2020270026 应用统计硕士 王姿文

---

1. Kernel Methods

Problem 1

1

已知$k(x, y) = 1 + x^T y,\ \forall X = \mathbb{R}^d,\ F = \mathbb{R}^{d+1}$，
根据$dot\ product\ of\ polynomials$和$kernel\ \ function$，

$$\varphi(x) = (1, x),\ \varphi(y) = (1, y),\ k(x, y) = (1, x)\begin{pmatrix} 1 \\ y \end{pmatrix} = 1 + x^T y$$

$$\Rightarrow \text{若让}d+1\text{，则}\varphi(x) = (1, \sqrt{2}x, x),\ \varphi(y) = (1, \sqrt{2}y, y),$$

$$k(x, y) = (1, \sqrt{2}x, x)\begin{pmatrix} 1 \\ \sqrt{2}y \\ y \end{pmatrix} = (1 + x^T y)^2$$

$$\Rightarrow \text{可以推出，}For\ General: k(x, y) = (1 + xy)^n,\ \forall X = \mathbb{R}$$

2

从理论来看，$k(x, y) = \langle \varphi(x), \varphi(y) \rangle$，故会是两个同样$function$的$matrix$做内积。
欲使$k(x, y) = xy - 1 = -1 + xy$，则可令$\varphi(x) = (i, x),\ \varphi(y) = (i, y)$

$$\Rightarrow k(x, y) = (i, x)\begin{pmatrix} i \\ y \end{pmatrix} = -1 + xy$$

然而上述推导与$k(x, y) = xy - 1,\ \forall X = \mathbb{R}$相矛盾，因为$i$是虚数，故$k(x, y) = xy - 1,\ \forall X = \mathbb{R}$不是$kernel function$

3

下证$k(x, y) = min(x, y),\ \forall X = [0, 1]$是$kernel\ \ function$:
欲证$min(x, y) = \langle \varphi(x), \varphi(y) \rangle$
$min(x, y) = \int \mathbb{1}_{(0,x)} \mathbb{1}_{(0,y)} = \langle \mathbb{1}_{(0,x)}, \mathbb{1}_{(0,y)} \rangle = \langle \varphi(x), \varphi(y) \rangle$
$\Rightarrow k(x, y) = min(x, y),\ \forall X = [0, 1]$是$kernel\ \ function$

Problem 2

1

已知

$\hat{w} = min\frac{1}{2} \parallel w \parallel^2 \;\; s.t. \;\; y_i w^T x_i \geq 1 \; \forall \; i = 1, \ldots, N$

$\Rightarrow \; Lagrangian \; function \; L(w, \alpha) = \frac{1}{2} \parallel w \parallel^2 - \sum_{i=1}^{N} \alpha_i (y_i w^T x_i - 1) \; \forall \; \alpha = (\alpha_1, \ldots, \alpha_N)^T \geq 0$

故可推出

$\hat{w} = min_{w \in \mathbb{R}^m, \xi \in \mathbb{R}^N} \frac{\lambda}{2} \parallel w \parallel^2 + \sum_{i=1}^{N} \xi_i \; s.t. \; \xi_i \geq 0, \; y_i w^T x_i \varphi(x_i) \geq 1 - \xi_i$

$\Rightarrow \; Lagrangian \; function \; L(w, \alpha) = \frac{\lambda}{2} \parallel w \parallel^2 + \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i (y_i w^T x_i \varphi(x_i) - 1 + \xi_i)$

$\forall \; \alpha = (\alpha_1, \ldots, \alpha_N)^T \geq 0$

2

已知$L(w, \alpha) = \frac{\lambda}{2} \parallel w \parallel^2 + \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i (y_i w^T x_i \varphi(x_i) - 1 + \xi_i) \; \forall \; \alpha = (\alpha_1, \ldots, \alpha_N)^T \geq 0$,
$k(x, y) = \langle \varphi(x), \varphi(y) \rangle$，欲推导$dual \; problem$，则先求$\hat{w}$和$\alpha$的范围:

- $\frac{d}{dw} L \mid_{\hat{w}} = 0 \;\Rightarrow\; \lambda w = \sum_{i=1}^{N} \alpha_i y_i x_i \varphi(x_i) \;\Rightarrow\; \hat{w} = \frac{1}{\lambda} \sum_{i=1}^{N} \alpha_i y_i x_i x_i^m$
  $\frac{d}{d\xi} L \mid_{\hat{\xi}} = 0 \;\Rightarrow\; 1 - \sum_{i=1}^{N} \alpha_i = 0 \;\Rightarrow\; \sum_{i=1}^{N} \alpha_i = 1 \;\Rightarrow\; 0 \leq \alpha \leq 1$
- $L(\hat{w}, \alpha) = \frac{\lambda}{2} \parallel \hat{w} \parallel^2 + \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i (y_i \hat{w}^T x_i \varphi(x_i) - 1 + \xi_i)$
  $= \frac{\lambda}{2} \parallel \frac{1}{\lambda} \sum_{i=1}^{N} \alpha_i y_i x_i x_i^m \parallel^2 + \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i (y_i (\frac{1}{\lambda} \sum_{i=1}^{N} \alpha_i y_i x_i x_i^m)^T x_i \varphi(x_i) - 1 + \xi_i)$
  $= \frac{\lambda}{2} \parallel \frac{1}{\lambda} \sum_{i=1}^{N} \alpha_i y_i x_i x_i^m \parallel^2 + \xi^T 1 + \alpha^T 1 - \xi^T \alpha - \sum_{i=1}^{N} \alpha_i (y_i (\frac{1}{\lambda} \sum_{i=1}^{N} \alpha_i y_i x_i x_i^m)^T x_i x_i^m)$
  $= \alpha^T 1 - \frac{1}{2} \alpha^T Y G Y \alpha \; \forall \; Y = diag(y_1, \ldots, y_N), \; G \in \mathbb{R}^{N \times N}, \; G_{ij} = x_i^t T x_j$

  $\Rightarrow \; \hat{\alpha} = max_{0 \leq \alpha \leq 1} \alpha^T 1 - \frac{1}{2} \alpha^T Y G Y \alpha$

3

$f(x) = sign(\hat{w}^T \varphi(x)) = sign(\frac{1}{\lambda} \sum_{i=1}^{N} \alpha_i y_i x_i \varphi(x_i)^T \varphi(x))$，
因為$\varphi(x)^T \varphi(x) = (x^T x)^d = k(x, x)$，所以$f(x) = sign(\frac{1}{\lambda} \sum_{i=1}^{N} \alpha_i y_i x_i x_i^T x)$

Problem 3

1

已知$min_{w\in\mathbb{R}^m}\frac{\lambda}{2}\mid\mid w\mid\mid^2+\sum_{i=1}^{N}(w^T\varphi(x_i)-y_i)^2$，欲求$\hat{w}$:

- 先令$\varphi(x_i)=x_i$来求解$\hat{w}$:
  $\Rightarrow min_{w\in\mathbb{R}^m}\frac{\lambda}{2}\mid\mid w\mid\mid^2+\sum_{i=1}^{N}(w^Tx_i-y_i)^2$
  $\Rightarrow \frac{d}{dw}=\lambda w+2\sum_{i=1}^{N}(w^Tx_i-y_i)x_i=0$
  $\Rightarrow \lambda w=2\sum_{i=1}^{N}(y_i-w^Tx_i)x_i$
  $\Rightarrow \lambda w=2\sum_{i=1}^{N}y_ix_i-2\sum_{i=1}^{N}w^Tx_ix_i$
  $\Rightarrow \hat{w}=(\lambda I+2\sum_{i=1}^{N}x_ix_i^T)^{-1}(2\sum_{i=1}^{N}y_ix_i)$
- 再把$x_i=\varphi(x_i)$代回去$\hat{w}$:
  $\hat{w}=(\lambda I+2\sum_{i=1}^{N}x_ix_i^T)^{-1}(2\sum_{i=1}^{N}y_ix_i)$
  $=(\lambda I+\varphi(x)\varphi(x)^T)^{-1}2\varphi(x)y$
  $=\varphi(x)2(\varphi(x)^T\varphi(x)+\lambda I)^{-1}y$
  $=\sum_{i=1}^{N}\alpha_i\varphi(x_i)\ \forall\ \alpha=(2(\varphi(x)^T\varphi(x)+\lambda I)^{-1}y)$得解

2

$f(x)=\hat{w}^T\varphi(x)$
$=\sum_{i=1}^{N}\alpha_i\varphi(x_i)^T\varphi(x)$
$=\sum_{i=1}^{N}\alpha_ix_i^Tx\ \forall\ \alpha=(2(\varphi(x)^T\varphi(x)+\lambda I)^{-1}y)$

Problem 4

1

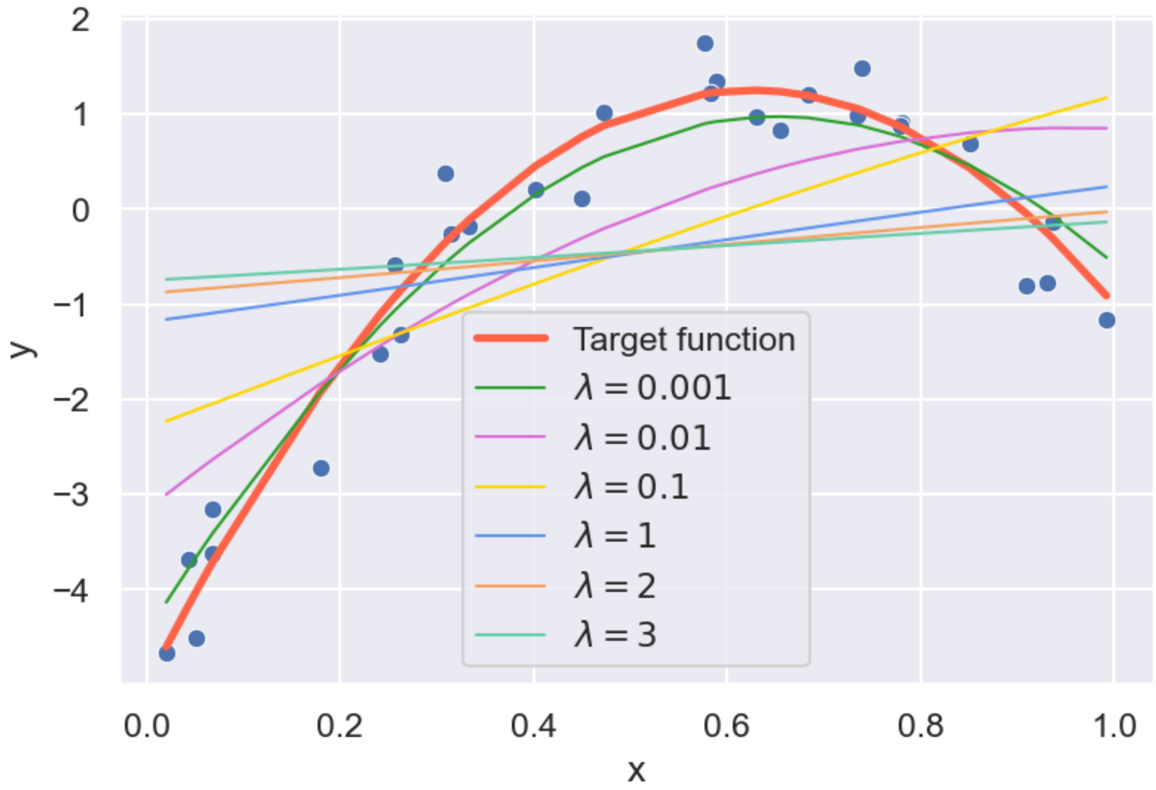
下表是生成的$training\ data\ set$

	x	ϵ	y
0	0.739020	1.107004	1.484793
1	0.333105	-0.150079	-0.173277
2	0.256019	0.846488	-0.589751
3	0.655249	-1.006529	0.832748
4	0.178938	-1.963550	-2.718955
5	0.068259	0.219462	-3.621587
6	0.020180	-0.159350	-4.666656
7	0.588464	0.307614	1.351688
8	0.472766	0.349717	1.019086
9	0.683708	0.009294	1.198571
10	0.068208	1.384748	-3.156372
11	0.402021	-0.608103	0.211242
12	0.930333	-1.306186	-0.764122
13	0.241439	-1.023469	-1.513287
14	0.307762	1.857230	0.382656
15	0.909157	-1.894218	-0.799611
16	0.051311	-1.212618	-4.500959
17	0.937331	0.442974	-0.133622
18	0.850742	0.629363	0.686397
19	0.780054	0.111900	0.910090
20	0.583185	-0.004178	1.220353
21	0.631064	-0.689769	0.973504
22	0.734816	-0.190518	0.980841
23	0.262222	-1.169472	-1.323510
24	0.042539	1.252769	-3.677074
25	0.450025	-1.629399	0.108378
26	0.576635	1.324379	1.742325
27	0.992353	-0.631145	-1.161629
28	0.314474	0.103586	-0.251388
29	0.779517	0.009098	0.871629

2

使用$kernel\ ridge\ regression$生成回归模型，以下设定$kernel = RBF$，并使用不同的$\lambda$值:

可以看出随着λ值上升，模型越趋和$Target\ Function$不像

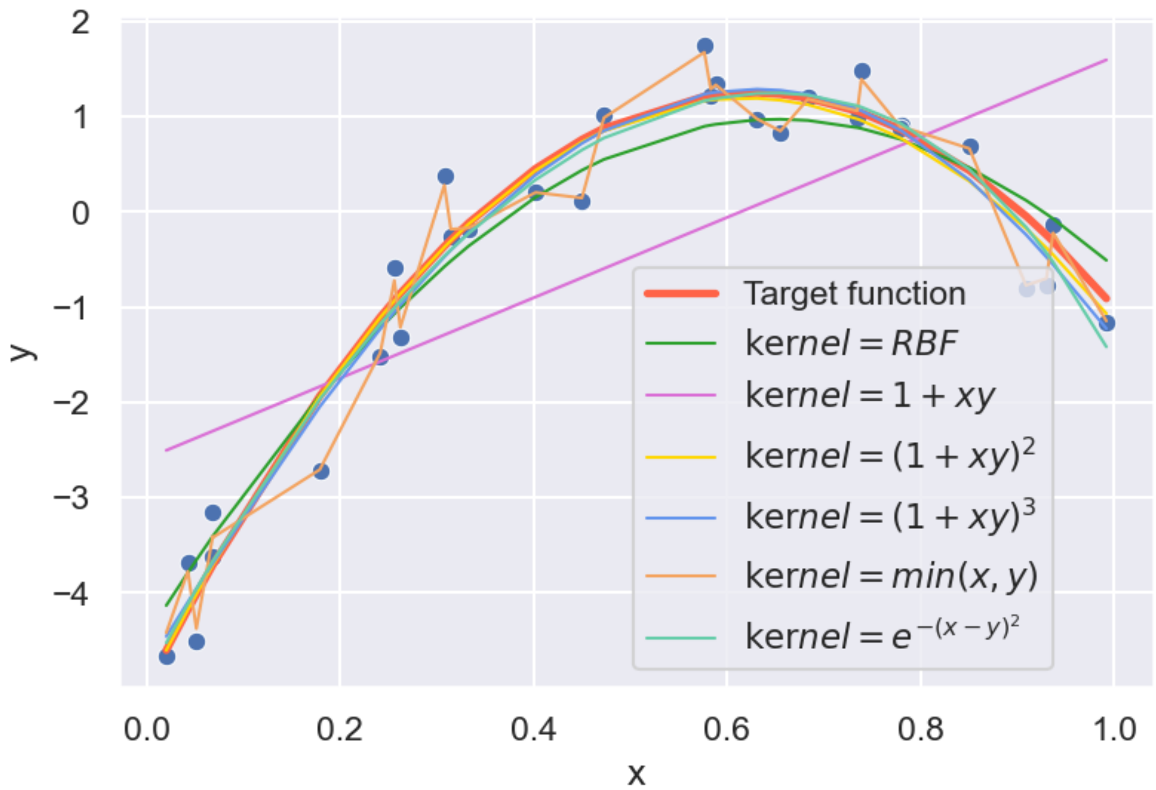


下表则为不同λ值的R^2 ，可以看出惩罚系数确实有用减少$overfitting$的问题

	λ	R^2
0	0.001	0.978979
1	0.010	0.741033
2	0.100	0.533162
3	1.000	0.288550
4	2.000	0.186369
5	3.000	0.136844

继续使用$kernel\ ridge\ regression$生成回归模型，以下设定$\lambda = 0.01$，并使用不同的$kernel$:

可以看出不同$kernel$确实会影响模型配适结果，也会有不同特性



下表则为不同$kernel$的R^2 ，可以看出不同$kernel$确实会影响模型

	kernel	R^2
0	RBF	0.978979
1	$1 + xy$	0.520176
2	$(1 + xy)^2$	0.998538
3	$(1 + xy)^3$	0.995929
4	$min(x, y)$	0.962126
5	$e^{(-(x-y)^2)}$	0.995126

---

2. Exponential Families

Problem 5

可以使用$MGF(Moment\ Generating\ Function)$来证明，首先下证$MGF$与$E(X)$、$Var(X)$、$Cov(X)$的关系：

- 已知$M_y(t) = E(e^{ty})$，
  $\frac{d}{dt}E(e^{ty} = E(e^{ty}y)) \Rightarrow M_y'(0) = E(Y)$
  $\frac{d}{dt}E(e^{ty}y)) = E(e^{ty}y^2)) \Rightarrow M_y''(0) = E(Y^2) \Rightarrow Var(Y) = M_y''(0) - (M_y'(0))^2$
  $\Rightarrow Cov(Y_i, Y_j) = M''_{y_i,y_j}(0) - (M'_{y_i,y_j}(0))^2$

- 现将$p(x|\eta) = h(x)exp(\eta^T T(x) - A(\eta))$代入$M_{T(x)}(t)$
  $\Rightarrow M_{T(x)}(t) = E(e^{tT(x)}) = \int e^{tT(x)}h(x)exp(\eta^T T(x) - A(\eta))dx$
  $= \int h(x)exp((\eta^T + t)T(x) - A(\eta + t))dxexp(A(\eta + t) - A(\eta))$
  $\Rightarrow M'_{T(x)}(t) = exp(A(\eta + t) - A(\eta))A'(\eta + t)$
  $\Rightarrow M'_{T(x)}(0) = A'(\eta)$
  $\Rightarrow M''_{T(x)}(t) = exp(A(\eta + t) - A(\eta))(A''(\eta + t) + (A'(\eta + t))^2)$
  $\Rightarrow M''_{T(x)}(0) = A''(\eta) + (A'(\eta))^2$
  $Thus,$
  $\frac{d}{d\eta_i}A(\eta) = E_{p(x|\eta)}[T_i(X)]$
  $\frac{d}{d\eta_i\eta_j}A(\eta) = Cov_{p(x|\eta)}[T_i(X)T_j(X)]$

Problem 6

1

已知$x_1, \ldots x_N \sim^{i.i.d} p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$为$Exponential\ Family$
$\forall\ T(x) = [x; vec(xx^T)],\ A(\eta) = \frac{1}{2}\mu^T\Sigma^{-1}\mu + log|\Sigma|,\ h(x) = (2\pi)^{-\frac{d}{2}}, \eta = [\Sigma^{-1}\mu; -\frac{1}{2}vec(\Sigma^{-1})]$
根据$Exponential\ Family$的性质，可以直接推出$\hat{\mu} = \frac{1}{N}\sum_N T_1(x_N) = \frac{1}{N}\sum_{i=1}^N x_i$

再使用MLE解出$l(\mu, \mathbf{\Sigma}|\mathbf{x_i}) = C + N2log|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^N tr[(\mathbf{x_i} - \mu)(\mathbf{x_i} - \mu)^{\mathbf{T}}\mathbf{\Sigma}^{-1}]$
$\Rightarrow \frac{d}{d\Sigma^{-1}}l(\mu, \mathbf{\Sigma}|\mathbf{x_i}) = \frac{N}{2}\sum_{i=1}^N(\mathbf{x_i} - \mu)(\mathbf{x_i} - \mu)$
$\Rightarrow \hat{\Sigma} = \frac{1}{N}\sum_{i=1}^N(\mathbf{x_i} - \hat{\mu})(\mathbf{x_i} - \hat{\mu})^T$，
根据MLE的不变性，$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^N(\mathbf{x_i} - \frac{1}{N}\sum_{j=1}^N xj)(\mathbf{x_i} - \frac{1}{N}\sum_{k=1}^N x_i)^T$

$\Rightarrow \hat{\mu} = \frac{1}{N}\sum_{i=1}^N x_i,\ \hat{\Sigma} = \frac{1}{N}\sum_{i=1}^N(\mathbf{x_i} - \frac{1}{N}\sum_{j=1}^N xj)(\mathbf{x_i} - \frac{1}{N}\sum_{k=1}^N x_i)^T$

2

这两个均是不偏估计，以下为证明：
$E(\hat{\mu}) = \frac{1}{N} N E(X) = E(X) = \mu$
$E(\hat{\Sigma}) = \frac{1}{N} N E(X - E(X))^T E(X - E(X)) = \Sigma$

再补充说明，$Exponential\ Family$最棒的特点就是它涵盖$T(x)$这个充分统计量，根据 $Rao - Blackwell\ THM$知道，在给定$T(x)$的条件下求不偏估计的话，则会达成UMVUE，因此这两者不只是UE更是UMVUE。

3

以下证明$E[||\ \hat{\mu}_{ML} - \mu\ ||]^2 = \frac{Tr\Sigma}{N}$:

$$E[||\ \hat{\mu}_{ML} - \mu\ ||]^2 = E(\mu_{ML}^{\wedge} - \mu)^2 = E(\bar{X} - \mu)^2 = E(\bar{X} - E(\bar{X}))^2 = Var(\bar{X}) = \frac{Var(X)}{N} = \frac{Tr\Sigma}{N}$$

4

(a)

以下证明$[Cov(X, Y)]^2 \le (VarX)(VarY)$:

令$h(t) = E((X - \mu_X) + t(Y - \mu_Y))^2 \ge 0,\ \forall t \in R$
$\Rightarrow h(t) = E((X - \mu_X))^2 + t^2 E(Y - \mu_Y)^2 + 2tE((X - \mu_X)(Y - \mu_Y)) \ge 0$
$\Rightarrow Var(X) + t^2 Var(Y) + 2tCov(X, Y) \ge 0$

已知$h(t)$是开口向上的函数，故$\Delta \le 0,\ \forall\ at^2 + bt + c = 0,\ \Delta = b^2 - 4ac$
$\Rightarrow (2Cov(X, Y))^2 - 4Var(X)Var(Y) \le 0$
$\Rightarrow [Cov(X, Y)]^2 \le (VarX)(VarY)$

(b)

以下证明$E_{p(x|\mu)}[\frac{d}{d\mu} logp(X|\mu)] = 0$：

已知$(x_1, x_2, \ldots, x_N) \sim^{i.i.d} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$

$\frac{d}{d\mu} logp(x|\mu) = \frac{1}{p(x|\mu)} \frac{d}{d\mu} p(x|\mu) = \frac{1}{p(x|\mu)} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \frac{x-\mu}{\sigma^2} = \frac{x-\mu}{\sigma^2}$

$\Rightarrow E_{p(x|\mu)}[\frac{d}{d\mu} logp(X|\mu)] = E_{p(x|\mu)}[\frac{X-\mu}{\sigma^2}] = \frac{1}{\sigma^2} E_{p(x|\mu)}[X - \mu] = \frac{1}{\sigma^2}(\mu - \mu) = 0$

(c)

以下证明$\frac{d}{d\mu}E_{p(x|\mu)}[\hat{\mu(X)}] = E_{p(x|\mu)}[\hat{\mu(X)}\frac{d}{d\mu}logp(X|\mu)] = Cov(\hat{\mu(X)}, \frac{d}{d\mu}logp(X|\mu))$:

- $\frac{d}{d\mu}E_{p(x|\mu)}[\hat{\mu(X)}] = \int \frac{d}{d\mu}[\hat{\mu(x)}p(x|\mu)]dx$
- $E_{p(x|\mu)}[\hat{\mu(X)}\frac{d}{d\mu}logp(X|\mu)] = \int \hat{\mu(x)}\frac{d}{d\mu}logp(x|\mu)p(x|\mu)dx = \int \frac{d}{d\mu}[\hat{\mu(x)}p(x|\mu)]dx$
- $Cov(\hat{\mu(X)}, \frac{d}{d\mu}logp(X|\mu)) = E_{p(x|\mu)}[\hat{\mu(X)}\frac{d}{d\mu}logp(X|\mu)] - E_{p(x|\mu)}(\hat{\mu(X)})E_{p(x|\mu)}[\frac{d}{d\mu}logp(X|\mu)]$
  $= E_{p(x|\mu)}[\hat{\mu(X)}\frac{d}{d\mu}logp(X|\mu)]$
  $= \int \frac{d}{d\mu}[\hat{\mu(x)}p(x|\mu)]dx$

$\Rightarrow \frac{d}{d\mu}E_{p(x|\mu)}[\hat{\mu(X)}] = E_{p(x|\mu)}[\hat{\mu(X)}\frac{d}{d\mu}logp(X|\mu)] = Cov(\hat{\mu(X)}, \frac{d}{d\mu}logp(X|\mu))$

(d)

以下证明$E_{p(x|\mu)}[||\ \hat{\mu} - \mu\ ||^2] = Var_p(x|\mu)[\hat{\mu(X)}] \geq \frac{(\frac{d}{d\mu}E_{p(x|\mu)}[\hat{\mu(X)}])^2}{E_{p(x|\mu)}[(\hat{\mu(X)}\frac{d}{d\mu}logp(X|\mu))^2]} = \frac{\sigma^2}{N}$:

- $E_{p(x|\mu)}[||\ \hat{\mu} - \mu\ ||^2] = E_{p(x|\mu)}(\hat{\mu} - \mu)^2 = E_{p(x|\mu)}(\hat{\mu} - E(\hat{\mu}))^2 = Var_p(x|\mu)[\hat{\mu(X)}]$
- 令$X = \hat{\mu(X)},\ Y = \frac{d}{d\mu}logp(X|\mu)$，代入$Cauchy - Schwarz\ inequality$
  $(Cov_{p(x|\mu)}(\hat{\mu(X)}, \frac{d}{d\mu}logp(X|\mu)))^2 \leq Var_p(x|\mu)[\hat{\mu(X)}]Var_p(x|\mu)[\frac{d}{d\mu}logp(X|\mu))]$
  $\Rightarrow Var_p(x|\mu)[\hat{\mu(X)}] \geq \frac{Cov_{p(x|\mu)}(\hat{\mu(X)}, \frac{d}{d\mu}logp(X|\mu)))^2}{Var_p(x|\mu)[\frac{d}{d\mu}logp(X|\mu))]}$
  由(b)和(c)可以得出，$\Rightarrow Var_p(x|\mu)[\hat{\mu(X)}] \geq \frac{(\frac{d}{d\mu}E_{p(x|\mu)}[\hat{\mu(X)}])^2}{E_{p(x|\mu)}[(\hat{\mu(X)}\frac{d}{d\mu}logp(X|\mu))^2]}$
- $\frac{(\frac{d}{d\mu}E_{p(x|\mu)}[\hat{\mu(X)}])^2}{E_{p(x|\mu)}[(\hat{\mu(X)}\frac{d}{d\mu}logp(X|\mu))^2]} = \frac{1}{N\frac{1}{\sigma^2}} = \frac{\sigma^2}{N}$

$\Rightarrow E_{p(x|\mu)}[||\ \hat{\mu} - \mu\ ||^2] = Var_p(x|\mu)[\hat{\mu(X)}] \geq \frac{(\frac{d}{d\mu}E_{p(x|\mu)}[\hat{\mu(X)}])^2}{E_{p(x|\mu)}[(\hat{\mu(X)}\frac{d}{d\mu}logp(X|\mu))^2]} = \frac{\sigma^2}{N}$

根据此题得证$\hat{\mu}_{ML}$会达成UMVUE。