A Proofs

In this supplement, we present proofs for "Local Clustering of Density Upper Level Sets". We begin by providing technical lemmas, before moving on to proving the main results of the paper.

Throughout, we will fix $\mathcal{A} \subset \mathbb{R}^d$ to be an arbitrary set. To simplify expressions, for the σ-expansion $\mathcal{A}_\sigma$, we will write the set difference between $\mathcal{A}_\sigma$ and the $(\sigma + r)$-expansion $\mathcal{A}_{\sigma+r}$ as

$$\mathcal{A}_{\sigma,\sigma+r} := \{x : 0 < \rho(x, \mathcal{A}_\sigma) \leq r\},$$

where $\rho(x, \mathcal{A}) = \min_{x' \in \mathcal{A}} \|x - x'\|$. .

For notational ease, we write

$$\text{cut}_{n,r} = \text{cut}(\mathcal{C}_\sigma[\mathbf{X}]; G_{n,r}),\ \mu_K = \mathbb{E}(\text{cut}_{n,r}),\ p_K = \frac{\mu_K}{\binom{n}{2}}$$
$$\text{vol}_{n,r} = \text{vol}(\mathcal{C}_\sigma[\mathbf{X}]; G_{n,r}),\ \mu_V = \mathbb{E}(\text{vol}_{n,r}),\ p_V = \frac{\mu_V}{\binom{n}{2}}$$

for the random variable, mean, and probability of cut size and volume, respectively.

A.1 Technical Lemmas

We state Lemma 1 without proof, as it is trivial. We formally include it mainly to comment on its (potential) suboptimality; for sets $\mathcal{A}$ with diameter much larger than σ, the volume estimate of Lemma 1 will be quite poor.

Lemma 1. *For any $\sigma > 0$ and the σ-expansion $\mathcal{A}_\sigma = \mathcal{A} + \sigma B$,*

$$\sigma B \subset \mathcal{A}_\sigma, \quad \text{and } \nu(\mathcal{A} + \sigma B) \leq \nu((1 + \sigma)\mathcal{A}) = (1 + \sigma)^d \nu(\mathcal{A}).$$

We will need to carefully control the volume of the expansion set using the above estimate; Lemma 2 serves this purpose.

Lemma 2. *For any $0 \leq x \leq 1/2d$,*

$$(1 + x)^d \leq 1 + 2dx.$$

The proof of Lemma 2 is based on approximation via Taylor series, and we omit it.

We will repeatedly employ Lemma 1 and Lemma 2 in tandem. As a first example, in Lemma 3, we use it to bound the ratio of $\nu(\mathcal{A}_\sigma)$ to $\nu(\mathcal{A}_{\sigma-r})$. This will be useful when we bound $\text{vol}(\mathcal{C}_\sigma)$.

Lemma 3. *For σ, $\mathcal{A}_\sigma$ as in Lemma 1, let $r > 0$ satisfy $r \le \sigma/4d$. Then,*

$$\frac{\nu(\mathcal{A}_\sigma)}{\nu(\mathcal{A}_{\sigma-r})} \le 2.$$

Proof. Fix $q = \sigma - r$. Then,

$$\begin{aligned}\nu(\mathcal{A}_\sigma) &= \nu(\mathcal{A}_{q+\sigma-q}) = \nu(\mathcal{A}_q + (\sigma - q)B) \\ &\le \nu(\mathcal{A}_q + \frac{(\sigma-q)}{q}\mathcal{A}_q) = \left(1 + \frac{\sigma - q}{q}\right)^d \nu(\mathcal{A}_q)\end{aligned}$$

where the inequality follows from Lemma 1. Of course, $\sigma - q = r$, and $\frac{r}{q} \le \frac{1}{2d}$ for $r \le \frac{1}{4d}$. The claim then follows from Lemma 2. □

A.2 Cut and volume estimates

Lemma 4. *Under the setup and conditions of Theorem 1, and for any $r < \sigma/2d$,*

$$\mathbb{P}(\mathcal{C}_{\sigma,\sigma+r}) \le 2\nu(\mathcal{C}_\sigma)\frac{rd}{\sigma}\left(\lambda_\sigma - \frac{r^\gamma}{\gamma+1}\right)$$

Proof. Recalling that f is the density function for $\mathbb{P}$, we have

$$\mathbb{P}(\mathcal{C}_{\sigma,\sigma+r}) = \int_{\mathcal{C}_{\sigma,\sigma+r}} f(x)dx \tag{A.1}$$

We partition $\mathcal{C}_{\sigma,\sigma+r}$ into slices, based on distance from $\mathcal{C}_\sigma$, as follows: for $k \in \mathbb{N}$,

$$\mathcal{T}_{i,k} = \left\{x \in \mathbb{R}^d : t_{i,k} < \frac{\rho(x, \mathcal{C}_\sigma)}{r} \le t_{i+1,k}\right\}, \quad \mathcal{C}_{\sigma,\sigma+r} = \bigcup_{i=0}^{k-1} \mathcal{T}_{i,k}$$

where $t_i = i/k$ for $i = 0, \dots, k-1$. As a result,

$$\int_{\mathcal{C}_{\sigma,\sigma+r}} f(x)dx = \sum_{i=0}^{k-1} \int_{\mathcal{T}_{i,k}} f(x)dx \le \sum_{i=0}^{k-1} \nu(\mathcal{T}_{i,k}) \max_{x \in \mathcal{T}_{i,k}} f(x).$$

We substitute

$$\nu(\mathcal{T}_{i,k}) = \nu(\mathcal{C}_\sigma + rt_{i+1,k}B) - \nu(\mathcal{C}_\sigma + rt_{i,k}B) := \nu_{i+1,k} - \nu_{i,k}.$$

where for simplicity we've written $\nu_{i,k} = \nu(\mathcal{C}_\sigma + rt_{i+1,k}B)$. This, in concert with the upper bound

$$\max_{x \in \mathcal{T}_{i,k}} f(x) \le \lambda_\sigma - (rt_{i,k})^\gamma,$$

which follows from (A1) and (A2), yields

$$\sum_{i=0}^{k-1} \nu(\mathcal{T}_{i,k}) \max_{x \in \mathcal{T}_{i,k}} f(x) \leq \sum_{i=0}^{k-1} \left\{ \nu_{i+1,k} - \nu_{i,k} \right\} \left(\lambda_\sigma - (rt_{i,k})^\gamma \right)$$
$$= \sum_{i=1}^{k} \underbrace{\nu_{i,k} \left([\lambda_\sigma - (rt_{i,k})^\gamma] - [\lambda_\sigma - (rt_{i-1,k})^\gamma] \right)}_{:=\Sigma_k} + \underbrace{\left(\nu_{k,k} \left[\lambda_\sigma - r^\gamma \right] - \nu_{1,k} \lambda_\sigma \right)}_{:=\xi_k} \tag{A.2}$$

We first consider the term Σ_k. Here we use Lemma 1 to upper bound

$$\nu_{i,k} \leq \text{vol}(\mathcal{C}_\sigma) \left(1 + \frac{rt_{i,k}}{\sigma} \right)^d$$

and so we can in turn upper bound Σ_k:

$$\Sigma_k \leq \text{vol}(\mathcal{C}_\sigma) r^\gamma \sum_{i=1}^{k} \left(1 + \frac{rt_{i,k}}{\sigma} \right)^d \left((t_{i-1,k})^\gamma - (t_{i,k})^\gamma \right). \tag{A.3}$$

This, of course, is a Riemann sum, and as the inequality holds for all values of k it holds in the limit as well, which we compute to be

$$\lim_{k \to \infty} \sum_{i=1}^{k} \left(1 + \frac{rt_{i,k}}{\sigma} \right)^d \left((t_{i-1,k})^\gamma - (t_{i,k})^\gamma \right) = \gamma \int_0^1 \left(1 + \frac{rt}{\sigma} \right)^d t^{\gamma - 1} dt$$
$$\overset{(i)}{\leq} \gamma \int_0^1 \left(1 + \frac{2drt}{\sigma} \right) t^{\gamma - 1} dt = \left(1 + \frac{\gamma 2dr}{\gamma + 1} \right).$$

where (i) follows from Lemma 2. We plug this estimate in to (A.3) and obtain

$$\lim_{k \to \infty} \Sigma_k \leq \text{vol}(\mathcal{C}_\sigma) r^\gamma \left(1 + \frac{\gamma 2dr}{\gamma + 1} \right).$$

We now provide an upper bound on ξ_k. It will follow the same basic steps as the bound on Σ_k, but will not involve integration:

$$\xi_k \overset{(ii)}{\leq} \nu(\mathcal{C}_\sigma) \left\{ \left(1 + \frac{r}{\sigma} \right)^d (\lambda - r^\gamma) - \lambda \right\}$$
$$\overset{(iii)}{\leq} \nu(\mathcal{C}_\sigma) \left\{ \left(1 + \frac{2dr}{\sigma} \right) (\lambda - r^\gamma) - \lambda \right\} = \nu(\mathcal{C}_\sigma) \left\{ \frac{2dr}{\sigma} (\lambda - r^\gamma) - r^\gamma \right\}.$$

where (ii) follows from Lemma 1 and (iii) from Lemma 2. The final result comes from adding together the upper bounds on Σ_k and ξ_k and taking the limit as $k \to \infty$. □

Lemma 5. *Under the setup and conditions of Theorem 1, and for any $r < \sigma/2d$,*

$$p_K \leq \frac{4\lambda\nu_d r^{d+1}\nu(\mathcal{C}_\sigma)d}{\sigma}\left(\lambda_\sigma - \frac{r^\gamma}{\gamma+1}\right)$$

Proof. We can write $\text{cut}_{n,r}$ as the sum of indicator functions,

$$\text{cut}_{n,r} = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{1}(x_i \in \mathcal{C}_{\sigma,\sigma+r})\mathbf{1}(x_j \in B(x_i,r)\cap\mathcal{C}_\sigma) \tag{A.4}$$

and by linearity of expectation, we can obtain

$$p_K = \frac{\mu_K}{\binom{n}{2}} = 2\cdot\mathbb{P}(x_i \in \mathcal{C}_{\sigma,\sigma+r}, x_j \in B(x_i,r)\cap\mathcal{C}_\sigma)$$

Writing this with respect to the density function f, we have

$$\begin{aligned} p_K &= 2\int_{\mathcal{C}_{\sigma,\sigma+r}} f(x)\left\{\int_{B(x,r)\cap\mathcal{C}_\sigma} f(x')dx'\right\}dx \\ &\leq 2\nu_d r^d\lambda\int_{\mathcal{C}_{\sigma,\sigma+r}} f(x)dx \end{aligned}$$

where the inequality follows from Assumption (A3), which implies that the density function $f(x') \leq \lambda$ for all $x' \in \mathcal{C}_\sigma \setminus \mathcal{C}$ (otherwise, x' would be in some $\mathcal{C}' \in \mathbb{C}_f(\lambda)$, which (A3) forbids). Then, upper bounding the integral using Lemma 5 gives the final result. □

Lemma 6. *Under the setup and conditions of Theorem 1,*

$$p_V \geq \lambda_\sigma^2\nu_d r^d\nu(\mathcal{C}_\sigma)$$

Proof. The proof will proceed similarly to Lemma 5. We begin by writing $\text{vol}_{n,r}$ as the sum of indicator functions,

$$\text{vol}_{n,r} = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{1}(x_i \in \mathcal{C}_\sigma)\mathbf{1}(x_j \in B(x_i,r)) \tag{A.5}$$

and by linearity of expectation we obtain

$$p_V = \frac{\mu_V}{\binom{n}{2}} = 2\cdot\mathbb{P}(x_i \in \mathcal{C}_\sigma, x_j \in B(x_i,r)).$$

Writing this with respect to the density function f, we have

$$\begin{aligned} p_V &= 2\int_{\mathcal{C}_\sigma} f(x)\left\{\int_{B(x,r)} f(x')dx'\right\}dx \\ &\geq 2\int_{\mathcal{C}_{\sigma-r}} f(x)\left\{\int_{B(x,r)} f(x')dx'\right\}dx \\ &\overset{(i)}{\geq} 2\lambda_\sigma^2\nu_d r^d\int_{\mathcal{C}_{\sigma-r}} f(x)dx \end{aligned}$$

where (i) follows from the fact that $B(x,r) \subset \mathcal{C}_\sigma$ for all $x \in C_{\sigma-r}$, along with the lower bound in Assumption (A1). The claim then follows from Lemma 3. $\square$

We now convert from bounds on p_K and p_V to probabilistic bounds on $\mathrm{cut}_{n,r}$ and $\mathrm{vol}_{n,r}$ in Lemmas 7 and 8. The key ingredient will be Lemma 9, Hoeffding's inequality for U-statistics; the proofs for both are nearly identical and we give only a proof for Lemma 7.

Lemma 7. *The following statement holds for any $\delta \in (0,1]$: Under the setup and conditions of Theorem 1,*

$$\frac{\mathrm{cut}_{n,r}}{\binom{n}{2}} \leq p_K + \sqrt{\frac{\log(1/\delta)}{n}} \tag{A.6}$$

with probability at least $1-\delta$.

Lemma 8. *The following statement holds for any $\delta \in (0,1]$: Under the setup and conditions of Theorem 1,*

$$\frac{\mathrm{vol}_{n,r}}{\binom{n}{2}} \geq p_V - \sqrt{\frac{\log(1/\delta)}{n}} \tag{A.7}$$

with probability at least $1-\delta$.

Proof of Lemma 7. From (A.4), we see that $\mathrm{cut}_{n,r}$, properly scaled, can be expressed as an order-2 U-statistic,

$$\frac{\mathrm{cut}_{n,r}}{\binom{n}{2}} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \phi_K(x_i, x_j)$$

where

$$\phi_K(x_i,x_j) = \mathbf{1}(x_i \in \mathcal{A}_{\sigma,\sigma+r})\mathbf{1}(x_j \in B(x_i,r) \cap \mathcal{A}_\sigma) + \mathbf{1}(x_j \in \mathcal{A}_{\sigma,\sigma+r})\mathbf{1}(x_i \in B(x_j,r) \cap \mathcal{A}_\sigma).$$

From Lemma 9 we therefore have

$$\frac{\mathrm{cut}_{n,r}}{\binom{n}{2}} \leq p_k + \sqrt{\frac{\log(1/\delta)}{n}}$$

with probability at least $1-\delta$. $\square$

A.3 Proof of Theorem 1

The proof of Theorem 1 is more or less given by Lemmas 5, 6, 7, and 8. All that remains is some algebra, which we take care of below.

Fix $\delta \in (0,1]$ and let $\delta' = \delta/2$. Noting that $\Phi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}]) = \frac{\mathrm{cut}_{n,r}}{\mathrm{vol}_{n,r}}$, some trivial algebra gives us the expression

$$\Phi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}]) = \frac{p_K + \left(\frac{\mathrm{cut}_{n,r}}{\binom{n}{2}} - p_K\right)}{p_V + \left(\frac{\mathrm{vol}_{n,r}}{\binom{n}{2}} - p_V\right)} \tag{A.8}$$

We assume (A.6) and (A.7) hold with respect to δ', keeping in mind that this will happen with probability at least $1-\delta$. Along with (A.8) this means

$$\Phi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}]) \leq \frac{p_K + \mathrm{Err}_n}{p_V - \mathrm{Err}_n}$$

for $\mathrm{Err}_n = \sqrt{\frac{\log(1/\delta')}{n}}$. Now, some straightforward algebraic manipulations yield

$$\begin{aligned}
\frac{p_K + \mathrm{Err}_n}{p_V - \mathrm{Err}_n} &= \frac{p_K}{p_V} + \left(\frac{p_K}{p_V - \mathrm{Err}_n} - \frac{p_K}{p_V}\right) + \frac{\mathrm{Err}_n}{p_V - \mathrm{Err}_n} \\
&= \frac{p_k}{p_V} + \frac{\mathrm{Err}_n}{p_V - \mathrm{Err}_n}\left(\frac{p_K}{p_V} + 1\right) \\
&\leq \frac{p_K}{p_V} + 2\frac{\mathrm{Err}_n}{p_V - \mathrm{Err}_n}.
\end{aligned}$$

By Lemmas 5 and Lemma 6, we have

$$\frac{p_K}{p_V} \leq \frac{4rd}{\sigma}\frac{\lambda}{\lambda_\sigma}\frac{\left(\lambda_\sigma - \frac{r^\gamma}{\gamma+1}\right)}{\lambda_\sigma}$$

Then, the choice of

$$n \geq \frac{9\log(2/\delta)}{\epsilon^2}\left(\frac{1}{\lambda_\sigma^2 \nu(\mathcal{C}_\sigma)\nu_d r^d}\right)^2$$

implies $2\frac{\mathrm{Err}_n}{p_V - \mathrm{Err}_n} \leq \epsilon$.

A.4 Concentration inequalities

Given a symmetric kernel function $k : \mathcal{X}^m \to \mathbb{R}$, and data $\{x_1, \ldots, x_n\}$, we define the *order-m U statistic* to be

$$U := \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \ldots < i_m \leq n} k(x_{i_1}, \ldots, x_{i_m})$$

For both Lemmas 9 and 10, let $X_1, \ldots, X_n \in \mathcal{X}$ be independent and identically distributed. We will additionally assume the order-m kernel function k satisfies the boundedness property $\sup_{x_1,\ldots,x_m} |k(x_1, \ldots, x_m)| \leq 1$.

Lemma 9 (Hoeffding's inequality for U-statistics.)**.** *For any $t > 0$,*

$$\mathbb{P}(|U - \mathbb{E}U| \geq t) \leq 2\exp\left\{-\frac{2nt^2}{m}\right\}$$

Further, for any $\delta > 0$, we have

$$U \leq \mathbb{E}U + \sqrt{\frac{m\log(1/\delta)}{2n}},$$
$$U \geq \mathbb{E}U - \sqrt{\frac{m\log(1/\delta)}{2n}}$$

each with probability at least $1 - \delta$.

B OTHER STUFF

Lemma 10 (Bernstein's inequality for U-statistics)**.** *Additionally, assume $\sigma^2 = \mathrm{Var}\left(k(X_1, \ldots, X_m)\right) < \infty$. Then for any $\delta > 0$,*

$$\mathbb{P}(U - \mathbb{E}U \geq t) \leq \exp\left\{-\frac{n}{2m}\frac{t^2}{\sigma^2 + t/3}\right\},$$

Moreover if $\sigma^2 \leq \mu/n$,

$$U \leq \mathbb{E}U \cdot \left(1 + \max\left\{\sqrt{\frac{2m\log(1/\Delta)}{\mu}}, \frac{2m\log(1/\Delta)}{3\mu}\right\}\right),$$
$$U \geq \mathbb{E}U \cdot \left(1 - \max\left\{\sqrt{\frac{2m\log(1/\Delta)}{\mu}}, \frac{2m\log(1/\Delta)}{3\mu}\right\}\right)$$

each with probability at least $1 - \Delta$.

Multiplicative bound: As $\tilde{k}(x_1, x_2)$ is the sum of two Bernoulli random variables with negative covariance (since $\mathbf{1}(x_i \in \mathcal{A}_{\sigma,\sigma+r})\mathbf{1}(x_j \in B(x_i, r) \cap \mathcal{A}_\sigma) = 1$ implies $\mathbf{1}(x_j \in \mathcal{A}_{\sigma,\sigma+r})\mathbf{1}(x_i \in B(x_j, r) \cap \mathcal{A}_\sigma) = 0$ and vice versa), we can upper bound $\mathrm{Var}\left(\tilde{k}(x_1, x_2)\right) \leq \widetilde{p}$, where we recall

$$\widetilde{p} = 2 \cdot \mathbb{P}\left(\mathbf{1}(x_1 \in \mathcal{A}_{\sigma,\sigma+r})\mathbf{1}(x_2 \in B(x_1, r) \cap \mathcal{A}_\sigma)\right)$$

From Lemma 10, we therefore have

$$\frac{\widetilde{\mathcal{E}}}{\binom{n}{2}} \leq \widetilde{p} + \max\left\{\sqrt{\frac{4\log(1/\Delta)\widetilde{p}}{n}}, \frac{4\log(1/\Delta)}{3n}\right\}$$

with probability at least $1 - \Delta$.

Multiplicative bound: The two terms on the right hand side are both distributed Bernoulli$(p/2)$. Moreover, since $\mathbf{1}(x_i \in A_\sigma) = 1$ implies $\mathbf{1}(x_j \in A_\sigma) = 0$, they have negative covariance. We can therefore upper bound $\mathrm{Var}(k'(x_i, x_j)) \leq p$, and so from Lemma 10, we have

$$\frac{\mathcal{V}}{\binom{n}{2}} \geq p - \max\left\{\sqrt{\frac{4\log(1/\Delta)p}{n}}, \frac{4\log(1/\Delta)}{3n}\right\}$$

with probability at least $1 - \Delta$.