ML HW4 math

r14921A13 鄭皓中

Problem 1

(a) $\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$

Constraint : $y_i - w^T x_i - b \le \varepsilon + \xi_i \Rightarrow y_i - w^T x_i - b - \varepsilon - \xi_i \le 0$

$w^T x_i + b - y_i \le \varepsilon + \xi_i \Rightarrow w^T x_i + b - y_i - \varepsilon - \xi_i \le 0$

$\xi_i \ge 0 \Rightarrow -\xi_i \le 0$

$$L(w, b, \xi, \alpha, \alpha^*, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i(y_i - w^T x_i - b - \varepsilon - \xi_i) + \sum_{i=1}^{m}\alpha_i^*(w^T x_i + b - y_i - \varepsilon - \xi_i) + \sum_{i=1}^{m}\beta_i\cdot(-\xi_i)$$

(b) $\nabla_w L = w - \sum_{i=1}^{m}\alpha_i x_i + \sum_{i=1}^{m}\alpha_i^* x_i = 0 \Rightarrow w = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)x_i$

$\frac{\partial L}{\partial b} = \sum_{i=1}^{m}\alpha_i(-1) + \sum_{i=1}^{m}\alpha_i^*\cdot 1 = 0 \Rightarrow \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0$

$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \alpha_i^* - \beta_i = 0 \Rightarrow \beta_i = C - \alpha_i - \alpha_i^*$

$$L = \frac{1}{2}w^T w - w^T\sum_{i=1}^{m}(\alpha_i - \alpha_i^*)x_i + b\sum_{i=1}^{m}(\alpha_i^* - \alpha_i) + \sum_{i=1}^{m}\xi_i(C - \alpha_i - \alpha_i^* - \beta_i) - \varepsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)y_i$$

$$= \frac{1}{2}w^T w - w^T w + 0 + 0 + \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)y_i - \varepsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^*)$$

$$= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i^T x_j + \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)y_i - \varepsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^*)$$

Dual problem : $\max_{\alpha_i, \alpha_i^*} -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)x_i x_j + \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)y_i - \varepsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^*)$

Constraint : $\sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0$, $0 \le \alpha_i + \alpha_i^* \le C$

(c) $\min_{b,\xi} C\sum_{i=1}^{m}\xi_i$ Constraint : $y_i - w^T x_i - b \le \varepsilon + \xi_i$, $w^T x_i + b - y_i \le \varepsilon + \xi_i$, $\xi_i \ge 0$

Constraint : $0 \le |y_i - (w^T x + b)| \le \varepsilon + \xi_i$, $\xi_i = 0$

$\Rightarrow \xi_i \ge |y_i - (w^T x_i + b)| - \varepsilon$, $\xi_i \ge 0$

$\Rightarrow \xi_i \ge \max(0, |y_i - (w^T x_i + b)| - \varepsilon)$

$$C\sum_{i=1}^{m}\xi_i = C\cdot\sum_{i=1}^{m}\left(\max(0, |y_i - (w^T x_i + b)| - \varepsilon)\right)$$

$$\min_{b,\xi} C\sum_{i=1}^{m}\xi_i \Rightarrow \bar{b} = \operatorname{argmin}_{b\in\mathbb{R}} C\sum_{i=1}^{m}\max(0, |y_i - (w^T x_i + b)| - \varepsilon)$$

(c). $e = y_i - (\bar{w}^T x_i + \bar{b})$

1. $|e| < \varepsilon \Rightarrow$ 不在 Constraint $|e| \le \varepsilon$ 的 boundary $\Rightarrow$ Lagrange multiplier $= 0$

$\bar{\alpha}_i = \bar{\alpha}_i^* = 0$, $\bar{\beta}_i \xi_i = 0 = (C - \bar{\alpha}_i - \bar{\alpha}_i^*)\xi_i \Rightarrow \xi_i = 0 \quad \forall i = 1, 2, \dots, m$

2. $e = \varepsilon \Rightarrow$ 满足 $e = \varepsilon$ 的 boundary 但 $-e \le \varepsilon$ 依然没有 active

$\Rightarrow \bar{\alpha}_i^* = 0$, $\xi_i = 0 \quad \forall i = 1, 2, \dots, m$.

由於 $\xi_i = 0 \Rightarrow \beta_i$ 不一定為 0 $\Rightarrow C - \alpha_i = \beta_i \ge 0 \Rightarrow 0 \le \bar{\alpha}_i \le C$

3. $e = -\varepsilon \Rightarrow$ 满足 $-e = \varepsilon$ 的 boundary, $e = \varepsilon$ 則沒有 active

$\Rightarrow \alpha_i = 0$, $\xi_i = 0 \quad \forall i = 1, 2, \dots, m$

由於 $\xi_i = 0 \Rightarrow \beta_i$ 不一定 $= 0 \Rightarrow C - \bar{\alpha}_i^* = \beta_i \ge 0 \Rightarrow 0 \le \bar{\alpha}_i^* \le C$

4. $e > \varepsilon \Rightarrow$ 引入 ξ_i 使得 $e = \varepsilon + \xi_i \Rightarrow -e < \varepsilon + \xi_i$ 的 constraint 是 inactive

$\Rightarrow \bar{\alpha}_i^* = 0$, 由於 $\xi_i > 0 \Rightarrow \sum_{i=1}^m \beta_i \xi_i = 0$ 且 $\beta_i \ge 0$ 得到 $\beta_i = 0$

$C - \bar{\alpha}_i = \beta_i = 0 \Rightarrow \bar{\alpha}_i = C$

$\Rightarrow \bar{\alpha}_i = C$, $\xi_i = e - \varepsilon$,

5. $e < -\varepsilon \Rightarrow$ 引入 $\xi_i > 0$ 使得 $-e = \varepsilon + \xi_i \Rightarrow e < \varepsilon + \xi_i$ inactive

$\Rightarrow \bar{\alpha}_i = 0$, 由於 $\xi_i > 0 \Rightarrow \sum_{i=1}^m \beta_i \xi_i = 0$ 且 $\beta_i \ge 0$ 得到 $\beta_i = 0$

$C - \bar{\alpha}_i^* = \beta_i = 0 \Rightarrow \bar{\alpha}_i^* = C$

$\Rightarrow \bar{\alpha}_i^* = C$, $\xi_i = -(e + \varepsilon)$

(d) (1) 由於 dual problem 只會到 $x_i^T x_j$ 影響 $\Rightarrow$ 可替換成 kernel func. $K(x_i, x_j) = x_i^T x_j$

From b. dual problem: $\max_{\alpha_i, \alpha_i^*} -\frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) + \sum_{i=1}^m y_i(\alpha_i - \alpha_i^*) - \varepsilon \sum_{i=1}^m (\alpha_i + \alpha_i^*)$

Constraint: $\sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0$, $0 \le \alpha_i + \alpha_i^* \le C$

決策函式 $f(x) = w^T x + b$, Suppose $w = \sum_{i=1}^m (\alpha_i - \alpha_i^*) x_i$

$f(x) = \sum_{i=1}^m (\alpha_i - \alpha_i^*) x_i^T x + b = \sum_{i=1}^m (\alpha_i - \alpha_i^*) K(x_i, x) + b$

截距 b. 假設一個 $0 < \alpha_k < C$ 的 support vector x_k

$\Rightarrow b = y_k - \varepsilon - \sum_{i=1}^m (\alpha_i - \alpha_i^*) x_i^T x_k = y_k - \varepsilon - \sum_{i=1}^m (\alpha_i - \alpha_i^*) K(x_i, x_k)$

f. b 也只受到 $x_i^T x$ 影響 $\Rightarrow$ SVR algorithm can be kernalized.

Problem 2.

1. $\min f(\rho, \mu, \xi) = \rho + \frac{1}{\nu}\sum_{i=1}^{N} C_i \xi_i$, $C_i > 0$, $0 < \nu < C_i$

Constraint $\begin{cases} g_{1,i}(\rho, \mu, \xi) = \|x_i - \mu\|^2 - \rho - \xi_i \le 0 \\ g_{2,i}(\rho, \mu, \xi) = -\xi_i \le 0 \end{cases}$ $\forall i = 1, 2, \cdots, N$.

$g_3(\rho) = -\rho \le 0$

$\alpha \longmapsto g_{1,i}$, $\beta \longmapsto g_{2,i}$, $\gamma \longmapsto g_3$

$$L(\rho, \xi, \mu, \alpha, \beta, \gamma) = \rho + \frac{1}{\nu}\sum_{i=1}^{N} C_i \xi_i + \sum_{i=1}^{N} \alpha_i (\|x_i - \mu\|^2 - \rho - \xi_i) + \sum_{i=1}^{N} \beta_i (-\xi_i) + \gamma(-\rho)$$

$$= \rho\left(1 - \sum_{i=1}^{N} \alpha_i - \gamma\right) + \frac{1}{\nu}\sum_{i=1}^{N} C_i \xi_i - \sum_{i=1}^{N} \alpha_i \xi_i - \sum_{i=1}^{N} \beta_i \xi_i + \sum_{i=1}^{N} \alpha_i \|x_i - \mu\|^2$$

$$= \rho\left(1 - \sum_{i=1}^{N} \alpha_i - \gamma\right) + \sum_{i=1}^{N} \left(\frac{1}{\nu} C_i - \alpha_i - \beta_i\right)\xi_i + \sum_{i=1}^{N} \alpha_i \|x_i - \mu\|^2$$

2. 證明 strong duality : ① f, g are convex. ② Slater's condition

① $f = \rho + \frac{1}{\nu}\sum_{i=1}^{N} C_i \xi_i$ $\Rightarrow$ linear $\Rightarrow$ convex.

$g_{1,i} = \|x_i - \mu\|^2 - \rho - \xi_i$: $\|x_i - \mu\|^2$, ρ, ξ_i convex $\Rightarrow$ $g_{1,i}$ convex.

$g_{2,i} = -\xi_i$: convex.

$g_3 = -\rho$: convex.

$\Rightarrow$ ① 成立.

② Slater's condition : $\exists (\hat{\rho}, \hat{\mu}, \hat{\xi})$ s.t. $g_{1,i}, g_{2,i}, g_3 < 0$.

Choose $\hat{\mu} = 0$ $\Rightarrow$ $\|x_i\|^2 - \rho - \xi_i < 0$ $\xi_i > 0$, $\rho > 0$

Choose $\hat{\xi}_i = 1$ $\forall i$ $\Rightarrow$ $\|x_i\|^2 - \rho - 1 < 0$, $1 > 0$, $\rho > 0$

Choose $\rho = \max \|x_i\|^2$ $\Rightarrow$ $\|x_i\|^2 - \max\|x_i\|^2 - 1 < 0$, $1 > 0$, $\max\|x_i\|^2 > 0$

$\Rightarrow$ $(\max\|x_i\|^2, 0, 1)$ satisfies $g_{1,i}, g_{2,i}, g_3 < 0$

$\Rightarrow$ Slater's condition 成立.

①, ② 皆成立 $\Rightarrow$ Strong duality 成立 $\Rightarrow$ duality gap $= 0$

3. $L(\rho, \mu, \xi, \alpha, \beta, \gamma) = \rho\left(1 - \sum_{i=1}^{N}\alpha_i - \gamma\right) + \sum_{i=1}^{N}\left(\frac{1}{\nu}C_i - \alpha_i - \beta_i\right)\xi_i + \sum_{i=1}^{N}\alpha_i \|x_i - \mu\|^2$

$\frac{\partial L}{\partial \rho} = 1 - \sum_{i=1}^{N}\alpha_i - \gamma = 0 \Rightarrow \sum_{i=1}^{N}\alpha_i = 1 - \gamma$

$\nabla_\mu L = \sum_{i=1}^{N} -2(x_i - \mu)\alpha_i = 0 \Rightarrow \mu^* = \frac{\sum_{i=1}^{N}\alpha_i x_i}{\sum_{i=1}^{N}\alpha_i}$

$\frac{\partial L}{\partial \xi_i} = \frac{C_i}{\nu} - \alpha_i - \beta_i = 0 \Rightarrow \beta_i = \frac{C_i}{\nu} - \alpha_i$

确保 Lower bound 存在. $\Rightarrow \frac{\partial L}{\partial \rho}, \frac{\partial L}{\partial \xi_i} = 0$, 否则 Lower bound $= -\infty$

$$\Rightarrow \theta(\alpha, \beta, \gamma) = \begin{cases} \sum_{i=1}^{N}\alpha_i \|x_i\|^2 - \frac{\|\sum_{i=1}^{N}\alpha_i x_i\|^2}{\sum_{i=1}^{N}\alpha_i}, & \text{if } \sum_{i=1}^{N}\alpha_i = 1-\gamma,\ \beta_i = \frac{C_i}{\nu} - \alpha_i \\ -\infty, & \text{otherwise.} \end{cases}$$

4. $\|\alpha\|_1 = \sum_{i=1}^{N}\alpha_i$, $\alpha_i = \|\alpha\|_1 \hat{\alpha}_i$

$\theta(\alpha, \beta, \gamma) = \|\alpha\|_1 \sum_{i=1}^{N}\hat{\alpha}_i \|x_i\|^2 - \|\alpha\|_1 \|\sum_{i=1}^{N}\hat{\alpha}_i x_i\|^2$

$\|\sum_{i=1}^{N}\hat{\alpha}_i x_i\|^2 = \left(\sum_{i=1}^{N}\hat{\alpha}_i x_i\right)^T\left(\sum_{i=1}^{N}\hat{\alpha}_i x_i\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\hat{\alpha}_i\hat{\alpha}_j x_i^T x_j$

$$\begin{cases} \theta(\alpha, \beta, \gamma) = \|\alpha\|_1\left(\sum_{i=1}^{N}\hat{\alpha}_i\|x_i\|^2 - \sum_{i=1}^{N}\sum_{j=1}^{N}\hat{\alpha}_i\hat{\alpha}_j x_i^T x_j\right) \\ \text{Constraint}: \quad \sum_{i=1}^{N}\alpha_i = 1-\gamma \Rightarrow \sum_{i=1}^{N}\alpha_i \le 1 \quad \text{since } \gamma \ge 0 \\ \qquad\qquad \frac{C_i}{\nu} - \alpha_i = \beta_i \ge 0 \Rightarrow 0 \le \alpha_i \le \frac{C_i}{\nu} \end{cases}$$

5. (a) $\bar{\mu} = \mu^*$ (in 3.) $\Rightarrow \sum_{i=1}^{N} -2(x_i - \bar{\mu})\alpha_i = 0 \Rightarrow \|\bar{\alpha}\|_1 \bar{\mu} = \sum_{i=1}^{N}\bar{\alpha}_i x_i$

(b) $\xi_i \ge 0$, $\xi_i \ge \|x_i - \bar{\mu}\|^2 - \rho \Rightarrow \xi_i \ge \max.(\|x_i - \bar{\mu}\|^2 - \rho, 0)$

$f = \rho + \frac{1}{\nu}\sum_{i=1}^{N}C_i\xi_i = \rho + \frac{1}{\nu}\sum_{i=1}^{N}C_i\left(\max(\|x_i - \bar{\mu}\|^2 - \rho, 0)\right)$

Hence, $\bar{\rho} = \arg\min_{\rho \ge 0}\left(\rho + \frac{1}{\nu}\sum_{i=1}^{N}C_i\left(\max(\|x_i - \bar{\mu}\|^2 - \rho, 0)\right)\right)$

(c) $J(\rho) = \rho + \frac{1}{\nu}\sum_{i=1}^{N}C_i\left(\max.(d_i^2 - \rho, 0)\right)$, $d_i^2 = \|x_i - \bar{\mu}\|^2$

$$\frac{\partial}{\partial\rho}\left(\max.(d_i^2 - \rho, 0)\right) = \begin{cases} -1, & d_i^2 > \rho \\ 0, & d_i^2 < \rho \\ [-1, 0], & d_i^2 = \rho \end{cases}$$

$\Rightarrow \partial J(\rho) = 1 - \frac{1}{\nu}\sum_{d_i^2 > \rho}C_i - \frac{1}{\nu}\sum_{d_i^2 = \rho}C_i[0, 1]$

$\bar{\rho}$ 是 min $\Rightarrow 0 \in \partial J(\rho) \Rightarrow 1 - \frac{1}{\nu}\sum_{d_i^2 > \bar{\rho}}C_i \ge 0$, $1 - \frac{1}{\nu}\sum_{d_i^2 \ge \bar{\rho}}C_i \le 0$

$$\Rightarrow \sum_{d_i^2 > \bar{\rho}}C_i \le \nu$$

$$\sum_{d_i^2 \ge \bar{\rho}}C_i \ge \nu$$

Consider set $A = \{\rho \geq 0 \mid \sum_{i: d_i^2 > \rho} C_i \leq \nu\}$

$\bar{\rho}$ satisfies $\sum_{i: d_i^2 > \bar{\rho}} C_i \leq \nu \Rightarrow \bar{\rho} \in A \Rightarrow \bar{\rho} \geq \min\left(\{\rho \geq 0 \mid \sum_{i: \|x_i - \mu\|^2 > \rho} C_i \leq \nu\}\right)$

Consider set $B = \{\rho \geq 0 \mid \sum_{i: d_i^2 > \rho} C_i < \nu\}$

$\forall \hat{\rho} \in B, \quad \partial J(\hat{\rho}) = 1 - \frac{1}{\nu} \sum_{i: d_i^2 > \hat{\rho}} C_i > 0$

$\Rightarrow J$ is increasing if $\hat{\rho} \in B \Rightarrow \bar{\rho} \leq \min\left(\{\rho \geq 0 \mid \sum_{i: \|x_i - \mu\|^2 > \rho} C_i < \nu\}\right)$

$\Rightarrow \min\left(\{\rho \geq 0 \mid \sum_{i: \|x_i - \mu\|^2 > \rho} C_i \leq \nu\}\right) \leq \bar{\rho} \leq \min\left(\{\rho \geq 0 \mid \sum_{i: \|x_i - \mu\|^2 > \rho} C_i < \nu\}\right)$

(d) ① $\|x_i - \bar{\mu}\|^2 - \bar{\rho} > 0 \Rightarrow \bar{\xi}_i \geq \|x_i - \mu\|^2 - \bar{\rho} > 0 \Rightarrow \xi_i$ 不滿足 $-\xi_i \leq 0$ 的 boundary.

$\Rightarrow \bar{\beta}_i = 0 \Rightarrow \alpha_i = \frac{C_i}{\nu} > 0 \Rightarrow \alpha_i$ 存在 $\Rightarrow \bar{\xi}_i = \|x_i - \bar{\mu}\|^2 - \bar{\rho}$

② $\|x_i - \bar{\mu}\|^2 - \bar{\rho} < 0$, Suppose $\xi_i > 0 \Rightarrow \xi_i$ 不滿足 $-\xi_i \leq 0$ 的 boundary $\Rightarrow \beta_i = 0$

$\Rightarrow \alpha_i = \frac{C_i}{\nu} > 0 \Rightarrow \alpha_i$ 存在 $\Rightarrow \bar{\xi}_i = \|x_i - \bar{\mu}\|^2 - \bar{\rho} < 0$ ⟶

Hence $\bar{\xi}_i = 0$

$\Rightarrow \bar{\xi}_i = \max\left(\|x_i - \bar{\mu}\|^2 - \bar{\rho}, 0\right)$

(e) From (d), we have $\alpha_i = \frac{C_i}{\nu}$ if $\|x_i - \bar{\mu}\|^2 > \bar{\rho}$

if $\|x_i - \bar{\mu}\|^2 < \bar{\rho} \Rightarrow \bar{\xi}_i = 0 > \|x_i - \bar{\mu}\|^2 - \bar{\rho} \Rightarrow \xi_i$ 不滿足 $g_{1,i}$ 的 boundary.

$\Rightarrow \bar{\alpha}_i = 0 \quad \forall i$

if $\|x_i - \bar{\mu}\|^2 = \bar{\rho} \Rightarrow \xi_i = 0 \Rightarrow$ 同時在 $g_{1,i}, g_{2,i}$ 的 boundary.

$\Rightarrow \bar{\beta}_i = \frac{C_i}{\nu} - \bar{\alpha}_i \geq 0 \Rightarrow 0 \leq \bar{\alpha}_i \leq \frac{C_i}{\nu}$

6. From 5.(e). For $\|x_i - \bar{\mu}\|^2 > \bar{\rho} \Rightarrow \bar{\alpha}_i = \frac{1}{n\nu}$, $\sum_{i=1}^{n} \bar{\alpha}_i = 1$

$\Rightarrow 1 = \sum_{i=1}^{n} \bar{\alpha}_i \geq \sum_{i: \|x_i - \bar{\mu}\|^2 > \bar{\rho}} \bar{\alpha}_i = n_{out} \cdot \frac{1}{n\nu} \Rightarrow \nu \geq \frac{n_{out}}{n}$

For $\|x_i - \bar{\mu}\|^2 = \bar{\rho} \Rightarrow 0 \leq \bar{\alpha}_i \leq \frac{1}{n\nu}$, $\|x_i - \bar{\mu}\|^2 < \bar{\rho} \Rightarrow \bar{\alpha}_i = 0$

$\Rightarrow \sum_{i=1}^{n} \bar{\alpha}_i = 1 = \sum_{i: \|x_i - \mu\|^2 > \bar{\rho}} \bar{\alpha}_i + \sum_{i: \|x_i - \mu\|^2 = \bar{\rho}} \bar{\alpha}_i \leq \frac{n_{out}}{n\nu} + \frac{n_{sv}}{n\nu}$

$\Rightarrow \nu \leq \frac{n_{out} + n_{sv}}{n} \Rightarrow \frac{n_{out}}{n} \leq \nu \leq \frac{n_{out} + n_{sv}}{n}$

至少有 ν 的 data 在 hypersphere 的 boundary 上或外面

至多有 ν 的 data 在 hypersphere 的外部

$\Rightarrow$ 這個 hypersphere 至少包含了 $1 - \nu$ 的 data

Problem 3.

(a) $C_t = f_t \odot C_{t-1} + (1 - f_t) \odot \tilde{C}_t$

$$\frac{\partial C_t}{\partial f_t} = C_{t-1} - \tilde{C}_t$$

$$\frac{\partial C_t}{\partial \tilde{C}_t} = 1 - f_t$$

$$\frac{\partial C_t}{\partial C_{t-1}} = f_t$$

$$\frac{\partial L}{\partial f_t} = \frac{\partial L}{\partial C_t} \odot \frac{\partial C_t}{\partial f_t} = \frac{\partial L}{\partial C_t} \odot (C_{t-1} - \tilde{C}_t)$$

$$\frac{\partial L}{\partial \tilde{C}_t} = \frac{\partial L}{\partial C_t} \odot \frac{\partial C_t}{\partial \tilde{C}_t} = \frac{\partial L}{\partial C_t} \odot (1 - f_t)$$

$$\frac{\partial L}{\partial C_{t-1}} = \frac{\partial L}{\partial C_t} \odot \frac{\partial C_t}{\partial C_{t-1}} = \frac{\partial L}{\partial C_t} \odot f_t$$

(b) ① $C_t \longrightarrow h_t \longrightarrow L$ ② $C_t \longrightarrow C_{t+1} \longrightarrow L$

① $h_t = o_t \odot \tanh(C_t)$

$$\frac{\partial h_t}{\partial C_t} = o_t \odot (1 - \tanh^2(C_t))$$

② $$\frac{\partial C_{t+1}}{\partial C_t} = f_{t+1}$$

$$\Rightarrow \frac{\partial L}{\partial C_t} = \frac{\partial L}{\partial h_t} \odot \frac{\partial h_t}{\partial C_t} + \frac{\partial L}{\partial C_{t+1}} \odot \frac{\partial C_{t+1}}{\partial C_t}$$

$$= \delta h_t \odot o_t \odot (1 - \tanh^2(C_t)) + \delta C_{t+1} \odot f_{t+1}$$

$$\frac{\partial L}{\partial f_t} = \delta C_t \odot (C_{t-1} - \tilde{C}_t) \quad \text{From (a)}$$

$z_{f,t} = W_f x_t + U_f h_{t-1} + b_f$, $f_t = \sigma(z_{f,t})$, $\sigma'(z) = \sigma(z)(1 - \sigma(z))$

$$\frac{\partial L}{\partial z_{f,t}} = \frac{\partial L}{\partial f_t} \odot \frac{\partial f_t}{\partial z_{f,t}} = \delta f_t \odot f_t (1 - f_t)$$

$z_{o,t} = W_o x_t + U_o h_{t-1} + b_o$, $o_t = \sigma(z_{o,t})$

$$\frac{\partial L}{\partial z_{o,t}} = \frac{\partial L}{\partial o_t} \odot \frac{\partial o_t}{\partial z_{o,t}} = \delta o_t \odot o_t (1 - o_t)$$

$z_{c,t} = W_c x_t + U_c h_{t-1} + b_c$, $\tilde{C}_t = \tanh(z_{c,t})$

$$\frac{\partial L}{\partial z_{c,t}} = \frac{\partial L}{\partial \tilde{C}_t} \odot \frac{\partial \tilde{C}_t}{\partial z_{c,t}} = \delta \tilde{C}_t \odot (1 - \tilde{C}_t^2)$$

(c) $$\frac{\partial z_{f,t}}{\partial W_f} = x_t^T, \quad \frac{\partial L}{\partial W_f} = \sum_t \left(\frac{\partial L}{\partial z_{f,t}} \cdot \frac{\partial z_{f,t}}{\partial W_f} \right) = \sum_t \delta z_{f,t} x_t^T$$

$$\frac{\partial z_{f,t}}{\partial U_f} = h_{t-1}^T, \quad \frac{\partial L}{\partial U_f} = \sum_t \left(\frac{\partial L}{\partial z_{f,t}} \cdot \frac{\partial z_{f,t}}{\partial U_f} \right) = \sum_t \delta z_{f,t} h_{t-1}^T$$

$$\frac{\partial z_{f,t}}{\partial b_f} = 1, \quad \frac{\partial L}{\partial b_f} = \sum_t \left(\frac{\partial L}{\partial z_{f,t}} \cdot \frac{\partial z_{f,t}}{\partial b_f} \right) = \sum_t \delta z_{f,t}$$

Problem 4.

(a) $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s = \prod_{s=1}^{t} (1 - \beta_s)$

$q(x_t | x_{t-1}) = \mathcal{N}(x_t ; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I)$

$x_t = \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\, \varepsilon_{t-1}$

$= \sqrt{\alpha_t}\, x_{t-1} + \sqrt{1-\alpha_t}\, \varepsilon_{t-1}$

$x_{t-1} = \sqrt{\alpha_{t-1}}\, x_{t-2} + \sqrt{1-\alpha_{t-1}}\, \varepsilon_{t-2}$

$\Rightarrow x_t = \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\, x_{t-2} + \sqrt{1-\alpha_{t-1}}\, \varepsilon_{t-2}\right) + \sqrt{1-\alpha_t}\, \varepsilon_{t-1}$

$= \sqrt{\alpha_t \alpha_{t-1}}\, x_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\, \varepsilon_{t-2} + \sqrt{1-\alpha_t}\, \varepsilon_{t-1}$

ε_{t-1}, ε_{t-2} : Gaussian Noise, $\varepsilon_{t-2}, \varepsilon_{t-1} \in \mathcal{N}(0, 1)$

$\sqrt{\alpha_t(1-\alpha_{t-1})}\, \varepsilon_{t-2} \in \mathcal{N}(0,\ \alpha_t - \alpha_t\alpha_{t-1})$ $\sqrt{1-\alpha_t}\varepsilon_{t-1} \in \mathcal{N}(0,\ 1-\alpha_t)$

$\Rightarrow \sqrt{\alpha_t(1-\alpha_{t-1})}\, \varepsilon_{t-2} + \sqrt{1-\alpha_t}\, \varepsilon_{t-1} \in \mathcal{N}(0,\ 1-\alpha_t\alpha_{t-1})$

$\Rightarrow$ Let $\sqrt{1-\alpha_t\alpha_{t-1}}\, \bar{\varepsilon}_{t-2} = \sqrt{\alpha_t(1-\alpha_{t-1})}\, \varepsilon_{t-2} + \sqrt{1-\alpha_t}\, \varepsilon_{t-1}$

Then $x_t = \sqrt{\alpha_t\alpha_{t-1}}\, x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\, \bar{\varepsilon}_{t-2}$

根據遞迴，我們可以把 x_t 以 x_0, ε 表示

$x_t = \sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_1}\, x_0 + \sqrt{1-\alpha_t\alpha_{t-1}\cdots\alpha_1}\, \varepsilon$, $\varepsilon \in \mathcal{N}(0, I)$

$= \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon$

$\mathbb{E}[x_t | x_0] = \sqrt{\bar{\alpha}_t}\, x_0$, $\mathrm{Var}(x_t | x_0) = (1-\bar{\alpha}_t) I$

$q(x_t | x_0) = \mathcal{N}(x_t ; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t) I)$

(b) $q(x_{t-1} | x_t, x_0) = \dfrac{q(x_t | x_{t-1}, x_0)\, q(x_{t-1} | x_0)}{q(x_t | x_0)}$

Markov property : $q(x_t | x_{t-1}, x_0) = q(x_t | x_{t-1})$

$\Rightarrow q(x_{t-1} | x_t, x_0) \propto q(x_t | x_{t-1})\, q(x_{t-1} | x_0)$

$q(x_t | x_{t-1}) = \mathcal{N}(x_t ; \sqrt{\alpha_t}\, x_{t-1}, \beta_t I)$

$q(x_{t-1} | x_0) = \mathcal{N}(x_{t-1} ; \sqrt{\bar{\alpha}_{t-1}}\, x_0, (1-\bar{\alpha}_{t-1}) I)$

考虑 exponential term (ignore $-\frac{1}{2}$)

$$L = \frac{(x_t - \sqrt{\alpha_t}x_{t-1})^2}{\beta_t} + \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0)^2}{1-\bar{\alpha}_{t-1}}$$

$$= \frac{x_t^2 - 2\sqrt{\alpha_t}x_t x_{t-1} + \alpha_t x_{t-1}^2}{\beta_t} + \frac{x_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}}x_0 x_{t-1} + \bar{\alpha}_{t-1}x_0^2}{1-\bar{\alpha}_{t-1}}$$

$$= x_{t-1}^2\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right) - 2\left(\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0\right)x_{t-1} + \frac{1}{\beta_t}x_t^2 + \frac{\bar{\alpha}_{t-1}}{1-\bar{\alpha}_{t-1}}x_0^2$$

$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I) \Rightarrow$ exponential term $= \frac{(x_{t-1}-\mu)^2}{\tilde{\beta}_t}$

$$\Rightarrow \frac{1}{\tilde{\beta}_t} = \frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}} = \frac{\alpha_t(1-\bar{\alpha}_{t-1})+\beta_t}{\beta_t(1-\bar{\alpha}_{t-1})} = \frac{\alpha_t - \bar{\alpha}_t + \beta_t}{\beta_t(1-\bar{\alpha}_{t-1})} = \frac{1-\bar{\alpha}_t}{(1-\bar{\alpha}_{t-1})\beta_t}$$

$$\Rightarrow \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

$$\frac{2\mu_t}{\tilde{\beta}_t} = 2\left(\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0\right)$$

$$\mu_t = \left(\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\right)\left(\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0\right)$$

$$= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\beta_t\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}x_0$$

$$= \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\beta_t\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}x_0$$

(c) $$\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, t)\right)$$

$$\mu_\theta(x_t, \hat{x}_0) = \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\beta_t\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}\cdot\frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta)$$

$$= \left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} + \frac{\beta_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\right)x_t - \frac{\beta_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta$$

$$= \frac{\alpha_t(1-\bar{\alpha}_{t-1})+\beta_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}x_t - \frac{\beta_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\sqrt{1-\bar{\alpha}_{t-1}}\,\epsilon_\theta$$

$$= \frac{\alpha_t+\beta_t-\bar{\alpha}_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}x_t - \frac{\beta_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta$$

$$= \frac{1}{\sqrt{\alpha_t}}x_t - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta$$

$$= \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\right)$$

(d) Show: $D_{KL}\left(q(x_{t-1}|x_t, x_0) \,\|\, p_\theta(x_{t-1}|x_t)\right) = \frac{1}{2\tilde{\beta}_t}\|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2 + \text{const}$

$q(x_{t-1}|x_t, x_0) \sim \mathcal{N}(\tilde{\mu}_t, \tilde{\beta}_t I)$

$p_\theta(x_{t-1}|x_t) \sim \mathcal{N}(\mu_\theta, \tilde{\beta}_t I)$

$D_{KL}\left(\mathcal{N}(\mu_1, \Sigma) \,\|\, \mathcal{N}(\mu_2, \Theta)\right) = \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)$

$D_{KL}\left(q_\theta(x_{t-1}|x_t, x_0) \,\|\, p_\theta(x_{t-1}|x_t)\right) = \frac{1}{2\tilde{\beta}_t}(\mu_t - \mu_\theta)^T(\mu_t - \mu_\theta)$

$= \frac{\|\mu_t - \mu_\theta\|^2}{2\tilde{\beta}_t}$

$\mu_t = \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\beta_t\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t} x_0 = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_0\right)$

$\mu_\theta = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta\right)$

$\mu_t - \mu_\theta = -\frac{\beta_t}{\sqrt{\alpha_t(1-\bar{\alpha}_t)}}(\varepsilon - \varepsilon_\theta)$

$\|\mu_t - \mu_\theta\|^2 = \frac{\beta_t^2}{\alpha_t(1-\bar{\alpha}_t)}\|\varepsilon - \varepsilon_\theta\|^2$

$D_{KL}(q \,|\, p_\theta) = \frac{\beta_t^2}{2\tilde{\beta}_t\alpha_t(1-\bar{\alpha}_t)}\|\varepsilon - \varepsilon_\theta\|^2$

$\Rightarrow -\text{ELBO}(\theta) = \text{const} + \sum_{t=2}^{T} E\left[\frac{\beta_t^2}{2\tilde{\beta}_t\alpha_t(1-\bar{\alpha}_t)}\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2\right] + E_q[-\log p_\theta(x_0|x_1)]$