On the Reliability of Clustering Stability in the Large Sample Regime - Supplementary Material

Ohad Shamir[†] **and Naftali Tishby**[†‡]
† School of Computer Science and Engineering
‡ Interdisciplinary Center for Neural Computation
The Hebrew University
Jerusalem 91904, Israel
{ohadsh,tishby}@cs.huji.ac.il

A Exact Formulation of the Sufficient Conditions

In this section, we give a mathematically rigorous formulation of the sufficient conditions discussed in the main paper. For that we will need some additional notation.

First of all, it will be convenient to define a scaled version of our distance measure $d_{\mathcal{D}}(\mathtt{A}_k(S_1), \mathtt{A}_k(S_2))$ between clusterings. Formally, define the random variable

$$d^m_{\mathcal{D}}(\mathtt{A}_k(S_1), \mathtt{A}_k(S_2)) := \sqrt{m} d_{\mathcal{D}}(\mathtt{A}_k(S_1), \mathtt{A}_k(S_2)) = \sqrt{m} \Pr_{\mathbf{x}\sim\mathcal{D}}\left(\operatorname*{argmax}_i f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x}) \neq \operatorname*{argmax}_i f_{\hat{\boldsymbol{\theta}}',i}(\mathbf{x})\right),$$

where $\theta, \theta' \in \Theta$ are the solutions returned by $\mathtt{A}_k(S_1), \mathtt{A}_k(S_2)$, and S_1, S_2 are random samples, each of size m, drawn i.i.d from the underlying distribution $\mathcal{D}$. The scaling by the square root of the sample size will allow us to analyze the non-trivial asymptotic behavior of these distance measures, which without scaling simply converge to zero in probability as $m \to \infty$.

For some $\epsilon > 0$ and a set $S \subseteq \mathbb{R}^n$, let $B_\epsilon(S)$ be the ϵ-neighborhood of S, namely

$$B_\epsilon(S) := \left\{\mathbf{x} \in \mathcal{X} : \inf_{\mathbf{y}\in S} \|\mathbf{x} - \mathbf{y}\|_2 \leq \epsilon\right\}.$$

In this paper, when we talk about neighborhoods in general, we will always assume they are uniform (namely, contain an ϵ-neighborhood for some positive ϵ).

We will also need to define the following variant of $d^m_{\mathcal{D}}(\mathtt{A}_k(S_1), \mathtt{A}_k(S_2))$, where we restrict ourselves to the mass in some subset of $\mathbb{R}^n$. Formally, we define the restricted distance between two clusterings, with respect to a set $B \in \mathbb{R}^n$, as

$$d^m_{\mathcal{D}}(\mathtt{A}_k(S_1), \mathtt{A}_k(S_2), B) := \sqrt{m} \Pr_{\mathbf{x}\sim\mathcal{D}}\left(\operatorname*{argmax}_i f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x}) \neq \operatorname*{argmax}_i f_{\hat{\boldsymbol{\theta}}',i}(\mathbf{x}) \wedge \mathbf{x} \in B\right). \quad (1)$$

In particular, $d^m_{\mathcal{D}}(\mathtt{A}_k(S_1), \mathtt{A}_k(S_2), B_{r/\sqrt{m}}(\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j}))$ refers to the mass which switches clusters, and is also inside an $r/\sqrt{m}$-neighborhood of the limit cluster boundaries (where the boundaries are defined with respect to $f_{\boldsymbol{\theta}_0}(\cdot)$). Once again, when S_1, S_2 are random samples, we can think of it as a random variable with respect to drawing and clustering S_1, S_2.

Conditions. *The following conditions shall be assumed to hold:*

1. ***Consistency Condition:*** *$\hat{\boldsymbol{\theta}}$ converges in probability (over drawing and clustering a sample of size m, $m \to \infty$) to some $\boldsymbol{\theta}_0 \in \Theta$. Furthermore, the association of clusters to indices $\{1, \ldots, k\}$ is constant in some neighborhood of $\boldsymbol{\theta}_0$.*

2. ***Central Limit Condition:*** *$\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a multivariate zero mean Gaussian random variable Z.*

3. ***Regularity Conditions:***

   (a) $f_{\boldsymbol{\theta}}(\mathbf{x})$ ***is Sufficiently Smooth:*** *For any $\boldsymbol{\theta}$ in some neighborhood of $\boldsymbol{\theta}_0$, and any $\mathbf{x}$ in some neighborhood of the cluster boundaries $\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j}$, $f_{\boldsymbol{\theta}}(\mathbf{x})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$, with a non-zero first derivative and uniformly bounded second derivative for any $\mathbf{x}$. Both $f_{\boldsymbol{\theta}_0}(\mathbf{x})$ and $(\partial/\partial\boldsymbol{\theta}) f_{\boldsymbol{\theta}_0}(\mathbf{x})$ are twice differentiable with respect to any $\mathbf{x} \in \mathcal{X}$, with a uniformly bounded second derivative.*

   (b) ***Limit Cluster Boundaries are Reasonably Nice:*** *For any two clusters i, j, $F_{\boldsymbol{\theta}_0,i,j}$ is either empty, or a compact, non-self-intersecting, orientable $n-1$ dimensional hyper-surface in $\mathbb{R}^n$ with finite positive volume, a boundary (edge), and with a neighborhood contained in $\mathcal{X}$ in which the underlying density function $p(\cdot)$ is continuous. Moreover, the gradient $\nabla(f_{\boldsymbol{\theta}_0,i}(\cdot) - f_{\boldsymbol{\theta}_0,j}(\cdot))$ has positive magnitude everywhere on $F_{\boldsymbol{\theta}_0,i,j}$.*

   (c) ***Intersections of Cluster Boundaries are Relatively Negligible:*** *For any two distinct non-empty cluster boundaries $F_{\boldsymbol{\theta}_0,i,j}, F_{\boldsymbol{\theta}_0,i',j'}$, we have that*

   $$\frac{1}{\epsilon}\int_{B_\epsilon(F_{\boldsymbol{\theta}_0,i,j}\cup F_{\boldsymbol{\theta}_0,i',j'})\cap B_\delta(F_{\boldsymbol{\theta}_0,i,j})\cap B_\delta(F_{\boldsymbol{\theta}_0,i',j'})} 1 d\mathbf{x} \quad , \quad \frac{1}{\epsilon}\int_{B_\epsilon(\partial F_{\boldsymbol{\theta}_0,i,j})} 1 d\mathbf{x}$$

   converge to 0 as $\epsilon, \delta \to 0$ (in any manner), where $\partial F_{\boldsymbol{\theta}_0,i}$ is the edge of $F_{\boldsymbol{\theta}_0,i,j}$.

   (d) ***Minimal Parametric Stability:*** *It holds for some $\delta > 0$ that*

   $$\Pr\left(d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2)) \neq d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}}(\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j}))\right) = O(r^{-3-\delta}) + o(1),$$

   where $o(1) \to 0$ as $m \to \infty$. Namely, the mass of $\mathcal{D}$ which switches between clusters is with high probability inside thin strips around the limit cluster boundaries, and this high probability increases at least polynomially as the width of the strips increase (see below for a further discussion of this).

The regularity assumptions are relatively mild, and can usually be inferred based on the consistency and central limit conditions, as well as the the specific clustering framework that we are considering. For example, condition 3c and the assumptions on $F_{\boldsymbol{\theta}_0,i,j}$ in condition 3b are fulfilled in a clustering framework where the clusters are separated by hyperplanes. As to condition 3d, suppose our clustering framework is such that the cluster boundaries depend on $\hat{\boldsymbol{\theta}}$ in a smooth manner. Then the asymptotic normality of $\hat{\boldsymbol{\theta}}$, with variance $O(1/m)$, and the compactness of $\mathcal{X}$, will generally imply that the cluster boundaries obtained from clustering a sample are contained with high probability inside strips of width $O(1/\sqrt{m})$ around the limit cluster boundaries. More specifically, the asymptotic probability of this happening for strips of width $r/\sqrt{m}$ will be exponentially high in r, due to the asymptotic normality of $\hat{\boldsymbol{\theta}}$. As a result, the mass which switches between clusters, when we compare two independent clusterings, will be in those strips with probability exponentially high in r. Therefore, condition 3d will hold by a large margin, since only polynomially high probability is required there.

B Proofs - General Remarks

The proofs will use the additional notation and the sufficient conditions, as presented in Sec. A.

Throughout the proofs, we will sometimes use the stochastic order notation $O_p(\cdot)$ and $o_p(\cdot)$ (cf. [8]), defined as follows. Let $\{X_m\}$ and $\{Y_m\}$ be sequences of random vectors, defined on the same probability space. We write $X_m = O_p(Y_m)$ to mean that for each $\epsilon > 0$ there exists a real number M such that $\Pr(\|X_m\| \geq M\|Y_m\|) < \epsilon$ if m is large enough. We write $X_m = o_p(Y_m)$ to mean that $\Pr(\|X_m\| \geq \epsilon\|Y_m\|) \to 0$ for each $\epsilon > 0$. Notice that $\{Y_m\}$ may also be non-random. For example, $X_m = o_p(1)$ means that $X_m \to 0$ in probability. When we write for example $X_m = Y_m + o_p(1)$, we mean that $X_m - Y_m = o_p(1)$.

C Proof of Proposition 1

By condition 3a, $f_{\boldsymbol{\theta}}(\mathbf{x})$ has a first order Taylor expansion with respect to any $\hat{\boldsymbol{\theta}}$ close enough to $\boldsymbol{\theta}_0$, with a remainder term uniformly bounded for any $\mathbf{x}$:

$$f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) = f_{\boldsymbol{\theta}_0}(\mathbf{x}) + \left(\frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{x})\right)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|). \tag{2}$$

By the asymptotic normality assumption, $\sqrt{m}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(1)$, hence $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(1/\sqrt{m})$. Therefore, we get from Eq. (2) that

$$\sqrt{m}\left(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) - f_{\boldsymbol{\theta}_0}(\mathbf{x})\right) = \left(\frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{x})\right)^\top (\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)) + o_p(1), \tag{3}$$

where the remainder term $o_p(1)$ does not depend on $\mathbf{x}$. By regularity condition 3a and compactness of $\mathcal{X}$, $(\partial/\partial\boldsymbol{\theta}) f_{\boldsymbol{\theta}_0}(\cdot)$ is a uniformly bounded vector-valued function from $\mathcal{X}$ to the Euclidean space in which Θ resides. As a result, the mapping $\hat{\boldsymbol{\theta}} \mapsto ((\partial/\partial\boldsymbol{\theta}) f_{\boldsymbol{\theta}_0}(\cdot))^\top \hat{\boldsymbol{\theta}}$ is a mapping from Θ, with the metric induced by the Euclidean space in which it resides, to the space of all uniformly bounded $\mathbb{R}^k$-valued functions on $\mathcal{X}$. We can turn the latter space into a metric space by equipping it with the obvious extension of the supremum norm (namely, for any two functions $f(\cdot), g(\cdot)$, $\|f - g\| := \sup_{\mathbf{x}\in\mathcal{X}} \|f(\mathbf{x}) - g(\mathbf{x})\|_\infty$, where $\|\cdot\|_\infty$ is the infinity norm in Euclidean space). With this norm, the mapping above is a continuous mapping between two metric spaces. We also know that $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a multivariate Gaussian random variable Z. By the continuous mapping theorem [8] and Eq. (3), this implies that $\sqrt{m}(f_{\hat{\boldsymbol{\theta}}}(\cdot) - f_{\boldsymbol{\theta}_0}(\cdot))$ converges in distribution to a Gaussian process $G(\cdot)$, where

$$G(\cdot) := \left(\frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\cdot)\right)^\top Z. \tag{4}$$

D Proof of Thm. 1

D.1 A High Level Description of the Proof

The full proof of Thm. 1 is rather long and technical, mostly due to the many technical subtleties that need to be taken care of. Since these might obscure the main ideas, we present here separately a general overview of the proof, without the finer details.

The purpose of the stability estimator $\hat{\eta}^k_{m,q}$, scaled by $\sqrt{m}$, boils down to trying to assess the "expected" value of the random variable $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$: we estimate q instantiations of $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$, and take their average. Our goal is to show that this average, taking $m \to \infty$, is likely to be close to the value $\widehat{\mathrm{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$ as defined in the theorem. The most straightforward way to go about it is to prove that $\widehat{\mathrm{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$ actually equals $\lim_{m\to\infty} \mathbb{E} d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$, and then use some large deviation bound to prove that $\sqrt{m}\ \hat{\eta}^k_{m,q}$ is indeed close to it with high probability, if q is large enough. Unfortunately, computing $\lim_{m\to\infty} \mathbb{E} d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$ is problematic. The reason is that the convergence tools at our disposal deals with convergence in distribution of random variables, but convergence in distribution does not necessarily imply convergence of expectations. In other words, we can try and analyze the asymptotic distribution of $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$, but the expected value of this asymptotic distribution is not necessarily the same as $\lim_{m\to\infty} \mathbb{E} d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$. As a result, we will have to take a more indirect route.

Here is the basic idea: instead of analyzing the asymptotic expectation of $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$, we analyze the asymptotic expectation of a different random variable, $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B)$, which was formally defined in Eq. (1). Informally, recall that $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$ is the mass of the underlying distribution $\mathcal{D}$ which switches between clusters, when we draw and cluster two independent samples of size m. Then $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B)$ measures the subset of this mass, which lies inside some $B \subseteq \mathbb{R}^n$. In particular, following the notation of Sec. A, we will pick B to be $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}}(\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j}))$ for some $r > 0$. In words, this constitutes strips of width $r/\sqrt{m}$ around the limit cluster boundaries. Writing the above expression for B as $B_{r/\sqrt{m}}$, we have that if r be large enough, then $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}})$ is equal to $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$ with very high probability over drawing and clustering a pair of samples, for any large enough sample size m. Basically, this is because the fluctuations of the cluster boundaries, based on drawing and clustering a random sample of size m, cannot be too large, and therefore the mass which switches clusters is concentrated around the limit cluster boundaries, if m is large enough.

The advantage of the 'surrogate' random variable $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}})$ is that it is *bounded* for any finite r, unlike $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$. With bounded random variables, convergence in distribution does imply convergence of expectations, and as a result we are able to calculate $\lim_{m\to\infty} \mathbb{E} d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}})$ explicitly. This will turn out to be very close to

$\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$ as it appears in the theorem (in fact, we can make it arbitrarily close to $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$ by making r large enough). Using the fact that $d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}})$ and $d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$ are equal with very high probability, we show that conditioned on a highly probable event, $\sqrt{m}\ \hat{\eta}^k_{m,q}$ is an unbiased estimator of $d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}})$, based on q instantiations, for any sample size m. As a result, using large deviation bounds, we get that $\sqrt{m}\ \hat{\eta}^k_{m,q}$ is close to $d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}})$, with a high probability which does not depend on m. Therefore, as $m \to \infty$, $\sqrt{m}\ \hat{\eta}^k_{m,q}$ will be close to $\lim_{m\to\infty} \mathbb{E} d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}})$ with high probability. By picking r to scale appropriately with q, our theorem follows.

For convenience, the proof is divided into two parts: in Subsec. D.2, we calculate $\lim_{m\to\infty} \mathbb{E} d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}})$ explicitly, while Subsec. D.3 executes the general plan outlined above to prove our theorem.

A few more words are in order about the calculation of $\lim_{m\to\infty} \mathbb{E} d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}})$ in Subsec. D.2, since it is rather long and involved in itself. Our goal is to perform this calculation without going through an intermediate step of explicitly characterizing the distribution of $d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}})$. This is because the distribution might be highly dependent on the specific clustering framework, and thus it is unsuitable for the level of generality which we aim at (in other words, we do not wish to assume a specific clustering framework). The idea is as follows: recall that $d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}})$ is the mass of the underlying distribution $\mathcal{D}$, inside strips of width $r/\sqrt{m}$ around the limit cluster boundaries, which switches clusters when we draw and cluster two independent samples of size m. For any $\mathbf{x} \in \mathcal{X}$, let $A_\mathbf{x}$ be the event that $\mathbf{x}$ switched clusters. Then we can write $d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}})$, by Fubini's theorem, as:

$$\mathbb{E} d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}}) = \sqrt{m}\mathbb{E}\int_{B_{r/\sqrt{m}}} \mathbf{1}(A_\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \int_{B_{r/\sqrt{m}}} \sqrt{m}\Pr(A_\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{5}$$

The heart of the proof is Lemma D.5, which considers what happens to the integral above inside a single strip near one of the limit cluster boundaries $F_{\boldsymbol{\theta}_0,i,j}$. The main body of the proof then shows how the result of Lemma D.5 can be combined to give the asymptotic value of Eq. (5) when we take the integral over all of $B_{r/\sqrt{m}}$. The bottom line is that we can simply sum the contributions from each strip, because the intersection of these different strips is asymptotically negligible. All the other lemmas in Subsec. D.2 develop technical results needed for our proof.

Finally, let us describe the proof of Lemma D.5 in a bit more detail. It starts with an expression equivalent to the one in Eq. (5), and transforms it to an expression composed of a constant value, and a remainder term which converges to 0 as $m \to \infty$. The development can be divided into a number of steps. The first step is rewriting everything using the asymptotic Gaussian distribution of the cluster association function $f_{\hat{\boldsymbol{\theta}}}(\mathbf{x})$ for each $\mathbf{x}$, plus remainder terms (Eq. (13)). Since we are integrating over $\mathbf{x}$, special care is given to show that the convergence to the asymptotic distribution is uniform for all $\mathbf{x}$ in the domain of integration. The second step is to rewrite the integral (which is over a strip around the cluster boundary) as a double integral along the cluster boundary itself, and along a normal segment at any point on the cluster boundary (Eq. (14)). Since the strips become arbitrarily small as $m \to \infty$, the third step consists of rewriting everything in terms of a Taylor expansion around each point on the cluster boundary (Eq. (16), Eq. (17) and Eq. (18)). The fourth and final step is a change of variables, and after a few more manipulations we get the required result.

D.2 Part 1: Auxiliary Result

As described in the previous subsection, we will need an auxiliary result (Proposition D.1 below), characterizing the asymptotic expected value of $d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}}(\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j}))$.

Proposition D.1. *Let $r > 0$. Assuming the set of conditions from Sec. A holds, $\lim_{m\to\infty} \mathbb{E} d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2), B_{r/\sqrt{m}}(\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j}))$ is equal to*

$$2\left(\frac{1}{\sqrt{\pi}} - h(r)\right) \sum_{1\le i<j\le k} \int_{F_{\boldsymbol{\theta}_0,i,j}} \frac{p(\mathbf{x})\sqrt{Var(G_i(\mathbf{x}) - G_j(\mathbf{x}))}}{\|\nabla(f_{\boldsymbol{\theta}_0,i}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,j}(\mathbf{x}))\|} d\mathbf{x},$$

where $h(r) = O(\exp(-r^2))$.

To prove this result, we will need several technical lemmas.

Lemma D.1. *Let S be a hypersurface in $\mathbb{R}^n$ which fulfill the regularity conditions 3b and 3c for any $F_{\boldsymbol{\theta}_0,i,j}$, and let $g(\cdot)$ be a continuous real function on $\mathcal{X}$. Then for any $\epsilon > 0$,*

$$\frac{1}{\epsilon}\int_{B_\epsilon(S)} g(\mathbf{x})d\mathbf{x} = \frac{1}{\epsilon}\int_S\int_{-\epsilon}^{\epsilon} g(\mathbf{x}+y\mathbf{n}_\mathbf{x})dyd\mathbf{x} + o(1), \tag{6}$$

where $\mathbf{n}_\mathbf{x}$ is a unit normal vector to S at $\mathbf{x}$, and $o(1) \to 0$ as $\epsilon \to 0$.

Proof. Let $B'_\epsilon(S)$ be a strip around S, composed of all points which are on some normal to S and close enough to S:

$$B'_\epsilon(S) := \{\mathbf{y} \in \mathbb{R}^n \ : \ \exists \mathbf{x} \in S, \exists y \in [-\epsilon,\epsilon], \mathbf{y} = \mathbf{x} + y\mathbf{n}_\mathbf{x}\}.$$

Since S is orientable, then for small enough $\epsilon > 0$, $B'_\epsilon(S)$ is diffeomorphic to $S \times [-\epsilon,\epsilon]$. In particular, the map $\phi : S \times [-\epsilon,\epsilon] \mapsto B'_\epsilon(S)$, defined by

$$\phi(\mathbf{x}, y) = \mathbf{x} + y\mathbf{n}_\mathbf{x}$$

will be a diffeomorphism. Let $D\phi(\mathbf{x}, y)$ be the Jacobian of ϕ at the point $(\mathbf{x}, y) \in S \times [-\epsilon,\epsilon]$. Note that $D\phi(\mathbf{x}, 0) = 1$ for every $\mathbf{x} \in S$.

We now wish to claim that as $\epsilon \to 0$,

$$\frac{1}{\epsilon}\int_{B_\epsilon(S)} g(\mathbf{x})d\mathbf{x} = \frac{1}{\epsilon}\int_{B'_\epsilon(S)} g(\mathbf{x})d\mathbf{x} + o(1). \tag{7}$$

To see this, we begin by noting that $B'_\epsilon(S) \subseteq B_\epsilon(S)$. Moreover, any point in $B_\epsilon(S) \setminus B'_\epsilon(S)$ has the property that its projection to the closest point in S is not a normal to S, and thus must be ϵ-close to the edge of S. As a result of regularity condition 3c for S, and the fact that $g(\cdot)$ is continuous and hence uniformly bounded in the volume of integration, we get that the integration of $g(\cdot)$ over $B_\epsilon \setminus B'_\epsilon$ is asymptotically negligible (as $\epsilon \to 0$), and hence Eq. (7) is justified.

By the change of variables theorem from multivariate calculus, followed by Fubini's theorem, and using the fact that $D\phi$ is continuous and equals 1 on $S \times \{0\}$,

$$\begin{aligned}
\frac{1}{\epsilon}\int_{B'_\epsilon(S)} g(\mathbf{x})d\mathbf{x} &= \frac{1}{\epsilon}\int_{S\times[-\epsilon,\epsilon]} g(\mathbf{x}+y\mathbf{n}_\mathbf{x})D\phi(\mathbf{x},y)d\mathbf{x}dy \\
&= \frac{1}{\epsilon}\int_{-\epsilon}^{\epsilon}\left(\int_S g(\mathbf{x}+y\mathbf{n}_\mathbf{x})D\phi(\mathbf{x},y)d\mathbf{x}\right)dy \\
&= \frac{1}{\epsilon}\int_{-\epsilon}^{\epsilon}\left(\int_S g(\mathbf{x}+y\mathbf{n}_\mathbf{x})d\mathbf{x}\right)dy + o(1),
\end{aligned}$$

where $o(1) \to 0$ as $\epsilon \to 0$. Combining this with Eq. (7) yields the required result. □

Lemma D.2. *Let $(g_m : \mathcal{X} \mapsto \mathbb{R})_{m=1}^\infty$ be a sequence of integrable functions, such that $g_m(\mathbf{x}) \to 0$ uniformly for all $\mathbf{x}$ as $m \to \infty$. Then for any $i, j \in \{1,\ldots,k\}, i \neq j$,*

$$\int_{B_{r/\sqrt{m}}(F_{\boldsymbol{\theta}_0,i,j})} \sqrt{m}g_m(\mathbf{x})p(\mathbf{x})d\mathbf{x} \to 0$$

as $m \to \infty$

Proof. By the assumptions on $(g_m(\cdot))_{m=1}^\infty$, there exists a sequence of positive constants $(b_m)_{m=1}^\infty$, converging to 0, such that

$$\left|\int_{B_{r/\sqrt{m}}(F_{\boldsymbol{\theta}_0,i,j})} \sqrt{m}g_m(\mathbf{x})p(\mathbf{x})d\mathbf{x}\right| \leq b_m \int_{B_{r/\sqrt{m}}(F_{\boldsymbol{\theta}_0,i,j})} \sqrt{m}p(\mathbf{x})d\mathbf{x}.$$

For large enough m, $p(\mathbf{x})$ is bounded and continuous in the volume of integration. Applying Lemma D.1 with $\epsilon = r/\sqrt{m}$, we have that as $m \to \infty$,

$$\begin{aligned} b_m\sqrt{m}\int_{B_{r/\sqrt{m}}(F_{\boldsymbol{\theta}_0,i,j})} p(\mathbf{x})d\mathbf{x} &= b_m\sqrt{m}\int_{F_{\boldsymbol{\theta}_0,i,j}}\int_{-r/\sqrt{m}}^{r/\sqrt{m}} p(\mathbf{x}+yn_{\mathbf{x}})dyd\mathbf{x} + o(1) \\ \leq b_m\sqrt{m}\frac{C}{\sqrt{m}} + o(1) &= b_m C + o(1) \end{aligned}$$

for some constant C dependant on r and the upper bound on $p(\cdot)$. Since b_m converge to 0, we have that the expression in the lemma converges to 0 as well. □

Lemma D.3. *Let (X_m) and (Y_m) be a sequence of real random variables, such that X_m, Y_m are defined on the same probability space, and $X_m - Y_m$ converges to 0 in probability. Assume that Y_m converges in distribution to a continuous random variable Y. Then $|\Pr(X_m \leq c) - \Pr(Y_m \leq c)|$ converges to 0 uniformly for all $c \in \mathbb{R}$.*

Proof. We will use the following standard fact (see for example section 7.2 of [4]): for any two real random variables A, B, any $c \in \mathbb{R}$ and any $\epsilon > 0$, it holds that

$$\Pr(A \leq c) \leq \Pr(B \leq c+\epsilon) + \Pr(|A-B| > \epsilon).$$

From this inequality, it follows that for any $c \in \mathbb{R}$ and any $\epsilon > 0$,

$$\begin{aligned} |\Pr(X_m \leq c) - \Pr(Y_m \leq c)| \leq &\Big(\Pr(Y_m \leq c+\epsilon) - \Pr(Y_m \leq c)\Big) \\ &+ \Big(\Pr(Y_m \leq c) - \Pr(Y_m \leq c-\epsilon)\Big) + \Pr(|X_m - Y_m| \geq \epsilon). \end{aligned} \quad (8)$$

We claim that the r.h.s of Eq. (8) converges to 0 uniformly for all c, from which the lemma follows. To see this, we begin by noticing that $\Pr(|X_m - Y_m| \geq \epsilon)$ converges to 0 for any ϵ by definition of convergence in probability. Next, $\Pr(Y_m \leq c')$ converges to $\Pr(Y \leq c')$ uniformly for all $c' \in \mathbb{R}$, since Y is continuous (see section 1 of [6]). Moreover, since Y is a continuous random variable, we have that its distribution function is uniformly continuous, hence $\Pr(Y \leq c+\epsilon) - \Pr(Y \leq c)$ and $\Pr(Y \leq c) - \Pr(Y \leq c-\epsilon)$ converges to 0 as $\epsilon \to 0$, uniformly for all c. Therefore, by letting $m \to \infty$, and $\epsilon \to 0$ at an appropriate rate compared to m, we have that the l.h.s of Eq. (8) converges to 0 uniformly for all c. □

Lemma D.4. $\Pr(\langle \mathbf{a}, \sqrt{m}(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) - f_{\boldsymbol{\theta}_0}(\mathbf{x}))\rangle < b)$ *converges to* $\Pr(\langle a, G(\mathbf{x})\rangle < b)$ uniformly *for any $\mathbf{x} \in \mathcal{X}$, any $\mathbf{a} \neq 0$ in some bounded subset of $\mathbb{R}^k$, and any $b \in \mathbb{R}$.*

Proof. By Eq. (3),

$$\sqrt{m}\left(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) - f_{\boldsymbol{\theta}_0}(\mathbf{x})\right) = \left(\frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{x})\right)^\top (\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)) + o_p(1).$$

Where the remainder term does not depend on $\mathbf{x}$. Thus, for any $\mathbf{a}$ in a bounded subset of $\mathbb{R}^k$,

$$\left\langle \mathbf{a}, \sqrt{m}\left(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) - f_{\boldsymbol{\theta}_0}(\mathbf{x})\right)\right\rangle = \left\langle \mathbf{a}\left(\frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{x})\right)^\top, \sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right\rangle + o_p(1), \quad (9)$$

Where the convergence in probability is uniform for all bounded $\mathbf{a}$ and $\mathbf{x} \in \mathcal{X}$.

We now need to use a result which tells us when is a convergence in distribution uniform. Using thm. 4.2 in [6], we have that if a sequence of random vectors $(X_m)_{m=1}^\infty$ in Euclidean space converge to a random variable X in distribution, then $\Pr(\langle \mathbf{y}, X_m\rangle < b)$ converges to $\Pr(\langle \mathbf{y}, X\rangle < b)$ uniformly for any vector $\mathbf{y}$ and $b \in \mathbb{R}$. We note that a stronger result (Thm. 6 in [2]) apparently allows us to extend this to cases where X_m and X reside in some infinite dimensional, separable Hilbert space (for example, if Θ is a subset of an infinite dimensional reproducing kernel Hilbert space in kernel clustering). Therefore, recalling that $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a random normal vector Z, we have that uniformly for all $\mathbf{x}, \mathbf{a}, b$,

$$\begin{aligned}\Pr\left(\left\langle \mathbf{a}\left(\frac{\partial}{\partial\boldsymbol{\theta}}f_{\boldsymbol{\theta}_0}(\mathbf{x})\right)^\top, \sqrt{m}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right\rangle < b\right) &= \Pr\left(\left\langle \mathbf{a}\left(\frac{\partial}{\partial\boldsymbol{\theta}}f_{\boldsymbol{\theta}_0}(\mathbf{x})\right)^\top, Z\right\rangle < b\right) + o(1) \\ &= \Pr\left(\langle \mathbf{a}, G(\mathbf{x})\rangle < b\right) + o(1) \qquad (10)\end{aligned}$$

Here we think of $\mathbf{a}((\partial/\partial\boldsymbol{\theta})f_{\boldsymbol{\theta}_0}(\mathbf{x}))^\top$ as the vector $\mathbf{y}$ to which we apply the theorem. By regularity condition 3a, and assuming $\mathbf{a} \neq 0$, we have that $\left\langle a((\partial/\partial\boldsymbol{\theta})f_{\boldsymbol{\theta}_0}(\mathbf{x}))^\top, Z\right\rangle$ is a continuous real random variable for any $\mathbf{x}$, unless $Z = 0$ in which case the lemma is trivial. Therefore, the conditions of Lemma D.3 apply: the two sides of Eq. (9) give us two sequences of random variables which converge in probability to each other, and by Eq. (10) we have convergence in distribution of one of the sequences to a fixed continuous random variable. Therefore, using Lemma D.3, we have that

$$\Pr\left(\left\langle \mathbf{a}, \sqrt{m}\left(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) - f_{\boldsymbol{\theta}_0}(\mathbf{x})\right)\right\rangle < b\right) = \Pr\left(\left\langle \mathbf{a}\left(\frac{\partial}{\partial\boldsymbol{\theta}}f_{\boldsymbol{\theta}_0}(\mathbf{x})\right)^\top, \sqrt{m}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right\rangle < b\right) + o(1), \qquad (11)$$

where the convergence is uniform for any bounded $\mathbf{a} \neq 0$, b and $\mathbf{x} \in \mathcal{X}$.

Combining Eq. (10) and Eq. (11) gives us the required result. □

Lemma D.5. *Fix some two clusters i, j. Assuming the expression below is integrable, we have that*

$$\begin{aligned}&2\int_{B_{r/\sqrt{m}}(F_{\boldsymbol{\theta}_0,i,j})} \sqrt{m}\Pr(f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x}) - f_{\hat{\boldsymbol{\theta}},j}(\mathbf{x}) < 0)\Pr(f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x}) - f_{\hat{\boldsymbol{\theta}},j} > 0)p(\mathbf{x})d\mathbf{x} \\ &= 2\left(\frac{1}{\sqrt{\pi}} - h(r)\right)\int_{F_{\boldsymbol{\theta}_0,i,j}} \frac{p(\mathbf{x})\sqrt{\mathrm{Var}(G_i(\mathbf{x}) - G_j(\mathbf{x}))}}{\|\nabla(f_{\boldsymbol{\theta}_0,i}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,j}(\mathbf{x}))\|}d\mathbf{x} + o(1)\end{aligned}$$

where $o(1) \to 0$ as $m \to \infty$ and $h(r) = O(\exp(-r^2))$.

Proof. Define $\mathbf{a} \in \mathbb{R}^k$ as $a_i = 1$, $a_j = -1$, and 0 for any other entry. Applying Lemma D.4, with $\mathbf{a}$ as above, we have that uniformly for all $\mathbf{x}$ in some small enough neighborhood around $F_{\boldsymbol{\theta}_0,i,j}$:

$$\begin{aligned}&\Pr(f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x}) - f_{\hat{\boldsymbol{\theta}},j}(\mathbf{x}) < 0) \\ &= \Pr\left(\sqrt{m}(f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,i}(\mathbf{x})) - \sqrt{m}(f_{\hat{\boldsymbol{\theta}},j}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,j}(\mathbf{x})) < \sqrt{m}(f_{\boldsymbol{\theta}_0,j}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,i}(\mathbf{x}))\right) \\ &= \Pr(G_i(\mathbf{x}) - G_j(\mathbf{x}) < \sqrt{m}(f_{\boldsymbol{\theta}_0,j}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,i}(\mathbf{x}))) + o(1).\end{aligned}$$

where $o(1)$ converges uniformly to 0 as $m \to \infty$.

Since $G_i(\mathbf{x}) - G_j(\mathbf{x})$ has a zero mean normal distribution, we can rewrite the above (if $\mathrm{Var}(G_i(\mathbf{x}) - G_j(\mathbf{x})) > 0$) as

$$\begin{aligned}&\Pr\left(\frac{G_i(\mathbf{x}) - G_j(\mathbf{x})}{\sqrt{\mathrm{Var}(G_i(\mathbf{x}) - G_j(\mathbf{x}))}} < \frac{\sqrt{m}(f_{\boldsymbol{\theta}_0,j}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,i}(\mathbf{x}))}{\sqrt{\mathrm{Var}(G_i(\mathbf{x}) - G_j(\mathbf{x}))}}\right) + o(1) \\ &= \Phi\left(\frac{\sqrt{m}(f_{\boldsymbol{\theta}_0,j}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,i}(\mathbf{x}))}{\sqrt{\mathrm{Var}(G_i(\mathbf{x}) - G_j(\mathbf{x}))}}\right) + o(1), \qquad (12)\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution function. Notice that by some abuse of notation, the expression is also valid in the case where $\mathrm{Var}(G_i(\mathbf{x}) - G_j(\mathbf{x})) = 0$. In that case, $G_i(\mathbf{x}) - G_j(\mathbf{x})$ is equal to 0 with probability 1, and thus $\Pr(G_i(\mathbf{x}) - G_j(\mathbf{x}) < \sqrt{m}(f_{\boldsymbol{\theta}_0,j}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,i}(\mathbf{x})))$ is 1 if $f_{\boldsymbol{\theta}_0,j}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,i}(\mathbf{x})) \geq 0$ and 0 if $f_{\boldsymbol{\theta}_0,j}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,i}(\mathbf{x})) < 0$. This is equal to Eq. (12) if we are willing to assume that $\Phi(\infty) = 1, \Phi(0/0) = 1, \Phi(-\infty) = 0$.

Therefore, we can rewrite the l.h.s of the equation in the lemma statement as

$$2\int_{B_{r/\sqrt{m}}(F_{\boldsymbol{\theta}_0,i,j})} \sqrt{m}\Phi\left(\frac{\sqrt{m}(f_{\boldsymbol{\theta}_0,i}(\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x}))}{\sqrt{\text{Var}(G_i(\mathbf{x})-G_j(\mathbf{x}))}}\right) \left(1-\Phi\left(\frac{\sqrt{m}(f_{\boldsymbol{\theta}_0,i}(\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x}))}{\sqrt{\text{Var}(G_i(\mathbf{x})-G_j(\mathbf{x}))}}\right)\right)+\sqrt{m}o(1)p(\mathbf{x})d\mathbf{x}.$$

The integration of the remainder term can be rewritten as $o(1)$ by Lemma D.2, and we get that the expression can be rewritten as:

$$2\int_{B_{r/\sqrt{m}}(F_{\boldsymbol{\theta}_0,i,j})} \sqrt{m}\Phi\left(\frac{\sqrt{m}(f_{\boldsymbol{\theta}_0,i}(\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x}))}{\sqrt{\text{Var}(G_i(\mathbf{x})-G_j(\mathbf{x}))}}\right) \left(1-\Phi\left(\frac{\sqrt{m}(f_{\boldsymbol{\theta}_0,i}(\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x}))}{\sqrt{\text{Var}(G_i(\mathbf{x})-G_j(\mathbf{x}))}}\right)\right)p(\mathbf{x})d\mathbf{x}+o(1). \quad (13)$$

One can verify that the expression inside the integral is a continuous function of $\mathbf{x}$, by the regularity conditions and the expression for $G(\cdot)$ as proven in Sec. C (namely Eq. (4)). We can therefore apply Lemma D.1, and again take all the remainder terms outside of the integral by Lemma D.2, to get that the above can be rewritten as

$$2\int_{F_{\boldsymbol{\theta}_0,i,j}}\int_{-r/\sqrt{m}}^{r/\sqrt{m}} \sqrt{m}\Phi\left(\frac{\sqrt{m}(f_{\boldsymbol{\theta}_0,i}(\mathbf{x}+y\mathbf{n}_\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x}+y\mathbf{n}_\mathbf{x}))}{\sqrt{\text{Var}(G_i(\mathbf{x}+y\mathbf{n}_\mathbf{x})-G_j(\mathbf{x}+y\mathbf{n}_\mathbf{x}))}}\right) \left(1-\Phi\left(\frac{\sqrt{m}(f_{\boldsymbol{\theta}_0,i}(\mathbf{x}+y\mathbf{n}_\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x}+y\mathbf{n}_\mathbf{x}))}{\sqrt{\text{Var}(G_i(\mathbf{x}+y\mathbf{n}_\mathbf{x})-G_j(\mathbf{x}+y\mathbf{n}_\mathbf{x}))}}\right)\right)p(\mathbf{x})dyd\mathbf{x}+o(1), \quad (14)$$

where $\mathbf{n}_\mathbf{x}$ is a unit normal to $F_{\boldsymbol{\theta}_0,i,j}$ at $\mathbf{x}$.

Inspecting Eq. (14), we see that y ranges over an arbitrarily small domain as $m\to\infty$. This suggests that we can rewrite the above using Taylor expansions, which is what we shall do next.

Let us assume for a minute that $\text{Var}(G_i(\mathbf{x})-G_j(\mathbf{x}))>0$ for some point $\mathbf{x}\in F_{\boldsymbol{\theta}_0,i,j}$. One can verify that by the regularity conditions and the expression for $G(\cdot)$ in Eq. (4), the expression

$$\frac{f_{\boldsymbol{\theta}_0,i}(\cdot)-f_{\boldsymbol{\theta}_0,j}(\cdot)}{\sqrt{\text{Var}(G_i(\cdot)-G_j(\cdot))}} \quad (15)$$

is twice differentiable, with a uniformly bounded second derivative. Therefore, we can rewrite the expression in Eq. (15) as its first-order Taylor expansion around each $\mathbf{x}\in F_{\boldsymbol{\theta}_0,i,j}$, plus a remainder term which is uniform for all $\mathbf{x}$:

$$\frac{f_{\boldsymbol{\theta}_0,i}(\mathbf{x}+y\mathbf{n}_\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x}+y\mathbf{n}_\mathbf{x})}{\sqrt{\text{Var}(G_i(\mathbf{x}+y\mathbf{n}_\mathbf{x})-G_j(\mathbf{x}+y\mathbf{n}_\mathbf{x}))}} = \frac{f_{\boldsymbol{\theta}_0,i}(\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x})}{\sqrt{\text{Var}(G_i(\mathbf{x})-G_j(\mathbf{x}))}}+\nabla\left(\frac{f_{\boldsymbol{\theta}_0,i}(\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x})}{\sqrt{\text{Var}(G_i(\mathbf{x})-G_j(\mathbf{x}))}}\right)y\mathbf{n}_\mathbf{x}+O(y^2).$$

Since $f_{\boldsymbol{\theta}_0,i}(\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x})=0$ for any $\mathbf{x}\in F_{\boldsymbol{\theta}_0,i,j}$, the expression reduces after a simple calculation to

$$\frac{\nabla(f_{\boldsymbol{\theta}_0,i}(\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x}))}{\sqrt{\text{Var}(G_i(\mathbf{x})-G_j(\mathbf{x}))}}y\mathbf{n}_\mathbf{x}+O(y^2).$$

Notice that $\nabla(f_{\boldsymbol{\theta}_0,i}(\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x}))$ (the gradient of $f_{\boldsymbol{\theta}_0,i}(\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x})$) has the same direction as $\mathbf{n}_\mathbf{x}$ (the normal to the cluster boundary). Therefore, the expression above can be rewritten, up to a sign, as

$$y\left\|\frac{\nabla(f_{\boldsymbol{\theta}_0,i}(\mathbf{x})-f_{\boldsymbol{\theta}_0,j}(\mathbf{x}))}{\sqrt{\text{Var}(G_i(\mathbf{x})-G_j(\mathbf{x}))}}\right\|+O(y^2).$$

As a result, denoting $\mathbf{s}(\mathbf{x}) := \nabla(f_{\boldsymbol{\theta}_0,i}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,j}(\mathbf{x}))/\sqrt{\text{Var}(G_i(\mathbf{x}) - G_j(\mathbf{x}))}$, we have that

$$\Phi\left(\frac{\sqrt{m}(f_{\boldsymbol{\theta}_0,i}(\mathbf{x}+y\mathbf{n}_\mathbf{x}) - f_{\boldsymbol{\theta}_0,j}(\mathbf{x}+y\mathbf{n}_\mathbf{x}))}{\sqrt{\text{Var}(G_i(\mathbf{x}+y\mathbf{n}_\mathbf{x}) - G_j(\mathbf{x}+y\mathbf{n}_\mathbf{x}))}}\right)\left(1-\Phi\left(\frac{\sqrt{m}(f_{\boldsymbol{\theta}_0,i}(\mathbf{x}+y\mathbf{n}_\mathbf{x}) - f_{\boldsymbol{\theta}_0,j}(\mathbf{x}+y\mathbf{n}_\mathbf{x}))}{\sqrt{\text{Var}(G_i(\mathbf{x}+y\mathbf{n}_\mathbf{x}) - G_j(\mathbf{x}+y\mathbf{n}_\mathbf{x}))}}\right)\right) \tag{16}$$

$$= \Phi\Big(\sqrt{m}\big(\|\mathbf{s}(\mathbf{x})\|y + O(y^2)\big)\Big)\Big(1-\Phi\Big(\sqrt{m}\big(\|\mathbf{s}(\mathbf{x})\|y + O(y^2)\big)\Big)\Big)$$

$$= \Phi\Big(\sqrt{m}\big(\|\mathbf{s}(\mathbf{x})\|y\big)\Big)\Big(1-\Phi\Big(\sqrt{m}\big(\|\mathbf{s}(\mathbf{x})\|y\big)\Big)\Big) + O(\sqrt{m}y^2). \tag{17}$$

In the preceding development, we have assumed that $\text{Var}(G_i(\mathbf{x}) - G_j(\mathbf{x})) > 0$. However, notice that the expressions in Eq. (16) and Eq. (17), without the remainder term, are both equal (to zero) even if $\text{Var}(G_i(\mathbf{x}) - G_j(\mathbf{x})) = 0$ (with our previous abuse of notation that $\Phi(-\infty) = 0, \Phi(\infty) = 1$). Moreover, since y takes values in $[-r/\sqrt{m}, r/\sqrt{m}]$, the remainder term $O(\sqrt{m}y^2)$ is at most $O(\sqrt{m}r/m) = O(r/\sqrt{m})$, so it can be rewritten as $o(1)$ which converges to 0 as $m \to \infty$.

In conclusion, and again using Lemma D.2 to take the remainder terms outside of the integral, we can rewrite Eq. (14) as

$$2\int_{F_{\boldsymbol{\theta}_0,i,j}}\int_{-r/\sqrt{m}}^{r/\sqrt{m}} \sqrt{m}\Phi\left(\sqrt{m}\|\mathbf{s}(\mathbf{x})\|y\right)\left(1-\Phi\left(\sqrt{m}\|\mathbf{s}(\mathbf{x})\|y\right)\right)p(\mathbf{x})dyd\mathbf{x} + o(1). \tag{18}$$

We now perform a change of variables, letting $z_\mathbf{x} = \sqrt{m}\|\mathbf{s}(\mathbf{x})\|y$ in the inner integral, and get

$$2\int_{F_{\boldsymbol{\theta}_0,i,j}}\int_{-r\|\mathbf{s}(\mathbf{x})\|}^{r\|\mathbf{s}(\mathbf{x})\|} \frac{1}{\|\mathbf{s}(\mathbf{x})\|}\Phi\left(z_\mathbf{x}\right)\left(1-\Phi\left(z_\mathbf{x}\right)\right)p(\mathbf{x})dz_\mathbf{x}d\mathbf{x} + o(1),$$

which is equal by the mean value theorem to

$$2\left(\int_{F_{\boldsymbol{\theta}_0,i,j}}\frac{p(\mathbf{x})}{\|\mathbf{s}(\mathbf{x})\|}d\mathbf{x}\right)\left(\int_{-r\|\mathbf{s}(\mathbf{x}_0)\|}^{r\|\mathbf{s}(\mathbf{x}_0)\|}\Phi\left(z_{\mathbf{x}_0}\right)\left(1-\Phi\left(z_{\mathbf{x}_0}\right)\right)dz_{\mathbf{x}_0}\right) + o(1) \tag{19}$$

for some $\mathbf{x}_0 \in F_{\boldsymbol{\theta}_0,i,j}$.

By regularity condition 3b, it can be verified that $\|\mathbf{s}(\mathbf{x})\|$ is positive or infinite for any $\mathbf{x} \in F_{\boldsymbol{\theta}_0,i,j}$. As a result, as $r \to \infty$, we have that

$$\int_{-r\|\mathbf{s}(\mathbf{x}_0)\|}^{r\|\mathbf{s}(\mathbf{x}_0)\|}\Phi\left(z_{\mathbf{x}_0}\right)\left(1-\Phi\left(z_{\mathbf{x}_0}\right)\right)dz_{\mathbf{x}_0} \longrightarrow \int_{-\infty}^{\infty}\Phi(z_{\mathbf{x}_0})(1-\Phi(z_{\mathbf{x}_0}))dz_{\mathbf{x}_0} = \frac{1}{\sqrt{\pi}}.$$

and the convergence to $1/\sqrt{\pi}$ is at a rate of $O(\exp(-r^2))$. Combining this with Eq. (19) gives us the required result.

□

Proof of Proposition D.1. We can now turn to prove Proposition D.1 itself. For any $\mathbf{x} \in \mathcal{X}$, let $A_\mathbf{x}$ be the event (over drawing and clustering a sample pair) that $\mathbf{x}$ switched clusters. For any $F_{\boldsymbol{\theta}_0,i,j}$ and sample size m, define $F^m_{\boldsymbol{\theta}_0,i,j}$ to be the subset of $F_{\boldsymbol{\theta}_0,i,j}$, which is at a distance of at least $m^{-1/4}$ from any other cluster boundary (with respect to $\boldsymbol{\theta}_0$). Formally,

$$F^m_{\boldsymbol{\theta}_0,i,j} := \left\{\mathbf{x} \in F_{\boldsymbol{\theta}_0,i,j} : \forall\left(\{i',j'\} \neq \{i,j\}, F_{\boldsymbol{\theta}_0,i',j'} \neq \emptyset\right), \inf_{\mathbf{y}\in F_{\boldsymbol{\theta}_0,i',j'}}\|\mathbf{x}-\mathbf{y}\| \geq m^{-1/4}\right\}.$$

Letting S_1, S_2 be two independent samples of size m, we have by Fubini's theorem that

$$\begin{aligned}
&\mathbb{E}d^m_{\mathcal{D}}(\mathtt{A}_{\mathtt{k}}(S_1), \mathtt{A}_{\mathtt{k}}(S_2), B_{r/\sqrt{m}}(\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j})) \\
&= \sqrt{m}\mathbb{E}_{S_1,S_2} \int_{B_{r/\sqrt{m}}(\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j})} \mathbf{1}(A_{\mathbf{x}}) p(\mathbf{x}) d\mathbf{x} = \int_{B_{r/\sqrt{m}}(\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j})} \sqrt{m}\Pr(A_{\mathbf{x}}) p(\mathbf{x}) d\mathbf{x} \\
&= \int_{B_{r/\sqrt{m}}(\cup_{i,j} F^m_{\boldsymbol{\theta}_0,i,j})} \sqrt{m}\Pr(A_{\mathbf{x}}) p(\mathbf{x}) d\mathbf{x} + \int_{B_{r/\sqrt{m}}(\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j} \setminus F^m_{\boldsymbol{\theta}_0,i,j})} \sqrt{m}\Pr(A_{\mathbf{x}}) p(\mathbf{x}) d\mathbf{x}.
\end{aligned}$$

As to the first integral, notice that each point in $F^m_{\boldsymbol{\theta}_0,i,j}$ is separated from any point in any other $F^m_{\boldsymbol{\theta}_0,i',j'}$ by a distance of at least $2m^{-1/4}$. Therefore, for large enough m, $B_{r/\sqrt{m}}(F^m_{\boldsymbol{\theta}_0,i,j})$ are disjoint for each i, j, and we can rewrite the above as:

$$\sum_{1\le i<j\le k} \int_{B_{r/\sqrt{m}}(F^m_{\boldsymbol{\theta}_0,i,j})} \sqrt{m}\Pr(A_{\mathbf{x}}) p(\mathbf{x}) d\mathbf{x} + \int_{B_{r/\sqrt{m}}(\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j} \setminus F^m_{\boldsymbol{\theta}_0,i,j})} \sqrt{m}\Pr(A_{\mathbf{x}}) p(\mathbf{x}) d\mathbf{x}.$$

As to the second integral, notice that the integration is over points which are at a distance of at most $r/\sqrt{m}$ from some $F_{\boldsymbol{\theta}_0,i,j}$, and also at a distance of at most $m^{-1/4}$ from some other $F_{\boldsymbol{\theta}_0,i',j'}$. By regularity condition 3c, and the fact that $m^{-1/4} \to 0$, it follows that this integral converges to 0 as $m \to \infty$, and we can rewrite the above as:

$$\sum_{1\le i<j\le k} \int_{B_{r/\sqrt{m}}(F^m_{\boldsymbol{\theta}_0,i,j})} \sqrt{m}\Pr(A_{\mathbf{x}}) p(\mathbf{x}) d\mathbf{x} + o(1) \tag{20}$$

If there were only two clusters i, j, then

$$\Pr(A_{\mathbf{x}}) = 2\Pr(f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x}) - f_{\hat{\boldsymbol{\theta}},j}(\mathbf{x}) < 0)\Pr(f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x}) - f_{\hat{\boldsymbol{\theta}},j} > 0).$$

This is simply by definition of $A_{\mathbf{x}}$: the probability that under one clustering, based on a random sample, $\mathbf{x}$ is more associated with cluster i, and that under a second clustering, based on another independent random sample, $\mathbf{x}$ is more associated with cluster j.

In general, we will have more than two clusters. However, notice that any point $\mathbf{x}$ in $B_{r/\sqrt{m}}(F^m_{\boldsymbol{\theta}_0,i,j})$ (for some i, j) is much closer to $F_{\boldsymbol{\theta}_0,i,j}$ than to any other cluster boundary. This is because its distance to $F_{\boldsymbol{\theta}_0,i,j}$ is on the order of $1/\sqrt{m}$, while its distance to any other boundary is on the order of $m^{-1/4}$. Therefore, if $\mathbf{x}$ does switch clusters, then it is highly likely to switch between cluster i and cluster j. Formally, by regularity condition 3d (which ensure that the cluster boundaries experience at most $O(1/\sqrt{m})$ fluctuations), we have that uniformly for any $\mathbf{x}$,

$$\Pr(A_{\mathbf{x}}) = 2\Pr(f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x}) - f_{\hat{\boldsymbol{\theta}},j}(\mathbf{x}) < 0)\Pr(f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x}) - f_{\hat{\boldsymbol{\theta}},j} > 0) + o(1),$$

where $o(1)$ converges to 0 as $m \to \infty$.

Substituting this back to Eq. (20), using Lemma D.2 to take the remainder term outside the integral, and using the regularity condition 3c in the reverse direction to transform integrals over $F^m_{\boldsymbol{\theta}_0,i,j}$ back into $F_{\boldsymbol{\theta}_0,i,j}$ with asymptotically negligible remainder terms, we get that the quantity we are interested in can be written as

$$\sum_{1\le i<j\le k} 2\int_{B_{r/\sqrt{m}}(F_{\boldsymbol{\theta}_0,i,j})} \sqrt{m}\Pr(f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x}) - f_{\hat{\boldsymbol{\theta}},j}(\mathbf{x}) < 0)\Pr(f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x}) - f_{\hat{\boldsymbol{\theta}},j} > 0) p(\mathbf{x}) d\mathbf{x} + o(1).$$

Now we can apply Lemma D.5 to each summand, and get the required result.

D.3 Part 2: Proof of Thm. 1

For notational convenience, we will denote

$$d^m_{\mathcal{D}}(r) := d^m_{\mathcal{D}}(\mathtt{A}_{\mathtt{k}}(S_1), \mathtt{A}_{\mathtt{k}}(S_2), B_{r/\sqrt{m}}(\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j}))$$

whenever the omitted terms are obvious from context. If $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D}) = 0$, the proof of the theorem is straightforward. In this special case, by definition of $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$ in Thm. 1 and Proposition D.1, we have that $d^m_{\mathcal{D}}(r)$ converges in probability to 0 for any r. By regularity condition 3d, for any fixed q, $\frac{1}{q}\sum_{i=1}^{q} d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S^1_i), \mathtt{A}_\mathtt{k}(S^2_i))$ converges in probability to 0 (because $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S^1_i), \mathtt{A}_\mathtt{k}(S^2_i)) = d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S^1_i), \mathtt{A}_\mathtt{k}(S^2_i), B_{r/\sqrt{m}}(\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j}))$ with arbitrarily high probability as r increases). Therefore, $\sqrt{m}\ \hat{\eta}^k_{m,q}$, which is a plug-in estimator of the expected value of $\frac{1}{q}\sum_{i=1}^{q} d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S^1_i), \mathtt{A}_\mathtt{k}(S^2_i))$, converges in probability to 0 for any fixed q as $m \to \infty$, and the theorem follows for this special case. Therefore, we will assume from now on that $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D}) > 0$.

We need the following variant of Hoeffding's bound, adapted to conditional probabilities.

Lemma D.6. *Fix some $r > 0$. Let $X_1, \ldots, X_q$ be real, nonnegative, independent and identically distributed random variables, such that $\Pr(X_1 \in [0, r]) > 0$. For any X_i, let Y_i be a random variable on the same probability space, such that $\Pr(Y_i = X_i | X_i \in [0, r]) = 1$. Then for any $\nu > 0$,*

$$\Pr\left(\left|\frac{1}{q}\sum_{i=1}^{q} X_i - \mathbb{E}[Y_1 | X_1 \in [0, r]]\right| \geq \nu \;\middle|\; \forall i, X_i \in [0, r]\right) \leq 2\exp\left(-\frac{2q\nu^2}{r^2}\right).$$

Proof. Define an auxiliary set of random variables $Z_1, \ldots, Z_q$, such that $\Pr(Z_i \leq a) = \Pr(X_i \leq a | X_i \in [0, r])$ for any i, a. In words, X_i and Z_i have the same distribution conditioned on the event $X_i \in [0, r]$. Also, we have that Y_i has the same distribution conditioned on $X_i \in [0, r]$. Therefore, $\mathbb{E}[Y_1 | X_1 \in [0, r]] = \mathbb{E}[X_1 | X_1 \in [0, r]]$, and as a result $\mathbb{E}[Y_1 | X_1 \in [0, r]] = \mathbb{E}[Z_1]$. Therefore, the probability in the lemma above can be written as

$$\Pr\left(\left|\frac{1}{q}\sum_{i=1}^{q} Z_i - \mathbb{E}[Z_i]\right| \geq \nu\right),$$

where Z_i are bounded in $[0, r]$ with probability 1. Applying the regular Hoeffding's bound gives us the required result. □

We now turn to the proof of the theorem. Let A^m_r be the event that for all subsample pairs $\{S^1_i, S^2_i\}$, $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S^1_i), \mathtt{A}_\mathtt{k}(S^2_i), B_{r/\sqrt{m}(\cup_{i,j} F_{\boldsymbol{\theta}_0,i,j})}) = d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S^1_i), \mathtt{A}_\mathtt{k}(S^2_i))$. Namely, this is the event that for all subsample pairs, the mass which switches clusters when we compare the two resulting clusterings is always in an $r/\sqrt{m}$-neighborhood of the limit cluster boundaries.

Since $p(\cdot)$ is bounded, we have that $d^m_{\mathcal{D}}(r)$ is deterministically bounded by $O(r)$, with implicit constants depending only on $\mathcal{D}$ and $\boldsymbol{\theta}_0$. Using the law of total expectation, this implies that

$$\begin{aligned}
&\left|\mathbb{E}[d^m_{\mathcal{D}}(r)] - \mathbb{E}[d^m_{\mathcal{D}}(r) | A^m_r]\right| \\
&= \left|\Pr(A^m_r)\mathbb{E}[d^m_{\mathcal{D}}(r) | A^m_r] + (1 - \Pr(A^m_r))\mathbb{E}[d^m_{\mathcal{D}}(r) | \neg A^m_r] - \mathbb{E}[d^m_{\mathcal{D}}(r) | A^m_r]\right| \\
&= \left|\left(1 - \Pr(A^m_r)\right)\left(\mathbb{E}[d^m_{\mathcal{D}}(r) | \neg A^m_r] - \mathbb{E}[d^m_{\mathcal{D}}(r) | A^m_r]\right)\right| \\
&\leq (1 - \Pr(A^m_r))O(r). \qquad (21)
\end{aligned}$$

For any two events A, B, we have by the law of total probability that $\Pr(A) = \Pr(B)\Pr(A|B) + \Pr(B^c)\Pr(A|B^c)$. From this it follows that $\Pr(A) \leq \Pr(B) + \Pr(A|B^c)$. As a result, for any

$\epsilon > 0$,

$$\begin{aligned}
&\Pr\left(\left|\sqrt{m}\,\hat{\eta}^k_{m,q} - \widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})\right| > \epsilon\right) \\
&\leq \Pr\left(\left|\frac{1}{q}\sum_{i=1}^{q} d^m_{\mathcal{D}}(\mathtt{A_k}(S^1_i), \mathtt{A_k}(S^2_i)) - \widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})\right| > \frac{\epsilon}{2}\right) \\
&+ \Pr\left(\left[\left|\sqrt{m}\,\hat{\eta}^k_{m,q} - \widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})\right| > \epsilon\right] \,\middle|\, \left[\left|\frac{1}{q}\sum_{i=1}^{q} d^m_{\mathcal{D}}(\mathtt{A_k}(S^1_i), \mathtt{A_k}(S^2_i)) - \widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})\right| \leq \frac{\epsilon}{2}\right]\right).
\end{aligned} \tag{22}$$

We will assume w.l.o.g that $\epsilon/2 < \widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})$. Otherwise, we can upper bound $\Pr\left(\left|\sqrt{m}\,\hat{\eta}^k_{m,q} - \widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})\right| > \epsilon\right)$ in the equation above by replacing ϵ with some smaller quantity ϵ' for which $\epsilon'/2 < \widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D},)$.

We start by analyzing the conditional probability, forming the second summand in Eq. (22). Recall that $\hat{\eta}^k_{m,q}$, after clustering the q subsample pairs $\{S^1_i, S^2_i\}_{i=1}^q$, uses an additional i.i.d sample S^3 of size m to empirically estimate $\sum_q d^m_{\mathcal{D}}(\mathtt{A_k}(S^1_i), \mathtt{A_k}(S^2_i))/\sqrt{m}q \in [0,1]$. This is achieved by calculating the average percentage of instances in S^3 which switches between clusterings. Thus, conditioned on the event appearing in the second summand of Eq. (22), $\hat{\eta}^k_{m,q}$ is simply an empirical average of m i.i.d random variables in $[0,1]$, whose expected value, denoted as v, is a strictly positive number in the range of $(\widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D}) \pm \epsilon/2)/\sqrt{m}$. Thus, the second summand of Eq. (22) refers to an event where this empirical average is at a distance of at least $\epsilon/(2\sqrt{m})$ from its expected value. We can therefore apply a large deviation result to bound this probability. Since the expectation itself is a (generally decreasing) function of the sample size m, we will need something a bit stronger than the regular Hoeffding's bound. Using a relative entropy version of Hoeffding's bound [5], we have that the second summand in Eq. (22) is upper bounded by:

$$\exp\left(-mD_{kl}\left[\frac{v+\epsilon/2}{\sqrt{m}}\middle|\middle|\frac{v}{\sqrt{m}}\right]\right) + \exp\left(-mD_{kl}\left[\max\left\{0, \frac{v-\epsilon/2}{\sqrt{m}}\right\}\middle|\middle|\frac{v}{\sqrt{m}}\right]\right), \tag{23}$$

where $D_{kl}[p||q] := -p\log(p/q) - (1-p)\log((1-p)/(1-q))$ for any $q \in (0,1)$ and any $p \in [0,1]$. Using the fact that $D_{kl}[p||q] \geq (p-q)^2/2\max\{p,q\}$, we get that Eq. (23) can be upper bounded by a quantity which converges to 0 as $m \to \infty$. As a result, the second summand in Eq. (22) converges to 0 as $m \to \infty$.

As to the first summand in Eq. (22), using the triangle inequality and switching sides allows us to upper bound it by:

$$\begin{aligned}
\Pr\Bigg(&\left|\frac{1}{q}\sum_{i=1}^{q} d^m_{\mathcal{D}}(\mathtt{A_k}(S^1_i), \mathtt{A_k}(S^2_i)) - \mathbb{E}[d^m_{\mathcal{D}}(r)|A^m_r]\right| \\
&\geq \frac{\epsilon}{2} - \left|\mathbb{E}[d^m_{\mathcal{D}}(r)|A^m_r] - \mathbb{E}[d^m_{\mathcal{D}}(r)]\right| - \left|\mathbb{E}d^m_{\mathcal{D}}(r) - \widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})\right|\Bigg)
\end{aligned} \tag{24}$$

By the definition of $\widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})$ as appearing in Thm. 1 , and Proposition D.1,

$$\lim_{m\to\infty} \mathbb{E}d^m_{\mathcal{D}}(r) - \widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D}) = O(h(r)) = O(\exp(-r^2)). \tag{25}$$

Using Eq. (25) and Eq. (21), we can upper bound Eq. (24) by

$$\begin{aligned}
\Pr\Bigg(&\left|\frac{1}{q}\sum_{i=1}^{q} d^m_{\mathcal{D}}(\mathtt{A_k}(S^1_i), \mathtt{A_k}(S^2_i)) - \mathbb{E}[d^m_{\mathcal{D}}(r)|A^m_r]\right| \\
&\geq \frac{\epsilon}{2} - (1 - \Pr(A^m_r))O(r) - O(\exp(-r^2)) - o(1)\Bigg),
\end{aligned} \tag{26}$$

where $o(1) \to 0$ as $m \to \infty$. Moreover, by using the law of total probability and Lemma D.6, we have that for any $\nu > 0$,

$$\begin{aligned}
&\Pr\left(\left|\frac{1}{q}\sum_{i=1}^{q} d_{\mathcal{D}}^m(\mathtt{A}_\mathtt{k}(S_i^1), \mathtt{A}_\mathtt{k}(S_i^2)) - \mathbb{E}[d_{\mathcal{D}}^m(r)|A_r^m]\right| > \nu\right) \\
&\leq (1 - \Pr(A_r^m)) * 1 + \Pr(A_r^m)\Pr\left(\left|\frac{1}{q}\sum_{i=1}^{q} d_{\mathcal{D}}^m(\mathtt{A}_\mathtt{k}(S_i^1), \mathtt{A}_\mathtt{k}(S_i^2)) - \mathbb{E}[d_{\mathcal{D}}^m(r)|A_r^m]\right| > \nu \middle| A_r^m\right) \\
&\leq (1 - \Pr(A_r^m)) + 2\Pr(A_r^m)\exp\left(-\frac{2q\nu^2}{r^2}\right). \qquad (27)
\end{aligned}$$

Lemma D.6 can be applied because $d_{\mathcal{D}}^m(\mathtt{A}_\mathtt{k}(S_i^1), \mathtt{A}_\mathtt{k}(S_i^2)) = d_{\mathcal{D}}^m(r)$ for any i, if A_r^m occurs.

If m, r are such that

$$\frac{\epsilon}{2} - (1 - \Pr(A_r^m))O(r) - O(\exp(-r^2)) - o(1) > 0, \qquad (28)$$

we can substitute this expression instead of ν in Eq. (27), and get that Eq. (26) is upper bounded by

$$(1 - \Pr(A_r^m)) + 2\Pr(A_r^m)\exp\left(-\frac{2q\left(\frac{\epsilon}{2} - (1 - \Pr(A_r^m))O(r) - O(\exp(-r^2))) - o(1)\right)^2}{r^2}\right). \qquad (29)$$

Let

$$g_m(r) := \Pr_{S_1, S_2 \sim \mathcal{D}^m}(d_{\mathcal{D}}^m(r) \neq d_{\mathcal{D}}^m(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))) \quad , \quad g(r) = \lim_{m\to\infty} g_m(r)$$

By regularity condition 3d, $g(r) = O(r^{-3-\delta})$ for some $\delta > 0$. Also, we have that $\Pr(A_r^m) = (1 - g_m(r))^q$, and therefore $\lim_{m\to\infty}\Pr(A_r^m) = (1 - g(r))^q$ for any fixed q. In consequence, as $m \to \infty$, Eq. (29) converges to

$$(1 - (1 - g(r)))^q) + 2(1 - g(r))^q \exp\left(-\frac{2q\left(\frac{\epsilon}{2} - (1 - (1 - g(r))^q)O(r) - O(\exp(-r^2))\right)^2}{r^2}\right). \qquad (30)$$

Now we use the fact that r can be chosen arbitrarily. In particular, let $r = q^{1/(2+\delta/2)}$, where $\delta > 0$ is the same quantity appearing in condition 3d. It follows that

$$\begin{aligned}
&1 - (1 - g(r))^q \leq qg(r) = O(q/r^{3+\delta}) = O\left(q^{1-\frac{3+\delta}{2+\delta/2}}\right) \\
&(1 - (1 - g(r))^q)O(r) = qg(r)O(r) = O\left(q^{1-\frac{2+\delta}{2+\delta/2}}\right) = O(q^{-\frac{\delta}{4+\delta}}) \\
&q/r^2 = q^{1-\frac{1}{1+\delta/4}} \\
&\exp(-r^2) = \exp(-q^{\frac{1}{1+\delta/4}}).
\end{aligned}$$

It can be verified that the equations above imply the validness of Eq. (28) for large enough m and q (and hence r). Substituting these equations into Eq. (30), we get an upper bound

$$O\left(q^{1-\frac{3+\delta}{2+\delta/2}}\right) + \exp\left(-2q^{1-\frac{1}{1+\delta/4}}\left(\frac{\epsilon}{2} - O\left(q^{-\frac{\delta}{4+\delta}}\right) - O\left(\exp(-q^{\frac{1}{1+\delta/4}})\right)\right)^2\right).$$

Since $\delta > 0$, it can be verified that the first summand asymptotically dominates the second summand (as $q \to \infty$), and can be bounded in turn by $o(q^{-1/2})$.

Summarizing, we have that the first summand in Eq. (22) converges to $o(q^{-1/2})$ as $m \to \infty$, and the second summand in Eq. (22) converge to 0 as $m \to \infty$, for any fixed $\epsilon > 0$, and thus $\Pr(|\sqrt{m}\,\hat{\eta}_{m,q}^k - \widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})| > \epsilon)$ converges to $o(q^{-1/2})$.

E Proof of Thm. 2 and Thm. 3

The tool we shall use for proving Thm. 2 and Thm. 3 is the following general central limit theorem for Z-estimators (Thm. 3.3.1 in [8]). We will first quote the theorem and then explain the terminology used.

Theorem E.1 (Van der Vaart)**.** *Let Ψ_m and Ψ be random maps and a fixed map, respectively, from a subset Θ of some Banach space into another Banach space such that as $m \to \infty$,*

$$\frac{\|\sqrt{m}(\Psi_m - \Psi)(\hat{\boldsymbol{\theta}}) - \sqrt{m}(\Psi_m - \Psi)(\boldsymbol{\theta}_0)\|}{1 + \sqrt{m}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|} \to 0 \tag{31}$$

in probability, and such that the sequence $\sqrt{m}(\Psi_m - \Psi)(\boldsymbol{\theta}_0)$ converges in distribution to a tight random element Z. Let $\boldsymbol{\theta} \mapsto \Psi(\boldsymbol{\theta})$ be Fréchet-differentiable at $\boldsymbol{\theta}_0$ with an invertible derivative $\dot{\Psi}_{\boldsymbol{\theta}_0}$, which is assumed to be a continuous linear operator[1]*. If $\Psi(\boldsymbol{\theta}_0) = 0$ and $\Psi_m(\hat{\boldsymbol{\theta}})/\sqrt{m} \to 0$ in probability, and $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_0$, then $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to $-\dot{\Psi}_{\boldsymbol{\theta}_0}^{-1} Z$.*

A Banach space is any complete normed vector space (possible infinite dimensional). A tight random element essentially means that an arbitrarily large portion of its distribution lies in compact sets. This condition is trivial when Θ is a subset of Euclidean space. Fréchet-differentiability of a function $f : U \mapsto V$ at $\mathbf{x} \in U$, where U, V are Banach spaces, means that there exists a bounded linear operator $A : U \mapsto V$ such that

$$\lim_{\mathbf{h} \to 0} \frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - A(\mathbf{h})\|_W}{\|\mathbf{h}\|_U} = 0.$$

This is equivalent to regular differentiability in finite dimensional settings.

It is important to note that the theorem is stronger than what we actually need, since we only consider finite dimensional Euclidean spaces, while the theorem deals with possibly infinite dimensional Banach spaces. In principle, it is possible to use this theorem to prove central limit theorems in infinite dimensional settings, for example in kernel clustering where the associated reproducing kernel Hilbert space is infinite dimensional. However, the required conditions become much less trivial, and actually fail to hold in some cases (see below for further details).

We now turn to the proofs themselves. Since the proofs of Thm. 2 and Thm. 3 are almost identical, we will prove them together, marking differences between them as needed. In order to allow uniform notation in both cases, we shall assume that $\phi(\cdot)$ is the identity mapping in Bregman divergence clustering, and the feature map from $\mathcal{X}$ to $\mathcal{H}$ in kernel clustering.

With the assumptions that we made in the theorems, the only thing really left to show before applying Thm. E.1 is that Eq. (31) holds. Notice that it is enough to show that

$$\frac{\|\sqrt{m}(\Psi_m^i - \Psi^i)(\hat{\boldsymbol{\theta}}) - \sqrt{m}(\Psi_m^i - \Psi^i)(\boldsymbol{\theta}_0)\|}{1 + \sqrt{m}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|} \to 0$$

for any $i \in \{1, \ldots, k\}$. We will prove this in a slightly more complicated way than necessary, which also treats the case of kernel clustering where $\mathcal{H}$ is infinite-dimensional. By Lemma 3.3.5 in [8], since $\mathcal{X}$ is bounded, it is sufficient to show that for any i, there is some $\delta > 0$ such that

$$\{\psi^i_{\hat{\boldsymbol{\theta}}, \mathbf{h}}(\cdot) - \psi^i_{\boldsymbol{\theta}_0, \mathbf{h}}(\cdot)\}_{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \le \delta, \mathbf{h} \in \mathcal{X}}$$

is a *Donsker class*, where

$$\psi^i_{\boldsymbol{\theta}, \mathbf{h}}(\mathbf{x}) = \begin{cases} \langle \boldsymbol{\theta}_i - \phi(\mathbf{x}), \phi(\mathbf{h}) \rangle & \mathbf{x} \in C_{\boldsymbol{\theta}, i} \\ 0 & \text{otherwise.} \end{cases}$$

Intuitively, a set of real functions $\{f(\cdot)\}$ from $\mathcal{X}$ (with any probability distribution $\mathcal{D}$) to $\mathbb{R}$ is called Donsker if it satisfies a uniform central limit theorem. Without getting too much into the details,

[1] A linear operator is automatically continuous in finite dimensional spaces, not necessarily in infinite dimensional spaces.

this means that if we sample i.i.d m elements from $\mathcal{D}$, then $(f(\mathbf{x}_1) + \ldots + f(\mathbf{x}_m))/\sqrt{m}$ converges in distribution (as $m \to \infty$) to a Gaussian random variable, and the convergence is uniform over all $f(\cdot)$ in the set, in an appropriately defined sense.

We use the fact that if $\mathcal{F}$ and $\mathcal{G}$ are Donsker classes, then so are $\mathcal{F} + \mathcal{G}$ and $\mathcal{F} \cdot \mathcal{G}$ (see examples 2.10.7 and 2.10.8 in [8]). This allows us to reduce the problem to showing that the following three function classes, from $\mathcal{X}$ to $\mathbb{R}$, are Donsker:

$$\{\langle \boldsymbol{\theta}_i, \phi(\mathbf{h})\rangle\}_{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|\leq\delta, \mathbf{h}\in\mathcal{X}} \quad , \quad \{\langle \phi(\cdot), \phi(\mathbf{h})\rangle\}_{\mathbf{h}\in\mathcal{X}} \quad , \quad \{\mathbf{1}_{C_{\boldsymbol{\theta},i}}(\cdot)\}_{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|\leq\delta}. \tag{32}$$

Notice that the first class is a set of bounded constant functions, while the third class is a set of indicator functions for all possible clusters. One can now use several tools to show that each class in Eq. (32) is Donsker. For example, consider a class of real functions on a bounded subset of some Euclidean space. By Thm. 8.2.1 in [3] (and its preceding discussion), the class is Donsker if any function in the class is differentiable to a sufficiently high order. This ensures that the first class in Eq. (32) is Donsker, because it is composed of constant functions. As to the second class in Eq. (32), the same holds in the case of Bregman divergence clustering (where $\phi(\cdot)$ is the identity function), because it is then just a set of linear functions. For finite dimensional kernel clustering, it is enough to show that $\{\langle \cdot, \phi(\mathbf{h})\rangle\}_{\mathbf{h}\in\mathcal{X}}$ is Donsker (namely, the same class of functions after performing the transformation from $\mathcal{X}$ to $\phi(\mathcal{X})$). This is again a set of linear functions in $\mathcal{H}^k$, a subset of some finite dimensional Euclidean space, and so it is Donsker. In infinite dimensional kernel clustering, our class of functions can be written as $\{k(\cdot, \mathbf{h})\}_{\mathbf{h}\in\mathcal{X}}$, where $k(\cdot,\cdot)$ is the kernel function, so it is Donsker if the kernel function is differentiable to a sufficiently high order.

The third class in Eq. (32) is more problematic. By Theorem 8.2.15 in [3] (and its preceding discussion), it suffices that the boundary of each possible cluster is composed of a finite number of smooth surfaces (differentiable to a high enough order) in some Euclidean space. In Bregman divergence clustering, the clusters are separated by hyperplanes, which are linear functions (see appendix A in [1]), and thus the class is Donsker. The same holds for finite dimensional kernel clustering. This will still be true for infinite dimensional kernel clustering, if we can guarantee that any cluster in any solution close enough to θ_0 in Θ will have smooth boundaries. Unfortunately, this does not hold in some important cases. For example, universal kernels (such as the Gaussian kernel) are capable of inducing cluster boundaries arbitrarily close in form to any continuous function, and thus our line of attack will not work in such cases. In a sense, this is not too surprising, since these kernels correspond to very 'rich' hypothesis classes, and it is not clear if a precise characterization of their stability properties, via central limit theorems, is at all possible.

Summarizing the above discussion, we have shown that for the settings assumed in our theorem, all three classes in Eq. (32) are Donsker and hence Eq. (31) holds. We now return to deal with the other ingredients required to apply Thm. E.1.

As to the asymptotic distribution of $\sqrt{m}(\Psi_m - \Psi)(\boldsymbol{\theta}_0)$, since $\Psi(\boldsymbol{\theta}_0) = 0$ by assumption, we have that for any $i \in \{1, \ldots, k\}$,

$$\sqrt{m}(\Psi_m^i - \Psi^i)(\boldsymbol{\theta}_0) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \Delta_i(\boldsymbol{\theta}_0, \mathbf{x}_j). \tag{33}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_m$ is the sample by which Ψ_m is defined. The r.h.s of Eq. (33) is a sum of identically distributed, independent random variables with zero mean, normalized by $\sqrt{m}$. As a result, by the standard central limit theorem, $\sqrt{m}(\Psi_m^i - \Psi^i)(\boldsymbol{\theta}_0)$ converges in distribution to a zero mean Gaussian random vector Y, with covariance matrix

$$V_i = \int_{C_{\boldsymbol{\theta}_0,i}} p(\mathbf{x})(\phi(\mathbf{x}) - \boldsymbol{\theta}_{0,i})(\phi(\mathbf{x}) - \boldsymbol{\theta}_{0,i})^\top d\mathbf{x}.$$

Moreover, it is easily verified that $\mathrm{Cov}(\Delta_i(\boldsymbol{\theta}_0, \mathbf{x}), \Delta_{i'}(\boldsymbol{\theta}_0, \mathbf{x})) = 0$ for any $i \neq i'$. Therefore, $\sqrt{m}(\Psi_m - \Psi)(\boldsymbol{\theta}_0)$ converges in distribution to a zero mean Gaussian random vector, whose covariance matrix V is composed of k diagonal blocks $(V_1, \ldots, V_k)$, all other elements of V being zero.

Thus, we can use Thm. E.1 to get that $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a zero mean Gaussian random vector of the form $-\dot{\Psi}_{\boldsymbol{\theta}_0}^{-1} Y$, which is a Gaussian random vector with a covariance matrix of the form $\dot{\Psi}_{\boldsymbol{\theta}_0}^{-1} V \dot{\Psi}_{\boldsymbol{\theta}_0}^{-1}$.

F Proof of Thm. 4

Since our algorithm returns a locally optimal solution with respect to the differentiable log-likelihood function, we can frame it as a Z-estimator of the derivative of the log-likelihood function with respect to the parameters, namely the score function

$$\Psi_m(\hat{\boldsymbol{\theta}}) = \frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial \boldsymbol{\theta}} \log(q(\mathbf{x}_i|\hat{\boldsymbol{\theta}})).$$

This is a random mapping based on the sample $\mathbf{x}_1, \ldots, \mathbf{x}_m$.

Similarly, we can define $\Psi(\cdot)$ as the 'asymptotic' score function with respect to the underlying distribution $\mathcal{D}$:

$$\Psi(\hat{\boldsymbol{\theta}}) = \int_{\mathcal{X}} \frac{\partial}{\partial \boldsymbol{\theta}} \log(q(\mathbf{x}|\hat{\boldsymbol{\theta}}))p(\mathbf{x})d\mathbf{x}.$$

Under the assumptions we have made, the model $\hat{\boldsymbol{\theta}}$ returned by the algorithm satisfies $\Psi_m(\hat{\boldsymbol{\theta}}) = 0$, and $\hat{\boldsymbol{\theta}}$ converges in probability to some $\boldsymbol{\theta}_0$ for which $\Psi(\boldsymbol{\theta}_0) = 0$. The asymptotic normality of $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is now an immediate consequence of central limit theorems for 'maximum likelihood' Z-estimators, such as Thm. 5.21 in [7].

References

[1] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[2] P. Billingsley and F. Topsøe. Uniformity in weak convergence. *Probability Theory and Related Fields*, 7:1–16, 1967.

[3] R. Dudley. *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1999.

[4] G. R. Grimmet and D. R. Stirzaker. *Probability and Random Processes*. Oxford University Press, 2001.

[5] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, Mar. 1963.

[6] R. R. Rao. Relations betwen weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, 33(2):659–680, June 1962.

[7] A. W. V. D. Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[8] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes : With Applications to Statistics*. Springer, 1996.