



**HAL**  
open science

## Local versions of sum-of-norms clustering

Alexander Dunlap, Jean-Christophe Mourrat

► **To cite this version:**

Alexander Dunlap, Jean-Christophe Mourrat. Local versions of sum-of-norms clustering. *SIAM Journal on Mathematics of Data Science*, 2022, 4 (4), pp.1250-1271. 10.1137/21M1448732 . hal-03469585

**HAL Id: hal-03469585**

**<https://hal.science/hal-03469585>**

Submitted on 24 Dec 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 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

Let  $d \in \mathbf{N}$  and let  $x_1, \dots, x_N \in \mathbf{R}^d$  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. This problem has a long history in the theoretical statistics and computer science literature, which we do not attempt to review here. We focus our attention on the “sum-of-norms clustering” method (also known as “convex clustering shrinkage” or “Clusterpath”) introduced in [19, 13, 16], which identifies clusters as the level sets of the minimizer of the convex functional

$$(y_1, \dots, y_N) \mapsto \frac{1}{N} \sum_{n=1}^N |y_n - x_n|^2 + \frac{\lambda}{N^2} \sum_{m,n=1}^N w(|x_m - x_n|) |y_m - y_n| \quad (1.1)$$

over  $(y_1, \dots, y_N) \in (\mathbf{R}^d)^N$ , for some “weight function”  $w$ . Here  $|\cdot|$  denotes the Euclidean norm. The point  $y_n$  is thought of as a “representative point” of the cluster to which  $x_n$  belongs, and so  $x_n$  and  $x_m$  belong to the same cluster if  $y_n = y_m$ . The first term of (1.1) is designed to keep the representative point of a cluster close to the points in that cluster (thus 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  $\lambda$  controls the relative strength of these two effects.

The present work will investigate 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. Following our previous work [10], for the purposes of mathematical analysis we consider the somewhat more general problem of clustering of measures. Thus, for a measure  $\mu$  on  $\mathbf{R}^d$  of compact support, we define the functional  $J_{\mu,\lambda,\gamma} : (L^2(\mu))^d \rightarrow \mathbf{R}$  by

$$J_{\mu,\lambda,\gamma}(u) := \int |u(x) - x|^2 d\mu(x) + \lambda\gamma^{d+1} \iint e^{-\gamma|x-y|} |u(x) - u(y)| d\mu(x) d\mu(y). \quad (1.2)$$

We note that (1.1) with  $w(r) = \gamma^{d+1}e^{-\gamma r}$  is obtained from (1.2) by setting  $\mu = \frac{1}{N} \sum_{n=1}^N \delta_{x_n}$ . The regime  $\gamma \downarrow 0$  with  $\lambda\gamma^{d+1}$  kept constant corresponds to the unweighted problem (i.e. with  $w \equiv 1$ ), which enjoys some good theoretical properties as discussed in, for example, [23, 22, 9, 18, 20, 15, 8, 14, 21]. However, the unweighted problem has the drawback that it fails to recover the clusters in the stochastic ball model [17] if the balls are too close together, as we showed in [10]. In the present work, we will show that this deficiency can be overcome if  $\gamma$  is chosen as an appropriate function of the number of points  $N$ . To be

more precise, if our dataset is the empirical distribution of  $N \gg 1$  points drawn from a continuous distribution whose support is the disjoint union of sufficiently nice closed sets, and if  $\gamma$  is chosen suitably in terms of  $N$ , then the minimizer of (1.2) will approximately recover the  $\mu$ -centroids of these sets.

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

$$\text{cent}_\mu(U) := \frac{1}{\mu(U)} \int_U x \, d\mu(x)$$

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

$$d' := \begin{cases} \infty & \text{if } d = 1, \\ \frac{4}{3} & \text{if } d = 2, \\ d & \text{if } d \geq 3. \end{cases} \quad (1.3)$$

Our main result is the following.

**Theorem 1.1.** *Let  $\mu$  be a probability measure on  $\mathbf{R}^d$  such that  $\text{supp } \mu = \bigcup_{\ell=1}^L \overline{U}_\ell$ , where  $U_1, \dots, U_L$  are bounded, effectively star-shaped (see Definition 1.2 below) open sets with Lipschitz boundaries, such that their closures  $\overline{U}_1, \dots, \overline{U}_L$  are pairwise disjoint. Assume that  $\mu$  admits a density with respect to the Lebesgue measure, and that this density is Lipschitz and bounded away from zero on  $\text{supp } \mu$ . Then there exist  $\lambda_c, C < \infty$  such that for every  $\lambda \geq \lambda_c$ , the following holds. Let  $(X_n)_{n \in \mathbf{N}}$  be a sequence of independent random variables with law  $\mu$ ,  $N \geq 1$  be an integer,  $\mu_N := \frac{1}{N} \sum_{n=1}^N \delta_{X_n}$ , and*

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

For every  $\gamma \geq 1$ , we have

$$\mathbf{E} \left[ \frac{1}{N} \sum_{\ell=1}^L \sum_{n \in A_N^{(\ell)}} |u_{\mu_N, \lambda, \gamma}(X_n) - \text{cent}_\mu(U_\ell)|^2 \right] \leq C \left( \gamma N^{-1/(d \vee 2)} (\log N)^{1/d'} + (1 + \lambda) \gamma^{-1/3} \right). \quad (1.4)$$

Now we define the technical condition used in the statement of the theorem.

**Definition 1.2.** For  $U$  a subset of  $\mathbf{R}^d$  and  $\varepsilon > 0$ , let  $U_\varepsilon$  be the  $\varepsilon$ -enlargement of  $U$ , namely

$$U_\varepsilon := \{x \in \mathbf{R}^d \mid \text{dist}(x, U) \leq \varepsilon\}.$$

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  $U_\varepsilon$  under the mapping  $x \mapsto x_* + (1 - C_*\varepsilon)(x - x_*)$  is contained in  $U$ .

For  $d \geq 2$ , optimizing the right-hand side of (1.4) suggests the optimal choice  $\gamma \simeq N^{3/(4d)}$ , in which case the mean-square error is of the order of  $N^{-1/(4d)}$ , up to logarithmic corrections. We do not know if the estimate in (1.4) 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 (1.4) 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 1.1 can be obtained if the weight  $r \mapsto e^{-\gamma r}$  is replaced by a truncated version  $r \mapsto e^{-\gamma r} \mathbf{1}_{r \leq \omega}$  for an appropriate choice of  $\omega$ ; 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 (1.2), and thus may lead to improvements in computational efficiency. (See [7] 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 1.1 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  $\gamma$  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.

An important step in the proof of Theorem 1.1, which is also of independent interest, concerns what happens as  $\gamma$  is taken to infinity. The factor  $\gamma^{d+1}$  in (1.2) was indeed chosen so that a limiting functional would arise, under appropriate conditions on  $\mu$ . Let  $U$  be a bounded open subset of  $\mathbf{R}^d$  and suppose that  $\text{supp } \mu = \overline{U}$ . Suppose furthermore that  $\mu$  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  $\text{BV}(U)$  the space of functions of bounded variation on  $U$ . (Some elementary properties of the space  $\text{BV}(U)$  are recalled in Section 2 below; see also [2].) If  $u \in (L^2(U) \cap \text{BV}(U))^d$ , then we can define

$$J_{\mu,\lambda,\infty}(u) := \int |u(x) - x|^2 d\mu(x) + c\lambda \int \rho(x)^2 d|Du|(x), \tag{1.5}$$

where

$$c := \int_{\mathbf{R}^d} e^{-|y|} |y \cdot e_1| dy. \tag{1.6}$$

We will see in Proposition 2.1 below that  $J_{\mu,\lambda,\infty}$  admits a unique minimizer  $u_{\mu,\lambda,\infty} \in (L^2(U) \cap \text{BV}(U))^d$ . In Theorem 4.1, we will then show in a quantitative sense that, if  $U$  is sufficiently regular and the density  $\rho$  is Lipschitz, then  $u_{\mu,\lambda,\gamma}$  converges to  $u_{\mu,\lambda,\infty}$  as  $\gamma$  tends to infinity.

The utility of the gradient functional (1.5) in the proof of Theorem 1.1 is apparent in Proposition 5.1 below. This proposition states that when  $\lambda$  is large enough, the minimizer of the gradient functional recovers the centroids of the connected components of the support of the measure  $\mu$ .

The gradient clustering functional (1.5) only makes sense for smooth measures. In order to show the convergence of the minimizers of the weighted clustering functionals (1.2) on empirical distributions, we need to relate the minimizers of the finite- $\gamma$  problem for empirical distributions to the minimizers of the finite- $\gamma$  problem for smooth distributions. We do this by proving a stability result with respect to the  $\infty$ -Wasserstein metric  $\mathcal{W}_\infty$ , which is Proposition 3.1 below. This works in combination with a quantitative Glivenko–Cantelli-type result for the  $\infty$ -Wasserstein metric proved in [12], 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 (1.2), which is done in Section 6.

**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  $u_{\tilde{\mu},\lambda,\gamma}$  as  $\tilde{\mu} \rightarrow \mu$  in the  $\infty$ -Wasserstein distance. In Section 4 we prove the convergence result for  $u_{\mu,\lambda,\gamma}$  as  $\gamma \rightarrow \infty$ . In Section 5 we show that the limiting functional  $u_{\mu,\lambda,\infty}$  recovers the centroids of the connected components of  $\text{supp } \mu$  as long as  $\lambda$  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 1.1.

**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  $\text{BV}(U)$  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  $u \in \text{BV}(U)$ , 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 [2, Proposition 1.47] that

$$|Du|(V) = \sup_{\phi} \int_V \phi \cdot dDu = \sup_{\phi} \sum_{i=1}^d \int_V \phi_i dD_i u, \quad (2.1)$$

where the supremum is over all  $\phi \in (\mathcal{C}_c(V))^d$  such that  $\|\phi\|_{L^\infty(V)} \leq 1$ , with the understanding that

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

When  $u \in (\text{BV}(U))^d$ , the gradient  $Du$  is a Radon measure taking values in the space of  $d$ -by- $d$  matrices. Identifying each such matrix with a vector of length  $d^2$ , 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 [2] 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

$$\frac{1}{2} (J_{\mu,\lambda,\gamma}(u+v) + J_{\mu,\lambda,\gamma}(u-v)) - J_{\mu,\lambda,\gamma}(u) \geq \int v^2 d\mu. \quad (2.2)$$

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  $\text{supp } \mu = \bar{U}$ . Suppose furthermore that  $\mu$  is absolutely continuous with respect to the Lebesgue measure on  $\bar{U}$  with a density  $\rho \in \mathcal{C}(\bar{U})$  that is bounded away from zero. Then for any  $\lambda > 0$ , the functional  $J_{\mu,\lambda,\infty}$  admits a unique minimizer  $u_{\mu,\lambda,\infty} \in L^2(U) \cap \text{BV}(U)$ .*

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

$$\lim_{k \rightarrow \infty} J_{\mu,\lambda,\infty}(u_k) = \inf J_{\mu,\lambda,\infty}. \quad (2.3)$$

Since  $\rho$  is bounded away from zero, the functional  $J_{\mu,\lambda,\infty}$  is coercive on  $(L^2(U) \cap \text{BV}(U))^d$ . By the Banach–Alaoglu theorem and [2, Theorem 3.23], by passing to a subsequence we can assume that there is a  $u \in (L^2(U) \cap \text{BV}(U))^d$  such that  $u_k \rightarrow u$  weakly in  $(L^2(U))^d$  and weakly- $*$  in  $(\text{BV}(U))^d$ . From the weak convergence in  $(L^2(U))^d$  we see that

$$\int |u(x) - x|^2 d\mu(x) \leq \liminf_{k \rightarrow \infty} \int |u_k(x) - x|^2 d\mu(x).$$

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

$$\begin{aligned} \int_U \rho(x)^2 \, d|Du|(x) &= \sup_{\phi} \int_U \rho(x)^2 \phi(x) \cdot dDu(x) \\ &\leq \liminf_{k \rightarrow \infty} \sup_{\phi} \int_U \rho(x)^2 \phi(x) \cdot dDu_k(x) \\ &= \liminf_{k \rightarrow \infty} \int_U \rho(x)^2 \, d|Du_k|(x), \end{aligned}$$

where the supremum is over all  $\phi \in (\mathcal{C}_c(U))^{d^2}$  such that  $\|\phi\|_{L^\infty(U)} \leq 1$ . 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).  $\square$

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

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

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

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

where we note that the right-hand side is the variance of a random variable distributed according to  $\mu$ , and in particular is independent of  $\lambda$  and  $\gamma$ .

### 3. STABILITY WITH RESPECT TO $\infty$ -WASSERSTEIN PERTURBATIONS OF THE MEASURE

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

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

where the infimum is taken over all couplings  $\pi$  of  $\mu$  and  $\nu$ . It is classical to verify that this infimum is achieved. We call any  $\pi$  achieving this infimum an  $\infty$ -optimal transport plan from  $\mu$  to  $\nu$ . In this section we prove that, for finite  $\gamma$ , the minimizer  $u_{\mu,\lambda,\gamma}$  is stable under  $\infty$ -Wasserstein perturbations of  $\mu$ .

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

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

*Proof.* Throughout the proof,  $\lambda$  and  $\gamma$  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  $\lambda$  or  $\gamma$ .) Let  $\pi$  be an  $\infty$ -optimal transport plan from  $\mu$  to  $\tilde{\mu}$ . We write the disintegration

$$d\pi(x, \tilde{x}) = d\nu(\tilde{x} | x) \, d\mu(x)$$

and define

$$\bar{u}(x) := \int u_{\tilde{\mu}}(\tilde{x}) \, d\nu(\tilde{x} \mid x).$$

We have

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

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

$$\begin{aligned} |u_{\tilde{\mu}}(\tilde{x}) - \tilde{x}|^2 &= |u_{\tilde{\mu}}(\tilde{x}) - x|^2 - |x - \tilde{x}|^2 - 2(u_{\tilde{\mu}}(\tilde{x}) - \tilde{x}) \cdot (x - \tilde{x}) \\ &\geq |u_{\tilde{\mu}}(\tilde{x}) - x|^2 - 3M|x - \tilde{x}|. \end{aligned} \quad (3.3)$$

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

$$|\tilde{y} - \tilde{x}| \leq 2W + |y - x|,$$

so

$$e^{-\gamma|\tilde{x}-\tilde{y}|} \geq e^{-2\gamma W} e^{-\gamma|y-x|}.$$

Thus we can write

$$\begin{aligned} &\iiint e^{-\gamma|\tilde{x}-\tilde{y}|} |u_{\tilde{\mu}}(\tilde{x}) - u_{\tilde{\mu}}(\tilde{y})| \, d\nu(\tilde{x} \mid x) \, d\mu(x) \, d\nu(\tilde{y} \mid y) \, d\mu(y) \\ &\geq e^{-2\gamma W} \iint e^{-\gamma|x-y|} \left( \iint |u_{\tilde{\mu}}(\tilde{x}) - u_{\tilde{\mu}}(\tilde{y})| \, d\nu(\tilde{x} \mid x) \, d\nu(\tilde{y} \mid y) \right) \, d\mu(x) \, d\mu(y) \\ &\geq e^{-2\gamma W} \iint e^{-\gamma|x-y|} |\bar{u}(x) - \bar{u}(y)| \, d\mu(x) \, d\mu(y), \end{aligned} \quad (3.4)$$

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

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

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

$$\inf J_\mu \leq J_\mu(\bar{u}) \leq e^{2\gamma W} (\inf J_{\tilde{\mu}} + 3MW) \leq \inf J_{\tilde{\mu}} + 3Me^{2\gamma W}W + (e^{2\gamma W} - 1)M^2, \quad (3.5)$$

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

$$|\inf J_{\tilde{\mu}} - \inf J_\mu| \leq 3Me^{2\gamma W}W + (e^{2\gamma W} - 1)M^2. \quad (3.6)$$

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

$$\begin{aligned} \int |\bar{u} - u_\mu|^2 \, d\mu &\leq 2(J_\mu(\bar{u}) - \inf J_\mu) \leq 2(\inf J_{\tilde{\mu}} - \inf J_\mu) + 6Me^{2\gamma W}W + 2(e^{2\gamma W} - 1)M^2 \\ &\leq 12Me^{2\gamma W}W + 4(e^{2\gamma W} - 1)M^2 \leq (M+1)^2 Q(\gamma\mathcal{W}_\infty(\mu, \tilde{\mu})) \end{aligned} \quad (3.7)$$

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

The remainder of the proof is very similar to the second half of the proof of [10, Proposition 5.3]. For each  $\varepsilon > 0$ , let  $\mu_\varepsilon$  be a measure on the ball  $B$ , absolutely continuous with respect to the Lebesgue measure, and such that

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

Since  $\mu_\varepsilon$  is absolutely continuous with respect to the Lebesgue measure, by [6, Theorems 5.5 and 3.2] there are maps  $T_\varepsilon$  and  $\tilde{T}_\varepsilon$  from  $\text{supp } \mu_\varepsilon$  to  $\text{supp } \mu$  and  $\text{supp } \tilde{\mu}$ , respectively, such that  $(\text{id} \times T_\varepsilon)_*(\mu_\varepsilon)$  is an  $\infty$ -optimal transport plan between  $\mu_\varepsilon$  and  $\mu$  and similarly  $(\text{id} \times \tilde{T}_\varepsilon)_*(\mu_\varepsilon)$  is an  $\infty$ -optimal transport plan between  $\mu_\varepsilon$  and  $\tilde{\mu}$ . We have

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

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

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

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

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

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

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

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

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

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  $\pi$  is an  $\infty$ -optimal transport plan. This follows by using (3.8) to note that

$$\text{ess sup}_{x \sim \mu_\varepsilon} |T_\varepsilon(x) - \tilde{T}_\varepsilon(x)| \leq \text{ess sup}_{x \sim \mu_\varepsilon} |T_\varepsilon(x) - x| + \text{ess sup}_{x \sim \mu_\varepsilon} |x - \tilde{T}_\varepsilon(x)| \leq \varepsilon + \mathcal{W}_\infty(\mu_\varepsilon, \tilde{\mu}),$$

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

#### 4. CONVERGENCE AS $\gamma \rightarrow \infty$

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



**Theorem 4.1.** *Assume that  $U = \text{supp } \mu$  is effectively star-shaped and has a Lipschitz boundary, and that the measure  $\mu$  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*

$$|\inf J_{\mu, \lambda, \infty} - \inf J_{\mu, \lambda, \gamma}| + \int |u_{\mu, \lambda, \infty} - u_{\mu, \lambda, \gamma}|^2 d\mu \leq C\gamma^{-1/3}. \quad (4.1)$$

*Proof.* Without loss of generality, assume that the point  $x_*$  in Definition 1.2 is the origin, and that the constant  $C_*$  appearing there is 1. We denote by  $\rho$  the density of  $\mu$  with respect to the Lebesgue measure. By [11, Theorem 5.4.1], we can and do extend  $\rho$  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  $\mu, \lambda$  fixed, and write  $u_\gamma = u_{\mu, \lambda, \gamma}$  and  $J_\gamma = J_{\mu, \lambda, \gamma}$ . The constant  $C$  may depend on  $\mu$  but not on  $\gamma$  or  $\lambda$ , and may change over the course of the argument. We let  $U_\varepsilon$  be the  $\varepsilon$ -enlargement of  $U$  as in Definition 1.2.

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

$$\tilde{u}_{\gamma, \varepsilon}(x) := u_\gamma((1 - \varepsilon)x),$$

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

$$u_{\gamma, \varepsilon}(x) := (\tilde{u}_{\gamma, \varepsilon} * \chi_\varepsilon)(x),$$

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

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

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

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

$$\begin{aligned} \int_{U_\varepsilon} |\tilde{u}_{\gamma, \varepsilon}(x) - x|^2 \rho(x) dx + \lambda \gamma^{d+1} \iint_{U_\varepsilon^2} e^{-\gamma|x-y|} |\tilde{u}_{\gamma, \varepsilon}(x) - \tilde{u}_{\gamma, \varepsilon}(y)| \rho(x) \rho(y) dx dy \\ \leq J_\gamma(u_\gamma) + C\varepsilon. \end{aligned} \quad (4.3)$$

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

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

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

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

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

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

$$J_\gamma(u_{\gamma,\varepsilon}) \leq J_\gamma(u_\gamma) + C\varepsilon. \quad (4.4)$$

Using (4.2), we can write

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

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

$$\int_U |u_{\gamma,\varepsilon}(x) - x|^2 d\mu(x) \leq \int_{U_\varepsilon} |\tilde{u}_{\gamma,\varepsilon}(x) - x|^2 \rho(x) dx + C\varepsilon. \quad (4.5)$$

We also have

$$\begin{aligned} \gamma^{d+1} \iint_{U^2} e^{-\gamma|x-y|} |u_{\gamma,\varepsilon}(x) - u_{\gamma,\varepsilon}(y)| \rho(x) \rho(y) dx dy \\ \leq \gamma^{d+1} \iint_{U^2} e^{-\gamma|x-y|} \left| \int_{\mathbf{R}^d} [\tilde{u}_{\gamma,\varepsilon}(x-z) - \tilde{u}_{\gamma,\varepsilon}(y-z)] \chi_\varepsilon(z) dz \right| \rho(x) \rho(y) dx dy \\ \leq \gamma^{d+1} \iint_{U^2} \int_{\mathbf{R}^d} e^{-\gamma|x-y|} |\tilde{u}_{\gamma,\varepsilon}(x) - \tilde{u}_{\gamma,\varepsilon}(y)| \chi_\varepsilon(z) \rho(x+z) \rho(y+z) dz dx dy \\ \leq \gamma^{d+1} \iint_{U_\varepsilon^2} e^{-\gamma|x-y|} |\tilde{u}_{\gamma,\varepsilon}(x) - \tilde{u}_{\gamma,\varepsilon}(y)| \left( \int_{\mathbf{R}^d} \chi_\varepsilon(z) \rho(x+z) \rho(y+z) dz \right) dx dy \\ \leq \gamma^{d+1} \iint_{U_\varepsilon^2} e^{-\gamma|x-y|} |\tilde{u}_{\gamma,\varepsilon}(x) - \tilde{u}_{\gamma,\varepsilon}(y)| \rho(x) \rho(y) dx dy + C\varepsilon, \end{aligned}$$

where in the last step we used (4.3), (2.5), and the fact that  $\rho$  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  $\varepsilon \in (0, 1]$ ,

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

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  $\|D^2 u_{\gamma,\varepsilon}\|_{L^\infty(\mu)} \leq C\varepsilon^{-2}$  to write

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

Since  $\rho$  is bounded and

$$\gamma^{d+1} \int_{\mathbf{R}^d} e^{-\gamma|x-y|} |x-y|^2 dy = \gamma^{-1} \int_{\mathbf{R}^d} e^{-|y|} |y|^2 dy, \quad (4.8)$$

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  $\rho$  is Lipschitz and  $\|Du_{\gamma,\varepsilon}\|_{L^\infty(\mu)} \leq C\varepsilon^{-1}$ , the difference between these two quantities is bounded by

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

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

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

Recalling (1.6), we see that the first term on the right-hand side above can be rewritten as

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

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

$$\gamma^{d+1} \int_U \int_{\mathbf{R}^d \setminus U} e^{-\gamma|x-y|} |x-y| dy dx \leq C\gamma^{-1}. \quad (4.10)$$

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,

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

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

*Step 4.* We show that

$$\int_{U_\varepsilon} |\tilde{u}_{\infty,\varepsilon}(x) - x|^2 \rho(x) dx + c\lambda \iint_{U_\varepsilon^2} \rho(x)^2 d|D\tilde{u}_{\infty,\varepsilon}|(x) \leq J_\infty(u_\infty) + C\varepsilon. \quad (4.11)$$

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

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

and from the fact that  $\rho$  is Lipschitz.

*Step 5.* We show that

$$J_\infty(u_{\infty,\varepsilon}) \leq J_\infty(u_\infty) + C\varepsilon. \quad (4.12)$$

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

$$\int_U |u_{\infty,\varepsilon}(x) - x|^2 d\mu(x) \leq \int_{U_\varepsilon} |\tilde{u}_{\infty,\varepsilon}(x) - x|^2 \rho(x) dx + C\varepsilon. \quad (4.13)$$

For the second term, we notice that by [2, Proposition 3.2], we have

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

and thus

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

where we used (4.11), (2.5), and the fact that  $\rho$  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

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

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

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

and estimate each of these integrals in turn. The second integral 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  $\rho$  is Lipschitz, the difference between these two quantities is bounded by

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

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

$$\begin{aligned} &\gamma^{d+1} \iint_{U^2} e^{-\gamma|x-y|} |Du_{\infty,\varepsilon}(x) \cdot (x-y)| \rho(x)^2 dx dy \\ &\leq \int_{\mathbf{R}^2} e^{-|y|} |y \cdot e_1| dy \int_U |Du_{\infty,\varepsilon}(x)| \rho(x)^2 dx = c \int_U |Du_{\infty,\varepsilon}(x)| \rho(x)^2 dx, \end{aligned}$$

where we recalled (1.6) in the last step. Thus we have

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

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

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

From this, we deduce that

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

and moreover that

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

By (2.4) and (4.17), we obtain

$$\int |u_{\gamma, \gamma^{-1/3}} - u_\infty|^2 d\mu \leq C\gamma^{-1/3}. \quad (4.18)$$

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

$$\int |u_{\gamma, \gamma^{-1/3}} - u_\gamma|^2 d\mu \leq C\gamma^{-1/3}. \quad (4.19)$$

Combining (4.16), (4.18), and (4.19) yields (4.1).  $\square$

## 5. PROPERTIES OF THE LIMITING FUNCTIONAL

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

**Proposition 5.1.** *Let  $\mu$  be a probability measure on  $\mathbf{R}^d$  satisfying the conditions of Theorem 1.1, so its support is the disjoint union of  $\overline{U}_1 \sqcup \dots \sqcup \overline{U}_L$ . There is a  $\lambda_c < \infty$  such that if  $\lambda \geq \lambda_c$ , then  $u_{\mu, \lambda, \infty}(x) = \text{cent}_\mu(U_\ell)$  for all  $x \in U_\ell$ ,  $\ell \in \{1, \dots, L\}$ .*

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

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

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

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

$$\psi|_{\partial U} \equiv 0. \quad (5.2)$$

We note that the problem (5.1)–(5.2) separates into  $dL$  problems, one for each  $j$  and  $\ell$ . Each problem can be solved by [5, Theorem 2.4] (which follows the approach introduced in [3, 4]). We have, for every  $v \in (L^2(U) \cap \text{BV}(U))^d$ ,

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

A minor variant of (2.1) takes the form

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

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

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

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  $\lambda_c = \|\psi\|_{L^\infty(U)}$ .  $\square$

## 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

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

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

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

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

In light of this statement, we define

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

Then (6.2) implies that

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

*Proof of Proposition 6.1.* Subtracting (1.2) from (6.1), we obtain

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

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

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

and similarly,

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

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

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

as claimed.  $\square$

## 7. PROOF OF THEOREM 1.1

In this section we prove Theorem 1.1. We first need a result from [12]. Recall the notation  $d'$  introduced in (1.3).

**Proposition 7.1.** *Let  $U \subseteq \mathbf{R}^d$  be a bounded, connected domain with Lipschitz boundary. Let  $\mu$  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  $\alpha \geq 1$ , there is a constant  $C < \infty$ , depending only on  $U$ ,  $\alpha$ , and  $\mu$ , such that the following holds. If  $(X_n)_{n \in \mathbf{N}}$  are independent random variables with law  $\mu$ , then for every integer  $N \geq 1$ ,*

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

*Proof.* For  $d \geq 2$ , this is a restatement of [12, Theorem 1.1]. For  $d = 1$ , the result can be obtained from the classical Kolmogorov-Smirnov quantitative version of the Glivenko-Cantelli theorem.  $\square$

Now we can prove Theorem 1.1. For a measure  $\mu$  on  $\mathbf{R}^d$  and a Borel set  $U$ , we denote by  $\mu \llcorner U$  the restriction of  $\mu$  to the set  $U$ .

*Proof of Theorem 1.1.* Recalling (6.3), it is clear that if  $\gamma$  is so large that

$$(d + 4/3)\gamma^{-1} \log \gamma \leq \min_{1 \leq \ell, \ell' \leq L} \text{dist}(U_\ell, U_{\ell'}), \quad (7.1)$$

then

$$\bar{u}_{\mu \llcorner U_\ell, \lambda, \gamma}(x) = \bar{u}_{\mu \llcorner U, \lambda, \gamma}(x), \quad \text{for all } x \in U_\ell, \quad (7.2)$$

and similarly

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

Also, we have by the definitions and Proposition 5.1 that there exists  $\lambda_c$  such that for every  $\lambda \geq \lambda_c$ ,

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

By (7.4) and Theorem 4.1, we have

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

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

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

Combining the last two displays, we see that

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

Using (6.4) again, this implies that

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

On the other hand, by Proposition 7.1, we have for each  $\ell$  that

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

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

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

Combining this with (7.5), we see that

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

Now summing over  $\ell$  and using (7.6) and the fact that the term inside the expectation on the left-hand side of (1.4) is bounded almost surely, we obtain (1.4).  $\square$

## 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] 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.
- [3] M. E. Bogovskii. 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.
- [4] M. E. Bogovskii. 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.
- [5] W. Borchers and H. Sohr. On the equations  $\text{rot } \mathbf{v} = \mathbf{g}$  and  $\text{div } \mathbf{u} = f$  with zero boundary conditions. *Hokkaido Math. J.*, 19(1):67–87, 1990.



- [6] T. Champion, L. De Pascale, and P. Juutinen. The  $\infty$ -Wasserstein distance: local solutions and existence of optimal transport maps. *SIAM J. Math. Anal.*, 40(1):1–20, 2008.
- [7] E. C. Chi and K. Lange. Splitting methods for convex clustering. *J. Comput. Graph. Statist.*, 24(4):994–1013, 2015.
- [8] E. C. Chi and S. Steinerberger. Recovering trees with convex clustering. *SIAM J. Math. Data Sci.*, 1(3):383–407, 2019.
- [9] J. Chiquet, P. Gutierrez, and G. Rigai. Fast tree inference with weighted fusion penalties. *J. Comput. Graph. Statist.*, 26(1):205–216, 2017.
- [10] A. Dunlap and J.-C. Mourrat. Sum-of-norms clustering does not separate nearby balls. Preprint, arXiv:2104.13753.
- [11] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.
- [12] N. García Trillos and D. Slepčev. On the rate of convergence of empirical measures in  $\infty$ -transportation distance. *Canad. J. Math.*, 67(6):1358–1383, 2015.
- [13] 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.
- [14] T. Jiang and S. Vavasis. On identifying clusters from sum-of-norms clustering computation. Preprint, arXiv:2006.11355.
- [15] 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.
- [16] 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.
- [17] A. Nellore and R. Ward. Recovery guarantees for exemplar-based clustering. *Inform. and Comput.*, 245:165–180, 2015.
- [18] 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.
- [19] 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.
- [20] P. Radchenko and G. Mukherjee. Convex clustering via  $l_1$  fusion penalization. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(5):1527–1546, 2017.
- [21] D. Sun, K.-C. Toh, and Y. Yuan. Convex clustering: Model, theoretical guarantee and efficient algorithm. *J. Mach. Learn. Res.*, 22:1–32, 2021.
- [22] K. M. Tan and D. Witten. Statistical properties of convex clustering. *Electron. J. Stat.*, 9(2):2324–2347, 2015.
- [23] 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