AI Manifesto: a Non-TLDR Approach

Batuhan Karaca

June 8, 2024

Contents

1 Introduction

1.1 Binding Information (TLDR)

This book aims to give intuition for some ML (machine learning) concepts that I thought of significance. I started writing it simply because of the fact that I was bored of reading TLDR textbooks without understanding anything at all. That is why, I tried to give information as concise and as clear as possible. This compound thingy is solely my research and notetaking work piled up over years. I am trying to teach the topic using mostly mathematical proofs, though I also provide examples and/or explanations. I hope this book will be helpful for people wanting a career in the field.

I am not claiming to have 100% accurate statements, though have I not seen any conflictions between my findings and the publications. I do not think any theory is perfect as well. They are good only because they just work in practice. Furthermore, I am not claiming this book to be a textbook material on its own. I only wanted to share (hopefully) helpful information for deep understanding of the objects using math, as every scientific finding has a mathematical background in my opinion. Nevertheless, you will see some beliefs of my own as I could not come up with a clear answer (encountering conflicting statements in publications to check thereafter), so I warped it with my thoughts (science huh). For example, the fact that dependence is a superset of correlation in part 3.1 is such a gibberish. However, every other piece starts to fit nicely on each other with this statement (mutual coherency for now). I hope you will come up with your own brilliant ideas utilizing this knowledge and sharing them for all. Public science and open source forever!

1.2 Conventions and Notations of the Book

It is not mandatory, but recommended to have some familiarity with the concepts (e.g. through courses) in linear algebra, calculus and probability theory.

Unless stated otherwise, these math notations apply.

- Capital letters represent matrices (or hypermatrices in general) and random variables.
- Lowercase symbols denote vectors and scalars.
- In some contexts, 0 and 1 may indicate a vector or matrix only composed of these scalars.
- Vectors are column vectors by default.
- Row vectors are denoted as their transpose with a superscript capital T (i.e. a^T).
- Scalar (or a row/column vector) of a matrix is denoted with corresponding lowercase symbol.
- Pre-superscript on a symbol of an array denotes the shape of the array (i.e. ${}^{M\times N}A$ is a matrix A with shape $M \times N$ and ${}^{M\times 1}a$ is a row vector a with shape $M \times 1$).
- Subscript of a lowercase variable symbol (i.e. $a_{ijk...}$) denotes the index of that variable in its corresponding matrix with capital letter. Furthermore, colon between indices indicates the ordered array (vector/matrix etc.) of variables whose indices are bounded between upper and lower bounds ordered (i.e. $a^T_{i:j} = [a^T_i, a^T_{i+1}, ..., a^T_j]$). In a similar fashion, this notation can be used for parameters of a function (i.e. $f(x^T_{i:j}) = f(x^T_i, x^T_{i+1}, ..., x^T_j)$).
- Subscript of an integral or sum symbol denoting a variable (i.e. $\sum_X$, $\int_X$) indicates that the operation spans the entire domain of that variable (i.e. X).
- Subscript of a probabilitiy distribution denotes conditional probability (i.e. $p_\theta(x) = p(x|\theta)$. Similarly, Subscript of a function denotes an additional parameter (i.e. $f_\theta(x) = f(x, \theta)$
- Subscript of an expected value symbol (i.e. $E_{p(X)}$) denotes the probability weight of the random variable given in its formula (i.e. $p(X)$).
- Aggregation operators such as min support indexing (i.e. $\min_k x_k = x_{\arg_k \min x_k}$)

1.3 Axioms and Theory

1.3.1 Linear Algebra

Scalars, Vectors and Matrices A scalar s is a real, or complex number. A vector $v = [s_1 s_2 ... s_n]^T$ is a 1D (1-dimensional) array of scalars. A matrix

$$M = \begin{bmatrix} v_1^T \\ v_2^T \\ ...v_m^T \end{bmatrix}$$

or $M = [v_1 v_2 ... v_m]^T$ is an, 1D (1-dimensional), array of vectors, 2D (2-dimensional) array of scalars.

For a matrix M, m_{ij} denotes the scalar in ith row and jth column.

Subscript T is called the *transpose*, which flips the array along its left-diagonal (a line from top-left to bottom-right) so that the rows become columns and vice-versa. Trivially transpose of the scalar is equal to itself. A square matrix (equal number of rows and columns), whose transpose is equal to itself is called *symmetric* (i.e. $M^T = M$ or $m_{ij} = m_{ji}$ for all i, j).

For arrays (vectors and matrices), addition (and subtraction) operates element-wise whereas multiplication (and division) has a special variant (see 1.3.5) beside the traditional element-wise multiplication.

Dot Product Dot product of two vectors $a \cdot b = {}^{1\times N}a^{T\ N\times 1}b = {}^{1\times 1}c$ is defined such that c is a scalar and

$$c = \sum_{n=1}^{N} a_n b_n \tag{1.3.1}$$

P-Norm P-Norm of a vector a

$$\|a\|_p = (\sum_i a_i^p)^{\frac{1}{p}} \tag{1.3.2}$$

$$\|a\|_1 \qquad \text{(Manhattan norm)} \tag{1.3.3}$$

$$\|a\|_2 = \|a\| \qquad \text{(Euclidean norm)} \tag{1.3.4}$$

Note that for the squared Euclidean, we will omit the braces and use $a^T a = \|a\|_2^2 = a^2$ since a can be scalar or vector depending on the context.

Matrix Multiplication Multiplication of two matrices ${}^{M\times K}A\,{}^{K\times N}B = {}^{M\times N}C$ is defined such that

$$c_{ij} = \sum_{k=1}^{K} a_{ik} b_{kj} \tag{1.3.5}$$

Definiteness of a Symmetric Matrix A symmetric matrix M is positive definite if and only if for any vector z

$$z^T M z > 0 \tag{1.3.6}$$

Similarly, M is semi-positive definite if and only if for any vector z

$$z^T M z \geq 0 \tag{1.3.7}$$

Similar definitions for negative and semi-negative definite matrices exist (just reversing the equality sign).

1.3.2 Equality Comparison for Minimum and Maximum Functions

Let $x \in \mathcal{R}$. Since it is easy, I leave the understanding of 1.3.8, 1.3.9 and 1.3.10 to the reader. Same approach applies to the maximum function.

$$\arg_x \min(x + y) = \arg_x \min(x) \text{ for any } y \in \mathcal{R} \tag{1.3.8}$$

$$\arg_x \min(yx) = \arg_x \min(x) \text{ for any } y \in \mathcal{R}^+ \ (y > 0) \tag{1.3.9}$$

$$\arg_x \min(yx) = \arg_x \max(x) \text{ for any } y \in \mathcal{R}^- \ (y < 0) \tag{1.3.10}$$

1.3.3 Probability Theory

Probability Given an event Ω_X with a random variable X, $p(x_i)$ denotes probability of X being equal to x_i. It is a value in the range [0,1]. For example, given $\Omega_X =$ "throwing a dice" and $X \in \{1, 2, 3, 4, 5, 6\}$ is the outcome number, $p(2)$ would be probability of getting a 2. The below is an essential property that all probabilities in the event space will sum up to one (hence Manhattan norm of a vector of probabilities, see 1.3.3).

$$\int_X p(X)dX = 1 \tag{1.3.11}$$

Joint Probability The probability $p(x_i, y_j)$ denotes multiple events occuring at the same time. Comma can be thought of as an *and* operator. Adding Y as whether the outcome is odd (1) or not (0), $p(X = 3, Y = 0) = 0$ is getting a 3 and an even outcome at the same time which is improbable. Getting a 3 and 5 at the same time is also improbable ($p(X = 3, X = 5) = 0$). Below is an essential formulation

$$\int_Y p(X, Y)dY = p(X) \tag{1.3.12}$$

Conditional Probability $p(x_i|y_j)$ denotes the probability of $X = x_i$ given $Y = y_j$. For example, $p(Y = 1|X = 3) = 1$ is certain. If we know the outcome is 3, the outcome is an odd number for sure. Below is an essential formulation

$$p(X, Y) = p(X|Y)p(Y) \tag{1.3.13}$$

Bayesian Formula

$$p(A, B) = p(B, A) \tag{1.3.14}$$

$$p(A|B)P(B) = p(B|A)P(A) \tag{1.3.15}$$

$$\underbrace{p(A|B)}_{\text{posterior}} = \frac{\overbrace{p(B|A)}^{\text{likelihood}}\overbrace{P(A)}^{\text{prior}}}{\underbrace{P(B)}_{\text{evidence}}} \tag{1.3.16}$$

$$p(B) = \int_A p(A, B)dA \tag{1.3.17}$$

$$= \int_A p(B|A)p(A)dA \tag{1.3.18}$$

$$\underbrace{p(A|B)}_{\text{posterior}} = \frac{\overbrace{p(B|A)}^{\text{likelihood}}\overbrace{P(A)}^{\text{prior}}}{\underbrace{\int_A p(B|A)p(A)dA}_{\text{evidence}}} \tag{1.3.19}$$

Independence By definition, if

$$p(A, B) = p(A)p(B) \tag{1.3.20}$$

then A and B are independent. Given 1.3.13 and 1.3.20

$$p(A|B) = p(A) \tag{1.3.21}$$

$$p(B|A) = p(B) \tag{1.3.22}$$

Conditional Dependence By definition, if

$$p(A|B, C) = p(A|C) \tag{1.3.23}$$

then A and B are conditionally independent given C. Given 1.3.13 and 1.3.23

$$p(A|B, C)p(B|C)p(C) = p(A, B|C)p(C) \tag{1.3.24}$$

$$p(A|B, C)p(B|C) = p(A, B|C) \tag{1.3.25}$$

$$p(A|C)p(B|C) = p(A, B|C) \tag{1.3.26}$$

1.3.4 Statistics

Mean (Expected Value)

$$\mu_X = E[X] = E_{p(X)}[X] = \mu_X = \int_X p(X) X dX \tag{1.3.27}$$

Covariance

$$\sigma_{XY} = Cov(X,Y) = E_{p(X,Y)}[(X-\mu_X)(Y-\mu_Y)] = E[(X-\mu_X)(Y-\mu_Y)] = \int_X \int_Y p(X,Y)(X-\mu_X)(Y-\mu_Y) dX dY \tag{1.3.28}$$

Variance

$$\sigma_X = Cov(X,X) = Var(X) = \int_X p(X)(X-\mu_X)^2 dX \tag{1.3.29}$$

(Shannon's) Entropy

$$H(p(X)) = -E_{p(X)}[\log p(X)] = -E_{p(X)}[\log p(X)] = -\int_X p(X) \log p(X) dX \tag{1.3.30}$$

Cross Entropy

$$H(p(X), q(X)) = -E_{p(X)}[\log q(X)] = -\int_X p(X) \log q(X) dX \tag{1.3.31}$$

KL Divergence

$$KL(p(X) \parallel q(X)) = E_{p(X)}[\log \frac{p(X)}{q(X)}] = \int_X p(X) \log \frac{p(X)}{q(X)} dX \tag{1.3.32}$$

Categorical Distribution Can be thought of as the general distribution

$$p(y) = \prod_j p(y=j)^{1\{y=j\}} \tag{1.3.33}$$

$$1\{expression\} = \begin{cases} 1 & expression \text{ evaluates to true} \\ 0 & \text{otherwise} \end{cases} \tag{1.3.34}$$

Multi-variate Normal Distribution The general equation below holds only when the covariance matrix Σ is positive definite, where k is the number of features/variables in vectors x and μ. Note when $k = 1$, x and μ become scalar.

$$\mathcal{N}(x; \mu, \Sigma) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)}{\sqrt{(2\pi)^k |\Sigma|}} \tag{1.3.35}$$

$|\Sigma|$ and $\det(\Sigma)$ are notations for the determinant of Σ. I is the identity matrix, the matrix all 0s except the diagonal that is all 1s. It is multiplicative identity element of the matrices (i.e. $AI = IA = A$)

The specific case for when the variance is constant (same for all features, i.e. $\Sigma = \sigma^2 I$)

$$\mathcal{N}(x; \mu, \sigma^2 I) = \frac{1}{(\sigma\sqrt{2\pi})^k} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 I} = \frac{1}{(\sigma\sqrt{2\pi})^k} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 I\right) \tag{1.3.36}$$

For 1.3.36, we will use the notation $\mathcal{N}(x; \mu, \sigma^2) = \mathcal{N}(x; \mu, \sigma^2 I)$, omitting the identity to simplify calculations. For standard normal distribution we also omit the variable term (i.e. $\mathcal{N}(0,1) = \mathcal{N}(x; 0, I)$) and use special notation (i.e. $\mathcal{N}(x; 0, I) = \varphi(x)$). Note that both 1.3.35 and 1.3.36 are Gaussian functions, so the terms Normal and Gaussian are used interchangeably sometimes.

Reparameterization Trick Used in prominent network architectures such as the VAE (variational autoencoder) and the diffusion models, this technique facilitates calculations, such as the derivatives that are required in backpropagation operations. It is not hard to see a one-to-one and onto (bijective) mapping $x = \mu + \sigma \cdot \epsilon$ since looking at the formula below. Derivation is omitted since it is trivially simple.

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sigma^k} \varphi\Big(\frac{x-\mu}{\sigma}\Big) \tag{1.3.37}$$

2 Linear Algebra Used In Learning and Optimization

2.1 Cosine Law

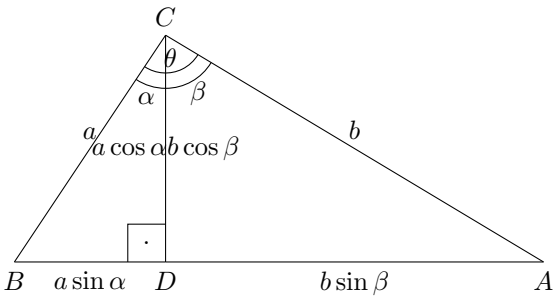

Figure 1:

Figure 1 shows a triangle with rules below

$$|BC| = a \tag{2.1.1}$$
$$|AC| = b \tag{2.1.2}$$
$$|BD| = a\sin\alpha \tag{2.1.3}$$
$$|AD| = b\sin\beta \tag{2.1.4}$$
$$|CD| = a\cos\alpha = b\cos\beta \tag{2.1.5}$$
$$|AB| = |BD| + |AD| \tag{2.1.6}$$
$$|CD| \perp |AB| \tag{2.1.7}$$
$$\angle BCD = \alpha \tag{2.1.8}$$
$$\angle DCA = \beta \tag{2.1.9}$$
$$\angle BCA = \theta = \angle BCD + \angle DCA \tag{2.1.10}$$

$$c = a\sin\alpha + b\sin\beta \tag{2.1.11}$$
$$c^2 = a^2\sin^2\alpha + b^2\sin^2\beta + 2ab\sin\alpha\sin\beta \tag{2.1.12}$$
$$= a^2\sin^2\alpha + b^2\sin^2\beta + 2ab[\cos\alpha\cos\beta - \cos(\alpha+\beta)] \text{ by trigonometric identities} \tag{2.1.13}$$
$$= a^2\sin^2\alpha + b^2\sin^2\beta + ab\cos\alpha\cos\beta + ab\cos\alpha\cos\beta - 2ab\cos(\alpha+\beta) \tag{2.1.14}$$
$$= a^2\sin^2\alpha + b^2\sin^2\beta + a^2\cos^2\alpha + b^2\cos^2\beta - 2ab\cos(\alpha+\beta) \tag{2.1.15}$$
$$= a^2(\sin\alpha^2 + \cos\alpha^2) + b^2(\sin^2\beta + \cos^2\beta) - 2ab\cos(\alpha+\beta) \tag{2.1.16}$$
$$= a^2 + b^2 - 2ab\cos(\alpha+\beta) \tag{2.1.17}$$
$$= a^2 + b^2 - 2ab\cos\theta \tag{2.1.18}$$

2.2 Relation Between Dot Product and Cosine

Assume

$$c = a - b \tag{2.2.1}$$

such that a, b and c are vectors. We define an operation (dot product) on vectors which has following properties

- commutativity (i.e. $a^T b = b^T a$)
- distributivity (i.e. $a^T(b+c) = a^T b + a^T c)$)
- dot product of a vector with itself is equivalent to squared Euclidean norm (length) (i.e. $a^T a = \|a\|^2$)

and many other properties of the scalar multiplication we know of (In fact, the well known dot product formula 1.3.1 satisfies all of these properties). These properties in mind we find

$$c^2 = (a-b)^T(a-b) \tag{2.2.2}$$
$$= a^T a + b^T b - 2a^T b \tag{2.2.3}$$
$$\tag{2.2.4}$$

Using 2.1.18 (note that the values shown there are scalar lengths)

$$a^T a + b^T b - 2a^T b = \|a\|^2 + \|b\|^2 - 2\|a\|\,\|b\|\cos\theta \tag{2.2.5}$$

$$a^T b = \|a\|\,\|b\|\cos\theta \tag{2.2.6}$$

Now for a given point p and normal vector a, every point x making a and $(x-p)$ perpendicular to each other ($\theta = \frac{\pi}{2}$)

$$a^T(x-p) = 0 \tag{2.2.7}$$

$$a^T x - a^T p = 0 \tag{2.2.8}$$

$$a^T x + b = 0 \tag{2.2.9}$$

indeed spans a hyperplane. Notice $b = -a^T p$ is a constant scalar. We have proven the hyperplane equation as well

2.3 Jensen's Inequality

Continuing from our results in previous part, we extract the last feature of x with its corresponding coefficient, to bring an axis of reference to our analysis

$$a^T x + b = 0 \tag{2.3.1}$$

$$a_{1:N-1}^T x + a_N y + b = 0 \tag{2.3.2}$$

Then we bring up a hypersurface with

$$y = f(x) \tag{2.3.3}$$

which intersects the plane in a region Ω such that their intersecting region Ω_c is the boundary of Ω. f is either convex (i.e. a bowl placed on a table in 3D, y increasing toward the sky) or concave (i.e. the bowl upside down in 3D) within Ω. Within Ω_c, we can take any set of points as column vectors organized as a matrix

$$X = \begin{bmatrix} x_0 & x_1 & \dots & x_M \end{bmatrix}$$

It is easy to see that X satisfies both 2.3.2 and 2.3.3. Let us multiply both sides with a vector of weights c whose Manhattan norm is 1 (sum of its items adds up to 1).

$$a_{1:N-1}^T X + (a_N y + b)1^T = 0 \tag{2.3.4}$$

$$a_{1:N-1}^T Xc + (a_N y + b)1^T c = 0 \tag{2.3.5}$$

$$a_{1:N-1}^T Xc + a_N y + b = 0 \tag{2.3.6}$$

Whenever only one element $c_i = 1$ (and trivially others are zero of course), we have

$$a_{1:N-1}^T x_i + a_N f(x_i) + b = 0 \tag{2.3.7}$$

$$a_{1:N-1}^T x_i c_i + a_N f(x_i) c_i + b c_i = 0 \tag{2.3.8}$$

$$\sum_i a_{1:N-1}^T x_i c_i + \sum_i a_N f(x_i) c_i + \sum_i b c_i = 0 \tag{2.3.9}$$

$$a_{1:N-1}^T \sum_i x_i c_i + a_N \sum_i f(x_i) c_i + b = 0 \tag{2.3.10}$$

$$\tag{2.3.11}$$

We have shown for a point $x_c = Xc$

$$y(Xc) = f(X)c = \sum_i c_i f(x_i) \tag{2.3.12}$$

We will further assume for all i, c_i is non-negative (i.e. $c_i \in [0,1]$)

$$\min_k x_{kj} \le x_{ij} \le \max_k x_{kj} \tag{2.3.13}$$

$$c_i \min_k x_{kj} \le c_i x_{ij} \le c_i \max_k x_{kj} \tag{2.3.14}$$

$$\sum_i c_i \min_k x_{kj} \le \sum_i c_i x_{ij} \le \sum_i c_i \max_k x_{kj} \tag{2.3.15}$$

$$\min_k x_{kj} \le \sum_i c_i x_{ij} \le \max_k x_{kj} \tag{2.3.16}$$

$$\tag{2.3.17}$$

We see that x_c lies within a subregion of Ω, hence by definition of convexity of surface,

$$f(Xc) = \begin{cases} \leq f(X)c & \text{if } f \text{ is convex} \\ \geq f(X)c & \text{if } f \text{ is concave} \end{cases} \tag{2.3.18}$$

2.4 Positive Semi-definiteness of a Matrix of the Form X^TX

Let A be a symmetric square matrix, and also $A = X^TX$. For any vector z,

$$z^TAz = z^TX^TXz = (Xz)^T(Xz) \geq 0 \tag{2.4.1}$$

Since squared norm is non-negative, $A = X^TX$ is positive semi-definite.

2.5 Existence of a (Trivially Unique) Orthonormal Basis of Eigenvectors for a Symmetric Diagonalizable Matrix

Let A be a diagonalizable symmetric square matrix. Let P be a square matrix such that it's ith column v_i is the ith eigenvector of A. Let Λ be the diagonal matrix whose ith diagonal λ_i is the eigenvalue corresponding to v_i. According to definition of eigenvalues/eigenvectors.

$$AP = P\Lambda \tag{2.5.1}$$
$$A = P\Lambda P^{-1} \tag{2.5.2}$$
$$\Lambda = P^{-1}AP \tag{2.5.3}$$
$$\lambda_i v_i = Av_i \tag{2.5.4}$$
$$\lambda_i v_i^T = v_i^TA^T \tag{2.5.5}$$
$$\lambda_i v_i^T v_j = v_i^TA^Tv_j \tag{2.5.6}$$
$$= v_i^TAv_j \tag{2.5.7}$$
$$= \lambda_j v_i^T v_j \tag{2.5.8}$$
$$(\lambda_i - \lambda_j)v_i^Tv_j = 0 \tag{2.5.9}$$

Since A is diagonalizable, $\lambda_i = \lambda_j$ if and only if $i = j$. Hence if $i \neq j$, $v_i^Tv_j = 0$ and P is orthogonal, hence $PP^T = D$ such that

$$D_{ij} = \begin{cases} |v_i|^2 & i = j \\ 0 & \text{otherwise} \end{cases} \tag{2.5.10}$$

We can have $P' = PD^{-\frac{1}{2}}$, and $\Lambda' = \Lambda D = D^{\frac{1}{2}}\Lambda D^{\frac{1}{2}}$. Then

$$A = P'\Lambda' P'^T \tag{2.5.11}$$

such that P' is orthonormal. We are going to reference P' and Λ' as P and Λ respectively. We have shown that any symmetric diagonalizable matrix has orthonormal basis of eigenvectors.

2.6 Relation Between Definiteness and Sign of the Eigenvectors

$$z^TAz = z^TP\Lambda P^Tz \tag{2.6.1}$$
$$= (P^Tz)^T\Lambda(P^Tz) \tag{2.6.2}$$
$$= \sum_i (P^Tz)_i^2\lambda_i \tag{2.6.3}$$
$$= \sum_i z_i'^2\lambda_i \tag{2.6.4}$$

The term is actually sum of eigenvalues weighted by values of z', which has the same norm as z due to the fact that P is orthonormal. For all z (hence z'), the weighted sum is non-negative if and only if A is semi-positive definite. Non-negativity in this case is guaranteed when all eigenvectors are non-negative. This approach also works for other types of definite matrices (we cannot say anything about indefinite matrices).

2.7 Relation Between Invertability and Definiteness

For any diagonalizable matrix A,

$$det(A) = det(P)det(\Lambda)det(P^{-1}) \tag{2.7.1}$$
$$= det(P)det(\Lambda)\frac{1}{det(P)} \tag{2.7.2}$$
$$= det(\Lambda) \tag{2.7.3}$$
$$= \prod_i \lambda_i \tag{2.7.4}$$

We see that the semi-definite matrices are not invertible since they have at least one zero eigenvalue. We have shown that semi-definite matrices are not invertible and definite matrices are invertible.

2.8 Definiteness of Hessian Matrix to Determine Curvature Near Critical Points

2.8.1 Quadratic Function

Let $q(x)$ be a quadratic function with ${}^{d\times d}A, {}^{d\times 1}x, {}^{d\times 1}b$ and a scalar c

$$q(x) = x^T Ax + x^T b + c \tag{2.8.1}$$

of the form where A is symmetric.

$$\nabla q(x) = (A + A^T)x + b \tag{2.8.2}$$
$$= 2Ax + b \tag{2.8.3}$$

The critical point $x^* = -\frac{1}{2}A^{-1}b$ where $\nabla q(x^*) = 0$.

$$q(x^* + \Delta x) = q(x^*) + \Delta x^T A\Delta x + 2\Delta x^T Ax^* + \Delta x^T b \tag{2.8.4}$$
$$= q(x^*) + \Delta x^T A\Delta x \tag{2.8.5}$$

We can deduce from previous results,

- $\forall\Delta x[q(x^*) < q(x^* + \Delta x)]$ if A is positive definite (x^* is minimum, minimization problem)
- $\forall\Delta x[q(x^*) > q(x^* + \Delta x)]$ if A is negative definite (x^* is maximum, maximization problem)

If A is indefinite, then depending on Δx, $q(x^*)$ can be less or greater than $q(x^*+\Delta x)$, which means x^* is a saddle point. Since semi-definite matrices are not invertible we cannot use $x^* = -\frac{1}{2}A^{-1}b$. If A is semi-definite, for some x, we have $q(x) = x^T b+c$ (a hyperplane with a $d+1$-dimensional normal having scalars of b and having 1 in the dimension of a reference axis, i.e. inclined-plane). If b is zero, then we have infinitely many critical points (a hyperplane with a normal along the reference axis, i.e. ground-plane).

Let d be an arbitrary vector.

$$\nabla q(x^* + Pd) = 2A(Pd + x^*) + b \tag{2.8.6}$$
$$= 2(P\Lambda P^T)(Pd + x^*) + b \tag{2.8.7}$$
$$= 2P\Lambda P^T Pd + 2Ax^* + b \tag{2.8.8}$$
$$= 2P\Lambda d \tag{2.8.9}$$
$$= \sum_i 2d_i\lambda_i v_i \tag{2.8.10}$$
$$\nabla q(x^* + Pd)^T\nabla q(x^* + Pd) = (2P\Lambda d)^T 2P\Lambda d \tag{2.8.11}$$
$$= 4d^T\Lambda P^T P\Lambda d \tag{2.8.12}$$
$$= 4d^T\Lambda^2 d \tag{2.8.13}$$
$$= \sum_i 4d_i^2\lambda_i^2 \tag{2.8.14}$$
$$= \bar{c}^2 \quad \text{For an arbitary scalar } \bar{c} \tag{2.8.15}$$
$$\sum_i \frac{d_i^2}{\frac{\bar{c}^2}{4\lambda_i^2}} = 1 \tag{2.8.16}$$

We have shown that contours (curves where gradient magnitudes/norms are equal) are ellipsoids.

$$\nabla q(x^* + \frac{\bar{c}}{2\lambda_i} v_i) = \bar{c} v_i \tag{2.8.17}$$

in a contour with multiplier $\bar{c}$. The eigenvector v_i is a unit direction along a principal semi-axis of the ellipsoid. Note that in order to reach to a contour with multiplier $\bar{c}$ from x^*, the magnitude/norm of the vector is $\frac{\bar{c}}{2\lambda_i}$. As λ_i increases, the magnitude/norm decreases, and vice-versa.

For any function, $f(x)$ its second order Tailor expansion approximates it in the neighbor of x.

$$f(x + \Delta x) \sim f(x) + \Delta x^T \nabla f(x) + \Delta x^T H(x) \Delta x \tag{2.8.18}$$

where $H(x) = \nabla^2 f(x)$ is Hessian of $f(x)$. Substituting A, b and c, this approximation can be used to determine the behavior of f near x in optimization algorithms such as gradient descent (where Δx is learning rate times the gradient). For example if $H(x)$ is semi-definite, this time we have a plateau, as plateau is a plane near x (ground or inclined).

2.8.2 Arbitrary Gaussian Function

Let $g(x) = he^{q(x)}$, with an arbitrary scalar h.

$$\nabla g(x) = \nabla q(x) g(x) \tag{2.8.19}$$

We see that the critical points of $g(x)$ are the same as of $q(x)$.

$$g(x^* + \Delta x) = he^{q(x^* + \Delta x)} \tag{2.8.20}$$

$$= he^{q(x^*) + \Delta x^T A \Delta x} \tag{2.8.21}$$

$$= e^{\Delta x^T A \Delta x} g(x^*) \tag{2.8.22}$$

Similarly deductions from part 2.8.1 can be made. Same as that part, If A is indefinite, then depending on Δx, $g(x^*)$ can be less or greater (saddle point). If A is semi-sefinite, for some infinitely many x, we have $g(x) = e^{x^T b + c}$ (inclined-plane). If b is zero, then we have infinitely many critical points (ground-plane). Otherwise, we have no critical points at all.

$$\nabla g(x^* + Pd) = \nabla q(x^* + Pd) g(x^* + Pd) \tag{2.8.23}$$

$$= (2P\Lambda d) e^{(Pd)^T A (Pd)} g(x^*) \tag{2.8.24}$$

$$= (2P\Lambda d) e^{d^T P^T A P d} g(x^*) \tag{2.8.25}$$

$$= (2P\Lambda d) e^{d^T \Lambda d} g(x^*) \tag{2.8.26}$$

$$\nabla g(x^* + Pd)^T \nabla g(x^* + Pd) = (\sum_i 4 d_i^2 \lambda_i^2) e^{2 d^T \Lambda d} g(x^*)^2 \tag{2.8.27}$$

$$= (\sum_i 4 d_i^2 \lambda_i^2) e^{\sum_i 2 d_i^2 \lambda_i} g(x^*)^2 \tag{2.8.28}$$

$$(\sum_i 4 d_i^2 \lambda_i^2) e^{\sum_i 2 d_i^2 \lambda_i} g(x^*)^2 = \bar{c}^2 \tag{2.8.29}$$

Note that after this point, I could not come with a rigorous proof, but used a computer. If we use density function of normal distribution which is a special case of $g(x)$ with $b = -2A\mu$ (x^* becomes μ), $c = \mu^T A \mu$, $h = \frac{1}{\sqrt{(2\pi)^n det(A)}} = \frac{1}{\sqrt{(2\pi)^n \prod_j \lambda_j}}$ (to have infinite integral equal 1 by definition of density functions), and inputting the values to a calculator (I used Desmos), the function behaves like an ellipsoid having eigenvectors along its principal semi-axes. please note that $A = -\frac{1}{2}\Sigma^{-1}$, where covariance matrix Σ needs to be positive definite (it can trivially be semi-positive definite in the limit for generalization). Therefore, A is negative definite, having all negative eigenvalues. As I increased the eigenvalues, the length of the principal semi-axis corresponding to the eigenvector has decreased. This analysis gives ellipsidicity information of the Gaussian that is fitted to a normally distributed sample.

3 Probability Theory and Statistics Concepts Used In Learning

3.1 Distinction Between Correlation and Dependence

We have seen the relationship between eigenvectors/values and the covariance matrix. Assume that we fitted a continuous normal distribution to an elliptic sample of data in order to estimate the most probable regions of occurence, the properties of the continuous distribution will also approximate the properties of the sample. If we squish the sample more, the data will be better approximated with a more squished Gaussian. Remember from the previous parts along the axis of compression,

the eigenvalue corresponding to that axis will become larger. At some point it will be the largest eigenvalue and the data will be approximated by a hyperplane with the normal being its eigenvector (1D hyperplane is a line). We say, such data has a *linear dependence*. However, dependence -relation between axes/dimensions/features in the data- is a more general concept and the covariance matrix only gives information about linear dependence. There can be infinitely many other types such as quadratic, cubic and so on. Correlation is a subset of dependence. When a number of random variables are independent of each other (i.e. mutual independence), they don't have any dependence whatsoever; hence they also become uncorrelated, but the reverse condition does not hold. In other words, independence implies uncorrelatedness but not vice-versa

3.2 Probabilistic Chain Rule Equations With Independent Random Variables

3.2.1 Conditional Mutual Independence

Now assume there exists an array of hidden and observed variables X and Y respectively. As a cliché example, X can be the raw pixel values of an image whereas Y is whether the image is a cat or not.

Further assume, for $i \neq j$, y_i, y_j are conditionally independent given x (1), and x_i, y_i are conditionally independent given x_j (2). You can think of the vector as the data and pairs (x, y) as points such that the point y_i is label of x_i. Using 1.3.23 and 1.3.26

$$p(y|x) = \prod_i p(y_i|x) \tag{3.2.1}$$

$$= \prod_i p(y_i|x_i) \tag{3.2.2}$$

If we use natural language, then we can say (as also proven above) y_i depends only on x_i. This is the foundational assumption that people build their learning models on. Note that x_i can be further divided into features x_{ij} that could be correlated among each other. Then we can further remove those features in the equation above, still having the same result $p(y|x)$.

People use dimension reduction algorithms such as PCA to find those correlated features for removal. PCA simply finds the line having eigenvector with the largest eigenvalue as its normal. Than merges the features affected by this eigenvector. One may iteratively drop the features until the desired dimension.

3.2.2 Markov Property and Markov Chain

Markov property is defined for a stochastic process when the future only depends on the present. Using the equation 1.3.23 along this definition gives

$$p(x_t|x_{t-1}, x_{t-2}, ..., x_0) = p(x_t|x_{t-1}) \tag{3.2.3}$$

Using the equation 1.3.13, we can recursively expand any joint probability as follows

$$p(x_{0:T}) = p(x_0, x_1, ..., x_T) \tag{3.2.4}$$

$$= p_\theta(x_0) \prod_{t=1}^{T} p(x_t|x_{t-1}, x_{t-2}, ..., x_0) \tag{3.2.5}$$

$$p(x_{1:T}|x_0) = \prod_{t=1}^{T} p(x_t|x_{t-1}, x_{t-2}, ..., x_0) \tag{3.2.6}$$

If we further assume the process involving p has markov property, then

$$p(x_{0:T}) = p_\theta(x_0) \prod_{t=1}^{T} p(x_t|x_{t-1}) \tag{3.2.7}$$

$$p(x_{1:T}|x_0) = \prod_{t=1}^{T} p(x_t|x_{t-1}) \tag{3.2.8}$$

3.3 Non-negativity of KL-Divergence

Negative of any function switches its convexity property (i.e. reversing the bowl on the table). Using the definition of KL-Divergence, the Jensen's inequality and the fact that logarithm is a concave function

$$\int_X p(X) \log \frac{p(X)}{q(X)} dX = \int_X p(X)(-\log) \frac{q(X)}{p(X)} dX \tag{3.3.1}$$

$$\geq -\log \int_X p(X) \frac{q(X)}{p(X)} dX \tag{3.3.2}$$

$$\geq -\log \int_X q(X) dX \tag{3.3.3}$$

$$\geq -\log \int_X q(X) dX \tag{3.3.4}$$

$$\geq -\log 1 \tag{3.3.5}$$

$$\geq 0 \tag{3.3.6}$$

Bear in mind that this was true because our assumptions of having a vector of all positive entries and Manhattan norm of 1 aligns with the probability distribution (Integral is a type of sum).

3.4 Monotonicity of KL-Divergence

A function f is *monotonically decreasing* if

$$f(x) \leq f(y) \longleftrightarrow x \geq y \tag{3.4.1}$$

similarly, there is also increasing monotonicity which is irrelevant at this point. We will show for any probability distributions $p(x)$ and $q(x)$, $KL(p(x) \parallel q(x))$ is monotonically decreasing in every direction. We can take derivatives and look at the behavior of the function for that. For simplicity, we use continuous indexing such that $p(x_i) = p_i$

$$\frac{\delta}{\delta q_j} \int_i p_i \log \frac{p_i}{q_i} di = \int_i p_i \frac{\delta}{\delta q_j} \log \frac{p_i}{q_i} di \tag{3.4.2}$$

$$= p_j \frac{-\frac{p_j}{q_j^2}}{\frac{p_j}{q_j}} \tag{3.4.3}$$

$$= -\frac{p_j}{q_j} \tag{3.4.4}$$

Let us keep p constant for now and express KL-divergence as a function of q (i.e. $KL(q)$). Because p has to sum up to 1, in at least one dimension k, 3.4.4 is always negative (but not zero), so KL-divergence is strictly decreasing in that dimension. Then KL-divergence is strictly decreasing in every dimension. In this regard having two points q_1 and q_2, $KL(q_1) = KL(q_2)$ if and only if $q_1 = q_2$, hence the only q that will reach the minimum of zero is $q = p$.

3.5 Cross-Entropy

3.5.1 General Derivation

In ML problems, we want to find the underlying distribution $q(y|x)$. Looking at the 1.3.19, we can find the posterior if the likelihood and the prior are *tractable* (their closed form can be found), which is not generally the case. Therefore, people tend to learn another distribution $p(y|x, \theta)$ with parameters θ. The ideal distribution $p(y|x, \theta)$ should be equal to $q(y|x)$, which is equivalent to saying that the difference $KL(q(y|x) \parallel p(y|x, \theta))$ should be zero. Then the optimum parameters

$$\hat{\theta} = \arg_\theta \min[KL(q(y|x) \parallel p(y|x, \theta))] \tag{3.5.1}$$

$$= \arg_\theta \min[KL(q(y|x) \parallel p(y|x, \theta)) + H(q(y|x))] \tag{3.5.2}$$

$$= \arg_\theta \min \left(E_{q(y|x)} \Big[\log \frac{q(y|x)}{p(y|x, \theta)} \Big] + E_{q(y|x)} \Big[-\log q(y|x) \Big] \right) \tag{3.5.3}$$

$$= \arg_\theta \min \left(E_{q(y|x)} \Big[-\log p(y|x, \theta) \Big] \right) \tag{3.5.4}$$

$$= \arg_\theta \min H(q(y|x), p(y|x, \theta)) \tag{3.5.5}$$

3.5.2 Relationship With Maximum Likelihood Estimator (MLE) for Supervised Tasks

Continuing from 3.5.4,

$$= \arg_\theta \min \sum_j -q(y=j|x) \log p(y=j|x,\theta) \tag{3.5.6}$$

We will assume q is known (see 1.3.33).

$$= \arg_\theta \min \sum_j -(1\{y=j\}) \log p(y=j|x,\theta) \tag{3.5.7}$$

$$= \arg_\theta \min \sum_j -\log p(y=j|x,\theta)^{1\{y=j\}} \tag{3.5.8}$$

$$\arg_\theta \max \big(\exp \big[-H(q(y|x), p(y|x,\theta)) \big] \big) = \arg_\theta \max \sum_j \exp \Big[\log p(y=j|x,\theta)^{1\{y=j\}} \Big] \tag{3.5.9}$$

$$= \arg_\theta \max \sum_j p(y=j|x,\theta)^{1\{y=j\}} \tag{3.5.10}$$

$$= \arg_\theta \max p(y|x,\theta) \text{ (see 1.3.33)} \tag{3.5.11}$$

$$= \hat{\theta}_{MLE} \tag{3.5.12}$$

3.6 Mean Squared Error

3.6.1 General Derivation

We assume $p(y|x,\theta) \sim \mathcal{N}(y; \mu_\theta(x), \sigma^2)$ and $q(y|x) \sim \mathcal{N}(y; \mu(x), \beta)$. We will obtain the result using KL-divergence. Deriving the equation below, Mr. Easy made our life easier.

$$p(x) \sim \mathcal{N}(x; \mu_p, \Sigma_p), q(x) \sim \mathcal{N}(x; \mu_q, \Sigma_q) \longleftrightarrow KL(p(x) \parallel q(x)) = \frac{1}{2} \left[(\mu_p - \mu_q)^T \Sigma_q^{-1} (\mu_p - \mu_q) + tr\left\{ \Sigma_q^{-1} \Sigma_p \right\} + \log \frac{|\Sigma_q|}{|\Sigma_p|} - k \right] \tag{3.6.1}$$

Vectors $x_i, y_i, \mu_i = \mu(x_i)$ and $\mu_\theta(x_i)$ share the same number k of dimensions/features. Substituting gives

$$\hat{\theta} = \arg_\theta \min[KL(q(y|x) \parallel p(y|x,\theta))] \tag{3.6.2}$$

$$= \arg_\theta \min \frac{1}{2} \left[\frac{1}{\sigma^2} \big(\mu - \mu_\theta(x)\big)^2 + \frac{\beta}{\sigma^2} + k \log \frac{\sigma^2}{\beta} - k \right] \tag{3.6.3}$$

$$= \arg_\theta \min \frac{1}{2\sigma^2} \big(\mu - \mu_\theta(x)\big)^2 \tag{3.6.4}$$

People may experiment with different constants $\frac{1}{2\sigma^2}$ and see whichever is the top performing as long as 1.3.8 and 1.3.9 are satisfied (it does for all due to the square as variance cannot be negative).

3.6.2 Relationship With Maximum Likelihood Estimator (MLE) for Supervised Tasks

Continuing from 3.5.12 with the assumption $p(y|x,\theta) \sim \mathcal{N}(y; \mu_\theta(x), \sigma^2)$; hence,

$$\hat{\theta}_{MLE} = \arg_\theta \max p(y|x,\theta) \tag{3.6.5}$$

$$= \arg_\theta \max \frac{1}{(\sigma\sqrt{2\pi})^k} \exp\Big(-\frac{1}{2\sigma^2} (\mu_\theta(x) - y)^2 \Big) \tag{3.6.6}$$

$$-\log \hat{\theta}_{MLE} = \arg_\theta \min \left[-\log \left(\frac{1}{(\sigma\sqrt{2\pi})^k} \exp\Big(-\frac{1}{2\sigma^2} (\mu_\theta(x) - y)^2 \Big) \right) \right] \tag{3.6.7}$$

$$= \arg_\theta \min[-k \log(\sigma\sqrt{2\pi}) + \frac{1}{2\sigma^2} (\mu_\theta(x) - y)^2] \tag{3.6.8}$$

$$= \arg_\theta \min \frac{1}{2} (\mu_\theta(x) - y)^2 \text{ after eliminating the constants} \tag{3.6.9}$$

Remember that we have a freedom when removing the constants as long as they comply with 1.3.8 and 1.3.9. In this regard, we did not eliminate $\frac{1}{2}$ for convenience, as taking derivatives will remove it anyways.

Notice that in 3.5.2, we are approximating a probability distribution (minimizing KL-divergence) between $\hat{y} = p(y|x,\theta)$ and an underlying distribution $q(y|x)$ with labels y (whose hot-encoding is the distribution itself). However, when substituting with the Gaussian, we see we are fitting a Gaussian by minimizing the distance of its mean to the datapoint y.

3.7 Loss Function as a Sum of Elements in Supervised Tasks

For supervised tasks, we found a relationship between MLE and the loss functions $\mathcal{L}_\theta(\hat{y}, y)$ in 3.5.2 and 3.6.2 for a single data point (x, y). What about a dataset? We start with an assumption again that the variables X and Y are independent as given in 3.2.1. Then

$$\hat{\theta}_{MLE} = \arg_\theta \max p_{model}(y|x, \theta) \tag{3.7.1}$$

$$= \arg_\theta \max \prod_i p_{model}(y_i|x_i, \theta) \tag{3.7.2}$$

$$-\log \hat{\theta}_{MLE} = \arg_\theta \min \sum_i -\log p_{model}(y_i|x_i, \theta) \tag{3.7.3}$$

$$= \arg_\theta \min \sum_i \mathcal{L}_\theta(\hat{y}_i, y_i) \tag{3.7.4}$$

If we multiply multiple single variable Gaussians, we reach the formula, in 1.3.36, having a diagonal covariance matrix. In this case, the features are uncorrelated ($Cov(x_i, x_j) = 0$ if $i \neq j$), hence mutually independent, not contradicting our claim in 3.1. In this case, we are also claiming $p(y|x, \theta) = \mathcal{N}(y; \mu_\theta(x), \sigma^2)$.

3.8 Expected Value of Multiple Variables

Here, we are proving that the expected value has the superposition property

$$E_{p(X,Y)}[X] = \int_X \int_Y p(X, Y) X dX dY \tag{3.8.1}$$

$$= \int_X \Big(\int_Y p(X, Y) dY \Big) X dY \tag{3.8.2}$$

$$= \int_X p(X) X dX \tag{3.8.3}$$

$$= E[X] \tag{3.8.4}$$

$$E[aX + bY] = E_{p(X,Y)}[aX + bY] \tag{3.8.5}$$

$$= aE_{p(X,Y)}[X] + bE_{p(X,Y)}[Y] \text{ due to linearity of the integral} \tag{3.8.6}$$

$$= aE_{p(X)}[X] + bE_{p(Y)}[Y] \tag{3.8.7}$$

$$= aE[X] + bE[Y] \tag{3.8.8}$$

3.9 Covariance of Multiple Variables

Here, we are deriving an equation similar to part 3.8

$$Cov(aX + bY, cZ + dT) = E[(aX + bY - \mu_{aX+bY})(cZ + dT - \mu_{cZ+dT})] \tag{3.9.1}$$

$$= E[(a(X - \mu_X) + b(Y - \mu_Y))(c(Z - \mu_Z) + d(T - \mu_T))] \tag{3.9.2}$$

$$= E[ac(X - \mu_X)(Z - \mu_Z) + ad(X - \mu_X)(T - \mu_T) + bc(Y - \mu_Y)(Z - \mu_Z) + bd(Y - \mu_Y)(T - \mu_T)] \tag{3.9.3}$$

$$= acE[(X - \mu_X)(Z - \mu_Z)] + adE[(X - \mu_X)(T - \mu_T)] + bcE[(Y - \mu_Y)(Z - \mu_Z)] + bdE[(Y - \mu_Y)(T - \mu_T)] \tag{3.9.4}$$

$$= acCov(X, Z) + adCov(X, T) + bcCov(Y, Z) + bdCov(Y, T) \tag{3.9.5}$$

We will use mathematical induction to prove the general case. assuming the equation

$$Cov(\sum_{i=1}^{m-1} a_i X_i, \sum_{i=1}^{n-1} b_i Y_i) = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} a_i b_j Cov(X_i, Y_j) \tag{3.9.6}$$

holds, it is trivial to see for

$$Cov(\sum_{i=1}^{m} a_i X_i, \sum_{i=1}^{n} b_i Y_i) = Cov(\sum_{i=1}^{m-1} a_i X_i + a_m X_m, \sum_{i=1}^{n-1} b_i Y_i + b_n Y_n) \tag{3.9.7}$$

$$= Cov(\sum_{i=1}^{m-1} a_i X_i, \sum_{i=1}^{n-1} b_i Y_i) + \underbrace{b_n Cov(\sum_{i=1}^{m-1} a_i X_i, Y_n)}_{c_1} + \underbrace{a_m Cov(X_m, \sum_{i=1}^{n-1} b_i Y_i)}_{c_2} + a_m b_n Cov(X_m, Y_n) \tag{3.9.8}$$

For c_1 and c_2, substituting the coefficients in 3.9.5, we can recursively expand the sum until its components as below.

$$= \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} a_i b_j Cov(X_i, Y_j) + \underbrace{b_n \sum_{i=1}^{m-1} a_i Cov(X_i, Y_n)}_{c_1} + \underbrace{a_m \sum_{j=1}^{n-1} b_j Cov(X_m, Y_j)}_{c_2} + a_m b_n Cov(X_m, Y_n) \tag{3.9.9}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j Cov(X_i, Y_j) \tag{3.9.10}$$

concluding the proof.

3.10 Variance of Multiple Variables

Using the fact that variance is just the *self-covariance*

$$Var(\sum_{i=1}^{n} a_i X_i) = Cov(\sum_{i=1}^{n} a_i X_i, \sum_{i=1}^{n} a_i X_i) \tag{3.10.1}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(X_i, X_j) \tag{3.10.2}$$

$$= \sum_{i=1}^{n} a_i^2 Cov(X_i, X_i) + \sum_{i=1}^{n}\sum_{j\neq i}^{n} a_i a_j Cov(X_i, X_j) \tag{3.10.3}$$

$$= \sum_{i=1}^{n} a_i^2 Var(X_i) + \sum_{i=1}^{n}\sum_{j\neq i}^{n} a_i a_j Cov(X_i, X_j) \tag{3.10.4}$$

$$= \sum_{i=1}^{n} a_i^2 Var(X_i) \text{ (assuming } X \text{ is uncorrelated)} \tag{3.10.5}$$

$$\tag{3.10.6}$$

3.11 Preference of Mean over Summation for Loss Functions in Supervised Tasks

Deriving 3.7.4 for a dataset, we see a pattern in which we try to minimize a sum. Using 3.10.5, for a sample $X_{1:n}$ with constant variance σ^2, the variance of the sample mean,

$$Var(\overline{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) \tag{3.11.1}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \tag{3.11.2}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \tag{3.11.3}$$

$$= \frac{1}{n^2} n\sigma^2 \tag{3.11.4}$$

$$= \frac{\sigma^2}{n} \tag{3.11.5}$$

$$Std(\overline{X}) = \sqrt{Var(\overline{X})} \tag{3.11.6}$$

$$= \frac{\sigma}{\sqrt{n}} \tag{3.11.7}$$

We realize that as number of samples increases, the mean of the sample approximates the true mean of the distribution better. That is why people usually use mean instead of sum. Hence new expression for the loss function becomes

$$\hat{\theta}_{MLE} = \frac{1}{n} \arg_\theta \min \sum_{i=1}^{n} \mathcal{L}_\theta(\hat{y_i}, y_i) \tag{3.11.8}$$

As n is a constant relative to the parameters, we are still maximizing the likelihood.

4 Gradients of Functions in Backpropagation

4.1 FC (Fully Connected) Layers

An FC layer implements the function on a set of matrices

$$^{N\times M}Y = {}^{N\times D}X\,{}^{D\times M}W + {}^{N\times 1}1\,{}^{1\times M}b \tag{4.1.1}$$

Using the definition of matrix multiplication in 1.3.5, we have

$$y_{nm} = \sum_{d=1}^{D} x_{nd}w_{dm} + b_m \tag{4.1.2}$$

Looking at this definition, we can find partial derivatives

$$\frac{\delta y_{nm}}{\delta x_{ij}} = \begin{cases} w_{jm} & if\ n = i \\ 0 & \text{otherwise} \end{cases} \tag{4.1.3}$$

$$\frac{\delta y_{nm}}{\delta w_{ij}} = \begin{cases} x_{ni} & if\ m = j \\ 0 & \text{otherwise} \end{cases} \tag{4.1.4}$$

$$\frac{\delta y_{nm}}{\delta b_{j}} = \begin{cases} 1 & if\ m = j \\ 0 & \text{otherwise} \end{cases} \tag{4.1.5}$$

Using the chain rule with a loss function gives

$$L_{x_{ij}} = \sum_{n=1}^{N}\sum_{m=1}^{M} L_{y_{nm}}\frac{\delta y_{nm}}{\delta x_{ij}} \tag{4.1.6}$$

$$= \sum_{m=1}^{M} L_{y_{nm}}w_{jm} \tag{4.1.7}$$

$$^{N\times D}L_X = {}^{N\times M}L_Y\,{}^{M\times D}W^T \tag{4.1.8}$$

$$L_{w_{ij}} = \sum_{n=1}^{N}\sum_{m=1}^{M} L_{y_{nm}}\frac{\delta y_{nm}}{\delta w_{ij}} \tag{4.1.9}$$

$$= \sum_{n=1}^{N} L_{y_{nm}}x_{ni} \tag{4.1.10}$$

$$^{D\times M}L_W = {}^{D\times N}X_T\,{}^{N\times M}L_Y \tag{4.1.11}$$

$$L_{b_j} = \sum_{n=1}^{N}\sum_{m=1}^{M} L_{y_{nm}}\frac{\delta y_{nm}}{\delta b_{j}} \tag{4.1.12}$$

$$= \sum_{n=1}^{N} L_{y_{nm}} \tag{4.1.13}$$

$$^{1\times M}L_b = {}^{1\times N}1\,{}^{N\times M}L_Y \tag{4.1.14}$$

5 Generative Models

5.1 Variational Autoencoder

5.2 Diffusion Models

5.2.1 What is a Diffusion Model

Diffusion models require two processes that adds noise to data for a number of steps T, (*Sampling steps* parameter in AUTOMATIC1111/stable-diffusion-webui and alike). The first one, *forward (diffusion) process*, makes data noisier by adding Gaussian noise T times, whereas the second, *reverse process*, tries to recover the original image from the result by adding noise again T times.

5.2.2 Forward Process

Let *forward process* be defined as a Markov Chain (see 3.2.8):

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{5.2.1}$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t) \tag{5.2.2}$$

Using 1.3.4 and substituting $\alpha = 1 - \beta_t$ we have

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_t \tag{5.2.3}$$

where $\epsilon_\tau \sim \mathcal{N}(0,1)$ is sampled from the standard normal distribution. Then

$$x_t = \Big(\sqrt{\prod_{\tau=1}^{t}\alpha_\tau}\Big)x_0 + \sum_{\tau=1}^{t-1}\Big(\sqrt{\prod_{i=\tau+1}^{t}\alpha_i - \prod_{i=\tau}^{t}\alpha_i}\Big)\epsilon_\tau + \sqrt{1-\alpha_t}\epsilon_t \tag{5.2.4}$$

Let us prove this argument using mathematical induction. When $t = 1$, it is easy to see equations 5.2.3 and 5.2.4 are the same. For an arbitrary t

$$x_{t-1} = \Big(\sqrt{\prod_{\tau=1}^{t-1}\alpha_\tau}\Big)x_0 + \sum_{\tau=1}^{t-2}\Big(\sqrt{\prod_{i=\tau+1}^{t-1}\alpha_i - \prod_{i=\tau}^{t-1}\alpha_i}\Big)\epsilon_\tau + \sqrt{1-\alpha_{t-1}}\epsilon_{t-1} \tag{5.2.5}$$

$$\sqrt{\alpha_t}x_{t-1} = \Big(\sqrt{\prod_{\tau=1}^{t}\alpha_\tau}\Big)x_0 + \sum_{\tau=1}^{t-2}\Big(\sqrt{\prod_{i=\tau+1}^{t}\alpha_i - \prod_{i=\tau}^{t}\alpha_i}\Big)\epsilon_\tau + \sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\epsilon_{t-1} \tag{5.2.6}$$

$$= \Big(\sqrt{\prod_{\tau=1}^{t}\alpha_\tau}\Big)x_0 + \sum_{\tau=1}^{t-1}\Big(\sqrt{\prod_{i=\tau+1}^{t}\alpha_i - \prod_{i=\tau}^{t}\alpha_i}\Big)\epsilon_\tau \tag{5.2.7}$$

$$\sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_t = \Big(\sqrt{\prod_{\tau=1}^{t}\alpha_\tau}\Big)x_0 + \underbrace{\sum_{\tau=1}^{t-1}\Big(\sqrt{\prod_{i=\tau+1}^{t}\alpha_i - \prod_{i=\tau}^{t}\alpha_i}\Big)\epsilon_\tau + \sqrt{1-\alpha_t}\epsilon_t}_{\Sigma_t} = x_t \tag{5.2.8}$$

Therefore, by induction we have proven that equation 5.2.4 is true. We should have an equation of the form

$$x_t = \mu + \sigma \cdot \epsilon_t \tag{5.2.9}$$

We want to calculate Σ_t. The information in part 3.11 gives some clues on what to do next. Recursively expanding the general case using **??**, we find

$$\mu_{\sum_{i=1}^{N} a_i X_i} = a_1\mu_{X_1} + \mu_{\sum_{i=2}^{N} a_i X_i} \tag{5.2.10}$$

$$= a_1\mu_{X_1} + a_2\mu_{X_2} + \mu_{\sum_{i=3}^{N} a_i X_i} \tag{5.2.11}$$

$$... \tag{5.2.12}$$

$$= \sum_{i=1}^{N} a_i\mu_{X_i} \tag{5.2.13}$$

We also found in the same part

$$Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 Var(X_i) \quad \text{(Assuming } X \text{ is uncorrelated)} \tag{5.2.14}$$

Narrowing down for the normal distribution gives

$$X_i \sim \mathcal{N}(x_i; \mu_i, \sigma_i^2) \text{ and } Y = \sum_{i=1}^{n} c_i X_i \longleftrightarrow Y \sim N(y; \sum_{i=1}^{n} c_i\mu_i, \sum_{i=1}^{n} c_i^2\sigma_i^2) \tag{5.2.15}$$

ϵ_t is already given to be normally distributed. Substituting $X_i = \epsilon_i$ and the corresponding coefficients

$$\sum_{\tau=1}^{t-1}(\prod_{i=\tau+1}^{t}\alpha_i - \prod_{i=\tau}^{t}\alpha_i) + (1-\alpha_t) = \prod_{i=2}^{t}\alpha_i - \prod_{i=1}^{t}\alpha_i + \prod_{i=3}^{t}\alpha_i - \prod_{i=2}^{t}\alpha_i + ... + \prod_{i=t-1}^{t}\alpha_i - \prod_{i=t-2}^{t}\alpha_i + \prod_{i=t}^{t}\alpha_i - \prod_{i=t-1}^{t}\alpha_i + (1-\alpha_t) \tag{5.2.16}$$

$$= \prod_{i=2}^{t}\alpha_i - \prod_{i=1}^{t}\alpha_i + \prod_{i=3}^{t}\alpha_i - \prod_{i=2}^{t}\alpha_i + ... + \prod_{i=t-1}^{t}\alpha_i - \prod_{i=t-2}^{t}\alpha_i + \alpha_t - \prod_{i=t-1}^{t}\alpha_i + 1 - \alpha_t \tag{5.2.17}$$

Note that similar elements cancel each other and we are left with $1 - \prod_{i=1}^{t}\alpha_i$. Substituting $\prod_{\tau=1}^{t}\alpha_\tau = \bar{\alpha}_t$.

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t \tag{5.2.18}$$

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, 1-\bar{\alpha}_t) \tag{5.2.19}$$

5.2.3 Reverse Process and the Loss Function

We found a closed form solution for the forward process function. However, reverse process will be learnt by an ML (machine learning) algorithm. The authors decided to use a type of CNN (Convolutional Neural Network), U-Net.

Let *reverse process* be defined as another Markov Chain (backward in time, see 3.2.7):

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T}p_\theta(x_{t-1}|x_t) \tag{5.2.20}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{5.2.21}$$

We want to minimize $KL(q(x_0) \parallel p_\theta(x_0))$. The formulation in 3.5.1 gives,

$$\arg_\theta \min KL(q(x_0) \parallel p_\theta(x_0)) = \arg_\theta \min H(q(x_0), p_\theta(x_0)) \tag{5.2.22}$$

Bear in mind that q is now the data distribution (generated by data) whereas p is the model distribution (generated by model). Our loss function becomes

$$\mathcal{L}_\theta(x_0) = H(q(x_0), p_\theta(x_0)) = E_{q(x_0)}[-\log p_\theta(x_0)] \tag{5.2.23}$$

which is intractable

$$-\log p_\theta(x_0) \leq -\log p_\theta(x_0) + KL(q(x_{1:T}|x_0) \parallel p_\theta(x_{1:T}|x_0)) \text{ due to 3.3} \tag{5.2.24}$$

$$\leq -\log p_\theta(x_0) + E_{q(x_{1:T}|x_0)}\Big[\frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0)}\Big] \tag{5.2.25}$$

$$\leq -\log p_\theta(x_0) + E_{q(x_{1:T}|x_0)}\Big[\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})/p_\theta(x_0)}\Big] \tag{5.2.26}$$

$$\leq -\log p_\theta(x_0) + E_{q(x_{1:T}|x_0)}\Big[\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\Big] + \log p_\theta(x_0) \tag{5.2.27}$$

$$\leq E_{q(x_{1:T}|x_0)}\Big[\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\Big] \tag{5.2.28}$$

$$E_{q(x_0)}[-\log p_\theta(x_0)] \leq E_{q(x_{0:T})}\Big[\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\Big] \text{ (Integrated both sides)} \tag{5.2.29}$$

$$\arg_\theta \min E_{q(x_0)}[-\log p_\theta(x_0)] \leq \arg_\theta \min E_{q(x_{0:T})}\Big[\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\Big] \tag{5.2.30}$$

We found another loss function which is an upper bound; hence, minimizing this function will minimize the original. This

function is tractable as shown below

$$\arg_\theta \min E_{q(x_{0:T})}\Big[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\Big] = \arg_\theta \min E_{q(x_{0:T})}\Big[\log \frac{\prod_{t=1}^T q(x_t|x_{t-1})}{p(x_T)\prod_{t=1}^T p_\theta(x_{t-1}|x_t)}\Big] \tag{5.2.31}$$

$$= \arg_\theta \min E_{q(x_{0:T})}\Big[-\log p(x_T) + \sum_{t=1}^T \log \frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)}\Big] \tag{5.2.32}$$

$$= \arg_\theta \min E_{q(x_{0:T})}\Big[-\log p(x_T) + \sum_{t=2}^T \log \frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)} + \log \frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\Big] \tag{5.2.33}$$

$$= \arg_\theta \min E_{q(x_{0:T})}\Big[-\log p(x_T) + \sum_{t=2}^T \log(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)} \cdot \frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}) + \log \frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\Big] \tag{5.2.34}$$

$$= \arg_\theta \min E_{q(x_{0:T})}\Big[-\log p(x_T) + \sum_{t=2}^T \log \frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)} + \sum_{t=2}^T \log \frac{q(x_t|x_0)}{q(x_{t-1}|x_0)} + \log \frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\Big] \tag{5.2.35}$$

$$= \arg_\theta \min E_{q(x_{0:T})}\Big[-\log p(x_T) + \sum_{t=2}^T \log \frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)} + \log \frac{q(x_T|x_0)}{q(x_1|x_0)} + \log \frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\Big] \tag{5.2.36}$$

$$= \arg_\theta \min E_{q(x_{0:T})}\Big[\log \frac{q(x_T|x_0)}{p(x_T)} + \sum_{t=2}^T \log \frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)} - \log p_\theta(x_0|x_1)\Big] \tag{5.2.37}$$

$$= \arg_\theta \min E_{q(x_{0:T})}\Big[\underbrace{KL(q(x_T|x_0) \parallel p(x_T))}_{L_T} + \underbrace{\sum_{t=2}^T KL(q(x_{t-1}|x_t,x_0) \parallel p_\theta(x_{t-1}|x_t))}_{L_{t-1}} \underbrace{-\log p_\theta(x_0|x_1)}_{L_0}\Big] \tag{5.2.38}$$

$$= \arg_\theta \min E_{q(x_{0:T})}\Big[\underbrace{\sum_{t=2}^T KL(q(x_{t-1}|x_t,x_0) \parallel p_\theta(x_{t-1}|x_t))}_{L_{t-1}} \underbrace{-\log p_\theta(x_0|x_1)}_{L_0}\Big] \tag{5.2.39}$$

$$= \mathcal{L}'(x_0) \tag{5.2.40}$$

We eliminated L_T term since it is constant with respect to parameters θ. As a side note, the authors assume $p(x_T) \sim \mathcal{N}(0,1)$ and $\beta_T = 1$ making $L_T \sim 0$.

We conditioned on x_0 at 5.2.34 since $q(x_{t-1}|x_t,x_0)$ is tractable and the proof is as follows

$$q(x_{t-1}|x_t,x_0)q(x_t|x_0)q(x_0) = q(x_t|x_{t-1},x_0)q(x_{t-1}|x_0)q(x_0) \tag{5.2.41}$$

$$q(x_{t-1}|x_t,x_0)q(x_t|x_0) = q(x_t|x_{t-1},x_0)q(x_{t-1}|x_0) \tag{5.2.42}$$

$$q(x_{t-1}|x_t,x_0) = q(x_t|x_{t-1},x_0)\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} \tag{5.2.43}$$

$$= q(x_t|x_{t-1})\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} \text{ due to Markov property} \tag{5.2.44}$$

$$= \mathcal{N}(x_t;\sqrt{1-\beta_t}x_{t-1},\beta_t)\frac{\mathcal{N}(x_{t-1};\sqrt{\bar{\alpha}_{t-1}}x_0,1-\bar{\alpha}_{t-1})}{\mathcal{N}(x_t;\sqrt{\bar{\alpha}_t}x_0,1-\bar{\alpha}_t)} \tag{5.2.45}$$

5.2.44 is apparent in 5.2.34. Doing the crazy calculations gives

$$q(x_{t-1}|x_t,x_0) = q(x_t|x_{t-1},x_0)\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} \tag{5.2.46}$$

$$\propto \exp\Big(-\frac{1}{2}\big(\frac{(x_t-\sqrt{\alpha_t}x_{t-1})^2}{\beta_t} + \frac{(x_{t-1}-\sqrt{\bar{\alpha}_{t-1}}x_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(x_t-\sqrt{\bar{\alpha}_t}x_0)^2}{1-\bar{\alpha}_t}\big)\Big) \tag{5.2.47}$$

$$= \exp\Big(-\frac{1}{2}\big(\frac{x_t^2 - 2\sqrt{\alpha_t}x_t x_{t-1} + \alpha_t x_{t-1}^2}{\beta_t} + \frac{x_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}}x_0 x_{t-1} + \bar{\alpha}_{t-1}x_0^2}{1-\bar{\alpha}_{t-1}} - \frac{(x_t-\sqrt{\bar{\alpha}_t}x_0)^2}{1-\bar{\alpha}_t}\big)\Big) \tag{5.2.48}$$

$$= \exp\Big(-\frac{1}{2}\big((\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}})x_{t-1}^2 - (\frac{2\sqrt{\alpha_t}}{\beta_t}x_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0)x_{t-1} + C(x_t,x_0)\big)\Big) \tag{5.2.49}$$

From 1.3.36, the Normal distribution has the form

$$\mathcal{N}(x;\mu,\sigma^2) \propto \exp\left(-\frac{1}{2}\Big(\frac{1}{\sigma^2}x^2 - \frac{2\mu}{\sigma^2}x + \frac{1}{\sigma^2}\mu^2\Big)\right) \tag{5.2.50}$$

Substituting gives

$$q(x_{t-1}|x_t,x_0) = \mathcal{N}(x_{t-1};\bar{\mu}_t(x_t,x_0),\bar{\beta}_t) \tag{5.2.51}$$
$$\bar{\beta}_t = 1/(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}) \tag{5.2.52}$$
$$= 1/(\frac{\alpha_t - \bar{\alpha}_t + \beta_t}{\beta_t(1-\bar{\alpha}_{t-1})}) \tag{5.2.53}$$
$$= 1/(\frac{1-\bar{\alpha}_t}{\beta_t(1-\bar{\alpha}_{t-1})}) \tag{5.2.54}$$
$$= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{5.2.55}$$
$$\bar{\mu}_t(x_t,x_0) = (\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0)\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{5.2.56}$$
$$= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 \tag{5.2.57}$$

Remembering 5.2.18, we have

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t) \tag{5.2.58}$$

Substituting gives

$$\bar{\mu}_t = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t) \tag{5.2.59}$$
$$= \frac{\sqrt{\alpha_t\bar{\alpha}_t}(1-\bar{\alpha}_{t-1}) + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_t)}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\sqrt{1-\bar{\alpha}_t}(1-\alpha_t)}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_t)}\epsilon_t \tag{5.2.60}$$
$$= \frac{\sqrt{\bar{\alpha}_{t-1}}(\alpha_t - \bar{\alpha}_t) + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_t)}x_t + \frac{\sqrt{1-\bar{\alpha}_t}(1-\alpha_t)}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}\epsilon_t \tag{5.2.61}$$
$$= \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\bar{\alpha}_t)}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_t)}x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\epsilon_t \tag{5.2.62}$$
$$= \frac{1-\bar{\alpha}_t}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\epsilon_t \tag{5.2.63}$$
$$= \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t\Big) \tag{5.2.64}$$

Let us focus on L_{t-1}.

$$\arg_\theta \min E_{q(x_{0:T})}\Big[\sum_{t=2}^{T} KL(q(x_{t-1}|x_t,x_0) \parallel p_\theta(x_{t-1}|x_t))\Big] = \arg_\theta \min \sum_{t=2}^{T} E_{q(x_{0:T})}\Big[KL(q(x_{t-1}|x_t,x_0) \parallel p_\theta(x_{t-1}|x_t))\Big] \tag{5.2.65}$$
$$= \arg_\theta \min \sum_{t=2}^{T} E_{q(x_0,x_{t-1},x_t)}\Big[KL(q(x_{t-1}|x_t,x_0) \parallel p_\theta(x_{t-1}|x_t))\Big] \tag{5.2.66}$$
$$\tag{5.2.67}$$

If we use a constant variance $\Sigma_\theta = \sigma_t^2$, we can use the *easy* derivation leading to 3.6.4, finding

$$\arg_\theta \min E_{q(x_{0:T})}\Big[\sum_{t=2}^{T} KL(q(x_{t-1}|x_t,x_0) \parallel p_\theta(x_{t-1}|x_t))\Big] = \arg_\theta \min \sum_{t=2}^{T} E_{\epsilon_t,x_t}\left[\frac{1}{2\sigma_t^2}(\bar{\mu}_t - \mu_\theta)^2\right] \tag{5.2.68}$$
$$= \arg_\theta \min \sum_{t=2}^{T} E_{\epsilon_t,x_t}\left[\frac{1}{2\sigma_t^2}\left(\frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t\Big) - \mu_\theta(x_t,t)\right)^2\right] \tag{5.2.69}$$

We cane parameterize μ_θ similar to μ_t as below

$$= \arg_\theta \min \sum_{t=2}^{T} E_{\epsilon_t, x_t} \left[\frac{1}{2\sigma_t^2} \left(\frac{1}{\sqrt{\alpha_t}} \Big(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t \Big) - \frac{1}{\sqrt{\alpha_t}} \Big(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \Big) \right)^2 \right] \tag{5.2.70}$$

$$= \arg_\theta \min \sum_{t=2}^{T} E_{\epsilon_t, x_t} \left[\frac{(1-\alpha_t)^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \big(\epsilon_t - \epsilon_\theta(x_t, t) \big)^2 \right] \tag{5.2.71}$$

$$= \arg_\theta \min \sum_{t=2}^{T} E_{\epsilon_t, x_0} \left[\frac{(1-\alpha_t)^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \Big(\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} \epsilon_t, t) \Big)^2 \right] \tag{5.2.72}$$

Now let us focus on L_0.

$$\arg_\theta \min[-\log p_\theta(x_0|x_1)] = \arg_\theta \min[-\log \mathcal{N}(x_0; \mu_\theta(x_1, 1), \sigma_1)] \tag{5.2.73}$$

$$= \arg_\theta \min \left[-\log \left(\frac{1}{(\sigma_1 \sqrt{2\pi})^k} \exp \Big[-\frac{1}{2} \Big(\frac{x_0 - \mu_\theta(x_1, 1)}{\sigma_1} \Big)^2 \Big] \right) \right] \tag{5.2.74}$$

$$= \arg_\theta \min \left[\frac{1}{2\sigma_1^2} \big(x_0 - \mu_\theta(x_1, 1) \big)^2 - \log \frac{1}{(\sigma_1 \sqrt{2\pi})^k} \right] \tag{5.2.75}$$

$$= \arg_\theta \min \left[\frac{1}{2\sigma_1^2} \big(x_0 - \mu_\theta(x_1, 1) \big)^2 \right] \tag{5.2.76}$$

$$= \arg_\theta \min \left[\frac{1}{2\sigma_1^2} \left(x_0 - \frac{1}{\sqrt{\alpha_1}} \Big(x_1 - \frac{1-\alpha_1}{\sqrt{1-\bar{\alpha}_1}} \epsilon_\theta(x_1, 1) \Big) \right)^2 \right] \tag{5.2.77}$$

$$= \arg_\theta \min \left[\frac{1}{2\sigma_1^2} \left(\frac{1}{\sqrt{\bar{\alpha}_1}} (x_1 - \sqrt{1-\bar{\alpha}_1} \epsilon_1) - \frac{1}{\sqrt{\alpha_1}} \Big(x_1 - \frac{1-\alpha_1}{\sqrt{1-\bar{\alpha}_1}} \epsilon_\theta(x_1, 1) \Big) \right)^2 \right] \text{ using 5.2.58} \tag{5.2.78}$$

$$= \arg_\theta \min \left[\frac{1}{2\sigma_1^2} \left(\frac{1}{\sqrt{\alpha_1}} (x_1 - \sqrt{1-\alpha_1} \epsilon_1) - \frac{1}{\sqrt{\alpha_1}} \big(x_1 - \sqrt{1-\alpha_1} \epsilon_\theta(x_1, 1) \big) \right)^2 \right] \tag{5.2.79}$$

Remember that $\bar{\alpha}_1 = \alpha_1$. Furthermore, because $\beta_1 = 1 - \alpha_1$ is a variance term, it cannot be negative. Then, $1 - \alpha_1 > 0$ and $\frac{1-\alpha_1}{\sqrt{1-\alpha_1}} = \sqrt{1-\alpha_1}$ but not $-\sqrt{1-\alpha_1}$. Continuing with these substitutions gives

$$= \arg_\theta \min \left[\frac{1}{2\sigma_1^2} \left(\frac{1}{\sqrt{\alpha_1}} (x_1 - \sqrt{1-\alpha_1} \epsilon_1) - \frac{1}{\sqrt{\alpha_1}} \big(x_1 - \sqrt{1-\alpha_1} \epsilon_\theta(x_1, 1) \big) \right)^2 \right] \tag{5.2.80}$$

$$= \arg_\theta \min \left[\frac{1-\alpha_1}{2\sigma_1^2 \alpha_1} \big(\epsilon_1 - \epsilon_\theta(x_1, 1) \big)^2 \right] \tag{5.2.81}$$

$$= \arg_\theta \min \left[\frac{(1-\alpha_1)^2}{2\sigma_1^2 \alpha_1 (1-\bar{\alpha}_1)} \big(\epsilon_1 - \epsilon_\theta(x_1, 1) \big)^2 \right] \tag{5.2.82}$$

$$\mathcal{L}'(x_0) = L_{t-1} + L_0 \tag{5.2.83}$$

$$= \arg_\theta \min \sum_{t=1}^{T} E_{\epsilon_t, x_0} \left[\frac{(1-\alpha_t)^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \Big(\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} \epsilon_t, t) \Big)^2 \right] \tag{5.2.84}$$