Notes on 'Adaptive Non-Parametric Regression With the K-NN Fused Lasso'

Alden Green

February 14, 2019

Let $\mathbf{X} = x_1, \ldots, x_n$ be sampled i.i.d from μ with density function $p(\cdot)$ over some subset $\mathcal{X}$ of Euclidean space, and suppose

$$y_i = f_0(x_i) + \epsilon_i, \quad \epsilon_i \overset{i.i.d}{\sim} SG(\sigma^2)$$

holds for some unknown f_0. Let $\widehat{\theta}$ be the solution to the *fused lasso*

$$\widehat{\theta} := \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \left\| y - \theta \right\|_2^2 + \lambda \left\| \nabla_G \theta \right\|_1 \right\}$$

where $\lambda > 0$ is a tuning parameter, and ∇_G is an oriented incidence matrix of the graph G.

The *K-NN-FL* estimator computes the fused lasso over the K-NN graph G_K of $\mathbf{X}$. The *ϵ-FL* estimator computes the fused lasso over the ϵ graph G_ϵ.

The assumptions required for Theorems 1 and 2 are as follows.

(a) For all $x \in \mathcal{X}$
$$0 < p_{\min} \leq p(x) \leq p_{\max} < \infty$$

(b) The base measure μ in $\mathcal{X}$ satisfies
$$r^d c_{1,d} \leq \mu(B_r(x)) \leq c_{2,d} r^d \qquad (\forall x \in \mathcal{X})$$

(c) There exists a homeomorphism (continuous bijection with continuous inverse) $h : \mathcal{X} \to [0,1]^d$ such that
$$L_{\min} d_{\mathcal{X}}(x, x') \leq \left\| h(x) - h(x') \right\|_2 \leq L_{\max} d_{\mathcal{X}}(x, x') \qquad (\forall x, x' \in \mathcal{X})$$

(d) g_0 is piecewise Lipschitz[1], meaning there exists a set $\mathcal{S} \subset (0,1)^d$ such that

  (a) $\nu(\mathcal{S}) = 0$.

[1] Technically, the requirement is slightly weaker than piecewise Lipschitz.

(b) $\mu\Big(h^{-1}\big(S_\epsilon \cup ([0,1]^d \setminus \Omega_\epsilon)\big)\Big) \leq C_{\mathcal{S}}\epsilon$

(c) There exists a positive constant L_0 such that if z and z' belong to the same connected component of $\Omega_\epsilon \setminus B_\epsilon(\mathcal{S})$, then

$$|g(z) - g(z')| \leq L_0 \|z - z'\|_2$$

where $\Omega_\epsilon = [0,1]^d \setminus B_\epsilon(\partial[0,1]^d)$.

Theorem 1. *Let $K \asymp \log^{1+2r} n$ for some $r > 0$, Then under Assumptions 1-3, with an appropriate choice of the tuning parameter λ, the K-NN-FL estimator $\widehat{\theta}$ satisfies*

$$\left\|\widehat{\theta} - \theta^\star\right\|_n^2 = O_{\mathbb{P}}\left(\frac{\log^{1+2r} n}{n} + \frac{\log^{1.5+r} n}{n}\|\nabla_{G_K}\theta^\star\|_1\right)$$

This upper bound also holds for ϵ-NN-FL if we replace $\|\nabla_{G_K}\theta^\star\|_1$ with $\|\nabla_{G_\epsilon}\theta^\star\|_1$.

Theorem 2. *Under Assumptions 1-5, with an appropriate choice of the tuning parameter λ, the K-NN-FL estimator $\widehat{\theta}$ satisfies*

$$\left\|\widehat{\theta} - \theta^\star\right\|_n^2 = \widetilde{O}_{\mathbb{P}}\left(\frac{1}{n^{1/d}}\right).$$

1 Proofs

To ease proofs, we will assume $\mathcal{X} = [0,1]^d$.

Construct $G_{lat} = (V_{lat}, E_{lat})$ a lattice graph with equal side lengths in $[0,1]^d$, where

$$V_{lat} = P_{lat}(N) := \left\{\left(\frac{i_1}{N} - \frac{1}{2N}, \cdots, \frac{i_d}{N} - \frac{1}{2N}\right) : i_1, \ldots, i_d \in \{1, \ldots, N\}\right\}$$

$$(z, z') \in E_{lat} \text{ if and only if } \|z - z'\| \leq \frac{1}{N}$$

where z and $z' \in P_{lat}(N)$.

Denoting $I = P_{lat}$, we define

$$P_I(x) = \text{argmin}\left\{\|x - z'\|_\infty, z' \in P_{lat}(N)\right\}$$

Then, let $C(z) = \left\{x \in [0,1]^d : z = P_I(x)\right\}$ be the collection of cells associated with the mesh $P_{lat}(N)$, noting that $\{C(z) : z \in P_{lat}(N)\}$ defines a partition over $[0,1]^d$.

Quantization. For a given $\theta \in \mathbb{R}^n$, the *quantization* $\theta_I \in \mathbb{R}^n$

$$(\theta_I)_i := \theta_j, \quad \text{where} \quad x_j = \underset{x_l, l \in [n]}{\operatorname{argmin}} \|P_I(x_i) - x_l\|_\infty$$

is constant over every cell $C(z)$. We now induce a signal in $\mathbb{R}^{N^d}$ corresponding to the elements in I. Let $\{z_1, \ldots, z_{N^d}\} = I$. Then we write

$$I_k = \{i \in [n] : P_I(x_i) = z_k\}$$

for $k = 1, \ldots, N^d$. Define $\theta^I \in \mathbb{R}^{N^d}$ by

$$(\theta^I)_k := \begin{cases} (\theta_I)_i, x_i \in I_k \\ 0, I_k = \emptyset \end{cases}$$

where we note that (θ^I) is well-defined since $(\theta_I)_i = (\theta_I)_j$ if x_i and x_j are both in I_k.

1.1 Controlling counts of mesh

Define the event Ω as: "If $x_i \in C(z_k)$ and $x_i \in C(z_l)$ for $z_k, z_l \in I$ with $\|z_k - z_l\|_2 \leq \frac{1}{N}$, then x_i and x_j are connected in the K-NN graph." Then,

Lemma 1. *Take Assumptions 1-3, and additionally assume that N in the construction of $G_{lat}(N)$ is chosen as*

$$N \geq \left\lceil \frac{3\sqrt{d}(2c_{2,d}p_{\max})^{1/d} n^{1/d}}{L_{\min} K^{1/d}} \right\rceil. \tag{1}$$

Then,

$$\mathbb{P}(\Omega) \geq 1 - n \exp(-K/3).$$

1.2 Bounding Empirical Process

Lemma 2.

1.3 Mesh embedding for K-NN graph

Lemma 3. *Fix N to satisfy (1), and let us assume that the event Ω from Lemma 1 holds. Denote $I = P_{lat}(N)$ to be the mesh. Then, for all $e \in \mathbb{R}^n$, it holds that*

$$\left|e^T(\theta - \theta_I)\right| \leq 2 \|e\|_\infty \|\nabla_{G_K} \theta\|_1, \qquad (\forall \theta \in \mathbb{R}^n)$$

Moreover,

$$\left\|D\theta^I\right\|_1 \leq \|\nabla_{G_K} \theta\|_1, \qquad (\forall \theta \in \mathbb{R}^n)$$

where D is the incidence matrix of G_{lat}.

Proof. Clearly

$$\langle \epsilon^T, \theta - \theta_I \rangle \leq \|\epsilon\|_\infty \cdot \|\theta - \theta_I\|_1$$

Then, for every $i = 1, \ldots, n$, the event Ω implies that there exists a $j \in [n]$ such that

$$(\theta_I)_i = \theta_j, \quad (x_i, x_j) \in E_{G_K}$$

and therefore

$$\|\theta - \theta_I\|_1 \leq \|\nabla_{G_K} \theta\|_1$$

□

1.4 Bounding empirical process

Lemma 4. *Conditional on the event Ω, we have that*

$$\langle \epsilon, \widehat{\theta}_I - \widehat{\theta}_I^\star \rangle \leq \max_{u \in I} \sqrt{|C(u)|} \left(\|\Pi \widetilde{\epsilon}\|_2 \left\| \widehat{\theta} - \theta^\star \right\|_2 + \left\|(D^\dagger)^T \widetilde{\epsilon}\right\|_\infty \left[\left\| \nabla_{G_K} \widehat{\theta} \right\|_1 + \|\nabla_{G_K} \theta^\star\|_1 \right] \right)$$

where $\widetilde{\epsilon}$ is an independent, mean-zero vector of subgaussian random variables, $|C(u)| := \sum_{i \in [n]} \mathbb{I}(x_i \in C(u))$, and $D^\dagger$ is the pseudoinverse of the incidence matrix D of G_{lat}.

Proof. Writing

$$\widetilde{\epsilon}_l = \left[\max_{u \in I} |C(u)| \right]^{-1/2} \sum_{x_j \in I_l} \epsilon_j$$

we have

$$\langle \epsilon, \widehat{\theta}_I - \theta_I^\star \rangle = \left[\max_{u \in I} |C(u)| \right]^{1/2} \langle \widetilde{\epsilon}, \widehat{\theta}^I - \theta^{\star, I} \rangle$$

Now, divide up $\widetilde{\epsilon}$ into

$$\widetilde{\epsilon} = P_1(\widetilde{\epsilon}) + P_{1^\perp}(\widetilde{\epsilon})$$

where P_1 is the projection onto the span of $\mathbf{1}$ the constant vector, and $P_{1^\perp}$ the projection onto the space orthogonal to $\mathbf{1}$. Note that $P_{1^\perp}(x) = (D^\dagger D)^T x$. Then, we have

$$\begin{aligned}
\langle P_{1^\perp}(\widetilde{\epsilon}), \widehat{\theta}^I - \theta^{\star, I} \rangle &= \langle (D^\dagger D)^T \widetilde{\epsilon}, \widehat{\theta}^I - \theta^{\star, I} \rangle \\
&= \langle (D^\dagger)^T \widetilde{\epsilon}, D(\widehat{\theta}^I - \theta^{\star, I}) \rangle \\
&\leq \left\|(D^\dagger)^T \widetilde{\epsilon}\right\|_\infty \left\| D(\widehat{\theta}^I - \theta^{\star, I}) \right\|_1 \quad \leq \left\|(D^\dagger)^T \widetilde{\epsilon}\right\|_\infty \left[\left\| \nabla_{G_K} \widehat{\theta} \right\|_1 + \|\nabla_{G_K} \theta^\star\|_1 \right]
\end{aligned}$$

where the last inequality follows from the triangle inequality and Lemma 3.

On the other hand,

$$\langle P_1(\widetilde{\epsilon}), \widehat{\theta}^I - \theta^{\star, I} \rangle \leq \|P_1(\widetilde{\epsilon})\|_2 \left\| \widehat{\theta}^I - \theta^{\star, I} \right\|_2$$

and so the desired result follows. □

1.5 Proof of Theorem 1

We begin with a basic inequality

$$\frac{1}{2}\left\|\widehat{\theta}-\theta^{*}\right\|_{n}^{2} \leq \frac{1}{n}\langle\epsilon,\widehat{\theta}-\theta^{\star}\rangle+\lambda_{n}\left(\|\nabla_{G}\theta^{*}\|_{1}-\left\|\nabla_{G}\widehat{\theta}\right\|_{1}\right)$$

We split up the empirical process,

$$\langle\epsilon,\widehat{\theta}-\theta^{\star}\rangle=\langle\epsilon,\widehat{\theta}-\widehat{\theta}_{I}\rangle+\langle\epsilon,\widehat{\theta}_{I}-\theta_{I}^{\star}\rangle+\langle\epsilon,\theta_{I}^{\star}-\theta^{\star}\rangle$$

Hereafter in the proof, we condition on the event Ω. Lemma 3 gives us bounds on the first and third terms

$$\langle\epsilon,\widehat{\theta}-\widehat{\theta}_{I}\rangle \leq 2\|\epsilon\|_{\infty}\cdot\left\|\nabla_{G_K}\widehat{\theta}\right\|_{1}$$
$$\langle\epsilon,\theta^{\star}-\theta_{I}^{\star}\rangle \leq 2\|\epsilon\|_{\infty}\cdot\|\nabla_{G_K}\theta^{\star}\|_{1}$$