ML2025 Fall Homework Assignment 4
Handwritten

Che-Wei Tsai
f13945039@ntu.edu.tw

Yi-Chen Lee
b12901024@ntu.edu.tw

Lu-Fang Chiang
b12901140@ntu.edu.tw

November 2025

Tools You Need to Know/Learn

- SVM, KKT conditions
- LSTM
- VAE, Diffusion model

You are expected to know (or learn through this homework) the definitions and usages of the above concepts, and are encouraged to discuss with TAs during TA hour if you encounter difficulties.

Homework Policy

- The homework is graded out of 100 points.
- The official submission deadline will follow the schedule announced on NTU COOL.
- Homework may be handwritten or typed (e.g. using LaTeX), but must be submitted in **PDF format**.
- If you discuss the homework with classmates, you should state their student IDs in your submission.
- Plagiarism is strictly prohibited. Serious violations will be dealt with according to NTU regulations.

Problem 1 (Support Vector Regression) (25 pts)

Suppose we are given a training set $\{(x_1, y_1), \cdots, (x_m, y_m)\}$, where $x_i \in \mathbb{R}^{(n+1)}$ and $y_i \in \mathbb{R}$. We would like to find a hypothesis of the form $f(x) = w^T x + b$. It is possible that no such function $f(x)$ exists to satisfy these constraints for all points. To deal with otherwise infeasible constraints, we introduce slack variables ξ_i for each point. The (convex) optimization problem is

$$\begin{aligned}
&\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i && && (1)\\
&\text{s.t. } y_i - w^T x_i - b \le \epsilon + \xi_i && i = 1, \ldots, m && (2)\\
&\quad\;\; w^T x_i + b - y_i \le \epsilon + \xi_i && i = 1, \ldots, m && (3)\\
&\quad\;\; \xi_i \ge 0 && i = 1, \ldots, m && (4)
\end{aligned}$$

where $\epsilon > 0$ is a given, fixed value and $C > 0$. Denote that $\xi = (\xi_1, \cdots, \xi_m)$.

(a) (3pts) Write down the Lagrangian for the optimization problem above. Consider the sets of Lagrange multiplier α_i, α_i^*, β_i corresponding to the (2), (3), and (4), so that the Lagrangian would be written as $\mathcal{L}(w, b, \xi, \alpha, \alpha^*, \beta)$, where $\alpha = (\alpha_1, \cdots, \alpha_m)$, $\alpha^* = (\alpha_1^*, \cdots, \alpha_m^*)$, and $\beta = (\beta_1, \cdots, \beta_m)$.

(b) (2pts) Derive the dual optimization problem. You will have to take derivatives of the Lagrangian with respect to w, b, and ξ

(c) (15 pts) Suppose that $(\bar{w}, \bar{b}, \bar{\xi})$ and $(\bar{\alpha}, \bar{\alpha^*}, \bar{\beta})$ are the optimal solutions to a primal and dual optimization problem, respectively.

Denote $\bar{w} = \sum_{i=1}^{m} (\bar{\alpha}_i - \bar{\alpha_i^*}) x_i$

(1) Prove that

$$\bar{b} = \arg\min_{b \in \mathbb{R}} C \sum_{i=1}^{m} \max(|y_i - (\bar{w}^T x_i + b)| - \epsilon, 0) \qquad (5)$$

(2) Define $e = y_i - \left(\overline{w}^T x_i + \bar{b}\right)$ Prove that

$$\begin{cases}
\bar{\alpha}_i = \bar{\alpha}_i^* = 0, & \bar{\xi_i} = 0, & \text{if } |e| < \epsilon \\
0 \le \bar{\alpha}_i \le C, & \bar{\xi_i} = 0, & \text{if } e = \epsilon \\
0 \le \bar{\alpha}_i^* \le C, & \bar{\xi_i} = 0, & \text{if } e = -\epsilon \\
\bar{\alpha}_i = C, & \bar{\xi_i} = e - \epsilon & \text{if } e > \epsilon \\
\bar{\alpha}_i^* = C, & \bar{\xi_i} = -(e + \epsilon) & \text{if } e < -\epsilon
\end{cases} \qquad (6)$$

(d) (5 pts)Show that the algorithm can be kernelized and write down the kernel form of the decision function. For this, you have to show that

(1) The dual optimization objective can be written in terms of inner products or training examples

(2) At test time, given a new x the hypothesis $f(x)$ can also be computed in terms of inner produce.

Problem 2 (Spherical one class SVM) (25 pts)

Suppose we aim to fit a hypersphere which encompasses a majority of data points $\mathbf{x}_1, ..., \mathbf{x}_N \in \mathbb{R}^M$ by considering the following optimization problem: (here $\boldsymbol{\mu}$ and each $\mathbf{x}_i$ are considered as column vectors)

$$\begin{array}{ll} \text{minimize} & R^2 + \frac{1}{\nu}\sum_{i=1}^N C_i\xi_i \\ \text{subject to} & \left.\begin{array}{l} \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \leq R^2 + \xi_i \\ \xi_i \geq 0 \end{array}\right\} \forall i \in [1, N] \\ & R \geq 0 \\ \text{variables} & R \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} = (\xi_1, ..., \xi_N) \in \mathbb{R}^N \end{array} \tag{7}$$

where $C_i > 0$ for each $i \in [1, N]$, and $0 < \nu < \sum_{i=1}^N C_i$. Let $\rho = R^2$ and rewrite (7) in the form of primal problem:

$$\begin{array}{ll} \text{minimize} & f(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = \rho + \frac{1}{\nu}\sum_{i=1}^N C_i\xi_i \\ \text{subject to} & \left.\begin{array}{l} g_{1,i}(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 - \rho - \xi_i \leq 0 \\ g_{2,i}(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = -\xi_i \leq 0 \end{array}\right\} \forall i \in [1, N] \\ & g_3(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = -\rho \leq 0 \\ \text{variables} & \rho \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} \in \mathbb{R}^N \end{array} \tag{8}$$

as well its Lagrangian dual problem:

$$\begin{array}{ll} \text{maximize} & \theta(\alpha, \beta, \gamma) = \inf_{\rho \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} \in \mathbb{R}^N} L(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma) \\ \text{subject to} & \alpha_i \geq 0, \beta_i \geq 0 \ \forall i \in [1, N] \\ & \gamma \geq 0 \\ \text{variables} & \alpha = (\alpha_1, ..., \alpha_N) \in \mathbb{R}^N, \beta = (\beta_1, ..., \beta_N) \in \mathbb{R}^N, \gamma \in \mathbb{R} \end{array} \tag{9}$$

1. (3 pts) Write down the Lagrangian function $L(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma)$ in explicit form of $\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma$.
2. (3 pts) Show that the duality gap between (8) and (9) is zero.
3. (3 pts) Derive $\theta(\alpha, \beta, \gamma)$ in explicit form of dual variables α, β, γ.
4. (3 pts) Show that the dual problem can be simplified as

$$\begin{array}{ll} \text{maximize} & \|\alpha\|_1 \left(\sum_{i=1}^N \hat{\alpha}_i\|\mathbf{x}_i\|^2 - \sum_{1 \leq i,j \leq N} \hat{\alpha}_i\hat{\alpha}_j\mathbf{x}_i^T\mathbf{x}_j\right) \\ \text{subject to} & \sum_{i=1}^N \alpha_i \leq 1 \\ \text{variables} & 0 \leq \alpha_i \leq \frac{C_i}{\nu}, i \in [1, N] \end{array} \tag{10}$$

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

5. (10 pts) Suppose $(\bar{\rho}, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\xi}})$ and $(\bar{\alpha}, \bar{\beta}, \bar{\gamma})$ are optimal solutions to problems (8) and (9), respectively.
   (a) Show that $\|\bar{\alpha}\|_1\bar{\boldsymbol{\mu}} = \sum_{i=1}^N \bar{\alpha}_i\mathbf{x}_i$.
   (b) Show that
   $$\bar{\rho} \in \arg min_{\rho \geq 0}\left(\rho + \frac{1}{\nu}\sum_{i=1}^N C_i \max(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \rho, 0)\right).$$
   (c) Show that
   $$\min\left\{\rho \geq 0 : \sum_{i:\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \rho} C_i \leq \nu\right\} \leq \bar{\rho} \leq \min\left\{\rho \geq 0 : \sum_{i:\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \rho} C_i < \nu\right\}. \tag{11}$$
   (d) Prove that $\bar{\xi}_i = \max\left(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \bar{\rho}, 0\right)$ for each $i \in [1, N]$.
   (e) Prove that
   $$\begin{cases} \bar{\alpha}_i = C_i/\nu & \text{, if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \bar{\rho} \\ \bar{\alpha}_i = 0 & \text{, if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 < \bar{\rho} \\ 0 \leq \bar{\alpha}_i \leq C_i/\nu & \text{, if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 = \bar{\rho} \end{cases}.$$

6. (3 pts) Suppose $C_i = 1/n$ for each $i \in [1, n]$. What is the physical meaning of ν?

Problem 3 (Coupled Input-Forget Gate LSTM) (20 pts)

The **Coupled Input-Forget Gate LSTM (CIFG-LSTM)** modifies the standard LSTM by coupling the input and forget gates, reducing the total number of parameters while maintaining the gating structure. Given input $x_t \in \mathbb{R}^{d_x}$, the previous hidden state $h_{t-1} \in \mathbb{R}^{d_h}$, and the previous cell state $c_{t-1} \in \mathbb{R}^{d_h}$, the CIFG-LSTM equations are defined as:

$$\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f), \\
i_t &= 1 - f_t, \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o), \\
\tilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c), \\
c_t &= f_t \odot c_{t-1} + (1 - f_t) \odot \tilde{c}_t, \\
h_t &= o_t \odot \tanh(c_t),
\end{aligned}$$

where $\sigma(\cdot)$ is the sigmoid function and $\odot$ denotes elementwise multiplication.

(a) (4 pts) Derive the partial derivatives of the cell state c_t with respect to the gates and intermediate states:

$$\frac{\partial c_t}{\partial f_t}, \quad \frac{\partial c_t}{\partial \tilde{c}_t}, \quad \frac{\partial c_t}{\partial c_{t-1}}.$$

Then, using these results, compute the gradient flow for $\frac{\partial L}{\partial f_t}$, $\frac{\partial L}{\partial \tilde{c}_t}$, and $\frac{\partial L}{\partial c_{t-1}}$ in terms of $\frac{\partial L}{\partial c_t}$.

(b) (10 pts) Let $\delta a = \frac{\partial L}{\partial a}$. Derive the following gradients that appear in BPTT for CIFG-LSTM:

$$\begin{aligned}
\frac{\partial L}{\partial c_t} &= \delta h_t \odot o_t \odot (1 - \tanh^2(c_t)) + \delta c_{t+1} \odot f_{t+1}, \\
\frac{\partial L}{\partial f_t} &= \delta c_t \odot (c_{t-1} - \tilde{c}_t), \\
\frac{\partial L}{\partial z_{f,t}} &= \delta f_t \odot f_t(1 - f_t), \\
\frac{\partial L}{\partial z_{o,t}} &= \delta o_t \odot o_t(1 - o_t), \\
\frac{\partial L}{\partial z_{c,t}} &= \delta \tilde{c}_t \odot (1 - \tilde{c}_t^2),
\end{aligned}$$

where $z_{f,t} = W_f x_t + U_f h_{t-1} + b_f$, $z_{o,t} = W_o x_t + U_o h_{t-1} + b_o$ and $z_{c,t} = W_c x_t + U_c h_{t-1} + b_c$.

(c) (6 pts) Show that

$$\frac{\partial L}{\partial W_f} = \sum_t \delta z_{f,t} x_t^T, \qquad \frac{\partial L}{\partial U_f} = \sum_t \delta z_{f,t} h_{t-1}^T, \qquad \frac{\partial L}{\partial b_f} = \sum_t \delta z_{f,t}.$$

Note that the same derivation applies to the output and candidate cell gates (W_o, U_o, b_o) and (W_c, U_c, b_c).

Problem 4 (From VAE to Diffusion Model) (30 pts)

In this problem, we explore how the idea of a Variational Autoencoder (VAE) can be extended to a more powerful generative framework — the **Diffusion Model**.

The key motivation is: why should we map an image distribution to a unit Gaussian latent in just one step, as VAEs do? An intuitive idea is to do it gradually — we slowly add Gaussian noise to an image through multiple steps until it becomes pure noise, where the latent variable approximately follows $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Then, we train a model to reverse this noising process, step by step, to recover the original image. This forward–reverse framework forms the foundation of modern diffusion models such as Denoising Diffusion Probabilistic Models (DDPMs).

The following introduces some basic mathematical definitions that you should follow to derive the results in this problem.

Markov Chain. A Markov chain is a sequence of random variables $\{\mathbf{x}_t\}_{t=0}^{T}$ satisfying the Markov property:

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots, \mathbf{x}_0) = p(\mathbf{x}_t \mid \mathbf{x}_{t-1}), \quad \forall t \in [1, T].$$

That is, the current state $\mathbf{x}_t$ depends only on the previous state $\mathbf{x}_{t-1}$, not on earlier ones.

Diffusion Model Structure. The diffusion model defines a pair of processes:

$$\text{(Forward): } q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}), \quad \text{(Reverse): } p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t),$$

where $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a standard Gaussian prior. The model is trained so that $p_\theta(\mathbf{x}_0)$ approximates the true data distribution $q(\mathbf{x}_0)$. A naive approach would be to use the Maximum Likelihood (ML) objective directly, but similar to the VAE, we instead adopt the **Evidence Lower Bound (ELBO)** technique for tractable optimization.

(a) (7 pts) **(Forward Process)** Suppose we define a Markov chain $\{\mathbf{x}_t\}_{t=0}^{T}$ that gradually adds Gaussian noise to the data:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\mathbf{x}_t; \sqrt{1-\beta_t}\, \mathbf{x}_{t-1}, \, \beta_t \mathbf{I}\big).$$

Show that we can express $\mathbf{x}_t$ in closed form as a noisy version of the original image $\mathbf{x}_0$:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\, \mathbf{x}_0, \, (1-\bar{\alpha}_t)\mathbf{I}\big), \qquad \bar{\alpha}_t := \prod_{s=1}^{t} (1-\beta_s).$$

Equivalently,

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\, \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\, \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Here the schedule $\{\beta_t\}$ satisfies $0 < \beta_t \ll 1$ and typically increases slowly with t, so that $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ decreases to 0 as $t \to T$; hence $\mathbf{x}_T \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Hint: Unroll the recursion of $\mathbf{x}_t$ and use independence of the noise at each step.

(b) (7 pts) **(Reverse Process)** In the reverse process, we wish to sample $\mathbf{x}_{t-1}$ given $\mathbf{x}_t$. Because the forward process is Gaussian with a fixed schedule $\{\beta_t\}$, the conditional

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$$

is Gaussian and can be derived in closed form:

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_{t-1}; \, \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \, \tilde{\beta}_t \mathbf{I}\big).$$

Show that the mean and variance are

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{1-\beta_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\,\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\,\beta_t}{1-\bar{\alpha}_t}\,\mathbf{x}_0, \quad \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\,\beta_t.$$

Hint: Use Bayes' rule and the Gaussian identity $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \;\propto\; q(\mathbf{x}_t \mid \mathbf{x}_{t-1})\, q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)$.

(c) (7 pts) **(Denoising Model Design)** In practice, we learn a parameterized reverse model

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1};\, \boldsymbol{\mu}_\theta(\mathbf{x}_t, t),\, \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)\big).$$

In this problem we fix the reverse variance to the true posterior variance,

$$\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \tilde{\beta}_t \mathbf{I},$$

so that only the mean is learned. A neural network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ is trained to predict the forward noise $\boldsymbol{\epsilon}$; equivalently, it predicts

$$\widehat{\mathbf{x}}_0(\mathbf{x}_t, t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\big(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\big)\,.$$

Show that choosing

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)$$

yields the posterior-form decoder

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1};\, \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \widehat{\mathbf{x}}_0),\, \tilde{\beta}_t \mathbf{I}\big),$$

i.e., the true posterior mean $\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0)$ with the unknown $\mathbf{x}_0$ replaced by $\widehat{\mathbf{x}}_0$. If $\boldsymbol{\epsilon}_\theta$ perfectly predicts $\boldsymbol{\epsilon}$, then $\widehat{\mathbf{x}}_0 = \mathbf{x}_0$ and $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ matches the true posterior.

(d) (9 pts) **(Connection to ELBO Optimization)** We maximize $\log p_\theta(\mathbf{x}_0)$ via the ELBO. Write

$$\log p_\theta(\mathbf{x}_0) = \log \int q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)\, \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}\, \mathrm{d}\mathbf{x}_{1:T} \geq\; \mathbb{E}_q\left[\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}\right] \;:=\; \mathrm{ELBO}(\theta).$$

Rearranging gives

$$-\mathrm{ELBO}(\theta) = \mathrm{const} + \sum_{t=2}^{T} \mathbb{E}_q\Big[\mathbb{D}_{\mathrm{KL}}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)\,\|\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\big)\Big] + \mathbb{E}_q\big[-\log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)\big].$$

Assuming the fixed variance above,

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\big(\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0),\, \tilde{\beta}_t \mathbf{I}\big), \quad p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\big(\boldsymbol{\mu}_\theta(\mathbf{x}_t, t),\, \tilde{\beta}_t \mathbf{I}\big),$$

show that, for each $t \geq 2$,

$$\mathbb{D}_{\mathrm{KL}}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)\,\|\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\big) = \frac{1}{2\tilde{\beta}_t}\,\big\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\big\|^2 + \mathrm{const}.$$

Using the noise parameterization $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$ and the identities derived in part (b),

show that this KL after expectation equals (up to an additive constant) the weighted noise-MSE

$$\mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), \boldsymbol{\epsilon} \sim \mathcal{N}(0,\mathbf{I})}\left[\frac{\beta_t^2}{2\,\tilde{\beta}_t\,\alpha_t\,(1-\bar{\alpha}_t)}\,\big\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\big(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon},\, t\big)\big\|^2\right]$$

and hence

$$-\mathrm{ELBO}(\theta) = \mathrm{const} + \sum_{t=2}^{T} \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), \boldsymbol{\epsilon} \sim \mathcal{N}(0,\mathbf{I})}\left[\frac{\beta_t^2}{2\,\tilde{\beta}_t\,\alpha_t\,(1-\bar{\alpha}_t)}\,\big\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\big\|^2\right] + \mathbb{E}_q\big[-\log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)\big].$$

Therefore, minimizing $-$ELBO is equivalent (up to known time-dependent weights) to minimizing the noise-prediction objective. In other words, the variational-inference view is consistent with the engineering intuition.

You may notice there is still a term for the final denoising step. In practice, this is treated as a simple reconstruction term in pixel space; for our purposes we ignore it and focus on the "KL turns into MSE" result.

Hint:

(1) For $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ in $\mathbb{R}^d$, the Gaussian KL-Divergence formula is

$$\mathbb{D}_{\mathrm{KL}} = \frac{1}{2}\left(\operatorname{tr}\!\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - d + \log\frac{\det \boldsymbol{\Sigma}_2}{\det \boldsymbol{\Sigma}_1}\right).$$

(2) From part (b), write both means via the noise parameterization:

$$\tilde{\boldsymbol{\mu}}_t = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right), \qquad \boldsymbol{\mu}_\theta = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right).$$

Remarks (optional reading).

(1) In the exact ELBO, each timestep carries a non-uniform weight determined by terms such as $\tilde{\beta}_t$; empirically, DDPM replaces this weighted sum with the simpler unweighted loss

$$\mathcal{L}_{\text{simple}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}} \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\!\left(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon},\, t\right)\right\|^2,$$

which often trains more stably.

(2) Why is the simple loss better? This relates to denoising score matching: the optimal $\boldsymbol{\epsilon}_\theta$ estimates the score of the noisy data, and via Tweedie's formula one can recover $\mathbf{x}_0$ from the predicted noise.

(3) A continuous-time viewpoint leads to SDE-based diffusion models; for simplicity, we do not cover the SDE formulation here. Curious students can look up denoising diffusion probabilistic models, score matching, and score-based generative models.