# Notes for Week 4/3/20 - 4/9/20

Alden Green

April 8, 2020

Suppose we observe data $(X_1, Y_1), \ldots, (X_n, Y_n)$ according to the following random design regression model. The design points $X_1, \ldots, X_n \in \mathcal{X}$ are independently sampled from a distribution $P$, and the responses

$$Y_i = f_0(X_i) + \varepsilon_i$$

where $f_0 : \mathcal{X} \to \mathbb{R}$ is the unknown regression function and $\varepsilon_i$ are independent $N(0,1)$ noise samples. Our goal is to test

$$\mathbf{H}_0 : f_0 = 0, \quad \mathbf{H}_a : f_0 \neq 0.$$

Our test statistic $T = \|\widehat{f}\|_n^2$ will be a plug-in estimator of the $L_2$-norm of $f_0$. The estimator $\widehat{f}$ of the function $f_0$ will be a truncated-series estimator, using eigenvectors of a graph Laplacian. We now define the graph $\bar{G}_{n,r}$ we will build over the design points $\mathbf{X} = X_1, \ldots, X_n$, which we term the *histogram lattice*.

**Histogram Lattice.** Let $0 < r < 1$, and let $M = 1/r$. Let

$$Z = \left\{ \frac{1}{M}\big(2m_1 - 1, \ldots, 2m_d - 1\big) : m \in [M]^d \right\}$$

be a set of evenly spaced grid points. (We will always assume $M$ is an integer, merely to simplify some notational and definitional details.) A natural graph associated with $Z$ is the lattice

$$\bar{G} = (Z, \bar{E}), \quad (z, z') \in \bar{E} \text{ if } \|z - z'\|_2 = r$$

Let $\Pi : [0,1]^d \to Z$ map points $x \in [0,1]^d$ to the nearest grid points $z \in Z$,

$$\Pi(x) := \operatorname*{argmin}_{z \in Z} \; \|z - x\|_2$$

The *histogram lattice* $\bar{G}_{n,r} = (\mathbf{X}, \bar{E}_{n,r})$ is a product graph induced by $\bar{G}$ and the mapping $\Pi$, with edges

$$(X_i, X_j) \in \bar{E}_{n,r} \text{ if } \Big(\Pi(X_i), \Pi(X_j)\Big) \in \bar{E}.$$

The matrix $\bar{L}_{n,r}$ is the graph Laplacian of $\bar{G}_{n,r}$ and $(\lambda_1, v_1), \ldots, (\lambda_n, v_n)$ are the eigenvector/eigenvalue pairs of $\bar{L}_{n,r}$, defined by the equation

$$\bar{L}_{n,r} v_k = \lambda_k v_k, \quad \|v_k\|_n^2 = 1.$$

For a positive integer $\kappa$, the Laplacian eigenmaps estimator $\widehat{f}_{\mathrm{LE}}$ is defined as

$$\widehat{f}_{\mathrm{LE}} := \sum_{k=1}^{\kappa} \langle Y, v_k \rangle_n v_k.$$

and the resulting test statistic is thus

$$\widehat{T}_{\mathrm{LE}} = \sum_{k=1}^{\kappa} \big(\langle Y, v_k \rangle_n\big)^2.$$

# 1 Approximation Error

We make some assumptions on the function $f_0$ and the distribution $P$.

(A1) $f_0 \in C^s(\mathcal{X}; B)$ for some $s > 0$. If $s > 1$, then $f$ is also compactly supported on a strict subset of $\mathcal{X}$.

(A2) $P$ admits a density $p$ with respect to the Lebesgue measure on $\mathbb{R}^d$. The density $p \in C(\mathcal{X}; p_{\max})$, for some $k > 0$.

**Theorem 1.** *Suppose assumptions (A1) and (A2) are satisfied for some $s \geq 1$, and $k = s - 1$. Then, there exists a constant $c$ such that*

$$\sum_{k=1}^{\kappa} \bigl(\langle v_k, f_0 \rangle_n\bigr)^2 \geq c \cdot \|f_0\|_2^2 - \kappa^{-2/d}$$

# 2 Proof of Theorem 1

Let $\bar{f}_0 : \mathbf{X} \to \mathbb{R}$ be the histogram estimate of $f_0$, defined as

$$\bar{f}_0(X) = \frac{1}{|Q(\Pi(X))|} \sum_{i=1}^{n} f_0(X_i) \cdot \mathbf{1}\Bigl\{\pi(X_i) = \pi(X)\Bigr\}$$

We shall proceed according to the following steps.

1. Under the assumption (A2), for any $(\log(n)/n)^{1/d} \leq r$, there exist constants $c$ and $C$ such that

$$c \cdot \min\Bigl\{nr^{d+2}\kappa^{2/d}, \deg_{\min}(\bar{G}_{n,r})\Bigr\} \leq \lambda_\kappa(\bar{G}_{n,r}) \leq C \cdot nr^{d+2}\kappa^{2/d} \tag{1}$$

   and

$$c \cdot nr^d \leq \deg_{\min}(\bar{G}_{n,r}) \leq C \cdot nr^d \tag{2}$$

   with probability at least $1 - $(???). Supposing (1) and (2) hold, and additionally $r < \kappa^{-1/d}$, then some straightforward algebra implies

$$cnr^{d+2}\kappa^{2/d} \leq \lambda_\kappa(\bar{G}_{n,r}) \leq \deg_{\min}(\bar{G}_{n,r}).$$

2. **Deterministic bounds:** Specifically as a consequence of the upper bound $\lambda_k(\bar{G}_{n,r}) \geq \deg_{\min}(\bar{G}_{n,r})$, we have that

$$\sum_{k=1}^{\kappa} \bigl(\langle v_k, f_0 \rangle_n\bigr)^2 = \sum_{k=1}^{\kappa} \bigl(\langle v_k, \bar{f}_0 \rangle_n\bigr)^2 \tag{3}$$

   For any $f \in \mathbb{R}^n$, we have

$$\|f\|_n^2 - \frac{f^T \bar{L}_{n,r}^s f}{n\lambda_\kappa(\bar{G}_{n,r})} \tag{4}$$

   Neither the equality nor the inequality follow from probabilistic reasoning (except through the reasoning used to establish $\lambda_k(\bar{G}_{n,r}) \geq \deg_{\min}(\bar{G}_{n,r})$). The equality follows from the product graph structure of $\bar{G}_{n,r}$. The inequality is a standard inequality used in the analysis of truncated-series estimators, and would hold for any graph $G$ on the vertices $\mathbf{X}$.

3. Under the assumptions (A1) and (A2), the graph Sobolev seminorm is upper bounded

$$\bar{f}_0^T \bar{L}_{n,r}^s \bar{f}_0 \leq n^{s+1} r^{s(d+2)} B^2$$

   with probability at least $1 - C_1 r^{-d} \exp\{-c_2 nr^d\}$.

4. The empirical norm of the histogram estimate $\bar{f}_0$ is lower bounded
$$\|\bar{f}_0\|_n^2 \geq c\|f_0\|_{L_2}^2$$
with probability at least $1 - ???$.

## 2.1 Step 1: Bounds on graph eigenvalues

## 2.2 Step 2: Deterministic bounds

### 2.2.1 Proof of (3).

The graph $\bar{G}_{n,r}$ is a certain type of product graph, which we term the (Alden product graph). We shall show that all product graphs of this type satisfy an equality similar to (3).

**(Alden product graph).** Let $G = ([\mathcal{M}], E)$ be a graph on $\mathcal{M} \geq 1$ vertices, and let $n_1, \ldots, n_{\mathcal{M}}$ each be positive integers; let $N = \sum_{i=1}^{\mathcal{M}} n_i$. The (Alden product) graph $G^{\boxdot}(G; n_1, \ldots, n_{\mathcal{M}})$ is defined as

$$G^{\boxdot}(G; n_1, \ldots, n_{\mathcal{M}}) = \left(\bigcup_{m=1}^{\mathcal{M}} \bigcup_{i=1}^{n_m} (m,i), \quad E^{\boxdot}\right), \quad \text{where } \Big((\ell,i),(m,j)\Big) \in E^{\boxdot} \text{ if } (\ell,m) \in E.$$

We will simply write $G^{\boxdot}$ when it is clear from context what $G$ and $n_1, \ldots, n_{\mathcal{M}}$ are. Lemma 1 characterizes some of the eigenvectors of $G^{\boxdot}$.

**Lemma 1.** *For any $m \in \mathcal{M}$, and any $i \neq j \in [n_m]$, the vector*

$$g\big[(m',k)\big] = \mathbf{1}\Big\{(m',k) = (m,i)\Big\} - \mathbf{1}\Big\{(m',k) = (m,j)\Big\} \tag{5}$$

*is an eigenvector of $G^{\boxdot}$, with eigenvalue $\deg\Big((m,i); G^{\boxdot}\Big)$.*

*Proof of Lemma 1.* content... □

We can see from Lemma 1 that any eigenvector $g$ of $G^{\boxdot}$ which does not satisfy (5) must instead be piecewise constant, i.e. $g\big[(m,i)\big] = g\big[(m,j)\big]$ for all $m \in [\mathcal{M}]$ and $i,j \in [n_m]$. For any such eigenvector, and for any $f : \bigcup_{m=1}^{\mathcal{M}} \bigcup_{i=1}^{n_m} (m,i) \to \mathbb{R}$, we have that

$$\sum_{m=1}^{\mathcal{M}} \sum_{i=1}^{n_m} f\big[(m,i)\big] g\big[(m,i)\big] = \sum_{m=1}^{\mathcal{M}} \sum_{i=1}^{n_m} \bar{f}\big[(m,i)\big] g\big[(m,i)\big]$$

for $\bar{f}\big[(m,i)\big] = (n_m)^{-1} \sum_{i=1}^{n_m} f\big[(m,i)\big]$. Corollary 1 follows immediately from this reasoning.

**Corollary 1.** *Let $k$ be an integer such that $\lambda_k(G^{\boxdot}) < \deg\Big((m,i); G^{\boxdot}\Big)$. Then the eigenvectors $v_1(G^{\boxdot}), \ldots, v_k(G^{\boxdot})$ are all piecewise constant. As a result, for any function $f$,*

$$\sum_{k=1}^{\kappa} \left(\sum_{m=1}^{\mathcal{M}} \sum_{i=1}^{n_m} f\big[(m,i)\big] v_k\big[(m,i)\big]\right)^2 = \sum_{k=1}^{\kappa} \left(\sum_{m=1}^{\mathcal{M}} \sum_{i=1}^{n_m} \bar{f}\big[(m,i)\big] v_k\big[(m,i)\big]\right)^2$$

We obtain the equality (3) by applying Corollary 1 to the graph $\bar{G}_{n,r}$.

## 2.3 Step 3: Upper bound on the graph Sobolev semi-norm

## 2.4 Step 4: Lower bound on the empirical norm of the histogram estimate