sketching $\to$ randomized LA

nonparametric estimation $\to$ matrix $\Rightarrow$ function

variance-reduced $\to$ specific choice of sketching

1. sketching

1.1 matrix sketching

suppose matrix $B \in \mathbb{R}^{n_1 \times n_2}$ low-rank

① construct sketch $B' = BS$, $S \in \mathbb{R}^{n_2 \times k}$

$k \ll n_1, n_2$

sketching matrix.

$\mathrm{Range}(BS) = \mathrm{Range}(B)$

② orthogonalize

$$[\underset{\downarrow\ \mathbb{R}^{n_1 \times k}}{Q}, R] = qr(B') \ / \ [\underset{\downarrow\ \mathbb{R}^{n_1 \times k}}{U}, \Sigma, V] = svd(B')$$

$$\tilde{B} = \underset{\text{rank-}k}{QQ^T} B \ / \ \tilde{B} = \underset{\text{rank-}k}{UU^T} B$$

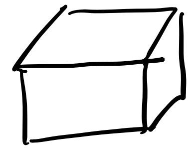
$\boxed{P = QQ^T / UU^T \ \text{projection}}$

$\|\tilde{B} - B\| \to$ small

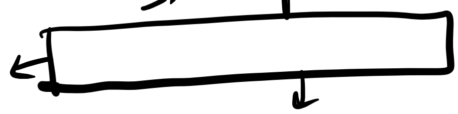
How to get P for continuous function ?

1.2 tensor sketching

$B \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}_{\text{index}(i_1, \cdots i_d)}$ $\rightarrow$ low-rank.



Step1: $B^{(1)} = B[i_1 ; i_2, i_3, \cdots i_d] \in \mathbb{R}^{n_1 \times \prod_{j=2}^{d} n_j}$



use matrix sketching for $B^{(1)}$

$S_1 : \mathbb{R}^{\prod n_j \times k_1}$

low-rank $\underline{Q_1} / U_1 \in \mathbb{R}^{n_1 \times k_1}$

low-rank.

$P_1 = Q_1 Q_1^T / P_1 = U_1 U_1^T$ $\rightarrow$ projection

Step2: repeat the procedure, $\forall j = 1, \cdots d$

$$B^{(j)} \underset{\text{permute}}{=} B[i_j, i_1, i_2, \cdots i_{j-1}, i_{j+1}, \cdots i_d]$$

$$\underset{\text{reshape}}{=} B[i_j ; i_1, i_2, \cdots i_{j-1}, i_{j+1}, \cdots i_d]$$

$$\in \mathbb{R}^{n_j \times \prod_{l \neq j} i_l}$$

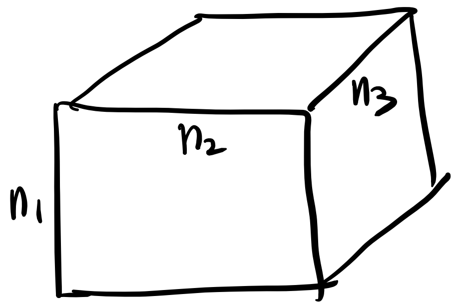
use matrix sketching for $B^{(j)}$ with S_j

get Q_j / U_j, projection $P_j = Q_j Q_j^T / U_j U_j^T$

Step3 : $\tilde{B} = B \times_1 P_1 \times_2 P_2 \times \cdots \times_d P_d$

$$(B \times_1 P_1)(i_1, \cdots, i_d) = \sum_j B(j, i_2, \cdots, i_d)\, P_1(j, i_1)$$

$B \in \mathbb{R}^{n_1 \times n_2 \times n_3}$:

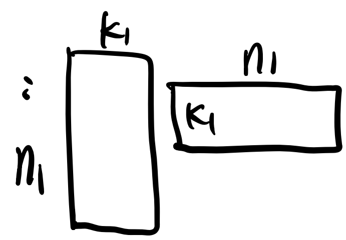


$P_1 \in \mathbb{R}^{n_1 \times n_1}$

$\downarrow$

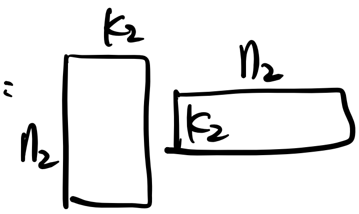
rank-k_1

P_1 :



$P_2 \in \mathbb{R}^{n_2 \times n_2}$

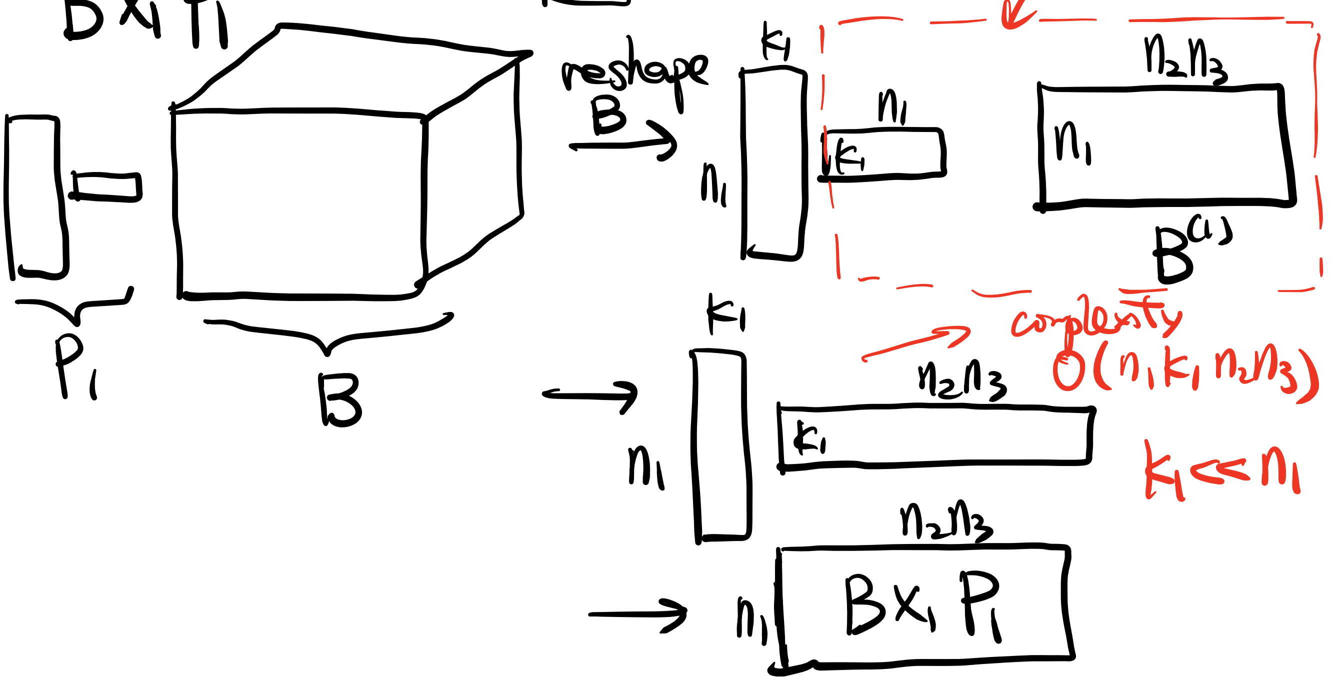
P_2 :

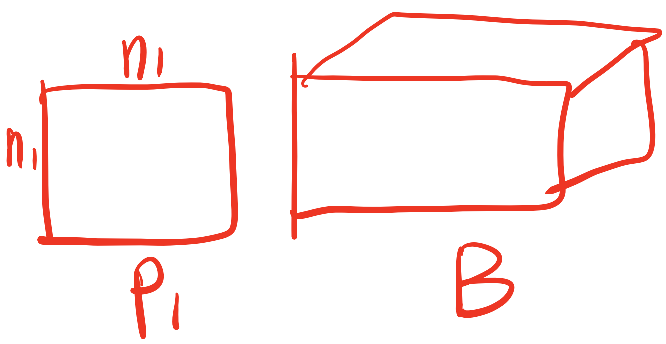


P_3 :

complexity $O(k_1 n_1 n_2 n_3)$

$B \times_1 P_1$

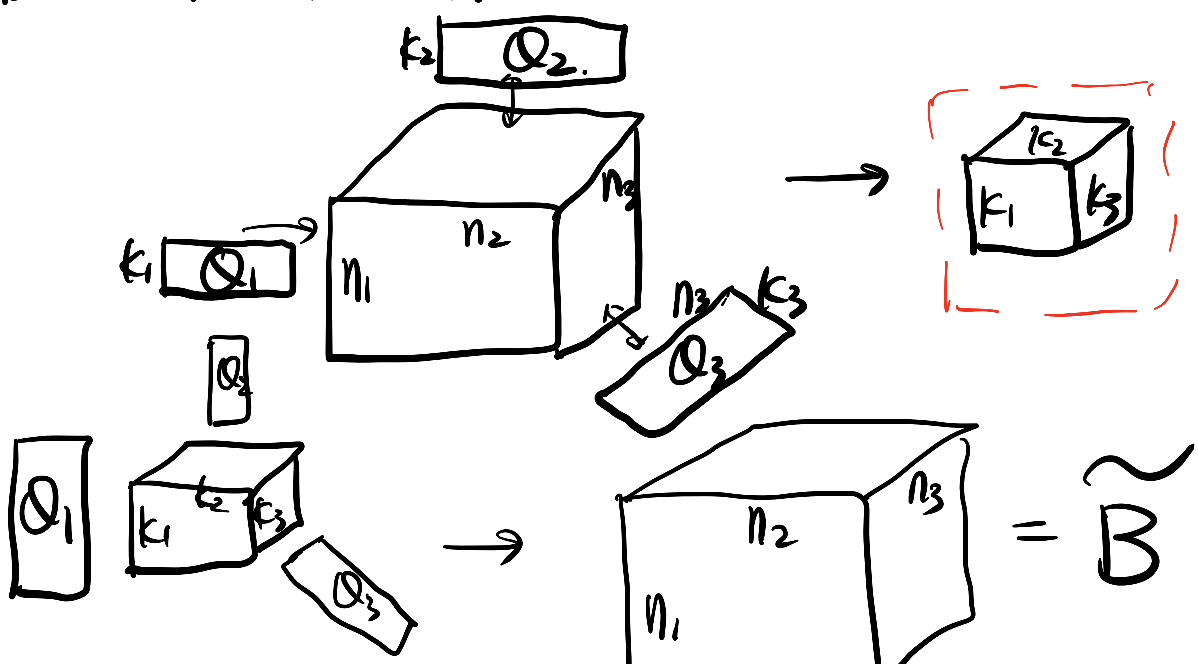




complexity

$O(n_1^2 n_2 n_3)$

$B \times_1 P_1 \times_2 P_2 \times_3 P_3$



$B \in \mathbb{R}^{n \times n \times \cdots \times n}$ $\quad O(n^d)$ $\quad k \ll n$

low-rank approximation. $O(dnk + k^d)$

2. Nonparametric estimation

2.1 Function representation

$\boxed{\text{Basis}}$

$f(x)$, $x \in \mathbb{R}^1$ $\quad \{\phi_i(x)\}$ orthonormal basis functions

Result:

$$f(x) \approx \sum_{i=1}^{m} a_i \phi_i(x)$$

if $m \to \infty$, $\|f - \sum a_i \phi_i\| \to 0$

$$a_i = \int f(x) \phi_i(x) \, dx$$

Ex: $\{\phi_i\}$: Legendre polynomial
Fourier basis

$f(x_1, x_2)$, 2-dim function $\{\phi_{i_1}(x_1)\}$, $\{\phi_{i_2}(x_2)\}$.

$$f(x_1, x_2) \approx \sum_{i_1, i_2 = 1}^{m} A_{i_1, i_2} \phi_{i_1}(x_1) \phi_{i_2}(x_2).$$

$A \in \mathbb{R}^{m \times m}$ coefficient matrix (sketching
$\downarrow$
low-rank approximation

$$\underline{\underline{A_{i_1, i_2}}} = \iint f(x_1, x_2) \phi_{i_1}(x_1) \phi_{i_2}(x_2) \, dx_1 dx_2$$

$f(x_1, \cdots x_d)$ d-dim function $\{\phi_{i_1}(x_1)\}$
$\vdots$
$\{\phi_{i_d}(x_d)\}$.

$$f(x_1,\cdots x_d) \approx \sum_{i_1,\cdots i_d=1}^{m} A_{i_1,i_2,\cdots i_d}\, \phi_{i_1}(x_1)\cdots\phi_{i_d}(x_d)$$

$A \in \mathbb{R}^{m\times m\times\cdots m}$ $\quad \underline{m}^d$) $\underline{\text{d-dim}}$, tensor

$$\boxed{A_{i_1,\cdots i_d} = \int\cdots\int f(x_1,\cdots x_d)\,\phi_{i_1}(x_1)\cdots\phi_{i_d}(x_d)\,dx_1\cdots dx_d}$$

2.2 $\boxed{\text{Density Estimation}}$. $f \xrightarrow{\text{prob}}$ density function p ($\underline{\text{pdf}}$)

Given $\{X_i\}_{i=1}^{N} \underset{\smile}{\text{i.i.d.}}\ p^*$, N: sample-size

Empirical density $\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_i}(x)$

$\hat{p} \to$ a random function based on X_i

$\int \delta_{X_i}(x)\,dx = 1$

Task: use some methods to get approximation of p^* based on $\hat{p}$

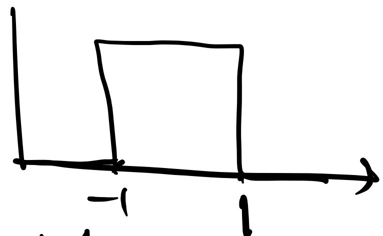
2.3 method kernel density estimation (KDE):

$$\hat{p}_h(x) = \frac{1}{N}\sum_{i=1}^{N} K_h(x-X_i) \leftarrow 1d.$$

$K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$, $K(\cdot)$ – kernel function.

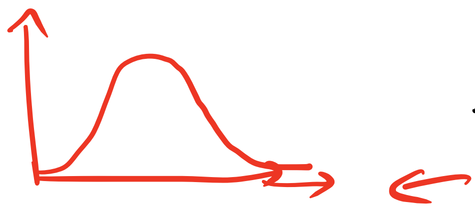
$\boxed{h \text{ — bandwidth.}}$

Ex: step function $k(x) = \frac{1}{2} I(x)$

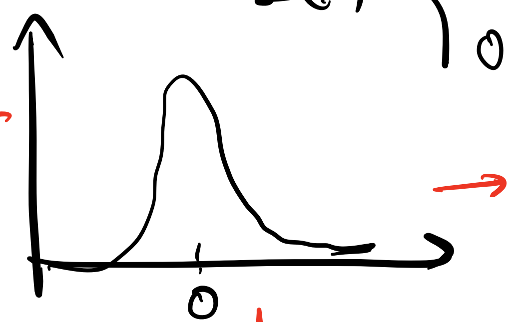


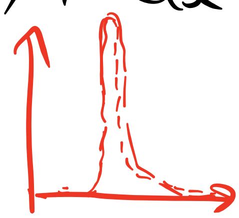
$I(x) = \begin{cases} 1, & x \in [-1, 1] \\ 0, & x \text{ else} \end{cases}$

Gaussian kernel: $k(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

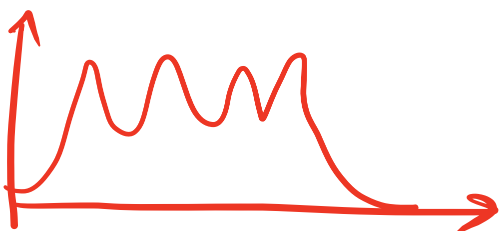


h large

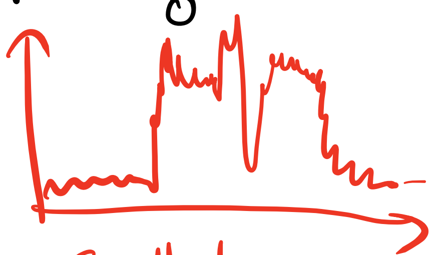




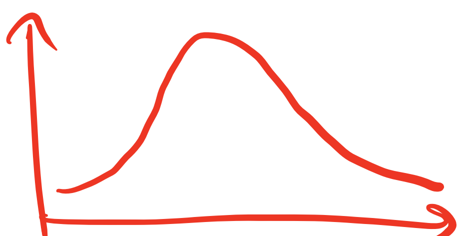
h small



True density



small-h.



large-h

high-d

$$\hat{P}_h(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h^d} k\left(\frac{\|x - X_i\|}{h}\right)$$

Error: bias error / variance error.

$$\|\mathbb{E}\hat{P}_h - p^*\| \qquad \mathbb{E}\|\hat{P}_h - \mathbb{E}\hat{P}_h\|^2$$

x

suppose p^* belongs to Holder class.

$$\Sigma(\beta, L) = \{ g : |D^s g(x) - D^s g(y)| \le L \|x-y\|, \text{ for all } s \text{ s.t. } |s| \le \beta - 1, \forall x, y \}$$

$$D^s = \frac{\partial^{s_1 + \cdots s_d}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}$$

$$\mathbb{E} \|\hat{P}_h - p^*\|^2 \le \text{bias}^2 + \text{Variance} \le c h^\beta + \frac{c}{N h^d}$$

$$h \simeq N^{-\frac{1}{2\beta + d}} \qquad \simeq \boxed{\left(\frac{1}{N}\right)^{\frac{2\beta}{2\beta + d}}}$$

$d \nearrow$, $\frac{2\beta}{2\beta+d} \to 0$

curse of dimensionality

Tensor: $\underline{\underline{n}}^d \leftarrow$ Low-rank.

keep error $\le \varepsilon$ tolerance

$$\left(\frac{1}{N}\right)^{\frac{2\beta}{2\beta+d}} \le \varepsilon$$

$$\frac{1}{\varepsilon} \le N^{\frac{2\beta}{2\beta+d}}$$

$$N \ge \left(\frac{1}{\varepsilon}\right)^{\left(1 + \frac{d}{2\beta}\right)}.$$

$d \nearrow$, $N \to$ exp in d. for fixed ε.

3. $\boxed{\text{Assumption 1:}}$ $x \in \Omega_1 \subset \mathbb{R}^{d_1}$, $y \in \Omega_2 \subset \mathbb{R}^{d_2}$

Let $A^*(x,y) : \Omega_1 \times \Omega_2 \to \mathbb{R}$ be a generic population function with $\|A^*\|_{L_2(\Omega_1 \times \Omega_2)} < \infty$.

Assume $A^*(x,y) = \sum_{\rho=1}^{r} \sigma_\rho \Phi_\rho^*(x) \Psi_\rho^*(y)$

r is const, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$

$\{\Phi_\rho^*(x)\}_{\rho=1}^r$, $\{\Psi_\rho^*(y)\}_{\rho=1}^r$ orthonormal function.

Ex: Mean-field model

$$p^*(x,y) = p_1^*(x) \cdot p_2^*(y)$$

$$p^*(z_1, \cdots z_d) = p_1^*(z_1) p_2^*(z_2) \cdot \cdots \cdot p_d^*(z_d).$$

Extension: mixed mean-field.

$$p^*(z_1, \cdots z_d) = \sum_{\rho=1}^{r} \tau_\rho\, p_{\rho,1}^*(z_1) p_{\rho,2}^*(z_2) \cdots p_{\rho,d}^*(z_d)$$

$\text{span}\{\Phi_\rho^*(x)\} = \text{Span}\{p_{\rho,1}^*(z_1)\}$, $\text{Span}\{\Psi_\rho^*(y)\}$
$= \text{Span}\{p_{\rho,2}^*(z_2) \cdots p_{\rho,d}^*(z_d)\}$

$\boxed{\text{rank at most } r}$

How to get $\underline{P}$? projection. Ex: Lengendre poly ... basis.

sketching: matrix $S \rightarrow$ function $\{w_j(\underline{y})\}$

$\boxed{\dim(\mathcal{L})}$

space $\mathcal{L} = \mathrm{Span}\{w_j(y)\}_{j=1}$

$B \cdot S$ ← matrix multi

① sketching stage: $\left\{ \int_{\Omega_2} \boxed{\hat{A}(x,y)} w_j(y)\, \underset{\text{integral}}{dy} \right\}_{j=1}^{\dim(\mathcal{L})}$

a function solely depends on x

$[BS(:,1), BS(:,2), \cdots BS(:,k)]$

② Estimation Stage: space $\mathcal{M} = \mathrm{Span}\{v_\mu(x)\}_{\mu=1}^{\dim(\mathcal{M})}$

basis function.

$\tilde{f}_j(x) = \underset{f \in \mathcal{M}}{\mathrm{argmin}} \left\| \boxed{\int_{\Omega_2} \hat{A}(x,y) w_j(y) dy} - \underline{f(x)} \right\|^2_{L_2(\Omega_1)}$

$\underline{\tilde{f}_j}(x) = \sum_{\mu=1}^{\dim(\mathcal{M})} B_{\mu,j} v_\mu(x)$,

$B_{\mu,j} = \left\langle \int_{\Omega_2} \hat{A}(x,y) w_j(y) dy,\ v_\mu(x) \right\rangle$

coefficient matrix

$B \in \mathbb{R}^{\dim(\mathcal{M}) \times \dim(\mathcal{L})}$

$= \boxed{\iint \hat{A}(x,y)\, v_\mu(x)\, w_j(y)\, dx\, dy}$

③ Orthogonalization do QR/SVD → $\underline{\underline{B}}$

$$[U, \Sigma, V] = \mathrm{svd}(B)$$

$U \to \mathbb{R}^{\dim(M)\times r}$ (truncate at rank-r).

$$\hat{\Phi}_\rho(x) = \sum_{\mu=1}^{\dim(M)} v_\mu(x)\, U_{\mu,\rho}\,, \quad \rho = 1, \cdots r.$$

projection operator $P_x(x, x') = \sum_{\rho=1}^{r} \hat{\Phi}_\rho(x)\, \hat{\Phi}_\rho(x')$

$\to UU^T \to U(x)\, U^T(x')$

Algorithm:

compute coefficient B:

$$B = \frac{1}{N}\sum_{i=1}^{N} v_\mu(x_i)\, w_\eta(y_i).$$

when $\hat{A}(x,y) = \frac{1}{N}\sum_{i=1}^{N} \delta_{(x_i, y_i)}(x, y)$

Function estimation. (2-d function).

$\dim(L_2) \to \frac{1}{G_r}$.

$\dim(M_2) \to$

① $\boxed{\{\hat{\Phi}^{(x)}_{\rho}(x)\}_{\rho=1}^{r}}$.

$\hat{A}(x,y) \to \begin{cases} \text{sketching space } \{w_\eta(x)\}_{\eta=1}^{\dim(L_2)} \\ \text{estimation space } \{v_\mu(y)\}_{\mu=1}^{\dim(M_2)} \end{cases}$

② $\xrightarrow{\text{Algorithm}}$ $\boxed{\{\hat{\Phi}^{(y)}_{\rho}(y)\}_{\rho=1}^{r}}$.

$\hat{A}$ is matrix

① projection: row $\underbrace{U_1 U_1^T \hat{A}}$

② projection: column $\hat{A} U_2 U_2^T$

③ Together $\tilde{A} = U_1 \boxed{U_1^T \hat{A} U_2} U_2^T$ $\to G$

③ coefficient core.

$$\underline{G_{\rho_1,\rho_2}} = \iint \hat{A}(x,y)\, \hat{\Phi}^{(x)}_{\rho_1}(x)\, \hat{\Phi}^{(y)}_{\rho_2}(y)\, dx\, dy$$

$\in \mathbb{R}^{r\times r}$ $\stackrel{\text{density}}{=} \frac{1}{N}\sum_{i=1}^{N} \underline{\hat{\Phi}^{(x)}_{\rho_1}(x_i)\, \hat{\Phi}^{(y)}_{\rho_2}(y_i)}$.

Represent function

$$\tilde{A}(x,y) = \sum_{\rho_1,\rho_2=1}^{r} \underline{G_{\rho_1,\rho_2}}\, \hat{\Phi}^{(x)}_{\rho_1}(x)\, \hat{\Phi}^{(y)}_{\rho_2}(y).$$

$$\tilde{A} = \hat{A} \times_1 \underline{P_1} \times_2 \underline{P_2}$$

Statistical analysis . $x \in \mathbb{R}^{d_1}, y \in \mathbb{R}^{d_2}$

Assumption 2: let M, L with $\dim(M) = m^{d_1}$,
$\dim(L) = l^{d_2}$,

$$\| A^* - A^* \times_x P_M \times_y P_L \|_{L_2}^2 = O\left(m^{-2\alpha} + l^{-2\alpha}\right)$$

α: parameter of space (containing A^*).

Ex: A^* in Sobolev ~~space~~ W_2^α : up to α derivative is bounded in L_2 norm

$$\|A^*\|_{W_2^\alpha}^2 = \left(\sum_{|s|=0}^{\alpha} \| D^s A^* \|_2^2 \right) < \infty$$

$$D^s = \frac{\partial^{s_1 + \cdots s_d}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}$$

$(s_1 + \cdots s_d = |s|)$ $x \in \mathbb{R}$

Intuition: basis in M : polynomial.

$f(x)$

$\dim(M)$: how many orthogonal polynomial

$\hookrightarrow$ degree of polynomial

to each dimensional

order of polynomial $\nearrow$.

Error $\searrow$.

$\dim(M) = m^{d_1}, x \in \mathbb{R}^{d_1}$

Ex: $\alpha = 2$

W_2^2, $f \in W_2^2$, $\|f\|_{L_2} < \infty$, $\|f'\|_{L_2} < \infty$,

$\|f''\|_{L_2} < \infty$ for $\forall x$

Assumption 3: for any test function $u(x,y)$,

① $\mathbb{E}_{\text{samples}} \langle \hat{A}, u \rangle = \langle A^*, u \rangle$ 1st moment ② $\sup_{\|u\|_{L_2} \le 1} \mathrm{Var}[\langle \hat{A}, u \rangle] = O\left(\frac{1}{N}\right)$

N: sample-size. 2nd moment.

$$\begin{cases} \langle \hat{A}, u \rangle = \iint \hat{A}(x,y)\, u(x,y)\, dx\, dy \\ \langle A^*, u \rangle = \iint A^*(x,y)\, u(x,y)\, dx\, dy \end{cases}$$

$\hat{A} = \frac{1}{N} \sum \delta_i$, ②: $\sup |A^*(x,y)| < \infty$

Theorem 1 (informal): suppose above assumptions hold,

$$\|\tilde{A} - A^*\|_{L_2}^2 = O\left(\frac{1}{N^{2\alpha/(2\alpha+d_1)}} + \frac{1}{N^{2\alpha/(2\alpha+d_2)}} \right)$$

$x \in \mathbb{R}^{d_1}$, $y \in \mathbb{R}^{d_2}$

with choice $m_1 \approx N^{\frac{1}{2\alpha+d_1}}$, $m_2 \approx N^{\frac{1}{2\alpha+d_2}}$

m_1: $\dim(M_1) \to$ variable x m_2: variable y.

sketching size

$l_1 = l_2 = C \sigma_r^{-\frac{1}{\alpha}}$. σ_r: r-th singular value of A^*

bias-variance trade-off.

KDE: $O\left(N^{-2\alpha/(2\alpha+d_1+d_2)}\right)$

Remark: slightly improvement for 2-d function.
huge improvement for high-d function

3.3 high-d function estimation.

Assumption 4: (modification of low-rank..

$$A^*(x_1, \cdots x_d) = \sum_{\rho=1}^{r_j} \sigma_{j,\rho} \Phi^*_{j,\rho}(x_j) \Psi^*_{j,\rho}(x_1, \cdots x_{j-1}, x_{j+1}, \cdots x_d)$$

for $j = 1, \cdots d$

sketching space: num of basis for x_j

$$M_j = \text{span}\{\phi_\mu(x_j)\}_{\mu=1}^{m}$$

$d-1$.

$$L_j = \text{span}\{\phi_{\eta_1}(x_1) \cdots \phi_{\eta_{j-1}}(x_{j-1}) \phi_{\eta_{j+1}}(x_{j+1}) \cdots \phi_{\eta_d}(x_d)\}_{\eta_1, \cdots \eta_d = 1}^{l_j}$$

$\underline{\dim(L_j) = l_j^{d-1}}$

$\left\{ \begin{array}{l} m \text{ could be large} \\ l_j \text{ small} \end{array} \right.$

$l_j \to$ sketching.

Thm2: (informal) Suppose above assumption holds, choose $m \simeq N^{\frac{1}{2\alpha+1}}$, $c_j = C\sigma_{j,r_j}^{-\frac{1}{\alpha}}$,

then

$$\|\tilde{A} - A^*\|_{L_2}^2 = O\left(\frac{1}{N^{2\alpha/(2\alpha+1)}}\right)$$

Remark1: intutive: $M_j \to x_j \in \mathbb{R}$.

d-1 dimensions are sketched. $\frac{1}{N^{2\alpha/(2\alpha+1)}}$

Remark2: we sketch d-1 dimension

$\dim(L_j) = l^{d-1}$

image l is small when A^* is good.