

# L1 Review

[Big  $O(\cdot)$ ]

$f(z) = O(g(z))$  as  $z \rightarrow z_0 \in \mathcal{R}$  if

$$\left| \frac{f(z)}{g(z)} \right| \leq M$$

for some  $M > 0$ , and for all  $z$  in neighborhood of  $z_0$ .

If  $z \rightarrow \infty$ , then there exists  $C > 0$  s.t. statement holds for all  $z > C$

[Small  $o(\cdot)$ ]

$f(z) = o(g(z))$  as  $z \rightarrow z_0 \in \mathcal{R}$  if

$$\lim_{z \rightarrow z_0} \frac{f(z)}{g(z)} = 0$$

[Taylor's Expansion]

Let  $f(\cdot)$  defined on  $[a, b]$  s.t. it has continuous  $(n+1)$ th order derivatives. Then for all  $x, x_0$  in  $[a, b]$

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2}{2!}f''(x_0) + \cdots + \frac{(x - x_0)^n}{n!}f^{(n)}(x_0) + R_n$$

where

$$R_n = \frac{(x - x_0)^{n+1}}{(n+1)!}f^{(n+1)}(\xi) = O(|x - x_0|^{n+1})$$

for some  $\xi \in (x, x_0)$  or  $(x_0, x)$

[Alternate Taylor]

Since  $f^{(n+1)}(\cdot)$  is bounded based on theorem condition

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \cdots + \frac{(x - x_0)^n}{n}f^{(n)}(x_0) + O(|x - x_0|^{n+1})$$

as  $x \rightarrow x_0$

[Multivariate Taylor expansion]

Let  $x = (x_1, x_2)^T, y = (y_1, y_2)^T$

$$f(x + y) = f(x) + y_1f_1(x) + y_2f_2(x) + R$$

$$R = \frac{1}{2}y_1^2f_{11}(\xi) + y_1y_2f_{12}(\xi) + \frac{1}{2}y_2^2f_{22}(\xi) = O(\|y\|^2)$$

and  $\xi = \alpha x + (1 - \alpha)(x + y)$  for some  $\alpha \in [0, 1]$

[Likelihood Inference]

$X_1, \dots, X_n$  be iid with  $f(x|\theta)$ , then likelihood of  $X_1 = x_1, \dots, X_n = x_n$  is

$$L(\theta) = \prod_{i=1}^n f(x_i|\theta)$$

Likelihood principle find  $\theta$  that maximises  $L(\theta)$ . Log-likelihood =  $\ell(\theta) = \log L(\theta)$ . Score function  $s(\theta) = \ell'(\theta)$

[Asymptotic Normality of MLEs]

[Convergence Order]

A root-finding method has convergence order  $\beta$  ( $\geq 1$ ) if

(a)  $\lim_{t \rightarrow \infty} \epsilon_t = 0$

(b)  $\lim_{t \rightarrow \infty} \frac{|\epsilon_{t+1}|}{\epsilon_t}^\beta = c$  for some  $c > 0$

When  $\beta = 1$ , we require  $c < 1$

[Matrix Digression]

Given  $y, z$  not orthogonal to each other, find symmetric matrix  $M$  s.t.  $y = Mz$

[[ Solution 1 ]]  $y^T z$  is scalar,  $M = \frac{yy^T}{y^T z}$

[[ Solution 2 ]] Given any symmetric matrix  $M_0$ , let  $v = y - M_0 z$ .  $M = M_0 + \frac{vv^T}{v^T z}$

[[ Solution 3 ]]  $M = M_0 - \frac{(M_0 z)(M_0 z)^T}{z^T M_0 z} + \frac{yy^T}{y^T z}$

## Optimisation

[Optimisation in Uni-variate: find  $x^*$  s.t.  $g'(x^*) = 0$ ]

[Bisection] Condition:  $g'(a) > 0, g'(b) < 0, g'(x)$  exist and continuous for all  $x \in (a, b)$   
 Let  $x_0 = (a + b)/2$ , set  $\tilde{a} = a, \tilde{b} = b, t = 0$   
 (1.1) If  $g'(x_{t-1}) > 0, X_t = (x_{t-1} + \tilde{b})/2, \tilde{a} = x_{t-1}$   
 (1.2) If  $g'(x_{t-1}) < 0, X_t = (\tilde{a} + x_{t-1})/2, \tilde{b} = x_{t-1}$   
 (2)  $t = t + 1$ , terminate when  $|x_t - x_{t-1}| < \epsilon$

[Modified Bisection] Instead of choosing the mid-point, we can choose

$$x_t = \frac{|g'(b)|}{|g'(a)| + |g'(b)|}a + \frac{|g'(a)|}{|g'(a)| + |g'(b)|}b$$

[Newton's Method]

$$x_{t+1} = x_t - \frac{g'(x_t)}{g''(x_t)}$$

[Fisher Scoring] Replace Hessian  $\ell''(\theta_t)$  in Newton method by  $-I(\theta_t)$

$$-I(\theta) = nE \left\{ \frac{d^2}{d\theta^2} \log f(X|\theta) \right\} = \sum_{i=1}^n \frac{d^2}{d\theta^2} \log f(x_i|\theta)$$

$$\theta_{t+1} = \theta_t + \frac{\ell'(\theta_t)}{I(\theta_t)}$$

[Secant Method] Approximate Hessian  $g''(x) = \lim_{y \rightarrow x} \frac{g'(y) - g'(x)}{y - x}$ , assuming update is small, i.e.  $|x_{t-1} - x_t| < \epsilon$

$$g''(x_t) \approx \frac{g'(x_{t-1}) - g'(x_t)}{x_{t-1} - x_t}$$

$$x_{t+1} = x_t - g'(x_t) \frac{x_t - x_{t-1}}{g'(x_t) - g'(x_{t-1})}$$

[Fixed-point Iteration] Let  $g'(a) > 0, g'(b) < 0$ . Assume  $\exists x^* \in [a, b], \epsilon \in (0, \frac{1}{2})$  s.t.

$(1 - \epsilon)(x^* - x) \geq g'(x) \geq \epsilon(x^* - x)$  for  $x < x^*$

$(1 - \epsilon)(x^* - x) \leq g'(x) \leq \epsilon(x^* - x)$  for  $x > x^*$

Then  $x_{t+1} = x_t + g'(x_t)$  converges to  $x^*$

[Optimisation in Multivariate]

[Newton's Method, Fisher scoring] Similar to single variable method, with  $g' = \nabla g, g'' = \nabla^2 g$

[Newton-like method] General form with  $-M_t$  a positive definite matrix

$$x_{t+1} = x_t - \alpha_t [M_t]^{-1} g'(x_t)$$

[Ascent Algorithm: Bracketing] Ascent algo: Control for  $\alpha_t$  s.t.  $g(x_{t+1}) \geq g(x_t)$

$$x_{t+1} = x_t + \alpha_t g'(x_t)$$

Bracketing:

(1) start with  $\alpha_t = 1$ , compute  $x_{t+1}$

(2) if  $g(x_{t+1}) < g(x_t)$ ,  $\alpha_t$  is too large and update  $\alpha_t = 1/2$

[Discrete Newton] Approximate Hessian  $g''$  by discrete version, with  $e_1 = (1, 0)^T, e_2 = (0, 1)^T$ , some small  $h_{ij} > 0$

$$M_{ij}^{(t)} = \frac{g_i(x_t + h_{ij}e_j) - g_i(x_t)}{h_{ij}}$$

To ensure symmetry, consider

$$N_{ij}^{(t)} = \frac{M_{ij}^{(t)} + M_{ji}^{(t)}}{2}$$

[Quasi-Newton] Estimate Hessian with  $g'(x_t) - g'(x_{t-1}) = M_t(x_t - x_{t-1})$ .

Consider  $y = g'(x_t) - g'(x_{t-1}), z = x_t - x_{t-1}, M_t = M_{t-1} + \frac{y y^T}{z^T z}$

If  $1/(v^T z) \leq 0, -M_0 \succ 0 \Rightarrow -M \succ 0$

If  $1/(v^T z) > 0, M_t = M_{t-1} + \alpha_t v v^T$  with  $\alpha_t > 0$  s.t.  $-M \succ 0$

[Gaussian-Newton] Model  $y_i = f(z_i, \theta) + \epsilon_i, \epsilon_i \sim N(0, \tau)$  iid, then  $\theta = (Z^T Z)^{-1} Z^T y$  (linear) else  
 $\theta_{t+1} = \theta_t + [A_t^T A_t]^{-1} A_t^T x_t$

[Nonlinear Gauss-Seidel] Restrict update to one co-ordinate at a time, find  $x_1^*, x_2^*$  s.t.  $g_1(x_1^*, x_2^*) = 0, g_2(x_1^*, x_2^*) = 0$   
 Iterate with  $g_1(x_1^{(t+1)}, x_2^{(t)}) = 0, g_2(x_1^{(t+1)}, x_2^{(t+1)}) = 0$

**L2: EM Optimization**

[EM] Want to solve  $\hat{\theta} = \arg \max \ell_X(\theta)$  with some missing data  $Z$ .  
Therefore, consider  $Y = (X, Z)$  complete data instead.  $\ell_Y(\theta) = \ell_X(\theta) + \ell_{Z|X}(\theta)$ .  
Solve for

$$Q(\theta|\theta^{(t)}) = E_{\theta^{(t)}} [\ell_Y(\theta)|X]$$

with (1) E-step: Compute  $Q(\theta|\theta^{(t)})$  (2) M-step: Maximise  $Q$  with respect to  $\theta$  and set  $\theta^{(t+1)} = \theta^*$   
Only requires:  $\ell_X(\theta^{(t+1)}) > \ell_X(\theta^{(t)})$  (generalised EM)

[EM for Canonical Exp Fam] Canonical Exp Fam has log-likelihood linear in missing data  $Z$  and observed data  $X$ . Check before solving (1) impute  $Z$  (2) estimate  $\theta^{(t+1)}$

$$\ell_Y(\theta) = c(Y) + d(\theta) + \sum_{j=1} p\theta_j Y_j$$

$$Q(\theta|\theta^{(t)}) = c(Y) + d(\theta) + \sum_{j=1}^p \theta_j E_{\theta^{(t)}}(Y_j|X)$$

[Var estimate of  $\hat{\theta}$ ] Fisher information  $I(\theta) = E_{\theta}[-\ell''_X(\theta)] = var_{\theta}(\ell'_X(\theta))$   
MLE asymptotic dist  $I(\theta)^{-1/2}(\hat{\theta} - \theta_0) \rightarrow N(0, I_K)$   
Fisher info for complete data  $i_Y(\theta) = i_X(\theta) + i_{Z|X}(\theta) \Rightarrow i_X = i_Y - i_{Z|X}$  (note the variance estimate  $\hat{\theta}$  is wrt to  $i_X$ )  
BS-MC estimate  $\hat{i}_Y(\theta) = -\frac{1}{m} \sum_{i=1}^m \ell''_{Y^{(k)}}(\theta)$ ,  $\hat{i}_{Z|X}(\theta) = -\frac{1}{m} \sum_{i=1}^m \ell''_{Z^{(k)}}(\theta)$

**Extended EM**

[MC-EM] Instead of calculating  $Q(\theta|\theta^{(t)})$  via integration, use MC instead.

[Expected Conditional Max] Instead of maximising  $\theta = (a, b)$  at once, maximise them sequentially  
(a)  $\max_a Q(a, b^{(t)}|\theta^{(t)})$  (b)  $\max_b Q(a^{(t+1)}, b|\theta^{(t)})$  (c)  $\theta^{(t+1)} = (a^{(t+1)}, b^{(t+1)})$

[EM Gradient] Instead of solving maximisation analytically, use gradient-based methods (e.g. Newton).  
 $\theta^{(t+1)} = \theta^{(t)} - Q''(\theta|\theta^t)^{-1}|_{\theta=\theta^t} \times Q'(\theta|\theta^t)|_{\theta=\theta^{(t)}}$

**EM Acceleration Methods**

[Convergence rate] EM est  $\hat{\theta}$  converge to  $\theta$  at linear rate, depending on fraction of observed information  $\rho(\theta) = \frac{i_X(\theta)}{i_Y(\theta)}$

[Aitken Acceleration] Use Newton method for optim (Quad rate) and estimate  $\ell_X(\theta)$  using EM with

$$\rho(\theta) = \frac{i_X(\theta)}{i_Y(\theta)} = 1 - \frac{i_{Z|X}(\theta)}{i_Y(\theta)}$$

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\theta_{EM}^{(t)} - \theta^{(t)}}{\rho(\theta^{(t)})}$$

[Quasi-Newton Acceleration] Avoid estimating  $\rho(\theta)$ ,  $\rho(\theta) \approx 1 - \frac{\theta_{EM}^{(t)} - \theta_{EM}^{(t-1)}}{\theta^{(t)} - \theta^{(t-1)}}$

$$\theta^{(t+1)} = \theta^{(t)} + (I - M^{(t)})^{-1}(\theta_{EM}^{(t)} - \theta^{(t)})$$

### L3: Numerical Integration

[Integration]

Objective: approximate  $\int_a^b f(x)dx$  numerically

Naive method: Divide  $[a, b]$  into  $n$  sub-intervals,  $x_i^*$  is the middle point of  $i$ th subinterval.

$$\int_a^b f(x)dx \approx \frac{b-a}{n} \sum_{i=1}^n f(x_i^*)$$

Improvement: for each of the sub-interval  $[x_i, x_{i+1}]$  add  $(m+1)$  nodes

[Trapezoidal Rule]

Choose 2 nodes ( $m=1$ ) in  $[x_i, x_{i+1}]$ . To approximate height  $I = \frac{\int_{x_i}^{x_{i+1}} f(x)dx}{x_{i+1}-x_i}$ . Area =  $(x_{i+1} - x_i) \times I$

$$\hat{I}_1 = \frac{f(x_0^*) + f(x_1^*)}{2}$$

Total area  $\int_a^b f(x)dx$ , with  $h = (b-a)/n$

$$\hat{T}(n) = h \sum_{i=1}^n \frac{f(x_i) + f(x_{i+1})}{2}$$

$$\hat{T}(n) - \int_a^b f(x)dx = O(n^{-2})$$

[Simpson Rule]

Choose 3 nodes ( $m=2$ ). Approximate height  $I$

$$\hat{I}_2 = \frac{1}{6}f(x_0^*) + \frac{4}{6}f(x_1^*) + \frac{1}{6}f(x_2^*)$$

Total area  $\int_a^b f(x)dx$ , with  $h = (b-a)/n$ ,  $x_i^* = (x_i + x_{i+1})/2$

$$\hat{S}(n) = h \sum_{i=1}^n \left\{ \frac{f(x_i)}{6} + \frac{2f(x_i^*)}{3} + \frac{f(x_{i+1})}{6} \right\}$$

$\hat{S}(n) - \int_a^b f(x)dx = O(n^{-4})$ , can be generalised to other polynomial order  $m$

[Gaussian Quadrature] Perfect est for polynomial order  $2m+1$  and below (or fn close enough) using  $2m+2$  points.

$$I = \int_a^b w(x)f(x)dx = \sum_{j=0}^m c_j f(x_j)$$

when  $a, b$  finite,  $w(x) = 1$ ; when  $a = 0, b = \infty$ ,  $w(x) = e^{-x}$ ; when  $a = -\infty, b = \infty$ ,  $w(x) = e^{-x^2/2}$

Requires solving for  $x_0, \dots, x_m$  and  $c_0, \dots, c_m$  ( $2m+2$  unknowns)

#### L4: Bootstrap

[Nonparametric]	Re-sample with replacement and estimate $E(f(X))$ with $\frac{1}{B} \sum_{b=1}^B f(X^{(b)})$
[Parametric]	First estimate $\hat{\theta}$ (e.g. with MLE) then generate samples from $F_{\hat{\theta}}(x)$ . require assumption on parametric form.
[BS techniques]	Paired BS: generate BS samples by pairing $Z_i = (x_i, y_i)$ BS residual: generate est $y_i^*$ by bootstrapping $\hat{e}_i^*$ Bias correction: $\text{bias} = \frac{1}{B} \sum_{k=1}^B (\hat{\theta}_k^* - \hat{\theta})$ , correct estimate with $\hat{\theta} - \text{bias}$
[BS Percentile CI]	90% BS CI for $\theta = (\hat{\theta}_{(5)}^*, \hat{\theta}_{(95)}^*)$ Only works well if $\hat{\theta} - \theta$ does not depend on $\theta$ and is symmetric about 0
[BS t CI]	Consider $\frac{\hat{\theta} - \theta}{\hat{\sigma}}$ instead, let $d_k^* = \frac{\hat{\theta}_k^* - \hat{\theta}}{\hat{\sigma}_k^*}$ , 90% CI for $\theta$ is $(\hat{\theta} - \hat{\sigma} d_{(95)}^*, \hat{\theta} - \hat{\sigma} d_{(5)}^*)$
[Balanced BS]	Reduce MC error from some observed $X_i$ are too frequently selected by chance. (1) Generate every $X_i$ exactly $B$ times. (2) Permute/re-order the samples (3) first $n$ is assigned to first BS sample
[Antithetic BS]	Reduce MC error by enforcing data pairing. (1) Generate $B$ data (2) second sample is replacing $X_{(k)}$ with $X_{(n-k+1)}$

## L5: Simulation and MC Integration

[MC integration]	Estimate $\mu = E[h(X)]$ , generate $X_i$ from $f(x)$ (known) $\hat{\mu}_{MC} = \frac{1}{n} \sum_{i=1}^n h(X_i)$ $\hat{\sigma}_{MC}^2 = \frac{1}{n-1} \sum_{i=1}^n [h(X_i) - \hat{\mu}_{MC}]^2$ MC estimate: $\hat{\mu}_{MC} \pm \hat{\sigma}/\sqrt{n}$
[Extract Simulation]	Simulate samples from $f(x)$ directly if $F^{-1}(U)$ exist and known, and is single-varaite (1) Generate $U \sim \text{Unif}(0, 1)$ (2) $X = F^{-1}(U)$ Known distributions such as Gaussian, Beta have special algorithm.
[Rejection Sampling]	Assume $f(x)$ can be computed easily, find proposal density $Y \sim g$ s.t. $f(x) \leq g(x)/\alpha$ for known $\alpha > 0$ If $\alpha f(Y)/g(y)$ is small, then algo is inefficient. (1) Generate $Y \sim g$ (2) Generate $U \sim \text{unif}(0, 1)$ (3) If $U \leq \alpha f(Y)/g(Y)$ , set $X = Y$ (4) Else, repeat (1-3) until succeed
[SIR]	Sampling Importance Resampling, with envelope function $g(x)$ Generate approximate distribution from $f(x)$ (previous 2 methods are exact). (1) Sample $Y_i, \cdot, Y_m$ from $g(x)$ (2) Calculate standardised importance weight $w(Y_1), \dots, w(Y_m)$ $w(Y_i) = \frac{f(Y_i)/g(Y_i)}{\sum_{j=1}^m f(Y_j)/g(Y_j)}$ (3) Resample $X_i$ from $Y_1, \dots, Y_m$ with probability $w(Y_1), \dots, w(Y_m)$
[Sequential MC]	Splitting high-dimensional task into sequence of simpler steps, each step updates the previous one. Goal: simulate $X_{1:t}^{(i)}, i = 1, \dots, n$ iid from $f(x_{1:t})$ (1) Sample $X_1 \sim g(x_1)$ . Let $w_1 = u_1 = f(x_1)/g(x_1)$ . set $t = 2, X_{1:t-1} = X_1$ (2) Sample $X_t = g(x_t X_{1:t-2})$ (3) Append $X_t$ to $X_{1:t-1}$ . Obtain $X_{1:t}$ (4) Let $u_t = f(X_t X_{1:t-1})/g(X_t X_{1:t-1})$ (5) Let $w_t = w_{t-1}u_t$ (6) Increase $t$ by 1 and return to step (2)
[SISR]	When $t$ increases $w_t^{(i)}$ may have large variability and reduce sampling efficiency. Effective sample size $\hat{N}_t = \frac{n}{1+cv_t^2}, cv_t^2 = \sum_{i=1}^n (w_t^{(i)} - \bar{w}_t)^2 / (n\bar{w}_t^2), \bar{w}_t = \sum_{i=1}^n w_t^{(i)} / n$ (1) When $\hat{N}_t$ is smaller than predetermined threshold, stop SIS (2) Resample $n$ sequences from $\{X_{1:t}^{(1)}, \dots, X_{1:t}^{(n)}\}$ with probability $\{w_t^{(1)}, \dots, w_t^{(n)}\}$ , set weight for new resampled seq as $1/n$ (3) Use resample sequences and weights as inputs for next step in SIS algo

## Variance Reduction

[Importance Sampling]	$\mu = E[h(X)] = \int h(x)w(x)g(x)dx, w(x) = \frac{f(x)}{g(x)}$ $\hat{\mu}_{IS} = \frac{1}{n} \sum_{i=1}^n h(X_i)w(X_i)$
[Antithetic Sampling]	Find two unbiased estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ that are negatively correlated

$$\hat{\mu}_{AS} = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$$

[Control Variates]	Generate 2 sets of samples $\{(X_i, Y_i)\}, \mu = E[h(X)], \theta = E(c(Y))$
--------------------	--

$$\hat{\mu}_{CV} = \hat{\mu}_{MC} + \lambda(\hat{\theta}_{MC} - \theta)$$

$$\text{with } \lambda_{\min} = -\frac{\text{cov}(h(X), c(Y))}{\text{var}(c(Y))}$$

[Rao-Blackwellization]	Remove randomness from some vectors by solving conditional expectation. Consider $X = (X_1, X_2), \mu = E(h(X)) = E[E(h(X) X_2)] = E(\tilde{h}(X_2))$
------------------------	--

$$\hat{\mu}_{RB} = \frac{1}{n} \sum_{i=1}^n \tilde{h}(X_{i2})$$

## L6: Markov Chain Monte Carlo

[MCMC] Objective: generate stationary distribution s.t.  $X_t \sim f(x) \Rightarrow X_{t+1} \sim f(x)$

[Independence Chains] Proposal distribution  $g(x)$ ,  $w(x) = f(x)/g(x)$

(1) Generate  $X_1 \sim g(x)$ , let  $t = 1$

(2) Generate  $Y \sim g(x)$ ,  $U \sim Unif(0, 1)$

(2.1) If  $U \leq w(Y)/w(X_t)$ ,  $X_{t+1} = Y$

(2.2) If  $U > w(Y)/w(X_t)$ ,  $X_{t+1} = X_t$

(3) Increase  $t$  by 1

(4) Repeat steps (2) and (3) to generate  $X_1, X_2, \dots$

Basically,

$$R(X_t, Y) = \frac{f(Y)g(X_t)}{f(X_t)g(Y)}$$

[Metropolis-Hasting] (1) Generate  $X_1$  from arbitrary initial distribution and set  $t = 1$

(2) Simulate  $Y \sim g(y|X_t)$

(3) Compute MH ratio  $R(X_t, Y)$

$$R(X_t, Y) = \frac{f(Y)g(X_t|Y)}{f(X_t)g(Y|X_t)}$$

(4) Generate  $U \sim Unif(0, 1)$ ,

(4.1) If  $U \leq R(X_t, Y)$ ,  $X_{t+1} = Y$

(4.2) Otherwise,  $X_{t+1} = X_t$

(5) Increase  $t$  by 1

(6) Repeat steps (2)-(5) to generate MC chain  $X_1, X_2, \dots$

[Gibbs Sampling] (1) Simulate  $X_1 = (X_{11}, X_{12})$  from arbitrary distribution, set  $t = 1$

(2) Simulate  $X_{t+1|1} \sim f_1(x_1|X_{t,2})$  and then simulate  $X_{t+1,2} \sim f_2(x_2|X_{t+1,1})$

(3) Increase  $t$  by 1 and repeat (2)

## L7: Non-parametric Density Estimation

[Measure of Performance]     ISE: Integrated squared error

$$ISE(\hat{f}(x)) = \int \left\{ \hat{f}(x) - f(x) \right\}^2 dx$$

MSE: mean squared error

$$MSE(\hat{f}(x)) = E \left[ \left\{ \hat{f}(x) - f(x) \right\}^2 \right] = \text{bias}^2\{\hat{f}(x)\} + \text{var}\{\hat{f}(x)\}$$

MISE: mean integrated squared error

$$\begin{aligned} MISE(\hat{f}(x)) &= E \left\{ ISE(\hat{f}(x)) \right\} = \int MSE(\hat{f}(x)) dx \\ &= \int \text{bias}^2\{\hat{f}(x)\} + \int \text{var}\{\hat{f}(x)\} \end{aligned}$$

[Naive Estimators]

$$\hat{f}_n(x) = \frac{\hat{F}_2(x+h) - \hat{F}_n(x-h)}{2h} = \frac{1}{2nh} (\# \text{ of } X_1, \dots, X_n \text{ in } (x-h, x+h])$$

Equivalently,

$$w(x) = I(|x| < 1) \frac{1}{2}$$

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n w\left(\frac{x - X_i}{h}\right)$$

[Kernel Density Estimators]      $h$  bandwidth - most important hyper-parameter,  $K(\cdot)$  kernel function,  $K_h(x) = K(y/h)/h$  bandwidth-rescaled kernel function

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i)$$

[Kernel Function]

Non-negative function  $K(\cdot)$  with following condition, usually a pdf

$$(1) \int_{-\infty}^{\infty} K(x) dx = 1$$

$$(2) \int_{-\infty}^{\infty} x K(x) dx = 0$$

$$(3) \int_{-\infty}^{\infty} x^2 K(x) dx = \sigma_k^2 > 0$$

Common kernel:

$$\text{Uniform: } K(t) = \frac{1}{2} I(|t| < 1)$$

$$\text{Gaussian (most popular): } K(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$$

$$\text{Epanechnikov (most popular): } K(t) = \max(0.75(1 - t^2), 0)$$

$$\text{Biweight } K(t) = \max(15/16(1 - t^2)^2, 0)$$

[Unbiased C-V]

UCV is a better approach than conventional Cross Validation

$$\min_h UCV(h) = \int \hat{f}_n^2(x) dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i,n}(x_i)$$