Chapter 13: Nonparametric Least Squares:

13.1

a) $G(t) = \mathbb{E}[(z-t)^2]$

$= \mathbb{E}[(z-\mathbb{E}[z])^2]$

$= \mathbb{E}[z^2] - 2\mathbb{E}[z]\mathbb{E}[\mathbb{E}[z]] + \mathbb{E}[\mathbb{E}[z]]^2$

$= \sigma^2$

(Tower Property, Law or Total Expectation)

$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}(X|Y)]$

b) (Equation 13.1) $\tilde{\mathcal{L}}_F := \mathbb{E}_{X,Y}[(Y-f^*(x))^2]$

$\tilde{\mathcal{L}}_F := \mathbb{E}_{X,Y}[(Y-f^*(x))^2]$

$= \mathbb{E}_{X,Y}[Y^2] - 2\mathbb{E}_{X,Y}[Y]\mathbb{E}[f^*(x)] + \mathbb{E}[f^*(x)]^2$

$= \mathbb{E}_{X,Y}[Y^2] - 2\mathbb{E}_{X,Y}[Y]\mathbb{E}[\mathbb{E}[Y|X=x]] + \mathbb{E}[\mathbb{E}[Y|X=x]]^2$

$= \mathbb{E}_{X,Y}[Y^2] - 2\mathbb{E}_{X,Y}[Y]\cdot\mathbb{E}_{X,Y}[Y] + \mathbb{E}[Y]^2$

$= \sigma_{XY}^2$

c) $\tilde{\mathcal{L}}_F - \tilde{\mathcal{L}}_{F^*} = \mathbb{E}_{X,Y}[(Y-f(X))^2] - \mathbb{E}[(Y-f^*(x))^2]$

$= \mathbb{E}_{X,Y}[(f(X)-f^*(x))^2]$

$= \|f-f^*\|_2^2$

13.2

$\mathbb{E}[\|f_\theta - f_{\theta^*}\|_n^2] = \mathbb{E}[\|f_\theta\|^2] - 2\mathbb{E}[\|f_\theta\|]\circ\mathbb{E}[\|f_{\theta^*}\|] + \mathbb{E}[\|f_{\theta^*}\|]^2$

$= \frac{\|X(\theta-\theta^*)\|_n^2}{n}$

$= \frac{\|X\|_n^2}{n}\sigma^2$

$= \text{rank}(X)\cdot\sigma^2/n$

Relationship/Test:

$\|X\|_n^2 = \left(\sum_{i=1}^{N}\sum_{j=1}^{m} X_{ij}^n\right)^{2/n} = \text{rank}(X)$

when $X_{ij} \geq 1$

13.3 (Equation 13.10) "Cubic Spline"

$$\hat{f} \in \underset{f}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \int_0^1 f''(x)^2 \, dx \right\}$$

Book's method: $\hat{f} \in \underset{f}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \int_0^1 f''(x)^2 dx \right\}$

$$= \underset{f}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \|f\|^2 \right.$$

where $f(x) = \langle f, K(\cdot, x) \rangle$

$$= \sum_{j=1}^{n} f \cdot \phi_i(x) \phi_j(x)$$

$$= \sqrt{n} K \alpha$$

$$\|f\|^2 = \frac{1}{n} \langle \sum \alpha K(\cdot, x), \sum \alpha K(\cdot, x) \rangle$$

$$= \alpha^T K \alpha$$

$$\hat{f} \in \underset{f}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x))^2 + \lambda_n \|f\|^2 \right\}$$

$$\in \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \sqrt{n} K \alpha)_2^2 + \lambda_n \alpha^T K \alpha \right\}$$

$$= 0$$

$$\hat{\alpha} = \frac{1}{(K + \lambda_n I)} \left(\frac{y}{\sqrt{n}} \right)$$

$$f(x) = \theta_0 + \theta_1 x + \sum_{i=1}^{n} \hat{\alpha}_i K(x, x)$$

$\boxed{\text{Alternative}}$ Method #1: A minimal function sums terms

$$\overset{*}{f}(x) = \hat{f}(x) + f(x)$$

Minimal Function — Original "Linear" — correction terms

$$= \theta_0 + \theta_1 x + \sum_{i=1}^{N} \alpha K(x, x)$$

$\boxed{\text{Alternative}}$

Method #2:

$$\hat{f} = \underset{\hat{f}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x))^2 + \lambda_n \int_0^1 f''(x)^2 dx \right\}$$

$$= \underset{f, \hat{\alpha}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x))^2 + \lambda_n \left(\sum_{i=1}^{n} \hat{\alpha} K \right)^2 \right.$$

$$= \underset{f, \hat{\alpha}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \frac{y_i^2}{n} - \frac{2}{n} \sum_{i=1}^{N} y_i f(x) + \sum_{i=1}^{N} \frac{f(x)^2}{n} + \lambda \left(\sum_{i=1}^{n} \hat{\alpha} K \right)^2 \right\}$$

$$= 0$$

$$f^*(x) = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$= \frac{\frac{2}{n} \sum_{i=1}^{N} y_i \pm \sqrt{\left(\frac{2}{n} \sum_{i=1}^{N} y_i\right)^2 - 4(1/n)\left(\sum_{i=1}^{N} \frac{y_i^2}{n} + (\lambda \sum_{i=1}^{n} \hat{\alpha} K)^2 \right)}}{2 \cdot (1/n)}$$

$$= y_i \pm \sqrt{\lambda \left(\sum_{i=1}^{n} \hat{\alpha} K \right)^2 / n}$$

$$= \theta_0 + \theta_1 x \pm \frac{1}{\sqrt{n}} \sum \hat{\alpha} K(x, x)$$

"Again, an assumption about a linear term."

b) $(\hat{\theta}, \alpha) = \underset{(\theta, \alpha) \in \mathbb{R}^2 \times \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{2n} \| y - X\theta - \sqrt{n} K \alpha \|_2^2 + \lambda_n \alpha^T K \alpha \right\}$

where $K \in \mathbb{R}^{n \times n}$ and $X \in \mathbb{R}^{n \times 2} = \begin{bmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$

$$f^*(\hat{\theta}, \alpha) = \theta_0 x_0 + \theta_1 x_1 + \sqrt{n} K \alpha$$

$$= X \cdot \theta - \sqrt{n} \sum_{i=1}^{n} \hat{\alpha} K(x, x)$$ from part a.

(Star Domain)

$x, y \in \mathbb{R}^n$

$[x,y] = \{tx + (1-t)y \,;\, 0 \le t \le 1\}$

$= x - (y-x)[0,1]$

(Star Set) A subset C, $x \in C$ and $x^* \in C$ for every

$[x^*, x] \subseteq C$, has a $x^* + (x - x^*)\alpha \in C$.

when $\alpha \in \mathbb{R}$

13.4.

a) $[x^*, x] = \{\underbrace{\alpha x + (1-\alpha) x^*}\,;\, 0 \le \alpha \le 1\}$

"The star domain is an affine relationship or weighted-function"

b) The definition is above in star set.

13.5. (Equation 13.71) $\frac{G_n(\delta;F)}{\delta} \le \frac{\delta}{2\sigma}$

Where $G(\delta;F) = \mathbb{E}_w\left[\sup_{\substack{g \in F \\ \|g\| \le \sigma}} \left|\frac{1}{n}\sum_{i=1}^{n} w_i\, g(x_i)\right|\right]$

a) If $\delta^2 = 4\sigma^2$, then $\frac{G_n(\delta,F)}{\delta} \le \frac{\delta}{2\sigma}$

$\le \frac{2\sigma}{2\sigma}$

≤ 1

b) If $\delta^2 \ge \min\left[1, \frac{8}{\pi}\frac{\sigma^2}{n}\right]$, then $\frac{G_n(\delta,F)}{\delta} \le \delta/2\sigma$

$$\leq \frac{\min\left[1, \sqrt{\frac{8}{n\pi}}\,\sigma\right]}{2\sigma}$$

$$\leq \min\left[\tfrac{1}{2\sigma}, \sqrt{\tfrac{2}{n\pi}}\right]$$

13.6.

a) $G_n(\delta; F^*(1)) := \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} w_i g(X_i)\right| \geq \delta\right]$

$$= \mathbb{P}\left[\left|\sum_{i=1}^{n} w_i g(X_i)\right| \geq n\delta\right]$$

$$= \mathbb{P}\left[\lambda\left|\sum_{i=1}^{n} w_i g(X_i)\right| \geq n\lambda\delta\right]$$

$$= \mathbb{P}\left[e^{1+\lambda\mathbb{E}\left[\left|\sum_{i=1}^{n} w_i g(X_i)\right|\right] + \frac{\lambda^2\mathbb{E}\left[\left|\sum_{i=1}^{n} w_i g(X_i)\right|^2\right]^2}{2} + \cdots} \geq e^{n\lambda\delta}\right]$$

$$\leq \mathbb{P}\left[e^{\frac{\lambda^2\mathbb{E}\left[\left|\sum_{i=1}^{n} w_i g_i(x)\right|^2\right]^2}{2}} \geq e^{n\lambda\delta}\right]$$

$$\operatorname*{argmin}_{\lambda}\left\{\frac{\lambda^2 \cdot \mathbb{E}\left[\left|\sum_{i=1}^{n} w_i g_i(x)\right|^2\right]^2}{2} - n\lambda\delta\right\} = 0$$

$$\lambda^* = n\delta / \mathbb{E}\left[\left|\sum_{i=1}^{n} w_i g_i(x)\right|^2\right]^2$$

$$= n\delta/\sigma^2$$

$$G_n(\delta; F^*(1)) \leq \mathbb{P}\left[e^{\frac{\lambda^2\mathbb{E}\left[\left|\sum_{i=1}^{n} w_i g_i(x)\right|^2\right]^2}{2}} \geq e^{n\lambda\delta}\right]$$

$$\leq \mathbb{P}\left[e^{n^2\delta^2/2\sigma^2} \geq e^{n^2\delta^2/\sigma^2}\right.$$

$$\leq \operatorname{tr}\left(e^{n^2\delta^2/2\sigma^2}\right) \cdot e^{-n^2\delta^2/\sigma^2}$$

$$\leq \operatorname{rank}(\sigma^2) \cdot e^{n^2\delta^2/2\sigma^2 - n^2\delta^2/\sigma^2}$$

$$\leq n \cdot e^{-n^2\delta^2/2\sigma^2}$$

$\boxed{\text{Note}}$

- The trace and rank coefficient derive from Theorem 6.12/6.15.
- The derivation fits a maximal limit to the trace of the largest eigenvalue.

$G(\sigma, F(1)) = \sigma$

$= \sqrt{\frac{2\sigma^2 \log n}{n}}$ ... for a $\mathbb{P}[X] = $ Normal whole $= 1$

$= c_1 \sqrt{\frac{\log n}{n}}$

$\|\hat{\theta} - \theta^*\|_2^2 = \sigma$

$= c_1' \sigma \sqrt{\frac{\log n}{n}}$ where $c_1 = \sqrt{2}\sigma$, $c_1' = \sqrt{2}$

b) From part a, $G(\sigma, F(1)) = \sqrt{\frac{2\sigma^2 \log n}{n}}$

$= \frac{c_2 \sigma \sqrt{\log n}}{n}$ when $c_2 = \sqrt{\frac{2}{n}}\sigma$

$\|\hat{\theta} - \theta^*\|_2^2 = \sqrt{\frac{2\sigma^2 \log n}{n}}$

$= \frac{c_2' \sigma^2 \log n}{n}$ when $c_2' = \sigma\sqrt{\frac{\log n}{n}}$

13.7. $\mathbb{P}_m = \{f_\theta : \mathbb{R}_n \to \mathbb{R}_n \mid \theta \in \mathbb{R}^m\}$ where $f_\theta(x) = \sum_{j=0}^{m-1} \theta_j x^j$

$$\mathbb{P}\left[\|\hat{f}-f^*\|_n^2 \geq c_0 \frac{\sigma^2 m \log n}{n}\right] = \mathbb{P}\left[\lambda\|\hat{f}-f^*\|_n^2 \geq c_0 \frac{\lambda\sigma^2 m \log n}{n}\right]$$

$$:= \mathbb{P}\left[e^{1+\lambda\mathbb{E}[\|\hat{f}-f^*\|_n^2] + \frac{\lambda^2\mathbb{E}[\|\hat{f}-f^*\|_n^4]^2}{2} + \cdots} \geq e^{c_0\frac{\lambda\sigma^2 m \log n}{n}}\right]$$

$$\leq \mathbb{P}\left[e^{\frac{\lambda^2\mathbb{E}[\|\hat{f}-f^*\|_n^4]^2}{2}} \geq e^{c_0\frac{\lambda\sigma^2 m \log n}{n}}\right]$$

$$\underset{\lambda}{\operatorname{argmin}}\left\{\lambda^2 \cdot \frac{\mathbb{E}[(\hat{f}-f^*)^2]^2}{2} - c_0\frac{\lambda\sigma^2 m \log n}{n}\right\} = 0$$

$$\lambda^* = \frac{c_0\sigma^2 m \log n}{n \cdot \mathbb{E}[(\hat{f}-f^*)^2]^2}$$

$$\mathbb{P}\left[\|\hat{f}-f^*\|_n^2 \geq c_0 \frac{\sigma^2 m \log n}{n}\right] \leq \mathbb{P}\left[e^{\frac{\lambda^2\mathbb{E}[(\hat{f}-f^*)_n^2]^2}{2}} \geq e^{c_0\frac{\lambda\sigma^2 m \log n}{n}}\right]$$

$$\leq \mathbb{P}\left[e^{\frac{c_0^2\sigma^4 m^2 \log^2 n \cdot \mathbb{E}[(\hat{f}-f^*)^2]^2}{n^2\mathbb{E}[(\hat{f}-f^*)^2]^4}} \geq e^{\frac{c_0^2\sigma^4 m^2 \log^2 n}{n^2\mathbb{E}[(\hat{f}-f^*)^2]^2}}\right]$$

$$\leq \operatorname{tr}\left(e^{\frac{2c_2 m \log n}{}}\right) \cdot e^{-c_2 m \log n}$$

$$\leq m \cdot e^{-c_2 m \log n}$$

$$\leq c_1 e^{-c_2 m \log n}.$$

<u>Notes</u>: Two approximations appear in the books derivation, 1) Taylor Expansion
2) Trace/Eigenvalue upper limit.

where $c_1 = m$, $c_2 = c_0 \left(\frac{\sigma}{n}\right)^2 m \log n$

13.8. $\mathbb{P}\left[\|\hat{f} - f^*\|_n^2 \geq c_0 \left(\frac{\sigma^2}{n}\right)^{4/5}\right]$

The book offers other representations with standard deviation [σ^2], Lipschitz constant [L] and a random constant [δ]. Past experiences involved a random constant or scalable-Lipschitz constant, rather than standard deviation, with manual and most probable fits. An example is an arbitrary intensity adjustment in an image where standard deviation is unknown without a reference in the image. A best guess becomes a decrease from maximum intensity toward a random intensity constant or slope-dependendenz constant, such as Lipschitz.

$$\mathbb{P}\left[\|\hat{f} - f^*\|_n^2 \geq c_0 \left(\frac{\sigma^2}{n}\right)^{4/5}\right] = \mathbb{P}\left[\|\hat{f} - f^*\|_n^2 \geq c_0 \left(\frac{\delta n}{n L^{9/5}}\right)^{1/3}\right]$$

$$:= \mathbb{P}\left[e^{1+\lambda \mathbb{E}[\|f-f^*\|_n^2]^2 + \frac{\lambda^2 \mathbb{E}[\|f-f^*\|_n^2]^2}{2} + \cdots} \geq e^{c_0 \lambda \left(\frac{\sigma^2}{n}\right)^{4/5}}\right]$$

$$\leq \mathbb{P}\left[e^{\lambda^2 \frac{\mathbb{E}[\|f-f^*\|_n^2]^2}{2}} \geq e^{c_0 \lambda \left(\frac{\sigma^2}{n}\right)^{4/5}}\right]$$

$$\underset{\lambda}{\operatorname{argmin}} \left\{ \lambda^2 \frac{\mathbb{E}[\|f-f^*\|_n^2]^2}{2} - c_0 \lambda \left(\frac{\sigma^2}{n}\right)^{4/5} \right\} = 0$$

$$\lambda^* = \frac{c_0 \left(\frac{\sigma^2}{n}\right)^{4/5}}{\mathbb{E}[\|f-f^*\|_n^2]^2}$$

$$\mathbb{P}\left[\|\hat{f}-f^*\|_n^2 \geq c_0 \left(\frac{\sigma^2}{n}\right)^{4/5}\right] \leq \mathbb{P}\left[e^{\frac{\lambda^2 \mathbb{E}[\|\hat{f}-f^*\|_n^2]^2}{2}} \geq e^{c_0 \lambda (\sigma^2/n)^{4/5}}\right]$$

$$\leq \operatorname{tr}\left(e^{\frac{c_0^2 \left(\frac{\sigma^2}{n}\right)^{2/5} \mathbb{E}[\|\hat{f}-f^*\|_n^4]^2}{2\,\mathbb{E}[\|\hat{f}-f^*\|^4]^4}}\right) \cdot e^{\frac{-c_0^2 \left(\frac{\sigma^2}{n}\right)^{8/5}}{\mathbb{E}[\|f-f\|^2]^2}}$$

$$\leq n \cdot e^{c_0^2 (\sigma^2/n)^{-1/5}}$$

$$\leq n \cdot e^{-c_0^2 (n/\sigma^2)^{1/5}}$$

$$\leq c_1 e^{-c_2 (n/\sigma^2)^{1/5}}$$

where $c_1 = n$; $c_2 = c_0^2$

13.9:

a) $$\underbrace{\frac{\sigma}{n}\left|\sum_{i=1}^{n} w_i \hat{\Delta}(x_i)\right|}_{\text{Optimal Mean squared error}} \leq \underbrace{\|\Delta\|_?^2}_{\text{Feasible Mean squared error}} + \underbrace{\|\Delta\|_?^2}_{\text{Error}}$$

The book suggests: $(\|\Delta\|_?^2 \leq t\delta) + (\|\Delta\|_?^2 \geq t\delta)$

I scoured the book for a "t" variable. With no hints, t represents a threshold. The value of t is unknown and arbitrary. I guess t-table and Students t-distribution for a threshold-cutoff. Cases involve size(d) and degrees of freedom
Knowledge prior to variance.

Proof by Exhaustion:

Case #1 : $\|\Delta\|_j^2 \le t \cdot \delta$, $\frac{\sigma}{n} \| \sum_{i=1}^{n} w_i g(x) \| \le d \cdot t \cdot \delta$

Case #2 : $\|\Delta\|_j^2 \ge t\delta$, $\frac{\sigma}{n} \| \sum_{i=1}^{n} w_i g(x) \| \le 2 \cdot \| t\delta \cdot \Delta \|$

$\le 2\sqrt{t\delta}\,\|\Delta\|$

$$\frac{\sigma}{n} \| \sum_{i=1}^{n} w_i g(x) \| \le d \cdot t \cdot \delta + 2\sqrt{t \cdot \delta}\,\|\Delta\|$$

b) Suppose $\sqrt{\sum_{j=1}^{n} \|g_j\|_n^2} \le \sqrt{K}\, \| \sum_{j=1}^{d} g_j \|_n$

$$\|\hat{f} - f^*\|_n^2 = \left(\sqrt{K}\, \| \sum_{j=1}^{d} g_j \|_n\right)^2$$

$$\le tr\left(K \cdot \| \sum_{j=1}^{d} g_j \|_n^2\right)$$

$$\le d \cdot K \cdot \delta_{n,max}^2$$

13.10 a) $\hat{f} = \min_{\theta \in \mathbb{R}^T} \left\{ \frac{1}{n} \|y - \Phi\theta\|_2^2 + \lambda_n \|\theta\|_2^2 \right\}$

$$= -\frac{2}{n}\Phi(y - \theta\Phi) + 2\lambda\theta$$

$$= 0$$

$$\theta^* = \frac{y\Phi}{n(\Phi^2 + \lambda I)}$$

b) $\inf_{f \in F(1:T)} \|f - f^*\|_2^2 = \inf \|\sum_{i=1}^{T} (\theta_m - \theta_m)\phi_m\|_2^2 + \inf \|\sum_{i=T+1}^{\infty} \theta_m^* \phi_m\|_2^2$

$$= \|\theta_{1:T}^* - \theta_{1:T}\|_2^2 + \|\theta_m^*\|_2^2$$

$$= \|\theta_m^*\|_2^2$$

13.11.

a) Function class:

$$F(1:T) = \left\{ \beta_0 + \sum_{i=1}^{\alpha} (\beta_m \phi_m + \bar{\beta}_m \phi_m) \,\middle|\, \beta_0 + \sum_{i=1}^{\alpha} (\beta_m \phi_m + \bar{\beta}_m \phi_m) \leq 1 \right\}$$

From the problem,

$$\int_0^1 f^{(\alpha)}(mx)^2 dx \leq R$$

$$m^{2\alpha} \int_0^1 f^{(\alpha)}(x)^2 dx \leq R$$

$$\int_0^1 f^{(\alpha)}(x)^2 dx \leq \frac{R}{m^{2\alpha}}$$

"Second term in the kernel problem from Chapter 12"

$$\sum (\beta_m^2 \phi_m + \beta_m^2 \phi_m) \leq \frac{R}{m^{2\alpha}}$$

$$\beta_m^2 + \bar{\beta}_m^2 \le \frac{cR}{m^{2\alpha}}$$

where c = constant

b) (Equation 13.47) $\|f - f^*\| = \beta_m^2 + \bar{\beta}_m^2$

$$\le \frac{cR}{M^{2\alpha}}$$ where c = arbitrary constant within a bound.