frac{\mu \llcorner U_\ell}{\mu(U_\ell)}, \frac{\mu_N \llcorner U_\ell}{\mu_N(U_\ell)}\right) + (1+\lambda)\gamma^{-1/3}\right).
\end{aligned}$$

Now summing over ℓ and using (7.6) and the fact that the term inside the expectation on the left-hand side of (1.5) is bounded almost surely, we obtain (1.5). □

References

[1] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[2] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine learning*, 75(2):245–248, 2009.

[3] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000.

[4] P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, and R. Ward. Relax, no need to round: Integrality of clustering formulations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, page 191–200, 2015.

[5] M. E. Bogovskiĭ. Solution of the first boundary value problem for an equation of continuity of an incompressible medium. *Dokl. Akad. Nauk SSSR*, 248(5):1037–1040, 1979.

[6] M. E. Bogovskiĭ. Solutions of some problems of vector analysis, associated with the operators div and grad. In *Theory of cubature formulas and the application of functional analysis to problems of mathematical physics*, volume 1980 of *Trudy Sem. S. L. Soboleva, No. 1*, pages 5–40, 149. Akad. Nauk SSSR Sibirsk. Otdel., Inst. Mat., Novosibirsk, 1980.

[7] W. Borchers and H. Sohr. On the equations $\operatorname{rot} \mathbf{v} = \mathbf{g}$ and $\operatorname{div} \mathbf{u} = f$ with zero boundary conditions. *Hokkaido Math. J.*, 19(1):67–87, 1990.

[8] T. Champion, L. De Pascale, and P. Juutinen. The ∞-Wasserstein distance: local solutions and existence of optimal transport maps. *SIAM J. Math. Anal.*, 40(1):1–20, 2008.

[9] E. C. Chi and K. Lange. Splitting methods for convex clustering. *J. Comput. Graph. Statist.*, 24(4):994–1013, 2015.

[10] J. Chiquet, P. Gutierrez, and G. Rigaill. Fast tree inference with weighted fusion penalties. *J. Comput. Graph. Statist.*, 26(1):205–216, 2017.

[11] A. De Rosa and A. Khajavirad. The ratio-cut polytope and K-means clustering. *ArXiv preprint arXiv:2006.15225*, 2020.

[12] A. Del Pia and M. Ma. K-median: exact recovery in the extended stochastic ball model. *ArXiv preprint arXiv:2109.02547*, 2021.

[13] A. Dunlap and J.-C. Mourrat. Sum-of-norms clustering does not separate nearby balls. Preprint, arXiv:2104.13753.

[14] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[15] N. García Trillos and D. Slepčev. On the rate of convergence of empirical measures in ∞-transportation distance. *Canad. J. Math.*, 67(6):1358–1383, 2015.

[16] T. Hocking, J. Vert, F. R. Bach, and A. Joulin. Clusterpath: an algorithm for clustering using convex fusion penalties. In L. Getoor and T. Scheffer, editors, *Proc. 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, page 745–752. Omnipress, 2011.

[17] T. Jiang and S. Vavasis. Certifying clusters from sum-of-norms clustering. Preprint, arXiv:2006.11355.

[18] T. Jiang, S. Vavasis, and C. W. Zhai. Recovery of a mixture of gaussians by sum-of-norms clustering. *J. Mach. Learn. Res.*, 21(225):1–16, 2020.

[19] F. Lindsten, H. Ohlsson, and L. Ljung. Clustering using sum-of-norms regularization: With application to particle filter output computation. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, page 201–204, 2011.

[20] S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[21] M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is NP-hard. In *WALCOM—Algorithms and computation*, volume 5431 of *Lecture Notes in Comput. Sci.*, page 274–285. Springer, Berlin, 2009.

[22] A. Nellore and R. Ward. Recovery guarantees for exemplar-based clustering. *Inform. and Comput.*, 245:165–180, 2015.

[23] C. H. Nguyen and H. Mamitsuka. On convex clustering solutions. *ArXiv preprint arXiv:2105.08348*, 2021.

[24] A. Panahi, D. P. Dubhashi, F. D. Johansson, and C. Bhattacharyya. Clustering by sum of norms: Stochastic incremental algorithm, convergence and cluster recovery. In D. Precup and Y. W. Teh, editors, *Proc. 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proc. Mach. Learn. Res.*, page 2769–2777, 2017.

[25] K. Pelckmans, J. De Brabanter, B. De Moor, and J. Suykens. Convex clustering shrinkage. In *Workshop on Statistics and optimization of clustering Workshop (PASCAL)*, 2005.

[26] D. Sun, K.-C. Toh, and Y. Yuan. Convex clustering: Model, theoretical guarantee and efficient algorithm. *J. Mach. Learn. Res.*, 22:1–32, 2021.

[27] K. M. Tan and D. Witten. Statistical properties of convex clustering. *Electron. J. Stat.*, 9(2):2324–2347, 2015.

[28] S. Vassilvitskii and D. Arthur. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2006.

[29] C. Zhu, H. Xu, C. Leng, and S. Yan. Convex optimization procedure for clustering: Theoretical revisit. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, page 1619–1627, 2014.

(A. Dunlap) Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 USA

E-mail address: alexander.dunlap@cims.nyu.edu

(J.-C. Mourrat) Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 USA; CNRS, Ecole Normale Supérieure de Lyon, Lyon, France

E-mail address: jean-christophe.mourrat@ens-lyon.fr