Machine Learning
Seoul National University

Homework #(**1**)
Haneul Choi

INSTRUCTIONS

- Anything that is received after the deadline will be considered to be late and we do not receive late homeworks. We do however ignore your lowest homework grade.
- Answers to every theory questions need to be submitted electronically on ETL. Only PDF generated from LaTex is accepted.
- Make sure you prepare the answers to each question separately. This helps us dispatch the problems to different graders.
- Collaboration on solving the homework is allowed. Discussions are encouraged but you should think about the problems on your own.
- If you do collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution.

1 Q1

1.1

Given X_i, w, b, it is possible to consider $\langle X_i, w \rangle + b$ as constant, and thus the following holds:

$$\begin{aligned} P(Y_i) &= P(\epsilon_i) && (1)\\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right) && (2)\\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_i - \langle X_i, w\rangle - b)^2}{2\sigma^2}\right) && (3) \end{aligned}$$

Therefore, the conditional distribution is

$$Y_i | X_i, w, b \sim \mathcal{N}(\langle X_i, w \rangle + b, \sigma^2) \quad (4)$$

1.2

Since given ϵ_is are i.i.d., $P(\mathbf{Y}|\boldsymbol{\beta}) = \prod_i P(Y_i|w,b)$ holds. Utilizing result from 1.1,

$$\begin{aligned} \log P(\mathbf{Y}|\boldsymbol{\beta}) &= \sum_i \log P(Y_i|\boldsymbol{\beta}) && (5)\\ &= -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_i \left(Y_i - \langle X_i', \boldsymbol{\beta}\rangle\right)^2 && (6) \end{aligned}$$

1.3

Since positive constants are negligible and $\boldsymbol{\beta}$ only affects the sum of squares in the loglikelihood, following holds.

$$\arg\max_{\boldsymbol{\beta}} \log P(\mathbf{Y}|\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \left(-\sum_i \left(Y_i - \langle X_i', \boldsymbol{\beta} \rangle \right)^2 \right) \tag{7}$$

$$= \arg\min_{\boldsymbol{\beta}} \left(\mathbf{Y} - \mathbf{X}'\boldsymbol{\beta} \right)^\top \left(\mathbf{Y} - \mathbf{X}'\boldsymbol{\beta} \right) \tag{8}$$

1.4

Hessian matrix of $f(\boldsymbol{\beta}) = \left(\mathbf{Y} - \mathbf{X}'\boldsymbol{\beta} \right)^\top \left(\mathbf{Y} - \mathbf{X}'\boldsymbol{\beta} \right)$ equals to $2\mathbf{X}'^\top\mathbf{X}'$. Since $\mathbf{X}'$ is full rank on column space, it can be shown that null space of $\mathbf{X}'$ and $\mathbf{X}'^\top\mathbf{X}'$ is both $\{0\}$, i.e. the square matrix ${\mathbf{X}'}^\top\mathbf{X}'$ is non-singular. This also implies that ${\mathbf{X}'}^\top\mathbf{X}'$ is actually a positive-definite matrix by the following:

$$\forall v \in \mathbb{R}^{d+1} \backslash \{0\}, \langle v, {\mathbf{X}'}^\top\mathbf{X}'v \rangle = \langle \mathbf{X}'v, \mathbf{X}'v \rangle > 0 \tag{9}$$

Therefore, f will be minimized by $\boldsymbol{\beta}$ which satisfies $\frac{\partial f}{\partial \boldsymbol{\beta}} = 0$, if exists.

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = 2(\mathbf{Y} - \mathbf{X}'\boldsymbol{\beta})^\top \frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}'\boldsymbol{\beta}) \tag{10}$$

$$= 2(\mathbf{Y} - \mathbf{X}'\boldsymbol{\beta})^\top (-\mathbf{X}') = 0 \iff {\mathbf{X}'}^\top\mathbf{X}'\boldsymbol{\beta} = {\mathbf{X}'}^\top\mathbf{Y} \tag{11}$$

As $\mathbf{X}'^\top\mathbf{X}'$ is invertible, the solution $\boldsymbol{\beta} = (\mathbf{X}'^\top\mathbf{X}')^{-1}{\mathbf{X}'}^\top\mathbf{Y}$ will minimize least squares and thus maximize the loglikelihood.

1.5

Given prior distribution, the distribution of Y_i can be derived as following:

$$X_i'^\top \boldsymbol{\beta}' \sim \mathcal{N}(0, X_i'^\top \eta^2 \mathbf{I}_{d+1} X_i') \tag{12}$$

$$\sim \mathcal{N}(0, \eta^2 X_i'^\top X_i') \tag{13}$$

Since $\boldsymbol{\beta}$ and ϵ_i are independent,

$$Y_i = \langle X_i', \boldsymbol{\beta} \rangle + \epsilon_i \sim \mathcal{N}(0, \eta^2 X_i'^\top X_i' + \sigma^2) \tag{14}$$

Now applying Bayes' rule, posterior distribution of $\boldsymbol{\beta}'|Y_i$ can be derived.

$$P(\boldsymbol{\beta}'|Y_i) = \frac{P(\boldsymbol{\beta}')P(Y_i|\boldsymbol{\beta}')}{P(Y_i)} \tag{15}$$

$$= \frac{1}{(2\pi)^{\frac{d+1}{2}}} \frac{(\eta^2 X_i^\top X_i/\sigma^2 + 1)^{\frac{1}{2}}}{\eta^{d+1}} \exp\left(-\frac{1}{2} \left(\frac{\|\boldsymbol{\beta}'\|^2}{\eta^2} + \frac{\|Y_i - \langle X_i', \boldsymbol{\beta} \rangle\|^2}{\sigma^2} - \frac{\|Y_i\|^2}{\eta^2 \|X_i'\|^2 + \sigma^2} \right) \right) \tag{16}$$

Let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$. Since $\boldsymbol{\beta}' \perp \boldsymbol{\epsilon}$, the distribution of $\mathbf{Y}$ can be derived as following:

$$\mathbf{X}'\boldsymbol{\beta} \sim \mathcal{N}(0, \eta^2 \mathbf{X}'\mathbf{X}'^\top) \tag{17}$$

$$\mathbf{Y} = \mathbf{X}'\boldsymbol{\beta} + \epsilon \sim \mathcal{N}(0, \eta^2 \mathbf{X}'\mathbf{X}'^\top + \sigma^2 \mathbf{I}_n) \tag{18}$$

The posterior distribution of $\boldsymbol{\beta}'|\mathbf{Y}$ can be derived applying Bayes' rule, as following:

$$P(\boldsymbol{\beta}'|\mathbf{Y}) = \frac{P(\boldsymbol{\beta}')P(\mathbf{Y}|\boldsymbol{\beta}')}{P(\mathbf{Y})} \tag{19}$$

$$= \frac{|\eta^2 \mathbf{X}'\mathbf{X}'^\top + \sigma^2 \mathbf{I}_n|^{\frac{1}{2}}}{(\sqrt{2\pi}\eta)^{d+1}\sigma^n} \exp\left(-\frac{1}{2}\left(\frac{\|\boldsymbol{\beta}'\|^2}{\eta^2} + \frac{\|\mathbf{Y} - \mathbf{X}'\boldsymbol{\beta}\|^2}{\sigma^2} - \mathbf{Y}^\top(\eta^2 \mathbf{X}'\mathbf{X}'^\top + \sigma^2 \mathbf{I}_n)^{-1}\mathbf{Y}\right)\right) \tag{20}$$

2 Q2

2.1

Set input sample size p 12, and the following algorithm will output an approximate Gaussian draw.

Algorithm 1 Approximated Gaussian Sample

```
procedure DRAW(s: uniform samples of size 12)
    z ← s_1 + ··· + s_12 − 6
    return z
```

As $s_1, \ldots, s_{12}$ are i.i.d. sample that follow $\mathcal{U}(0,1)$, z can be rewritten as below

$$z = \sqrt{12}\left[\frac{\frac{s_1 + \cdots + s_{12}}{12} - 0.5}{\sqrt{\frac{1}{12}}}\right] \tag{21}$$

$$= \sqrt{p}\left[\frac{\bar{s}_p - \mathbb{E}[u]}{\sqrt{\mathbb{V}[u]}}\right] \tag{22}$$

where $u \sim \mathcal{U}(0,1)$, and the distribution of z will be close to Gaussian, according to CLT.

2.2

Check Figure 1.

2.3

Check Figure 2.

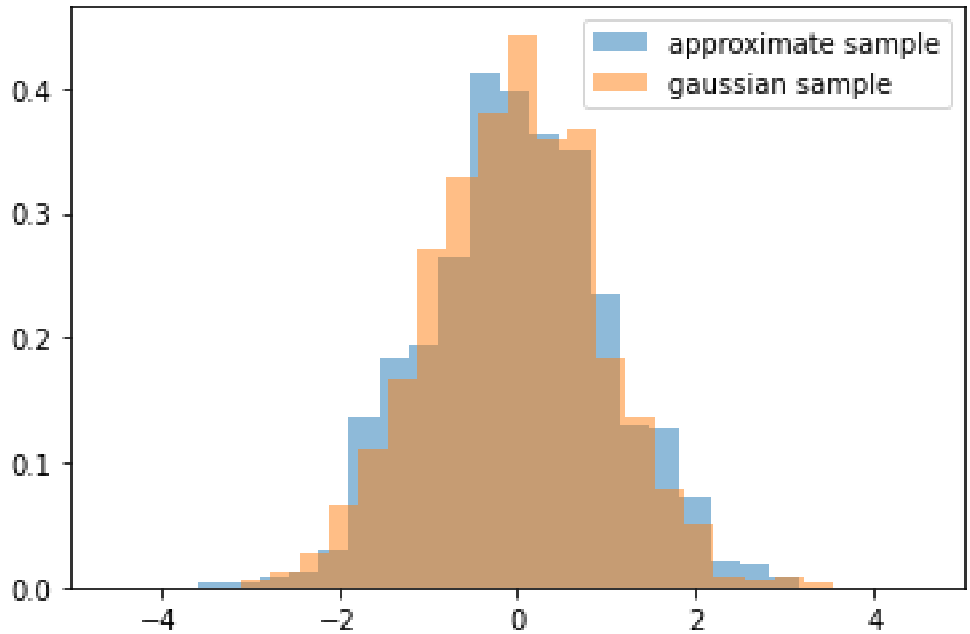

Figure 1: Result of histograms.

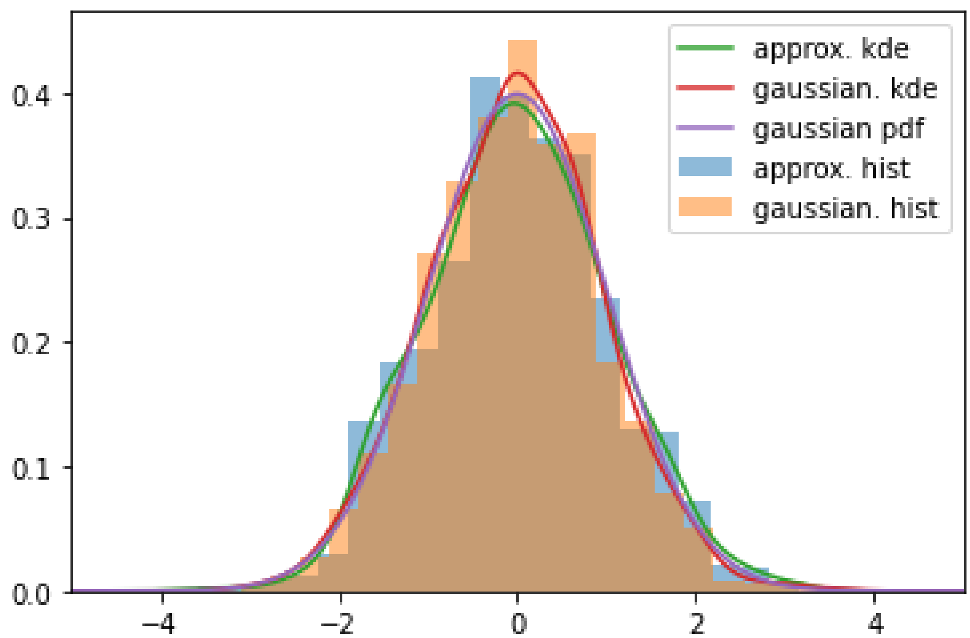

Figure 2: Result of KDEs, histograms, and true density. KDE bandwith was chosen using Silverman's method.